Prof. Dr. Hüner Şencan

SOSYAL VE DAVRANIŞSAL ÖLÇÜMLERDE

GÜVENİLİRLİK VE GEÇERLİLİK

seçkin

Sosyal ve Davranışsal Ölçümlerde

Güvenilirlik ve Geçerlilik

Prof. Dr. Hüner Şencan

İÜ İşletme Fakültesi
Davranış Bilimleri Ana Bilim Dalı
Öğretim Üyesi

seçkin
Ankara, 2005

FOTOKOPİ KİTABI ÖLDÜRÜR!

1000 adet basılmış olan bir kitap, eğer 2000 adet basılmış olsa %20, 3000 adet basılmış olsa %30 daha düşük bir fiyatla satışa sunulabilecektir.

Çeşitli nedenlerle kitabın korsan fotokopisini çeken, çektiren ve kullananlar sadece suç işlememekte, aynı zamanda ülkemizin kültürel ve bilimsel yönden gelişmesine ciddi bir şekilde zarar vermektedirler.

Eğer ülkemizde bilim ve kültürün gelişmesini ve kitapların ucuzlamasını gerçekten istiyorsak, korsan fotokopi çekilmesinin önlenmesi çabalarına hepimiz katkıda bulunmalıyız.

Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik
Prof. Dr. Hüner Şencan

ISBN: 975 347 884 4
Dewey No: 371.26

Birinci Baskı: Ankara, Ocak 2005

Copyright © Seçkin Yayıncılık Sanayi ve Ticaret AŞ
Bu kitabın Türkiye'deki tüm yayın hakları Seçkin Yayıncılık Sanayi ve Ticaret AŞ'ye aittir. Eserin hiçbir bölümü, önceden yazılı izin alınmaksızın fotokopi veya diğer baskı yöntemleriyle çoğaltılamaz. Eğitim ve tanıtım amaçlı referans göstererek yapılacak kısmî alıntılarda makul yararlanım ölçüsü aşılamaz.

Sayfa Tasarımı: Seçkin Yayıncılık
Kapak Tasarımı: Seçkin Yayıncılık

Yayın ve Dağıtım:
Merkez: Seçkin Yayıncılık, Sağlık Sokak No: 19/B 06410 Sıhhiye, Ankara
Tel: (0312) 435 30 30, Faks: (0312) 435 24 72

Şube: Ankara Adliye Sarayı K Blok Zemin Kat Sıhhiye, Ankara Tel: (0-312) 309 52 48
Şube: Abide-i Hürriyet Caddesi Arzu Pasajı No: 229/9 Şişli, İstanbul Tel: (0212) 234 34 77

Ağ: www.seckin.com.tr
E-posta: seckin@seckin.com.tr

Baskı: Sözkesen Matbaacılık, Tel: (0312) 395 21 10

Senih ve Sevde'ye...

YAZAR HAKKINDA

Hüner Şencan, 03.06.1955 tarihinde Kırklareli'nin Babaeski ilçesinde doğdu. İlk öğrenimi için Alpullu'da Şeker İlköğretim Okuluna ve orta öğrenimi için Kırklareli Lisesine devam etti. Lisans eğitimini Erzurum Atatürk Üniversitesi İşletme Fakültesinde tamamladı. Bu fakültenin Üretim Yönetimi Bölümü'nden 1980 yılında mezun oldu. Lisansüstü eğitimi için, 1981 yılında İÜ İşletme Fakültesinde Personel Yönetimi ve Endüstri İlişkileri Bölümü'ne kayıt oldu. Bu bölümden *Büro Personelinin Başarı Değerlendirmesi* isimli teziyle yüksek lisans derecesini elde etti. Daha sonra aynı fakültede İşletme ve Personel, Yönetim-Organizasyon Bölümü'nde doktora öğrencisi oldu. Doktora öğrencisi iken 1982 yılında Davranış Bilimleri Ana Bilim Dalı'na araştırma görevlisi olarak girdi. *Yönetici Stresi ve Kişilik* konusunda doktora tezi hazırladı ve tezini başarıyla savunarak 1986 yılında bilim doktoru unvanını aldı. Yaptığı bilimsel araştırmalara dayalı eser incelemesi ve sözlü savunma sınavıyla 1990 yılında doçent oldu. İÜ İşletme Fakültesinin açmış olduğu bir program çerçevesinde 1992 yılında kısa bir süre Azerbaycan'da ve 1994 yılında da altı ay süre ile ABD'de ziyaretçi öğretim üyesi olarak bulundu. Yirmi yılda bir düzenlenen Habitat uluslar arası konferansına onanmış üye olarak katıldı. İşletmelerde insan davranışlarının incelenmesine yönelik olarak değişik konularda araştırmalar yaptı ve 1997 yılında profesörlüğe yükseltildi.

Halen İÜ İşletme Fakültesi Davranış Bilimleri Ana Bilim Dalı'ndaki öğretim üyeliği görevini sürdürmekte olan Şencan; Davranış Bilimleri, Örgütsel Davranış, Halkla İlişkiler, Psikoteknik, İletişim Teknikleri, Araştırma Yöntem Bilimi, Davranışsal Araştırmalarda Bilgisayar Uygulamaları, Uluslar Arası Kalite Sistemleri, Davranışsal İstatistik, Sağlık Kuruluşlarında İlişkilerin Yönetimi ve Örgüt Geliştirme gibi dersleri vermektedir.

Şencan'ın, bağımsız olarak ve diğer yazarlarla birlikte hazırladığı *Tez Yazım Kılavuzu*, *İşletmeciler ve İktisatçılar İçin Rapor Yazımı*, *Orta Ölçekli İşletmeler ve Bürokrasi*, *Türk İnsanının Yaşadığı Konutlara ve Mekanlara İlişkin Eğilimler*, *Sosyal Psikoloji*, *İşletmelerde Eğitim İhtiyacı Analizi* ve *Bilimsel Yazım* isimli kitapları bulunmaktadır. Şencan, evli ve dört çocuk sahibidir.

İÇİNDEKİLER

ÖN SÖZ

Arkamızda bıraktığımız yirminci yüzyılı, kısaca *bilim çağı* olarak isimlendirilebiliriz. Bu yüzyılda ortaya çıkan bilimsel ve teknolojik gelişmeler önceki yüzyıllarda kat edilen mesafenin onlarca katı büyüklüğündedir ve söz konusu gelişmelerin tamamı bilimsel ölçümlere dayalı olarak ortaya çıkmıştır. Ölçümler, bilim adamlarının kuramsal bir varsayımı test etmek istemeleri veya uygulayıcıların daha yararlı ve etkili sonuçlara ulaşma amacı çerçevesinde yapılmıştır. Bilimsel bilginin geometrik artış hızıyla gelişmesi ölçümlerin güvenilir ve geçerli olmasıyla sağlanmıştır. İnsanlar elde edilen sonuçlara güvenmişler, olumlu sonuçlar almışlar neticede yaşanabilir ve sürdürülebilir bir yaşam çevresi oluşturmuşlardır.

Bilim dallarının tümünün gelişmesi; yapılan dikkatli, titiz, güvenilir ve geçerli ölçümlere dayanır. Ölçümlerin güvenilirliği konusu, öncelikle fizik bilimlerinde ve mühendislik branşlarında gündeme gelmiştir. Teknolojinin çağımızda olağanüstü başarılara imza atmasının altında bütünüyle güvenilirlik ve geçerlilik çalışmaları yatar. Daha sonraları sosyal bilimciler ve davranış bilimcileri de doğa ve fizik bilimcilerinden etkilenerek toplumsal ve davranışsal olguların ölçülmesinde, incelenmesinde ve araştırılmasında güvenilirlik ve geçerlilik konusunu gündemlerine almaya başlamışlardır. Sosyal ve davranışsal bilimlerde; psikolojiden sosyolojiye, eğitim bilimlerinden iktisada ve işletme bilimine kadar pek çok alanda ulaşılan sonuçlardan önce ölçüm araçlarının, ölçüm verilerinin veya ölçüm uygulamalarının güvenilirliği ve geçerliliği tartışma konusu olmaya başlamıştır. Sosyal bilimlerde ölçüm konusu özellikle psikoloji, sosyoloji ve eğitim bilimlerinde önem kazanmıştır. Dünyanın ABD, İngiltere, Kanada gibi gelişmiş ülkelerinde kamuoyunun yanlış yönlendirilmemesi, ölçüm prosedürlerinin belirli bir standarda kavuşturulması ve gerçek anlamda bilimsel gelişme sağlanabilmesi için "ölçme ve değerlendirme standartları" oluşturulmuştur. ABD'de 1940'lı yıllardan itibaren geliştirilmeye başlanan "Psikoloji ve Eğitimde Test Standartları" yaklaşımını bu çerçevede değerlendirebiliriz.

Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik adını verdiğim bu kitabı yazma amacım; iş hayatında çalışan ve işletme-iktisat araştırmalarıyla veya sosyal-eğitsel-psikolojik araştırmalarla ilgilenen uygulayıcı-

lara, üniversitelerde üst düzeyde eğitim gören yüksek lisans ve doktora öğrencilerine ve son olarak üniversite öğretim elemanlarına çok fazla ve değişik nitelikteki kaynaklar yerine belirli bazı bilgileri tek bir eser içinde toplu-ca sunmaktır. Yaptığım bilimsel araştırmalarda ve ölçümlerde sadece güvenilirlik ve geçerlilik analizleri için masamda her zaman beş altı kitap, makale bulundurma zorunluluğu hissediyordum. Kaldı ki bu kaynaklar dahi yeterli olmuyordu. O nedenle böyle bir eseri hazırlamakla araştırmacıların ve ölçüm yapan bilim adamlarının işini bir ölçüde kolaylaştırmak istedim.

Günümüzde bilimsel bilgi birikimi o denli artmıştır ki, İnternet'e rağmen bir taraftan bu bilgilere erişim zorluğu yaşarken, diğer taraftan elde edilen bilgilerin genişliği karşısında zamansızlık, alanı kucaklama güçlüğü ve asıl odak noktasını kaybetme veya ondan uzaklaşma rizikosu gibi tehditlerle karşı karşıya kalmaktayız. Ölçüm yapan bilim adamının veya uygulayıcının esas amacı bir sorunu çözmek, bir varsayımı doğrulamak, bir olguyu doğru bir şekilde saptamaktır. Ölçümlerin güvenilirlik ve geçerliliği konusu amaç değil, araçtır. Aracın amaç haline geldiği durumlarda, ölçüm verilerinin güvenilirlik ve geçerliliği için sonsuz döngü veya kısır döngü sarmalına dolanmamız hiç olmayacak bir iş değildir. Bu nedenle "mükemmel bilgiye" ulaşma ütopyasından kaçınarak, bilim alanlarındaki yaygın teamüllerin ve mutabakatların oluşturduğu uygulamaların izlenmesi daha doğrudur. Bu mutabakatlar da sürekli gelişme halinde olduğundan bilim adamlarının ilgili literatürü sürekli takip etmeleri gerekmektedir.

Bu kitapta güvenilirlik ve geçerlilik konuları ele alınırken disiplinler arası bir yaklaşımdan hareket edilmiştir. Konu bir taraftan araştırma yöntem bilimini ve istatistiği ilgilendirmekte diğer taraftan matematiği, psikometriyi, ölçme ve değerlendirme alanını kapsamaktadır. Yayımlanmış araştırma yöntem bilimi kitaplarında güvenilirlik ve geçerlilik konusuyla ilgili olarak istatistiksel tekniklere çok fazla girilmemiş, korelasyon analizi, varyans analizi gibi belirli temel istatistikî tekniklerle alfa ve KR-20 gibi matematiksel formüller verilerek az sayıda hesaplama yöntemi üzerinde durulmuştur. Bu kitapta ise, incelendiğinde de görülecektir ki, okurlara güvenirlik ve geçerlilik analizleri için geniş bir test portföyü sunulmuştur. Önemli olan, bu portföyün içerdiği testlerin hangi koşullarda uygulanacağının bilinmesidir. Bunun için uygulayıcı okumalı, araştırma veya ölçme tasarımını gözden geçirmeli ve mümkünse konusunu bir meslektaşıyla veya danışman hocasıyla tartışmalı, uygulayacağı testleri ondan sonra belirlemelidir. Kitapta güvenilirlik ve geçerlilik konusu disiplinler arası bir yaklaşımla ele alınırken spesifik kullanıcıların ihtiyaçları yerine daha genel bir yaklaşım tarzı benimsenmiştir. O nedenle eğitim sektöründe faaliyet gösteren ölçme ve değerlendirme uzmanlarının ders konularına ilişkin soruların güvenilirlik ve geçerlilik analizleri yüzeysel bir biçimde incelenmiştir. Belirli bilim dal-

larında daha ayrıntılı analizlerin yapılması gereklidir. Okuyucu bu kitabı eğitim bilimleri, psikoloji, sosyoloji, sağlık bilimleri gibi kendi bilim disiplini açısından incelediğinde kafasındaki soruların hepsine yanıt bulamayabilir. Belirli bilim disiplini alanlarında veya güvenilirlik ve geçerliliğin belirli özel tekniklerinde daha ayrıntılı bilgi edinmek isteyen okurlara ek kaynak araştırması yapmalarını öneririz.

Güvenilirlik ve geçerlilik konusunun 1900'lü yıllardan başlayan ve günümüze kadar gelen uzun bir tarihi vardır. Bu uzun tarihi geçmişi içinde başlıca iki teorik yaklaşım ortaya çıkmıştır. Bunlardan birincisi klasik ölçüm kuramı ve ikincisi ise modern ölçüm kuramıdır. Modern ölçüm kuramı birincisini iptal etmemiş ona eklemlenmiştir. Modern ölçüm kuramı, Cronbach'ın çalışmalarıyla 1960'lı yıllarda *genellenebilirlik kuramı* adı altında çıkış yapmış ve daha sonra test yerine test maddelerinin güvenilirlik ve geçerliliğini konu edinen *madde-yanıt kuramıyla* gelişme göstermiştir. Günümüzde ölçüm çalışmalarında her iki kuramın da etkisi görülmektedir. Türkiye'deki ölçüm çalışmalarının güvenilirlik ve geçerlilik analizlerinde *genellenebilirlik* ve *madde yanıt kuramları* daha az kullanılmıştır. Bunun nedeni karmaşık hesaplama yöntemleri nedeniyle dünyada da görece daha az kullanılıyor olmasıdır. Bu kitapta yeri gelen bölümlerde modern ölçüm kuramına göre yapılan güvenilirlik analizlerine de yer verilmiştir. Ancak bu bölümlerde konu kapsamlı ve ayrıntılı bir şekilde işlenmemiştir. Güvenilirlik analizlerini genellenebilirlik veya madde yanıt kuramına göre yapmak isteyen bilim adamlarının ve diğer araştırmacıların ek kaynak araştırması yapmaları gerekir. Konuya ilişkin olarak bu kitapta sunulan bilgiler, bir yüksek lisans ve doktora öğrencisinin ihtiyaçlarını giderecek niteliktedir.

Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik kitabı üniversiteye devam eden lisans öğrencileri için hazırlanmış bir ders kitabı değildir. Bu kitaptan yararlanacak okurların araştırma yöntem bilimi, temel istatistiksel teknikler ve matematiksel eşitlik çözümleri konusunda önceden bilgi sahibi oldukları varsayılmıştır. Kitapta daha çok güvenilirlik ve geçerlilik konusunun anlamı, bilimsel tartışmalar, yaşanan zorluklar, teknikler ve yöntemler *geniş bir gözden geçirme* kapsamında ele alınmıştır. Buradaki amacımız sadece pratik çözümler getiren bir kitap ortaya koymak değil, konun genişliğini, önemini ve kapsamını açığa çıkarmaktır. Güvenilirlik ve geçerlilik analizleri, teknikleri incelenirken aynı zamanda bu tekniklerin değerlendirilmesi yapılmış avantaj ve dezavantajları ortaya konmuştur. Ancak hiç örnek vermeme gibi bir anlayıştan da uzak durulmuştur. İyi bilinen istatistikî tekniklerin dışında, az kullanılan bir çok istatistiksel testin hesaplanma biçimine ilişkin örnekler verilerek okuyucuların bu tekniklere aşina olmaları amaçlanmıştır. Kitabın bir diğer özelliği, gerek güvenilirlik ve gerekse geçerlilik analizlerinde geliştirilen son tekniklere de yer verilmesidir.

Bu tekniklerin tanıtılmasında büyük ölçüde İnternet kaynaklarına başvurulmuştur. Söz konusu tekniklerin bir bölümü henüz kitap olarak yayımlanmadığından bazı tartışmalı konuları içeriyor olabilir. Kitap olarak yayımlanan diğerlerinin bir bölümüne ise kısıtlı zaman dilimi içinde ulaşılamamıştır

Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik kitabının yazımında sık başvurduğum ve yararlandığım birkaç kitabın ismini vermeden geçemeyeceğim. Bunlardan birincisi artık bir klasik sayılan Kerlinger'in (1973) *Foundations of Behavioral Research* isimli çalışmasıdır. Diğeri, psikolojik testler konusunda bilgi veren Cronbach'ın *Essentials of Psychological Testing* başlıklı kitabıdır. Paul Kline'nin (1980) *Handbook of Psychological Testing* isimli kitabı yararlandığım bir diğer eser olmuştur. Bu kitapların dışında son gelişmeler, yaklaşımlar ve teknikler konusunda büyük ölçüde İnternet'teki kamuoyuna açık sitelerden ve şifreli veri tabanlarından yararlandım.

Kitabın bazı bölümleri değerli hocam İstanbul Kültür Üniversitesi Rektörü Prof. Dr. Tamer Koçel tarafından gözden geçirilmiştir. Çeşitli konuları kendisine danıştığım ve fikirlerinden yararlandığım hocama özellikle teşekkür ediyorum. Ayrıca yurt dışından Prof. Dr. Subkoviak ulaşamadığım makalelerini e-posta ile göndererek bu eserin bazı bölümlerinin daha iyi hazırlanmasına katkı yapmıştır. Yine kitabın bazı bölümlerini göndererek görüş ve önerileri aldığım ODTÜ Sosyoloji Bölümü öğretim üyelerinden Prof. Dr. Yusuf Ziya Özcan'a teşekkür etmek isterim. Son olarak bizlerin yetişmesinde emeği geçen hocam Prof. Dr. Atilla Baransel'i rahmetle anıyorum. Ana bilim dalı başkanımız Prof. Dr. İlhan Erdoğan'ın üzerimizde önemli emekleri vardır. Kürsümüzün değerli elemanlarından Prof. Dr Erdal Tekarslan'ın moral desteği ise bu kitabın satırları arasına sinmiştir. Eserin gün ışığına çıkmasında doğrudan veya dolaylı olarak katkı sağlayan diğer kürsü arkadaşlarım ve meslektaşlarımın hepsine teşekkür ederim.

Bir kitabın yazarı, kendi eksiklerini herkesten daha iyi görür. Bu kitabın da eksiklikleri vardır. Bilim ve teknolojide olduğu gibi basılan kitapların da iyileşmesi belirli bir zaman süresi geçtikten sonra yeniden basılan baskılarla ortaya çıkabiliyor. Okuyucularımdan gelecek öneri ve uyarıların daha sonraki baskılar için bana ışık tutacağına inanıyorum. Bu düşüncelerle kitabın lisans üstü eğitim yapan öğrencilerimize, araştırmacılara ve üniversitelerdeki diğer öğretim elemanlarına yararlı olmasını dilerim.

Prof. Dr. Hüner Şencan

Avcılar, İstanbul

TABLOLAR LİSTESİ

Tablolar

ŞEKİLLER LİSTESİ

Şekiller

KISALTMALAR LİSTESİ

A	Ayırma İndeksi
APA	American Psychological Association, (Amerikan Psikoloji Derneği)
BYB	Bilgi, Yetenek, Beceri
ÇDVA	Çok Değişkenli Varyans Analizi
ÇYVA	Çok Yönlü Varyans Analizi
DDF	Diferansiyel Deste Fonksiyonu
DMF	Diferansiyel Madde Fonksiyonu
DTF	Diferansiyel Test Fonksiyonu
EB	Etki Büyüklüğü
FPG	Fark Puanlarının Güvenilirliği
FSH	Farklılıkların Standart Hatası
GA	Güven Aralığı
KFA	Keşfedici Faktör Analizi
KİK	Küme İçi Korelasyon Analizi
KİU	Konu İçeriği Uzmanları
KRT	Kriter Referanslı Test
KTK	Klasik Test Kuramı
MBF	Madde Bilgi Fonksiyonu
MO	Maksimum Olasılık
MÖE	Madde Özellikleri Eğrisi
MYF	Madde Yanıt Fonksiyonu
MYK	Madde-Yanıt Kuramı
NRT	Norm Referanslı test
OSH	Ortalamanın Standart Hatası
ÖSH	Ölçümün Standart Hatası

ÖSV	Ölçümün Standart Varyansı
S-B	Spearman – Brown (Formülü)
SH	Standart Hata
SPSS	Sosyal Bilimler İçin İstatistiksel Analiz Paket Programı
SS	Standart Sapma
TBF	Test Bilgi Fonksiyonu
TFA	Teyit Edici Faktör Analizi
TGP	Tahminî Gerçek Puan
TYVA	Tek Yönlü Varyans Analizi
VŞF	Varyans Şişme Faktörü

SİMGELER LİSTESİ

Ana kütle değerleri Grek harfleriyle; örneklem verileri ise ülke alfabesindeki harflerle gösterilir. Örneğin, μ ana kütle ortalamasını gösterirken $\overline{X}$ örneklem ortalamasını temsil eder. Bu kitapta sık kullanılan istatistiksel simgeler aşağıdaki gibi belirlenmiştir.

Simge	*Adı*	*Anlamı*
A		Ayırma indeksi
α	Alfa	Tip I hata oranı; iç tutarlılık güvenilirlik katsayısı
β	Beta	Tip II hata oranı; standartlaştırılmış regresyon katsayısı
C	Contingency	Kontenjans katsayısı
d		Etki büyüklüğü simgesi
Cov_{XY}, σ_{XY}	Kovaryans	Ortak varyans, ortak değişkenlik
Δ	Delta	Glass, delta etki büyüklüğü simgesi
γ , G	Gamma	Etki büyüklüğü
χ	Ki	Ki-kare istatistik testinin simgesi (χ^2)
ε	Epsilon	Ana kütlede tesadüfî hata
γ	Gamma	Bir araştırma modelinde sabit etkileri göstermek için kullanılır, etki büyüklüğü simgesi
η	Eta katsayısı	Eta katsayısı, TYVA tasarımında sorumlu tutulan varyans yüzdesi (η^2); regresyonda R^2 değerine benzer; doğrusal dağılıma sahip olmayan iki değişken arasındaki korelasyon katsayısı
η^2	Eta kare	Etki büyüklüğü ölçüsü; doğrusal dağılıma sahip olmayan iki değişken arasındaki ortak varyans
K		Kappa

K_a		Ağırlıklı Kappa (sıralı ölçek verileri için uygulanır)
μ	Mü	Ana kütle ortalaması (örneklem ortalaması üzerindeki bir çizgiyle gösterilir)
M, Ort, $\overline{X}$		Örneklem ortalaması
Med	Medyan	Medyan değerinin simgesi
Mod	Mod	Mod değerinin simgesi
ν	Nu	Çoğunlukla serbestlik derecesine işaret eder
π	Pi	Olasılık, tesadüfî katsayı modellerinde katsayı anlamında kullanılır
φ	Phi [fi]	Phi korelasyon katsayısı simgesi.
θ	Theta	Gizli kavramsal yapı, madde-yanıt kuramında bireylerin yetkinlik konumları
$P(X)$		Olasılık
ρ	Rho [ro]	Ana kütle korelasyon simgesi (örneklem korelasyon katsayısı r ile gösterilir)
ρ_K, ρ_I	Rho [ro]	Ana kütle küme içi korelasyon katsayısı (Intraclass)
ρ_x^2	Rho [ro] kare	Güvenilirlik indeksi
ρ_{nis}		Nokta-iki serili korelasyon simgesi
q		MYK'de gizli özellik simgesi
P		Ölçülen özelliğin ana kütledeki beklenen oranı
R		Küme içi korelasyon katsayısı, çoklu korelasyon katsayısı
R_k		Rasch ölçüm modelinde kişi güvenilirliği. Ölçüm değişkeninde kişilerin ne ölçüde iyi bir şekilde farklılaştırılabildiğini gösterir.
$R_{a.bcd}$		Çoklu korelasyon katsayısı
r		Pearson korelasyon katsayısı
r^2		Belirlilik katsayısı
r_λ		Korelasyona dayalı ağırlıklandırılmış (lambda) etki büyüklüğü simgesi

r_b^2		Birleşik korelasyona dayanan belirlilik katsayısı
r_b		Birleşik korelasyon katsayısı
r_y , r_{td}		Ters dönüştürülmüş etki büyüklüğü veya yansız etki büyüklüğü tahmin değeri
r_{tet}	Tetrachoric	Tetrakorik korelasyon katsayısı
$r_{ab.c}$		Kısmî korelasyon katsayısı
σ	Sigma	Ana kütle standart sapma simgesi
ϕ	phi	Gerçek ikili phi katsayısı
SS, $\hat{\sigma}$	Şapkalı sigma	Örneklem için standart sapma simgesi
Spearman ρ		Spearman sıra korelasyonu
σ^2	Sigma kare	Ana kütle varyans değeri
V, Var, $\hat{\sigma}^2$, S^2		Örneklem varyans değeri
τ	Tau	Tesadüfî etki modellerinde varyans bileşeni, Kendall tau simgesi
τ_a , ta	Tau a	Kendall tau a simgesi
τ_b , tb	Tau b	Kendall tau b simgesi
τ_c , tc	Tau c	Kendall tau c simgesi
ω	Omega	ω^2 TYVA modellerinde etki büyüklüğü tahminini gösteren simge
R	Varyasyon katsayısı	Çoklu korelasyon
R^2	Belirlilik katsayısı	Açıklanan veya sorumlu tutulan varyansın yüzdesi
$R_{a.bcd}$	Çoklu korelasyon katsayısı	A ve b, c, ve d gibi bir dizi değişken arasındaki korelasyon katsayısı
r	Değişkenlik katsayısı	Korelasyon (aynı zamanda iki serili korelasyon, Pearson ürün-moment korelasyonu olarak da isimlendirilir)
r_{nis}		Nokta-iki serili korelasyon katsayısı
r_{is}		İki serili korelasyon katsayısı
r_t		Tetrakorik korelasyon katsayısı

F		F-testi
t		t-testi
k		Regresyon modellerinde tahmin göstergelerinin sayısı
sd		Serbestlik derecesi
r_{sb}		Spearman-Brown düzeltme katsayısı
G		Genellenebilirlik kuramı
X		Gözlem değerleri
$x = X - \overline{X}$		Sapma, ortalamadan olan farklar (küçük harfler sapmaları gösterir)
x^2		Ortalamadan farkların karesi
Σx^2		Kareler toplamı (kısa ifadelendirme)
$\Sigma x^2 / n$		Kareler ortalaması (kısa ifadelendirme)
i, j, k, l, m, n		İndisler için kullanılan harfler
v, w, x, y, z		Değişkenler için kullanılan harfler
p, q		İki serili değişkenler için kullanılan harfler
Q		Varyansların homojenliği simgesi
z'	Fisher z'	Korelasyon katsayısı r'nin standart z puanlarına dönüştürülmesi
Z_r	Fisher z'	Fisher z'ye dönüştürülmüş korelasyon katsayısı
Δ	Effect size	Glass delta etki büyüklüğü

GİRİŞ

Bilim, yöntem açısından çok iyi tasarlanarak sorulmuş bir soruyla başlamamıştır ki, herkesi tatmin edici mükemmel bir cevapla sona ersin.

John Tukey

Üretilen bilgilerin bilimsel bir nitelik kazanması doğru olmasına ve bu bilgilerin her defasında yapılan gözlem ve deneylerle kanıtlanmasına bağlıdır. Belirli bir varsayımın test edildiği, değişkenler arasında nedensellik ilişkisi kurulduğu araştırma verileri eğer güvenilirlik ve geçerlilik analizlerine dayanıyorsa güven verir. Güvenilirlik ve geçerlilik analizleri yapılmadan herhangi bir araştırmanın analiz sonuçlarını tablolaştırmak, bu araştırmayla ilgili yorum yapmak, bir hipotezi kabul veya ret etmek doğru değildir.

Bilimsel ölçüm ve araştırmalarda güvenilirlik ve geçerlilik kavramları birlikte kullanılmıştır. Güvenilirlik ve geçerlilik, herhangi bir şeyin *uygun* ve *sağlam* olduğu hakkında bize bilgi verir. Bir ayakkabıyı düşünelim. Bu ayakkabı ayak ölçülerimize göre yapılmışsa uygundur. Daha büyük veya daha küçük numaralar uygun değildir, çünkü bu ayakkabılarla yürüyemeyiz veya uzun mesafeler kat edemeyiz. Öte yandan bu ayakkabı aynı zamanda sağlam olmak zorundadır. Bu ayakkabıyı tekrar tekrar giyebilmeliyiz. Ayakkabı eğer basit bir yapıştırıcı kullanılarak üretilmişse uzun süre giyemeyiz. Basit gibi görülecek bu örnekle vurgulamak istediğimiz, ayakkabının hem *uygun* hem de *sağlam* olması gerektiğidir. Uygunluk, geçerlilik; sağlamlık ise güvenilirliktir. Bir araştırmadan elde edilen verilerin geçerli olması, ölçüm amacına uygun olmasına bağlıdır. Araç, ölçüm amacıyla ilgili olmayan maddeleri içermemeli ve söz konusu maddeler aynı zamanda ölçüm amacı konusunda yetersiz kalmamalıdır.

Bir araştırmanın bilimsel açıdan güçlü olması büyük ölçüde hatalardan arındırılmasına bağlıdır. Hatalardan arındırma, güvenilirlikle ilgilidir. Deneyimsiz araştırmacılar çok sayıda hata yaparlar. Hatalar büyük ölçüde

araştırmanın veya ölçümün tasarımından kaynaklanır. Bir araştırmanın tasarımı yanlışsa, değişkenler arasındaki ilişkiler dikkatsiz bir şekilde tanımlanmışsa bir çok hata ile karşılaşmak kaçınılmaz hale gelir. Örneğin, araştırmanın kaç örneklem üzerinde yapılacağı, örneklem büyüklüğü ve örneklem hatasının ne olacağı, örnekleme sürecinde hangi yönetimin seçileceği doğru bir şekilde belirlenmemişse, *tesadüfî örnekleme* yerine *kolayda örnekleme* yöntemi seçilmişse, ölçüm aracı bu kitapta ele alınan bir takım testlerden geçirilmemişse toplanan verilerin ana kütle için güvenilirlik ve geçerliliği kuşkulu hale gelir.

Bir araştırmanın güçlü olmasını sağlayan ikinci faktör geçerliliktir. Geçerlilik, ölçüm amacına uygunluk ve ölçüm yapılan ana kütleye genelleme yapabilme anlamına gelir. Bilimsel araştırmalar genellenebilirlik özellikleri açısından farklı düzeylerde değerlendirilir. Örnek kütleyi tanımlamaya yönelik aksiyon ve vak'a araştırmalarının genellenebilirlik özellikleri düşüktür. Bu nedenle söz konusu araştırmaların bilimsel bilgi birikimine olan katkıları daha azdır. Böyle olunca, araştırma verilerinin olguyu tanımlama uygunluğu açısından her tür araştırma için kapsamlı geçerlilik analizleri yapmaya gerek yoktur. Geçerlilik analizlerinde; (a) süreç, (b) ölçüm aracı ve (c) toplanan veriler üzerinde durulur. Tarihsel, tanımlayıcı ve sonuç çıkarıcı araştırmaların her üçünde de *sürecin geçerliliği* önem kazanırken sadece sonuç çıkarıcı araştırmalarda ölçüm aracı ve toplanan verilerin istatistiksel geçerliliği ön plana çıkar. Sosyal bir olgunun belirli bir zaman kesitindeki fotoğrafını çekmeyi amaçlayan tanımlayıcı araştırmalarda değişik nitelikte *anket formları* kullanılır. Araştırmacı anket formunun hangi bölümleri için istatistiksel geçerlilik, hangi bölümleri için ise yargısal geçerlilik analizleri yapacağını, hangi bölümleri için ise buna gerek olmadığını önceden belirlemelidir. Örneğin, anket formunun *demografik değişkenleri* için istatistiksel geçerlilik analizleri yapılmaz. Arka plandaki gizli bir olguyu araştıran çoklu derecelere sahip ölçekler için ise yine belirli düzeydeki geçerlilik analizleri yapılır. Tanımlayıcı araştırmalarda toplanan verilerin geçerliliği, araştırma sürecinin bilimsel gereklere uygun olması ve kullanılan soru formunun yüzey ve içerik geçerliliğine sahip olmasıyla ölçülür. Tanımlayıcı araştırmalarda arka plandaki gizli kavramsal yapıların ortaya çıkarılması gibi bir amaç güdülmediğinden *kriter* ve *yapısal geçerlilik* analizleri yapılamaz. Aynı şekilde vak'a araştırmalarında da istatistiksel geçerlilik analizleri yapılmaz. Bilim adamı, vak'a araştırmalarında sadece araştırma sürecinin geçerliliği ile kullanılan ölçüm aracının yüzey ve içerik geçerliliği konuları üzerinde odaklanır. Bilim adamının daha başlangıç aşamasında yaptığı araştırma türü

ile geçerlilik analizi gereksinimi arasında bir bağ kurması ve araştırmanının tasarımını bu çerçevede gerçekleştirmesi gerekir. Tüm araştırma / ölçüm çalışmaları bittikten, istatistiksel analizler yapıldıktan sonra "bir de faktör analizi yöntemiyle yapısal geçerliliği test edelim" anlayışı yöntem bilimi açısından geçersizdir. Araştırmacı ölçüm verileri için uygulayacağı güvenilirlik ve geçerlilik analizlerini daha baştan planlanmalı ve araştırma süreci içinde bu planları sürekli olarak gözden geçirerek gerekli değişiklikleri yapmalıdır.

Bilimsel ölçümler değişik şekillerde yapılır. Deneysel araştırmalarla bilgi toplama, gözlem aracılığıyla ölçüm yapma, mülakat aracılığıyla bilgi toplama, kişisel bildirim araçlarıyla bilgi toplama bunların başlıcalarıdır. Sosyal ve davranışsal bilimlerde daha çok gözlem, mülakat ve kişisel bildirim (anket) araçlarının kullanıldığı görülür. Kişisel bildirim araçlarıyla çok değişik konularda bilgi toplanır. Bunların başlıcaları aşağıdaki gibidir:

1. Düşünceler, kanaatler.
2. İnançlar, ideolojiler.
3. Deneyimler.
4. Zeka.
5. Algılar.
6. Davranışlar.
7. Fiziksel güç.
8. Psikolojik güç.
9. İlgiler.
10. Kişilik.
11. Yetenek[1] ve beceriler.
12. Bilgi.
13. Tutumlar.
14. Tepkiler.
15. Duyusal yeterlilikler.
16. Diğerleri.

Kişisel bildirim araçları genellikle dolaylı ölçüm imkanı sağlar. Fiziksel ölçümlerde olduğu gibi bu araçlardan kişinin gücü, boyu, ağırlığı gibi tartışmasız somut bilgiler elde edilemez. Araştırmacı bu ölçüm araçlarını

[1] Yetenek: Başarma veya başarılı olma potansiyeli. Yetenekler, zihinsel veya fiziksel alanı kapsayabilir. Beceri: Deneyim ve tekrarlarla elde edilmiş kazanılmış özellikler. Yetenek, potansiyel güçtür. Beceri ise, somut gözlemlenebilir kanıtlara dayanır.

kullanarak arka planda gözükmeyen, fakat kendisini belli eden belirli kavramsal yapıları veya düşünceleri ortaya çıkarmaya çalışır. Bu tür ölçümlerde hata yapma olasılığı yüksektir. Çünkü üzerinde araştırma yapılan kişiler kendilerini çok iyi tanımıyor olabilirler, bazen kişilerin araştırılan konu hakkındaki bilgileri yetersizdir, bazen de araştırma yapılan kişiler tarafsız veya dürüst olmayabilirler. Araştırmacı eğer ölçüm aracını ve araştırma uygulamasının tasarımını bilgileri yansız toplayacak bir şekilde yapılandırmamışsa güvenilir bilgiler elde edemez. Ölçüm aracı, aynı ana kütledeki farklı gruplara veya aynı yetenek düzeyindeki diğer kişilere uygulandığında benzer sonuçlar vermiyorsa güvenilir değildir.

Her ölçme yönteminin güvenilirlik ve geçerlilik incelemesi kendine özgü özel prosedürleri gerektirir. Araştırmacıya düşen görev, bu prosedürlerin ayrıntılarını öğrenmek ve literatürde konuyla ilgili son araştırma bulgularını takip etmektir. Bilimsel alanda her geçen gün yeni uygulamalar, modeller ve yeni hesaplama yöntemleri geliştirilmektedir. Araştırmacı, klasik hesaplama yöntemleriyle modern hesaplama yaklaşımları arasında nerede duracağını iyi tespit etmelidir. Klasik hesaplama yöntemlerinin dar sınırları arasına sıkışıp kalmak araştırmacıya bir açılım sağlamaz. Bu nedenle, henüz yaygınlık kazanmamış hesaplama yöntemlerinden uzak kalmamalı, fakat bütünüyle bu hesaplama yöntemlerine de dayanmamalıdır.

Bu kitapta güvenilirlik ve geçerlilik konusu, sosyal ve davranışsal bilimlerin değişik alanlarında araştırma yapan bilim adamlarının temel bilgi gereksinimini karşılamaya yönelik olarak hazırlanmıştır. Her bir araştırmanın kendine özgü özel uygulamaları olabilir. Bilim adamları bu kitapta ele alınan güvenilirlik ve geçerlilik analizlerini belirli bir karma içinde uygulayabilirler. Test ve ölçek geliştiren bir araştırmacı güvenilirlik ve geçerlilik analizine ilişkin bilgileri örnekleme süreci ve istatistiksel analiz bilgileriyle bütünleştirmek durumundadır. Çünkü, bilim disiplinleri karmaşık bir ağ şeklinde iç içe geçmiş durumdadır. Geçerlilik ve güvenilirlik analizlerini yapan bilim adamı; ölçme ve değerlendirme, istatistiksel analiz, matematiksel analiz, soyut model geliştirme ve kavram haritası çıkarma bilgisine birlikte sahip olmalıdır. Bu bilim disiplinlerinden birinin payı bazen artarken bazen de azalabilir. Araştırma probleminin ortaya koyduğu gerçeklere göre bilim adamı kendi yönelimini deneme yanılma, sezgiler, önceki araştırma sonuçları ve ulaşmak istediği hedefleri göz önünde bulundurarak tespit eder. Hiçbir araştırmacı *şablon modellerden* hareket etmemelidir. Nihaî amaç, matematiksel ve istatistiksel işlemleri başarıyla uygulamak değil, yaşamın kendisinde var olan *gerçekleri* ortaya çıkarmaktır. Gerçek, başarılı bir şekilde ortaya konduğu

ölçüde istatiksel ve matematiksel işlemler anlam kazanır. Bu kitabı okuyan araştırmacılar, bilim adamları güvenilirlik ve geçerlilik analiziyle ilgili işlemleri bir *araç çantası* olarak değerlendirmelidirler. Güvenilirlik ve geçerlilik analizi yöntemlerinin çok ayrıntılı bir şekilde ve çok büyük örneklemlerde uygulanmış olduğu iddiası, "hayatın gerçeklerini ortaya çıkarma" olgusunun önüne geçmemelidir. Gerçekler, uzun yıllar sonunda defalarca yapılan araştırmalarla belirli bir oranda doğrulandığı zaman ortaya çıkar. Bir araştırmacıya düşen görev, bu uzun tarihsel süreç içinde gerçeğin ortaya çıkması için kendi misyonunu yerine getirmektir.

Öğrencilere uygulanan bir test sınavı, işe eleman seçerken adaylara uygulanan bir test bataryası, bir süper markette tüketicilerin davranışlarının gözlenmesi, çalışanlara uygulanan bir iş tatmini ölçeği, yöneticilere uygulanan bir gerilim ölçeği veya liselerden mezun olan öğrencilere uygulanan bir ilgi envanteri *sağlam* veya *güçlü* olmak zorundadır. Çünkü, insanlar bu ölçüm araçlarının sonuçlarına bakarak *karar* vereceklerdir. Verilen kararların isabet derecesi, toplanan verilerin güvenilir ve geçerli olmasına bağlıdır. Güvenilirlik ve geçerliliği yüksek ölçüm verileri, insanları amaçlarına ulaştırır. Ölçüm verileri güvenilir ve geçerli değilse emek, zaman ve parasal maliyetlerle karşılaşmak kaçınılmazdır.

Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik adını verdiğimiz kitabımızda konular iki kısım ve 15 bölüm içinde ele alınmıştır. Güvenilirlik kısmı on dört ve geçerlilik ise bir bölüm halinde düzenlenmiştir. Önümüzdeki yıllarda geçerlilik bölümünü genişleterek kitabı iki ayrı cilt halinde yayımlamayı düşünmekteyiz. Güvenilirlik kısmının birinci bölümde güvenilirlik kavramına ilişkin tanımlar ve güvenilirlik kavramının kapsamı ele alınmıştır. İkinci bölümde ölçüm düzeyleri, ölçüm araçları ve derecelendirme güvenilirliği üzerinde durulmuştur. Üçüncü bölümde güvenilirlik analizi yöntemleri, güvenilirlik indeksi ve güvenilirlik katsayıları ayrıntılı bir analize tâbi tutulmuştur. Dördüncü bölümde girdi kalitesinin değerlendirilmesi için veri taraması konusuna değinilmiştir. Beşinci bölümde Cronbach alfa ve güvenilirlik analizleri; altıncı bölümde ise güvenilirlik ve korelasyon analizleri konuları irdelenmiştir. Yedinci bölümde varyans analizi ve güvenilirlik konusu işlenmiş, sekizinci bölümde, bu kez faktör analizinin güvenilirlikle olan bağıntısı üzerinde durulmuştur. Dokuzuncu bölümde ölçümün standart hatası ve onuncu bölümde ise kriter referanslı testlerin güvenilirliği konusuna değinilmiştir. On birinci bölümde niteliksel araştırmalarda güvenilirlik konusu incelenmiş, on ikinci bölümde özel test uygulamalarında güvenilirlik analizlerinin yapılma biçimleri üzerinde durulmuştur.

On üçüncü bölümde güvenilirliği iyileştirme çalışmaları konusu ele alınmış ve on dördüncü bölümde güvenilirlik analizi yazılımları tanıtılmıştır. Kapsamlı bir bölüm olarak düzenlenen on beşinci bölümde ise sadece geçerlilik konusu incelenmiştir. Geçerlilik konusu da en az güvenilirlik konusu kadar kapsamlı ve geniştir. Ancak bu aşamada kitabın yayımlanmasının daha fazla gecikmemesi için geçerlilik konusu kapsamlı tek bir bölüm halinde verilmiştir.

Güvenilirlik ve geçerlilik konuları; araştırma yöntem bilimi, istatistik, matematik ve psikometri ile ortak bir ara kesitte yer alır. Konular ele alınırken bazen ayrıntılara girilmiş, bazen de okuyucuların ilgili kaynaklara başvurmaları önerilmiştir. Her bilim dalının, diğer bilim dallarıyla ortak bir kesişim alanı olmasına karşılık, daha fazla ileriye gidilmeyen bir sınır çizgisi de vardır. Durulan bu sınır çizgisi görecelidir. Bilim adamının ilgisi, kazanımları, okuyucuyu götürmek istediği yer ve pratik hayattaki ihtiyaçlar ortak kesişim alanının sınırlarını daraltabilir veya genişletebilir. Okuyucuların beklentileriyle yazarın amacı her zaman tam olarak örtüşmez. Bu nedenle her kitap eksiktir. Bu kitapta, güvenilirlik ve geçerlilik konusu çoklu disipliner anlayışla ele alınırken lisans üstü düzeyinde eğitime sahip kişilerin ihtiyaçları göz önünde bulundurulmuştur.

GÜVENİLİRLİK

Ölçme bir nesneye, olguya, tutuma ait *özelliği* sayısallaştırmak veya sayılabilir simgelerle göstermektir. Ölçümde nesnenin, olgunun veya belirli bir tutumun boyutları, miktarı, derecesi, sayısı veya oranı rakamlarla gösterilir. Sonuçta bu rakamlara bakarak yorum yapar veya karar veririz. Bilimsel ölçümler, belirli bir olguyla ilgili olarak ya *kriter, norm, ipsatif değerleri* oluşturmaya veya önceden belirlenmiş olan *norm* ve *kriter değerlerle* karşılaştırma yapmaya yöneliktir. Ölçümlerde, eğer önceden saptanmış bir *kriter değer* veya *norm*[a] *değeri* varsa güvenilirliği hesaplamak nispeten kolaydır. Bu tür kriter değerlerin bulunmadığı hallerde ise ölçüm verilerinin güvenilirlik sorunu gündeme gelir. Güvenilirlik, herhangi bir ölçüm işleminin *sine qua non*'ıdır.

Ölçüm işlemlerinde ve bilimsel araştırmalarda güvenilirlik konusuna değişik açılardan yaklaşmak mümkündür. *Bilimsel araştırmanın güvenilirliği* daha geniş bir kavramdır. Ölçüm olgusunun dışında seçilen metodolojinin, modelin ve örnekleme yönteminin doğru ve uygun olması anlamına gelir. Bilimsel araştırmalardaki "ölçme" işlemi ise metodolojinin sınırları önceden belirlenmiş dar bir alanını oluşturur. Bu kitapta esas olarak *ölçümün güvenilirliği* konusu üzerinde durduk. Günümüzde ölçümün güvenilirliği konusuna farklı iki bakış açısıyla bakmak gerekmektedir. Bunlardan birincisi klasik ölçüm kuramının bakış açısı ve ikincisi ise, modern ölçüm kuramının bakış açısıdır. Klasik ölçüm kuramında *güvenilirlik* kav-

---

[a] Norm değerleri. Standartlaştırılmış başarı testlerinde bir kişinin elde ettiği puanının yüksek veya düşük olup olmadığını belirlemek için kullanılan, o kişiyi temsil etme kabiliyetine sahip, en yakın ana kütle veya alt örnek kütle grubundan elde edilmiş verilerin değişik formüllere göre hesaplanmış olan karşılaştırma ölçütleridir. Norm değerleri norm gruplarına göre oluşturulur. Norm grupları; nüfus sayımı bilgileri çerçevesinde yaş, cinsiyet, coğrafi bölge, meslek, sosyo-ekonomik durum, etnik köken, kurum, sınıf veya grup faktörleri dikkate alınarak belirlenir. Norm değerleri, standart olmayan ve standart olmak üzere iki grupta ele alınır. Standart olmayan normlar; aritmetik ortalama, standart sapma ve yüzdelik değerleridir. Standart norm değerlerinin sık kullanılanları ise; *z* puanları, *T* puanları, standart dokuz puanları, standart dokuz yüzdelik dilimi puanları ve standart on puanlarıdır.

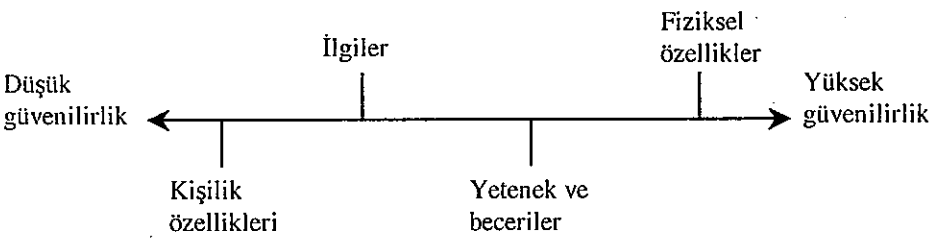
ramıyla bir bakışta dört farklı şey anlatılmak istenir: (a) bir ölçüm aracındaki maddelerin aynı kavramsal *yapıyı* hatasız bir biçimde ölçmesi, (b) farklı zamanlarda yapılan ölçüm sonuçlarının aynı çıkması, (c) bir ölçüm aracına ait sonuçların aynı kavramsal yapıyı ölçen diğer ölçüm araçlarının sonuçlarıyla tutarlı olması, (ç) farklı gözlemciler tarafından yapılan ölçüm/değerlendirme sonuçlarının benzer çıkması. Buna göre klasik ölçüm kuramında güvenilirlik tek bir cümle ile, "test veya ölçek sonuçlarının kavramsal yapıya ilişkin olguyu doğru bir şekilde ortaya çıkarması; ölçüm aracı farklı yerlerde, farklı zamanlarda ve aynı ana kütleden seçilen farklı örnek kütlelerde uygulandığında benzer sonuçlar vermesi" olarak tanımlanabilir.

Modern ölçüm kuramında ise güvenilirlik, örneklemden bağımsız olarak maddeye verilen "yanıtın fonksiyonu"dur. Diğer bir deyişle, daha sonraki bölümlerde ayrıntılı olarak ele alacağımız, maddenin *hedeflenen bilgi fonksiyonunu* gerçekleştirme derecesine bağlıdır. Bir maddenin zorluk derecesi için "ideal" veya "doğru" bir orandan söz edilemez. Bu oran, testi alan gruptaki kişilerin ortalama yetenek düzeylerine göre belirlenir. Herhangi bir ölçüm maddesinin bir test içinde yer alması; güçlük, ayırt edecilik gibi belirli parametrelerin tahmin edilmesine veya maddenin kalibre edilmesine bağlıdır. Bir testte sadece kalibre edilmiş olanlar güvenilir maddelerdir. Modern ölçüm kuramında bir test kalibre edilmiş maddelerden oluşturulmuşsa güvenilir bir ölçüm aracıdır.

Güvenilirlik kavramının tanımı ve sonuçta yapılacak güvenilirlik analizleri ölçüm modelinin dışında, ölçme işleminin türüne, ölçülen olguya veya ölçüm aracına bağımlıdır. Uygulanan ölçüm aracının bir ölçek, bir indeks, bir zihinsel yetenek testi, çoktan seçmeli bilgi testi olmasına veya ölçüm işleminin gözlemci değerlendirmelerine dayanmasına göre yapılacak güvenilirlik tanımı ve güvenilirlik analizleri de değişir. Araştırmanın niteliksel veya niceliksel içerikli olması seçilecek güvenilirlik analizi yöntemini etkiler. Ölçümün soyut veya soyut konularla ilgili olması da kesinlik derecesi üzerinde etkilidir.

İnsanlara ait psikolojik özelliklerin ölçüm sonuçlarının güvenilirlik tahminleri her zaman birbirine eşit çıkmaz. Fiziksel veya fizyolojik özelliklere ait ölçüm sonuçları ne denli güçlü ise, bireylerin kişilik ve psikolojik içsel durumlarını saptamaya yönelik ölçüm sonuçları güvenilirlik açısından o ölçüde zayıftır (*bk.*, Şekil 1-1). Bununla birlikte bilim adamları bu tür ölçümlerin güvenilirlik kanıtlarını araştırmaktan vazgeçemezler. Yıllar içinde güvenilirlik analizlerini tekrarlayarak meta analizi yöntemiyle sağ-

lıklı bir güvenilirlik değerine ulaşmaya çalışırlar. Güvenilirlik analizleri, bir kere uygulanıp biten bir işlem değil, devam eden bir süreçtir.



Şekil 1-1. Beşerî özelliklere ait ölçümlerin güvenilirlik düzeyleri.

"Giriş" bölümünde de belirtildiği gibi, güvenilirlikle geçerlilik arasında yakın bir ilişki vardır. Güvenilirlik, aynı zamanda yapısal geçerliliğin bir yönünü oluşturur. Bilim adamları güvenilirliği geçerliliğin alt sınır değeri olarak isimlendirmişlerdir. Güvenilir bir testin[a] sonuçları aynı zamanda geçerli olabilir, fakat kesin bir şekilde *geçerlidir* yargısını ileri süremeyiz. Alt sınırın sağlanmış olması üst sınırın da sağlandığı anlamına gelmez. Öte yandan yüksek güvenilirlik katsayısı, elde edilen bilimsel sonuçların çok iyi olduğunun garantisi olarak görülmemelidir. Güvenilirlik ön koşuldur, "...bilimsel sonuçlar yüksek güvenilirlik olmadan asla güven vermez."[1] Güvenilirliğin önemi, "APA İstatistiksel Sonuç Çıkarma Çalışma Komitesi" üyeleri tarafından da vurgulanmış ve komite bilim adamlarına araştır-

[a] Literatürde *test* kavramı değişik şekillerde tanımlanmıştır. Bir tanıma göre test, belirli bir kavramsal alandan örnek olarak seçilen davranışların standart bir prosedüre göre ölçülmesi ve söz konusu davranışlara puan verilmesidir. Geniş kapsamlı bu tanımda kontrol listeleri, ölçekler, gözlemler, bilgi ve yetenek ölçümleri test kapsamında değerlendirilmiştir. Test sonuçları sınıflandırma, seçme, yerleştirme, teşhis koyma, tedavi biçimini planlama, program değerlendirme, bilgilenme ve araştırma yapma amacıyla kullanılabilir. Türkçede test sözcüğü daha çok başarı veya başarısızlığı belirlemeye yönelik olarak geliştirilen ölçüm araçları için kullanılmıştır. Durum tespitine, özellikleri saptamaya, bilgilenmeye ve sınıflandırma yapmaya yönelik ölçüm araçları için ise *ölçek* sözcüğü tercih edilmiştir. Aslında ölçeklerden de bir kişinin başarılı olup olmayacağı hakkında bilgi edinilebilir. *Test* ve *ölçek* kavramlarıyla yakından ilgili üçüncü kavram *ölçüm* sözcüğüdür. Ölçüm kavramı değişik şekilde elde edilen çok sayıda maddenin bileşik değerini yansıtır. Bu kitapta sık aralıklarla "ölçek/test/ölçüm" şeklindeki ifadelendirme biçimi tercih edilmiştir. Okuyucu bu tür ifadeleri, ölçüm aracının veya ölçme işleminin niteliğine göre "ölçek, test veya ölçüm uygulaması" şeklinde düşünebilir.

manın odak noktası psikometrik bir çalışma olmasa dahi analizi yapılan verilerin güvenilirlik katsayılarının raporlanmasını önermiştir.[2] Güvenilirlik tek başına yeterli değildir, ayrıca geçerlilik analizleri de yapılmalıdır. Bir ölçüm aracının içerdiği maddelerin kavramsal yapıyı ne ölçüde temsil ettiği belirlenirken güvenilirlik ile geçerliliğin özellikle iç içe girmiş olduğu görülür. Güvenilirlikle geçerlilik, sadece belirli standartlarla ayrılabilir. *İyi–kötü*, *geçti–kaldı*, *başarılı–başarısız* gibi değerlendirmeler, yargılar eğer önceden belirlenmiş bir kriter veya standarda göre yapılmıyorsa geçerlilikle güvenilirlik arasındaki sınır belirsizdir (Moss, 1994).[3] Örneğin tutum ve görüşlerin derlendiği ölçekler bu gruba girer. Bu tür ölçümlerde *iç tutarlılık* açısından güvenilirlikle geçerlilik arasındaki sınırları kesin çizgilerle ayırmak zordur. Bu nedenle bazı araştırmacılar *güvenilirlik-geçerlilik* ifadesini birleşik tek bir kelime gibi kullanırlar. Cronbach, güvenilirlikle geçerlilik kelimelerini bütünleştirerek bu iki kavramı tek bir kelime olarak "genellenebilirlik" kavramıyla ifade etmeye çalışmıştır. Güvenilirlikle geçerlilik kelimelerini birlikte tanımlayan bir diğer kavram İngilizcedeki *relevancy* terimidir. Relevancy, *güvenilirlik ve geçerliliğin çarpımı* olarak tanımlanabilir. Dilimize *ilgililik* şeklinde çevrilebilecek bu kelime Türkçede istenen vurguyu ve anlam bütünlüğünü tam olarak sağlamamaktadır.

Güvenilirlik ve geçerlilik kavramları birlikte kullanıldığında aynı zamanda *standardizasyon* anlamına gelir. Bir test, indeks veya ölçeğin değişik ana kütlelerde geçerlilik ve güvenilirlik analizlerinin yapılmasıyla ve uygun güvenilirlik katsayılarına ulaşılmasıyla o test veya ölçek aynı zamanda standartlaştırılmış olur. Standardizasyon, "belirli bir süre" için ve "belirli ana kütlelerde veya örneklemlerde" test veya ölçeğin güvenle uygulanabilecek hale getirilmesi işlemidir.

Görüldüğü gibi, *genellenebilirlik*, *ilgililik* ve *standardizasyon* terimleriyle güvenilirlik ve geçerlilik kavramları tekleştirilmeye çalışılmıştır. Okuyucu, güvenilirlik kavramını sadece bağımsız bir yapı olarak değil, belirli tür ölçümlerde geçerlilikle bağlantılı geniş içerikli bir kavram olarak değerlendirmelidir.

TANIMI, KAPSAMI

Bu bölümde değişik güvenilirlik tanımları ve yaklaşımları üzerinde durularak, kavramın farklı yönlerine değinilmiştir. Bilim adamlarının tanımları temel aldıkları modele, ölçüm aracının niteliğine, kullanım amacına ve elde edilen sonuçların genelleme hedefine göre değişebilmektedir.

TANIM VE YAKLAŞIMLAR

Bilim Dalları ve Güvenilirlik

Güvenilirlik terimi, sosyal ve davranış bilimlerine fizik bilimlerinden geçmiştir. Fizik bilimlerinde güvenilirlik, bir makinenin ve o makineyi kullanan gözlemcinin ortak özelliğidir. Tekrarlanan gözlemlerle nesnenin / makinenin özelliği, gerçek durumu hakkında belli bir yargıya varılır. Operatör tarafından okunan göstergelerdeki varyans değişkenliğinin düşük olması makine-operatör güvenilirliğini ortaya koyar.[4] Bu çerçevede bir endüstri mühendisine göre güvenilirlik, üretimde hata oranının veya başarısızlık oranın düşük çıkmasıdır. Tanımda ürünün kalitesi ve değişmezlik ön planda tutulmuştur. Yirminci yüzyılın ilk yarısındaki psikometrisyenler güvenilirlik kavramını mümkün olduğu kadar fiziksel bilimlere yaklaştırarak ele almışlardır. Bu nedenle sosyal bilimcilerin güvenilirlik tanımlarında fizik bilimlerinin izleri görülür. Örneğin, bir endüstri psikologu için güvenilirlik ölçülen tutumların, davranışların veya kişiliğin arka planında yatan gizli kavramsal yapıların her defasında başarılı bir şekilde ortaya çıkarılma derecesidir. Bir sosyologa göre güvenilirlik, ölçüm sonuçlarının farklı ana kütlelerde veya aynı ana kütleye ait farklı örnek kütlelerde benzer sonuçlar vermesidir. Sosyologlar belirli sonuçlara ulaşmak için gizli faktörleri ortaya çıkarmak yerine, vak'aları karşılaştırmak için tipolojiler[a] ve hiyerarşik düzenlemelere dayanan taksonomiler[b] oluşturmaya çalışırlar. Bir *insan*

[a] Tipoloji (typologies). İkili karşılaştırma modelleri.

[b] Taksonomi (taxonomies). Kavram ve yapıların hiyerarşik bir düzen içinde sıralanarak sınıflandırılmasıdır.

kaynakları yöneticisine göre güvenilirlik, psikoteknik/psikometrik[a] test sonuçlarının uygulandığı farklı zaman dilimlerinde benzer sonuçlar vermesi ve iş yaşamında personelin göstereceği performansı doğru tahmin etmesidir. Görüldüğü gibi, bilim adamlarının ve uygulayıcıların güvenilirlik kavramına getirdikleri tanımlar kendi disiplinlerinin yaklaşımlarını ortaya koyar. Ancak bunu yaparken, bilim adamları fiziksel bilimlerdeki tanımın özünü oluşturan "gerçek durumu" saptama düşüncesinden hiçbir zaman vazgeçmemişlerdir.

Testin Güvenilirliğine Karşı Verilerin Güvenilirliği

Güvenilirlik, belirli bir amaçla kullanılan test veya ölçeğin sözel veya şekilsel içeriğine ilişkin değil, *verilerine* aittir. Pek çok tanımı veya yorumu olsa da güvenilirlik sadece test edilen gruptan toplanan *verilerle* ilgilidir. Bu konuda pek çok bilim adamı güvenilirliğin verilere ait olduğunu bildirmişlerdir (Crocker ve Algina, 1986; Gronlund ve Linn, 1990; Meier ve Davis, 1990; Pedhazur ve Schmelkin, 1991; Thompson, 1994; Vacha-Haase, 1998; aktaran Ackerman).[5]

Buna göre verilerin güvenilirliği, aynı ana kütleden seçilecek başka örneklemlerde aynı yöntemle, aynı prosedür uygulanarak yapılacak başka ölçümlerde benzer sonuçların elde edilme olasılığıdır. Daha sonraki uygulamalarda belirlenen koşullardan sapmalar varsa, elde edilen yeni veriler için eski güvenilirlik katsayısı ile ana kütleye genelleme yapılamaz. Test ve ölçüm araçlarının bizatihi kendileri güvenilir olmadıklarından her ölçüm uygulamasından sonra güvenilirlik analizleri yeniden yapılmalıdır.

Yüksek lisans ve doktora tezlerinde güvenilirlik analizleri, sadece pilot araştırma sırasında elde edilen ölçüm verilerine dayandırılmamalıdır. Pilot araştırma verilerinde yapılan güvenilirlik analizleri bir ön yordama niteliğindedir. Tezlerdeki güvenilirlik analizleri *esas araştırma* sonuçlarına dayalı olarak yapılır ve elde edilen bilgiler "Bulgular" başlığı altında sunulur. Ret veya kabul edilen hipotezlerin doğruluğu, *esas araştırma* sonuçlarının güvenilir olmasına bağlıdır.

---

[a] Ülkemizde psikometrik ve psikoteknik kelimelerinin her ikisi de aynı anlamda kullanılmaktadır. Psikoteknik kelimesi Fransızcadan psikometrik kelimesi ise İngilizceden dilimize geçmiştir. Bu kitapta tek bir kavram tercih edilmemiş bazen psikoteknik ve bazen de psikometrik kelimeleri kullanılmıştır. Bilimde kavram birliği önemlidir. Ancak dilimize geçen yabancı kavramların Türkçe karşılığı olarak değişik bilim adamlarının farklı terimleri kullanmaları sonucunda bazı olgular sanki farklı imiş gibi algılanabilmektedir. Yanlış algılamayı azaltmak ve okuyucunun her iki kavrama da aşina olmasını sağlamak için böyle bir yönteme başvurulmuştur.

Gerçek Puanın Ortaya Çıkarılması

Güvenilirlik, "gerçeğin ortaya çıkarıldığı" iddialarına kanıt bulma çabasıdır. Bu anlamda güvenilirlik, McDonald'ın tanımıyla *gerçek puan varyansının gözlem puanları varyansına* olan oranıdır (aktaran Fichman, 2003).[6] Diğer bir deyişle, üzerinde ölçüm yapılan kişilerin bilgi, yetenek, beceri ve tutumlarının gerçeğe yakın ölçülme derecesiyle ilgili bir tahmin değeridir. Bir bilim adamının ölçüm verilerinin ,85 güvenilirlik katsayısına sahip olduğunu söylemesi; verilerdeki değişkenliğin %85 oranında *gerçek varyansı*, %15 oranında ise *hata varyansını* yansıttığı anlamına gelir. Bu katsayıyı, gözlem puanlarıyla hipotetik gerçek puanlar arasındaki ilişki olarak da tanımlayabiliriz.

Teknik Yöntem Olarak Güvenilirlik

Literatürdeki güvenilirlik tanımlarında daha çok teknik hesaplama yöntemleri üzerinde odaklanılmıştır. Tutum ölçeklerinin gelişmesine önemli ölçüde katkı sağlamış olan R. Likert (1932) güvenilirliği teknik açıdan ele alarak *bir ölçeğin iki yarısı arasındaki korelasyonla* ve ayrıca *toplam puanla madde puanları arasındaki korelasyon* ortalamalarının yüksek çıkması olgusuyla açıklamıştır.[7] Konuya teknik açıdan yaklaşan bir başka bilim adamı olan L.L. Thurstone (1928), güvenilirliği tutarlılıkla ve tutarlılığı da "ilişkisizlik" (irrelevance) terimiyle eş değerde görmüştür. Ona göre ölçek maddeleri öyle bir şekilde yapılandırılmalı ki, belirli bir maddeyi sadece o görüşün taraftarları, yandaşları değil, tam aksi istikamette görüşlere sahip olanlar bile işaretleyebilmelidirler. Eğer ölçek maddelerinde bu koşul sağlanmışsa o ölçek iç tutarlılığa sahiptir ve güvenilir olarak değerlendirilir.[8] Bu koşulu sağlayamayan maddelerin toplam puanla olan korelasyonları düşüktür ve bu maddeler "ilişkisiz" veya "belirsiz" olarak değerlendirilir. Güvenilirliği artırmak için kişileri "farklılaştırmayan" bu maddelerin ölçekten çıkarılması gerekir. L. Guttman (1944) ise, teknik hesaplamaya

dayalı olarak güvenilirliği "üretilebilirlik" veya "yeniden tekrarlanabilirlik" kavramı ile açıklamıştır.[a]

Sonuçların Genellenebilirliği

Klasik test kuramına göre güvenilirlik, örneklem verilerinin daha geniş olan ve daha fazla eleman içeren ana kütleye genellenebilmesini gerektirir. L.J. Cronbach'a (1970) göre güvenilirlik "genellenebilirlik"tir. Ölçüm veya gözlemle elde ettiğimiz puanlar evrendeki puanların dağılımıyla yakından ilgili ise, bu puanların *doğru* ve *güvenilir* olduğunu söyleriz. Gözlem puanları aynı zamanda birbiriyle uyuşuyorsa *tutarlı* olduklarından ve *çok az hata varyansı* içerdiğinden söz ederiz. Literatürde çoğunlukla "güvenilirlik" kavramı kullanılmakla birlikte Cronbach bu olguyu *genellenebilirlik* kavramıyla tanımlamış ve *genellenebilirlik* katsayısının her bir ana kütle için ayrıca hesaplanması gerektiğini belirtmiştir.[9]

Sonuçların Tutarlılığı

Psikometri[b] konusunda uzman olan bir başka bilim adımı J.P. Peter'e (1979) göre güvenilirlik, "bir ölçümün hatalardan arındırılmış olması ve tekrar yapılan ölçümlerde tutarlı, istikrarlı sonuçlar vermesidir."[10] Bu ta-

---

[a] Guttman ölçekleri, tek boyutlu ölçüm aracı grubunda değerlendirilir. Bu yöntemde alt sırada yer alan ve önceki tüm maddeleri kapsayan son maddeye *Evet* cevabını veren yanıtlayıcı ondan önceki bütün maddeleri kabul etmiş veya onlara olumlu cevap vermiş sayılır. İlk madde daha uzak bir tutumu, son madde ise daha yakın ve olumlu bir tutumu gösterir. Guttman ölçeklerinde üretilebilirlik katsayısı ,90 olmalıdır. Üretilebilirlik katsayısı, 1 – (Hata sayısı / Toplam cevap sayısı) formülüne göre belirlenir. Bu yöntemde *mantıklı ve tutarlı cevaplar* kabul edilebilir yanıtlar olarak görülür. Mantıksal açıdan tutarsız olan cevaplar ise hata olarak değerlendirilir. Örneğin 10 kişi 3 maddeden oluşan bir Guttman ölçeğini işaretlemişlerse toplam 30 işaretleme yapılmış olur. Kişilerin puanlarının belirlenmesinde şöyle bir yöntem uygulanır. Diyelim ki elimizde üç maddeden oluşan bir ölçek var. Katılımcılar bu ölçeği +,–,– ; +,+,–; +,+,+ ve –,–,– şeklinde cevaplandırmışlarsa bu cevaplar mantıklı cevaplardır. Uygulamada +,–,– işaretlemesine 1; +,+,– işaretlemesine 2; +,+,+ işaretlemesine 3 puan ve –,–,– işaretlemesine ise 0 puan verilir. İşaretlemede -+-, +--+, --+, ve -++ şeklindeki yanıtlar hataları gösterir. Üretilebilirliğin sağlanması için pilot araştırma sırasında tespit edilen hatalı maddelerin veya ifadelerin ölçekten çıkarılması gerekir. Son uygulamada yine hatalı işaretlemeler yapılmışsa bu maddelere en üstte yer alan + işareti dikkate alınarak puan verilir. Örneğin -+- = 2, +--+ = 3, --+ = 3 ve -++ = 3 şeklinde puanlanır. Ancak --+ işaretlemesinin 3 olarak puanlanması yanlış olabilir. Ölçeği dolduran kişi bu işaretlemede muhtemelen hata yapmıştır. Bu işaretleme biçiminin 0 puanıyla kodlanması daha doğru olur. Bu konuda *bk.*, "Indices and Scales [İndeks ve Ölçekler]," <http://www-sociology.sbs.umass.edu/ courses/ soc210/ Indices_ and_Scales.htm.> (26.09.2002).

[b] Psikometri kavramı, Ivo Molenaar tarafından "psikoloji biliminin gelişmesine hizmet etmeyi amaçlayan *matematiksel istatistik* biçiminde tanımlanmıştır.

nımdaki "tutarlı sonuçlar" vermesi olgusunu nasıl anlamak gerekir? Örneğin, bir ölçek veya test farklı düzlemlerde ve farklı örneklemlerde uygulandığında farklı sonuçlar elde edilebilir. Öyle olunca da sonuçlar tutarsız gözükecektir. Çünkü söz konusu testin veya ölçeğin birden fazla güvenilirlik rakamları ortaya çıkacaktır. Ancak buradaki tutarlılık, aynı yerel düzlemdeki uygulamalar ve benzer örnek kütle verileri için söz konusudur. Bütün güvenilirlik çalışmaları, ölçümlerin yinelenmesini gerektirir. Aynı ana kütleye ait değişik örneklemlerde yapılan yinelenmiş ölçümlerde sonuçların birebir aynı çıkması gerekmez. Sonuçlar karşılaştırılabilir veya birbiriyle telif edilebilir olmalıdır. Karşılaştırılabilir sonuçlara dayalı olarak yapılacak hipotez testlerinde farklılığın istatistiksel olarak anlamlı olmadığı görülür. İstatistiksel olarak anlamlı çıkmaması tutarlılıktır. Tutarlılık veya istikrarlılık, "bir gruba ait test/ölçüm sonuçlarının aynı gruba farklı zamanlarda uygulandığında benzer çıkması ve bu nedenle testin kişilere güvenle uygulanabilmesidir (Berkowitz, Wolkowitz, Fitch, ve Kopriva, 2000, aktaran Rudner)."[11]

Sonuçların Kesin ve Tam Doğru Olması

Güvenilirlik, aynı zamanda *kesinlik derecesi* olarak ele alınmıştır. Kesinlik, ölçüm verilerinin doğru olması veya doğru çıkmasıdır. Bilim adamı, kesinliğe ulaşmaya çalışır. *Tahmin edilebilirlik, anlaşılırlık* ve *mantıklı olma* kesinlikle ilgilidir. Amerika Birleşik Devletleri'nde Amerikan Psikoloji Derneği (American Psychological Association – APA), Amerikan Eğitim Araştırmaları Derneği (American Educational Research Association – AERA) ve Eğitimde Ölçme ve Değerlendirme Ulusal Konseyi (National Council on Measurement in Education – NCME) gibi kurumların koordinasyonu altında geliştirilen *Eğitsel ve Psikolojik Test 'Standartları'*nın 1985 sürümünde[a] güvenilirlik yaklaşık olarak şu şekilde tanımlanmıştır:[12] Güvenilirlik, bir ölçüme ait genel puanların ve alt boyutlara ait toplam puanların doğru, sağlam ve güçlü olduğunu belirlemeye yönelik bir tahmin değeridir. Yorum yapılırken güvenilirlik değeri ile birlikte *ölçümün standart hatası* da ayrıntılı bir biçimde verilmelidir ki test kullanıcısı elde edilen puanların kullanım amacına uygun olup olmadığı konusunda belirli bir yargıya varabilsin. Eğer, ölçümün standart hatası yüksekse o değişkene ait veriler daha az güvenilirdir. Standardın 1999 sürümünde ise güvenilirliğin tanımı için *tekrarlanan ölçümler arasındaki tutarlılığa* atıf yapılmıştır.[13]

---

[a] *Standartlar* kitabı ilk kez 1966 yılında yayımlanmıştır. Daha sonra 1974, 1985 ve 1999 yıllarında gözden geçirilerek eserin yeni baskıları yapılmıştır.

Test Standartları tanımında genel ve alt boyutlara ait toplam puanların *doğru*, *sağlam* ve *güçlü* olması gibi temalar üzerinde durulmuştur. Öte yandan ölçümün standart hatası (ÖSH) ile ilgili değerlerin testi alan kişilere bildirilip bildirilmeyeceği konusuna değinilmemiştir.[a] *Eğitsel ve Psikolojik Test Standartları*'nda son yıllara kadar bu terimin anlamı üzerinde önemli bir değişiklik yapılmamış ve güvenilirlik, geçerliliğin alt sınır değeri olarak görülmüştür. Bunun anlamı güvenilirliğin geçerliliğe ilişkin belli ölçüde bir fikir verdiği fakat geçerliliği tam karşılamadığıdır. P. Moss (1994), geçerlilik olmaksızın güvenilirlikten söz etmeyi anlamsız bir kavram olarak nitelendirmiştir (aktaran Syverson ve Barr).[14]

İstatistiksel Güvenilirlik

Güvenilirlik kelimesinin bir diğer anlamı, elde edilen sonuçların istatistiksel olarak anlamlı olmasıdır. Bilim adamı araştırma tasarımına göre, bazen istatistiksel olarak anlamlı bir farklılık çıkmasını beklerken bazen de sonuçların istatistiksel olarak anlamlı çıkmamasını isteyebilir. *İstatistiksel güvenilirlik*, sonuçların şansa bağlı olmadığına işaret eder. Çıkan sonuçlar tesadüf eseri değildir. Olasılık (anlamlılık) değeri $p < ,05$ ise sonuçlar arasında istatistiksel olarak anlamlı bir farklılık olduğuna; $p > ,05$ ise sonuçlar arasında istatistiksel olarak anlamlı bir farklılık bulunmadığına hükmedilir. İkincisinde, ölçüm tekrar yapılsa bile %95 ihtimalle aynı sonuçlar elde edilecek demektir. İstatistiksel olarak anlamlı çıkma, "istatistiksel olarak güvenilir" bulunma anlamına gelir.[15]

İstatistiksel güvenilirlik, aynı zamanda belirli bir anlamlılık düzeyinde ölçüm ortalaması veya oranını gösteren rakama ait "güven aralığı değerlerinin" tahmin edilmesidir. Örneğin, yapılan bir ölçüm sonucunda çalışanların %67'sinin stres düzeyleri yüksek çıkmıştır. Ancak istatistiksel güvenilirlik açısından, "gerçek değer" %55 ilâ %83 arasında değişebilir. Güven aralığı değerleri başlıca iki faktörden etkilenir. Birincisi ölçülmek istenen özelliğin ana kütlede bulunma oranına ilişkin tahmin değerinin ne olduğudur. Tahmin değeri %50 gibi bir rakama geldiği ölçüde güven aralığı genişler. İkincisi ise örneklem hacmidir. Örneklem hacmi büyüdüğü ölçüde

---

[a] ABD'de Buckley Yasası'na göre (Buckley Amendment) kişilerin kendileri, anne ve babaları veya vâsileri uygulanan akademik standart testlerin sonuçlarını görme veya bu konuda bilgi alma hakkına sahiptirler. APA standartlarına göre de test sonuçları hakkında bilgi almak isteyen kişilere gerekli bilgiler verilir. Sadece verilen bilgilerin müşteri tarafından yanlış kullanılması, yorumlanması veya kişinin kendisine zarar vermesi ihtimalinin bulunduğu durumlarda bu bilgilerin verilmesinden kaçınılabilir.

güven aralığı daralır. Bu nedenle istatistiksel güvenilirlik n, p, q değerleri ve α anlamlılık düzeyiyle yakından ilgilidir.

Hatalardan Arındırma Anlamında Güvenilirlik

Güvenilirlik, ölçümün ne denli tesadüfî hatalardan arındırılmış olduğuna bağlıdır. Test veya ölçek verilerinin birbirinden bağımsız olan tesadüfî hatalardan arındırılmış olması güvenilirliği artırır. Tesadüfî hata en alt düzeyde ise teste/ölçeğe ait maddelerin puanları veya farklı zamanlarda uygulanan test toplam puanları birbiriyle daha fazla tutarlı olur. Ölçümde maddelere ait hata varyanslarının en aza indirilmesi *gerçek değeri* ortaya çıkarır.

Kalite Anlamında Güvenilirlik

İş hayatına yönelik olarak değişik amaçlı test ve ölçekleri üreten ticarî piyasada güvenilirlik "kalite" anlamında kullanılır. Test ve ölçek geliştiren merkezlerin, şirketlerin yetkilileri *güvenilirlik* kavramını politik bir anlayışla, ölçüm aracının "kaliteli" olduğunu belirtmek için kullanırlar. Bu anlayışta güvenilir bir test veya ölçek, müşteri beklentilerine uygun, arzulanan sonuçları elde etme kapasitesi yüksek ölçüm aracıdır. Kalite anlamındaki güvenilirlik kavramında, pratik yarar ve kullanışlılık ön plana çıkar.

Güvenilirlikle Geçerlilik Arasındaki İlişkiler

Güvenilirlik, bir yönüyle yapısal geçerlilikle ilgilidir. Bir test veya ölçekte güvenilirliği belirlemek üzere maddelerin iç tutarlılığından söz edildiğinde aynı zamanda "yapısal geçerlilik" konusu gündeme gelir. Ancak bu anlamdaki güvenilirlik, tek başına geçerliliği belirlemede yeterli olmaz. Dikkat edilirse bilimsel tezlerde ve makalelerde geçerlilikten çok güvenilirlik rakamları hesaplanmış ve verilmiştir. Bunun nedeni güvenilirlik katsayılarının doğrudan verilere bağlı olarak incelenmesi ve iç tutarlılığın nispeten daha kolay hesaplanmasıdır. Ancak bu bilgi ve bulgular ölçümün geçerliliği için yeterli değildir. Ölçümün güvenilir olması ön koşuldur, fakat yeterli değildir. Daha önce de belirtildiği gibi herhangi bir ölçümde güvenilirlik ve geçerlilik birlikte aranır.

Geçerliliğin çok yönlü olarak, (kavramsal tanımlamalar ve istatistiksel analizlerle) incelenmesinin tersine güvenilirlik esas olarak istatistiksel analizlere dayanır. Geçerlilikte mantık yürütmeyle de bazı sonuçlara ulaşılabi-

lirken güvenilirlikte korelasyon analizi, yapısal eşitlik modeli[a], varyans analizi, faktör analizi gibi istatistiksel tekniklerle, bu amaçla geliştirilmiş bulunan KR-20, alfa indeks değeri, Spearman-Brown gibi matematiksel formüller kullanılır. Belirli bir noktaya kadar güvenilirlik ve geçerlilik birlikte artış gösterir, fakat güvenilirliğin belirli bir noktasından sonra (yaklaşık olarak ,96'nın üzerinde) geçerlilik düşmeye başlar. Çünkü böyle bir durumda test maddeleri birbirini tekrarlayan bir özelliğe sahip olur.

TARİHSEL GELİŞİMİ

Güvenilirlik analizlerinin tarihi onar yıllık dönemler halinde ele alınıp incelenirse, söz konusu dönemlerin aynı zamanda test ve ölçek geliştirme çalışmalarının tarihi olduğu görülür. Bu bölümde, devrimsel keskin dönüşümleri ifade etmese de, inceleme kolaylığı sağlaması nedeniyle olgunun tarihsel gelişimine onar yıllık inceleme kesitleriyle yaklaşılmıştır.

1910-1920 Dönemi

Güvenilirlik konusundaki ilk çalışmalar Charles Spearman tarafından yapılmıştır. Spearman, *Correlation Calculated From Faulty Data* [Hatalı Verilere Dayalı Korelasyon Hesaplamaları] (1910) isimli makalesinde bir ölçeğin/testin güvenilirliğini tespit etmek için bir yarısının diğer yarısıyla korelasyonunun araştırılmasını veya bir testin paralel nitelikte başka bir formla korelasyonunun incelenmesini önermiştir. İlk güvenilirlik analizleri, daha çok ölçeğin iki yarısı arasında yapılan korelasyon analizlerine dayanıyordu.

1920-1930 Dönemi

C. Spearman 1920'li yıllarda arkadaşı W. Brown ile birlikte bir testin iki yarısı arasındaki korelasyonun ölçeğin/testin tamamını kapsaması için bugün *Spearman –Brown formülü* olarak bilinen hesaplama yönteminin geliştirilmesini sağlamıştır. Bu dönemde T.L. Kelley (1921, 1923) güvenilirliği, "bir testin ölçmek istediği şeyi ölçme derecesi" olarak tanımlamıştır ki, bu

---

[a] Yapısal eşitlik modeli. Gözlem değişkenleri ile gizli değişkenler arasındaki ilişkileri inceleyen istatistiksel çözümleme yöntemidir. *Yapısal model* gizli değişkenler arasındaki ilişkileri ele alırken *ölçüm modeli* gözlem değişkenleri ile gizli değişkenler arasındaki ilişkileri ele alır ve bu değişkenleri birbirine yaklaştırmaya çalışır. Yapısal eşitlik modeli aynı zamanda "çoklu özellik-çoklu yöntem analizinin" bir uzantısı niteliğindedir ve psikolojik ölçümlerde gerçek varyansı ölçüm hataları varyansından kurtarmaya çalışır. Bu konuda *bk.*, <http://www.uoregon.edu/~jwildes/SEM_Jen_Wildes.pdf> (23.10.2002)

gün bu yaklaşım güvenilirlik değil *geçerlilik* olarak bilinmektedir.[16] Kelley ayrıca, "güvenilirlik indeksi" adını verdiği gerçek puan ile gözlem puanları arasındaki korelasyona dayanan özel bir hesaplama yöntemi geliştirmiştir.

1930-1940 Dönemi

Tarihî süreç içinde 1930'lu yıllara gelindiğinde Spearman'ın önerdiği *çift faktör* kuramı psikolojik test bataryalarının yapısal özelliğini açıklamakta yetersiz kalmış ve psikologlar *testlerin çok boyutluluğu* üzerinde durmaya başlamışlardı. Bu yıllarda güvenilir bir test oluşturmak için madde analizi yönteminin yeterli olamayacağı anlaşılmıştır.[17] Louis Leon Thurstone, *Vectors of Mind* [Zihnin Boyutları] isimli kitabında (1935) yeteneklerin ancak *çoklu faktörle* açıklanabileceği düşüncesini geliştirmiştir. Ona göre çok sayıda faktörden oluşan bir uzayda, bu faktörlerin döndürülmesiyle birlikte tek bir faktör elde edilebilir ve kavramsal yapı *tek faktör* kuramı ile açıklanabilirdi. Harold Hotelling 1933'te Pearson'un çalışmalarından esinlenerek *temel bileşenler analizi*[a] yöntemini tanıtmıştır. Bu yıllarda psikometri araştırmaları iki yönde gelişme seyri göstermiştir. Bazı bilim adamları Spearman'ın "çift faktör" yaklaşımı üzerinde çalışmaya devam etmişler ve ikinci gruptaki bilim adamları ise daha modern olarak nitelendirdikleri Thurstone'nun "çoklu faktör" yaklaşımını benimsemişlerdir. İngiliz psikologu Godfrey Thomson (1939), bu konularda Spearman ile 20 yıla yakın bir süre bilimsel tartışmalar içinde bulunmuş, *faktör analizi* ile *temel bileşenler analizini* bir araya getirmeye yönelik çalışmalar yapmıştır.[18]

Güvenilirlik çalışmalarında önemli bir dönüm taşı, Kuder–Richardson 20 formülünün geliştirilmesidir. G.F. Kuder ve M.W. Richardson 1937 yılında yeni geliştirdikleri güvenilirlik analizi yöntemlerini tanıtmışlardır. Bu bilim adamlarının amaçları, *yarıya bölme* ve *Spearman–Brown* yönteminde karşılaşılan güçlükleri yenmekti. Geliştirdikleri pek çok formülden kendi adlarıyla anılan *Kuder-Richardson 20* formülü literatürde daha çok kabul görmüştür. Kuder ve Richardson bu formülde maddelerin korelasyon katsayılarının ve maddelerin standart sapmalarının eşit olduğunu varsaymışlardır. Eğer maddelerin standart sapmaları eşit değilse hesaplanan gü-

[a] Temel bileşenler analizi – TBA (Principle compenent analysis – PCA). Ölçek veya test maddelerinin işaret ettiği temel faktörleri ortaya çıkarmayı hedefleyen istatistiksel analiz yöntemi. İlgisiz değişkenleri bir kenara ayırıp birbirleriyle yüksek derecede ilişkili maddeleri bir araya getirerek faktörler olarak ortaya çıkarır. Gizli faktörlerin ortaya çıkarılmasına yönelik diğer teknikler; kümeleme (cluster) analizi, karşılık (correspondence) analizi, çok boyutlu ölçekleme analizi (multidimensional scaling) ve ortak faktör (factor) analizi yöntemleridir.

venilirlik katsayısının gerçeği olduğundan daha düşük göstereceğini belirtmişlerdir.

1940-1950 Dönemi

Kuder-Richardson formülü 1940'lı yıllardan 1950'li yıllara kadar etkili olmuştur. Bu dönemde başka bir psikometriysen Louis Guttman (1945) güvenilirlik kavramının *bir çok değişkenin toplamına ait olduğunu* bildirmiştir. Ona göre güvenilirlik tek tek maddelere ait bir olgu değildir; güvenilirlik maddeler arasındaki ilişkilere dayanır. Guttman farklı durumlar için, altı değişik *alt güvenilirlik sınırı* belirlemiştir. Söz konusu güvenilirlik katsayılarının her biri tek bir denemeden veya uygulamadan elde ediliyordu. Birden fazla zamanda deneme/ölçüm yapmaya gerek bulunmadığından verilerin güvenilirliğini hesaplamak nispeten daha kolaydı.[19] Yine bu dönemde Charles Mosier'in (1943) *bileşik değişken* olarak ölçeğin güvenilirlik analizi yaklaşımı; Godfrey H. Thomson (1940), E.A. Peel (1947) ve B.F. Green (1950) gibi bilim adamlarının geliştirdikleri başka hesaplama yöntemleri güvenilirlik analizi yaklaşımlarını zenginleştirmiştir.

1950-1960 Dönemi

Yirminci yüzyılın ikinci yarısından itibaren, psikometri konusunda önemli bir uzman olan Lee Joseph Cronbach, Kuder – Richardson yöntemini eleştirerek "KR-20 formülünde güvenilirliğin tutucu bir biçimde belirlenmiş olduğunu, değerin olması gerektiğinden daha düşük hesaplandığını ve bu düşüklüğün hangi oranda olduğunun bilinmediği" görüşünü ortaya atmıştır. Cronbach bilimsel araştırmalarda "tutucu" yaklaşımın önemli olduğunu, fakat KR-20 formülündeki hesaplamanın güçlük yarattığını bildirmiştir. Cronbach, 1951 yılında güvenirlilik için *alfa* formülünü tanıtmış, bu formülün aynı zamanda Kuder – Richardson 20 formülü olduğunu belirtmiş ve alfa güvenilirliğini muhtemel bütün yarı güvenilirlik rakamlarının ortalaması olduğunu ileri sürmüştür. Ancak bir süre sonra, *alfa değerinin* birden fazla boyutlu ölçeklerde *ağırlık kullanılmadan hesaplanan toplam puanlar* için yeterince güçlü olmadığı anlaşılmıştır. Bunun üzerine Cronbach, arkadaşları N. Rajaratnam ve G.C. Gleser ile birlikte 1963'te alfa katsayısıyla ilgili görüşlerini yeniden şekillendirmişlerdir. Cronbach, belirlediği yeni formülde güvenilirliği "genellenebilirlik" olarak açıklamıştır.[20]

Yine bu dönemde, Paul F. Lazarsfeld "gizli yapı kuramını" (latent structure analysis) tanıtmış ve iki şıklı yanıtlar için faktör analizi yöntemi-

nin uygulanması konusunda bazı yenilikler getirmiştir. Yirminci yüzyılın ikinci yarısından itibaren, klasik test kuramının yanında modern test kuramına ilişkin modeller ve istatistiksel analiz yöntemleri geliştirilmeye başlanmıştır. Klasik test kuramında, kişilerin yetenek düzeyleri ne olursa olsun herkese eşit uzunlukta ve eşit sürelî test uygulamasının yerini modern test kuramında kişilerin kendi yetkinlik düzeylerine uygun test maddeleriyle farklı süreler içinde sınanması biçiminde gerçekleşen test uygulamalarına geçilmeye başlanmıştır.

1960-1970 Dönemi

Bu dönemde, M. Novick ve C. Levis (1967) *bileşik ölçümlerde* Cronbach alfa için gerekli ve yeterli şartların neler olduğunu belirlemişler ve bu koşulları *tau eşitliği* (τ) olarak isimlendirmişlerdir. Tau eşitliği, ölçek maddelerinin gerçek değerleri arasında tam bir korelasyon bulunması ve maddelerin varyanslarının eşit olması anlamına gelmektedir (Lord ve Novick, 1974, aktaran Ferligoj).[21] Böyle olunca gözlem puanlarının kovaryanslarının da eşit çıkması beklenir. Eğer maddeler farklı gerçek varyanslara sahipse ve maddeler tek boyutlu değilse gerçek güvenilirliği tam olarak tespit edemeyiz. Bu nedenle Novick ve Levis, Cronbach alfanın çok boyutlu ölçeklerde düşük değerlikli bir güvenilirlik ölçüsü olduğunu söylemişlerdir. Maddeler eğer türdeş değilse ve ölçekte birden fazla boyut varsa *düşük değerlikli* bir güvenilirlik katsayısı ölçüm sonuçlarını sağlıklı bir şekilde yorumlamaya imkan vermez.[22]

Danimarkalı matematikçi Georg Rasch bu dönemde (1960) *Probabilistic Models for Some Intelligence and Attainment Tests* [Bazı Zeka ve Başarı Testleri için Olasılık Modelleri], isimli eseriyle ham puanların yerine lojistik değerleri temel alan yeni bir ölçüm modeli geliştirmiştir. Aynı dönemde Birnbaum madde-yanıt kuramı için "madde bilgi fonksiyonu" kavramını tanıtmıştır.

1970-1980 Dönemi

Tarihte 1970'li yıllara gelindiğinde sosyal bilimlerde kavramsal yapı araştırmalarının ve sosyal nitelikteki ölçüm çalışmalarının arttığı gözlenir. Bu yıllarda Goodman *maksimum olasılık* hesaplama yöntemini geliştirmiş ve bu yöntemi gizli yapı/küme analizlerinde kullanmıştır. Yine, A.E. Maxwell (1971) faktör ağırlıklarının *maksimum benzerlik tahmini*[a] isimli he-

[a] Maximum likelihood estimation.

saplama yöntemi üzerinde çalışmalar yapmıştır. K.G. Jöreskog 1971'de varyanslarının farklı olduğundan şüphelenilen maddeler için *Konjenerik Ölçüm Analizi (KÖA)*[a] adını verdiği başka bir katsayı hesaplama yöntemi geliştirmiştir. Konjenerik ölçümlerde maddeler ikili olarak gerçek korelasyon değerlerine sahip olabiliyordu, fakat maddelerin varyansları yine farklı çıkıyordu.[23] Jöreskog'un geliştirdiği KÖA'nın alfa değerine göre avantajı, maddelerin gerçek varyanslarında gözlemlenen büyük farklılıkların güvenilirlik katsayısını düşürmemesiydi. Fakat bu modelde de tek boyutluluğun aranması gerekiyordu. Konjenerik ölçüm analizinin bir diğer önemli avantajı geliştirilen modelin uygunluğunu değerlendirmeye veya test etmeye imkan vermesiydi. Eğer veriler geliştirilen konjenerik modele iyi uyuyorsa tahmin edilen güvenilirlik gerçek güvenilirlik rakamına yakın çıkıyordu.

D. Armor 1974'te Cronbach'ın alfa indeks değeri yaklaşımını ve madde analizi yöntemini eleştirmiş, kötü maddelerin ayıklanmasının gerçekte ölçeğin güvenilirliğini arttırmadığını ileri sürmüştür. Öte yandan P.H. Jackson ve C.C. Agunwamba 1977'de test maddeleri *yaklaşık olarak tau eşitliğine* sahip olmadığı zaman alt sınır güvenilirliğini belirlemeye yönelik olarak *Maksimum Düşük Değerlilik Güvenilirliği (MDDR)*[b] (Greatest Lower Bound Reliability) yaklaşımını geliştirmişlerdir.

1980-1990 Dönemi

Bilim adamları 1980'li yıllarda alfa güvenilirlik katsayısının düşük değerlikli olması, bileşik (kompozit)[c] ölçümlerde alfa katsayısı, kovaryans yapılar gibi konularda değişik araştırmalar yapmışlardır. Örneğin J. Fleishman ve J. Benson (1987) ölçüm modellerini değerlendirmek ve ölçeklerin güvenilirliğini belirlemek için K.G. Jöreskog ve Sörbom tarafından 1983 yılında geliştirilen LISREL modellerini kullanmışlardır.[24] 1988 yılında

---

[a] Konjenerik Ölçüm Analizi (Analysis of Congeneric Measures – ACM): Klasik test kuramı içinde araştırmacıya en az kısıtlama getiren bir yaklaşımdır. Bu modelde aynı olguyu ölçen test/ölçek maddelerinde sadece *gerçek değerler* arasındaki korelasyonun yüksek olmasına önem verilir. Bu nedenle konjenerik ölçümlerde maddelerin hata varyansları, gerçek puan ortalamaları ve gerçek puan varyansları eşit olmayabilir.

[b] Jackson, P.W. ve Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika*, 42, 567-578.

[c] Kompozit ölçek veya kompozit puan: Üç veya daha fazla maddenin bir araya getirilerek toplam puanlarının alınmasıdır. Kompozit ölçek, aynı zamanda boyut veya faktör anlamında da kullanılır. Böyle bir durumda çok faktörlü/boyutlu ölçekler kompozit ölçüm aracı olarak değerlendirilir.

Van der Linden ve Boekkiooi-Tamminga paralel test oluşturmada minimizasyon ve maksimizasyon yöntemlerini geliştirmişlerdir. 1989 yılında ise Adema ve Van der Linden test güvenilirliğini maksimize edecek modeller üzerinde çalışmalar yapmışlardır. Yine bu dönemde R. Wilcox (1992) Cronbach tarafından geliştirilen alfa katsayısının içerdiği varsayımlara herhangi bir eleştiri getirmeden *Güçlü Cronbach Alfa Katsayısı* adını verdiği yeni bir model geliştirmeye yönelik incelemeler yapmıştır.

1990-2000 Dönemi

Moss (1994), bu dönemde güvenilirlik ve geçerlilik ilişkisini analiz etmeye yönelik araştırmalar yapmıştır. Ona göre, güvenilirlik olmadan geçerlilik olabilir. Öğrencilerin tekrarlanan sınavlarda farklı sonuçlar almaları ölçümün güvenilir olmadığı anlamına gelmez; tersine bu sonuç, olayı daha geniş çerçevede açıklamak için deneysel yeni girdilerin araştırılması gerektiği anlamına gelir.[25] Moss, güvenilirliğin yapısal geçerliliğin bir yönü olduğunu belirtmiştir. Öte yandan B. Reinhardt (1996) küçük ölçekli simülasyonlarda alfa katsayısının negatif çıkabileceğini göstermiştir. Cronbach alfa değerinin sınırlarını ele alan araştırmalar halen devam etmektedir. W. Hofstee (1999) alfa ve alfanın türevlerini döndürülmüş ve döndürülmemiş *temel bileşenler analizi* ile test etmiştir.

2000'li Yıllar

Son yıllarda Lauri Tarkkonen (1987) ve Kimmo Vehkalahti (2000), kendi güvenilirlik hesaplama yöntemlerini geliştirmişlerdir, ancak bu yöntemler literatürde henüz yaygınlık kazanmamıştır.[26] Cronbach alfa değerinin varsayımları tek boyutlu ölçekler için geçerli iken Tarkkonen'in güvenilirlik yaklaşımının çok boyutlu/faktörlü modeller için de geçerli olduğu bildirilmiştir. Tarkkonen tarafından geliştirilen güvenilirlik katsayısı Cronbach alfa katsayısında olduğu gibi, kendi adıyla "Tarkkonen alfa değeri" olarak isimlendirilmiştir.

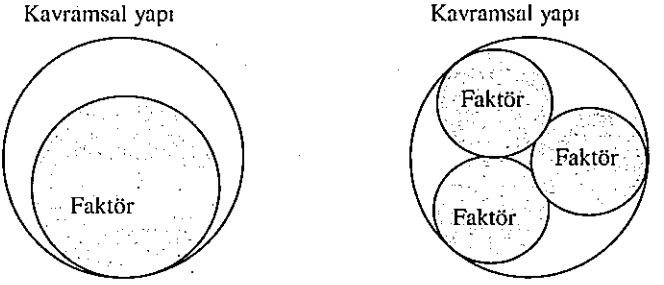
KAPSAMI

Klasik test kuramında güvenilirliğin kapsamı, bileşenlerinin içeriği ile ilgilidir. Güvenilirliğin temel bileşenleri beş başlıkta ele alınır. Bunlar; *iç tutarlılık, istikrarlılık, temsil edicilik, eş değerlilik* ve *nesnellik*tir.

İç Tutarlılık

İç tutarlılık, ölçek veya test içindeki maddelerin belirli bir kavramsal yapıya "birlikte asılması"dır. Bu kavramsal "yapı" kendi içinde tek veya birden fazla boyutu içerebilir. Diğer bir deyişle iç tutarlılık, ister tek boyutlu isterse çok boyutlu olsun ölçek maddelerinin birbiriyle ilişkili olarak aynı *yapıyı* ölçüyor olmasıdır. Ölçekteki her bir madde, ölçülmek istenen kavramsal yapıyı bir şekilde temsil ediyor olmalıdır. Bazı maddeler kavramsal yapıyla büyük ölçüde ilgili iken, diğer bazı maddeler ise kavramsal yapıyla nispeten daha az ilgili olabilir. Düşük ilgiye sahip maddelerin çıkarılmasıyla ölçeğin iç tutarlılığı ve dolayısıyla güvenilirliği artar. Böylece ölçekte sadece kavramsal yapıyla yüksek derecede ilgili olan maddeler kalmış olur. Kullanılan ölçek veya test ölçülmek istenen *kavramsal yapıyı* büyük ölçüde temsil etme özelliği kazanır.

Bir test veya ölçeğin tek veya çok boyutlu (faktörlü) olup olmayacağı literatürde tartışmalı bir konudur. Psikometrik ölçümlerin öncülerinden olan Thurstone ve Guttman ölçeklerin sadece tek bir vasfı ölçmesini diğer bir deyişle tek boyutlu olmasını ölçümün evrensel özelliği olarak tanımlamışlardır. Fakat daha sonraki yıllarda tek boyutluluk eski önemini yitirmiştir. Günümüzde az sayıda psikometrici (P.E. Levy 1973, R.P. McDonald 1999, J. Hattie 1985) hâlâ tek boyutluluğu savunur. Psikometricilerin önemli bir bölümü iyi bir şekilde yapılandırılmış, yüksek iç tutarlılığa sahip ölçeklerin çok boyutlu veya çok faktörlü olabileceği görüşündedirler (*bk.*, Şekil 1-2).



Şekil 1-2. Kavramsal yapı-faktör ilişkisi.

İç tutarlılık, çoğunlukla alfa değeri ile saptanır. Alfa güvenilirlik değeri, ölçeğin *türdeşlik indeksi* değildir.[27] Diğer bir deyişle bu katsayı, bütün

maddelerin sadece tek bir boyutu/faktörü ölçtüğünü göstermez. Maddelerin kavramsal yapıya ilişkin olarak birbirleriyle yüksek derecede korelasyona sahip olmalarıyla belirli faktörler altında türdeş olarak gruplandırılmaları farklı iki olgudur. Özellikle belirli faktörlere ait maddeler türdeştir. Bir ölçeğin tamamında eğer alt boyutlar varsa tüm maddeler türdeş olmayabilir. Testin veya ölçeğin tamamına ait iç tutarlılık, maddeler arasındaki korelasyonlarla veya maddeler arasındaki varyansla ilgilidir. Maddeler heterojen de olsa eğer birbiriyle yüksek derecede ilişkili ise yüksek alfa değeri elde edilir. Tutarlılık, maddelerin kavramsal yapıdan ayrı düşmemesi; cevaplayıcıların ölçek maddelerine birbiriyle tutarlı ve anlamlı cevaplar verme derecesidir. Farklı boyutlara/faktörlere ait maddeler kendi aralarında yüksek derecede ilişkili ise sonunda Cronbach alfa değeri de yüksek çıkar. Bununla birlikte alt boyutlar (faktörler) arasındaki korelasyonun her zaman yüksek çıkması gerekmez; önemli olan husus, maddeler arasındaki korelasyonun yüksek çıkmasıdır.[28]

İstikrarlılık

İstikrarlılık, ölçüm sonuçlarının aynı ve farklı koşullarda (zaman, yer, prosedür vb.) kararlılık göstermesi ve değişmemesidir. İstikrarlılığın gerçekleşebilmesi için testin belirli bir zaman geçtikten sonra veya başka bir yerde aynı örnek kütleyle uygulandığında benzer sonuçları vermesi gerekir. Bilim dünyasında davranış bilimcileri genel çizgileriyle iki grupta değerlendirilir. Birinci gruptaki bilim adamları geliştirdikleri *özellik teorileriyle* insanların düşünce ve davranışlarında oldukça kalıcı tutumlara sahip oldukları tezini savunurlar. Onlara göre kişiliğin oluşumu insanların istikrarlı davranışlarıyla açıklanabilir. Davranış bilimcisinin görevi oldukça istikrarlı olan bu davranışları, düşünce ve tepki biçimlerini ortaya çıkarmaktır. İkinci grup bilim adamları ise, değişen *durum ve koşulları* ön plana çıkarırlar. Onlara göre insanlar davranışlarında çok fazla tutarlı değildirler. Aradan belirli bir süre geçtikten sonra insanlar farklı bir şekilde düşünebilirler, değişebilirler veya yeni davranışları kazanabilirler. İnsanlar değişkendir, bazen çevrelerinde olup bitenlere çok fazla dikkat edip kendilerini yenilerlerken, bazen de bütünüyle kayıtsız kalıp kendi iç dünyalarıyla ilgilenirler. Durum ve koşulların insanları büyük ölçüde etkileyeceğine inanan davranış bilimcileri yapılan *ölçümlerin zaman içinde istikrarlı olup olmadığına* fazla önem vermezler.

Literatürdeki söz konusu kuramsal eğilimler dikkate alındığında, ölçek verilerinin istikrarlılığını, ölçeğin niteliğine göre bazen değişme etkisinin

ortaya çıkmaması için çok uzun olmayan zaman aralıklarında ve bazen de oldukça uzun zaman aralıklarında değerlendirmek gerekir. Kanaat ve düşünceler çok çabuk değişebilir. Zekanın, kişiliğin gelişimi ve değişimi ise daha yavaştır. Bu nedenle düşünce ve kanaatlerin belirlenmesinde istikrarlılık ölçümleri bir iki hafta gibi kısa zaman aralıklarında yapılırken, kişilik ve zeka ölçümleri altı ay, bir yıl, iki yıl gibi çok daha uzun zaman aralıklarında test edilir.

Temsil Edicilik

Ölçeğin/testin temsil edicilik özelliği, aynı ana kütleye ait farklı örneklemlerde uygulandığında benzer sonuçlar vermesidir. Temsil edicilik güvenilirliğini belirlemek için bir ana kütleye ait alt gruplarda araştırma yapılır. Testin temsil edicilik özelliği etnik yapı, cinsiyet, yaş dağılımı, sosyoekonomik durum, eğitim, okul, sınıf gibi faktör grupları için ayrı ayrı belirlenir.

Bir testin veya ölçeğin aynı ana kütleye ait farklı örneklemler için benzer değerleri vermesinde dahi çözülmemiş sorunlar vardır. Bunun için her bir özellikli örneklem gurubu için *ana kütle çerçevesi* yeniden tanımlanmalı ve güvenilirlik hesaplamaları buna göre yeniden yapılmalıdır. Çünkü aynı ana kütle içindeki farklı grupların norm değerleri de farklıdır. Göçmenler, engelli vatandaşlar, eğitim düzeyi düşük kişiler, kırsal kesimde yaşayanlar ve kentlerde yaşayanların algıları, yetenekleri veya olayları değerlendirmeleri farklı olabilir. Bu nedenle bu gruplarda yapılacak güvenilirlik analizlerinin sonuçları da farklı çıkar.

Yeni geliştirilen bir test/ölçek için *keşfedici faktör analizi* ile arka plandaki gizli yapıyı, faktörü veya faktörleri ortaya çıkarmak ve daha sonra bu faktörlerin her biri için iç tutarlılık analizleri yapmak anlamlı görünürken geliştirilmiş bulunan bir testin farklı gruplarda uygulanması halinde bu faktörler geçerliliğini, temsil edicilik özelliğini yitirebilir. Örneğin, diğer gruplarda faktör yapısı ve faktörleri temsil eden maddeler değişebilir. Keşfedici faktör analizinin gruplar arasındaki varyansı açıklayacak bir hipotez testi önermemesi nedeniyle bu teknik zayıf olarak değerlendirilmiştir. Onun yerine gruplar arasındaki varyansı daha iyi açıklayıp belirlenen hipotezi test etmeye imkan sağlayan *teyit edici faktör analizinin* kullanılması önerilmiştir.[29]

Klasik test kuramında temsil edicilik özelliği önemli bir öge olarak vurgulanırken modern test yaklaşımlarından *madde-yanıt kuramında* (MYK) test sonuçlarının ve dolayısıyla güvenilirlik derecelerinin araştırma

yapılan örneklemden bağımsız olduğu görüşü temel alınır. Testin yerine maddenin temsil ediciliği ön plana çıkar ve bu temsil edicilik maddeyi yanıtlayan kişilerin ortalama yetenek düzeyiyle birlikte düşünülür.

Eş Değerlilik

Benzerliği, eşit sonuçlara ulaşmayı veya eş değer kavramsal yapılara sahip olmayı gerektirir. Eş değerlilik kavramına faklı biçimlerde yaklaşılabilir. Bunlardan birincisi, yaklaşık olarak aynı zamanda uygulanan iki veya daha fazla testin benzer sonuçlar vermesidir. Test veya ölçeklerin benzer sonuçlar verebilmesi için her ikisinin de aynı kavramsal yapıyı ölçmesi gerekir. Ölçekler aynı zamanda benzer alt boyutlara sahip olmalıdır. Bu ölçeklerden biri araştırmacının geliştirdiği bir ölçek iken, diğeri başka bir bilim adamının geliştirdiği veya yabancı dilden Türkçeye çevrilerek uyarlanmış bir ölçek/test olabilir. Eş değerlilik korelasyon katsayılarıyla, özet istatistik analizi sonuçlarıyla veya boyutların benzer olup olmadığını belirlemek için faktör analizi yöntemiyle tespit edilir. Eğer örneklem hacminde 100 – 200 kadar kişi varsa faktöriyel eşitliği belirlemek için keşfedici faktör analizi veya daha uygunu *çok boyutlu ölçekleme analizi* (ÇBÖA) kullanılır.

Eş değerliliği test etmeye yönelik ikinci uygulama, testin rasgele oluşturulan iki yarısı arasında benzerlik olup olmadığını belirlemektir. İki yarı arasındaki korelasyon katsayısı yüksekse test/ölçek eş değerlilik kriterini karşılıyordur.

Nesnellik

Nesnellik, gözlemciler arasındaki değerlendirme güvenilirliğidir.[30] Farklı *değerlendiricilerin* veya farklı *gözlemcilerin* aynı kişilerle ilgili olarak benzer puanları vermeleridir. Nesnellik, derecelendirme ölçeklerinde aranan önemli bir özelliktir. Değerlendiricilerin tarafsız, objektif olabilmeleri için test uygulama koşullarının standartlaştırılması ve puan verme talimatının bulunması gerekir. Testin nerede, nasıl ve kimler tarafından uygulanacağı ve puanlamanın nasıl yapılacağı konusunda tam bir açıklık olmalıdır. Buna göre nesnellik aşağıdaki faktörleri içerir:

1. Test veya ölçüm yapmada uygulama benzerliği.
2. Puanlamanın nasıl yapılacağı konusunda açıklık.
3. Puanlama konusunda kişisel yetenek ve beceri.
4. Gözlemcilerin verdikleri puanlardaki tutarlılık.

Belirlenen şartlara uygun olarak iki veya daha fazla değerlendirici bir etkinliği veya kişiyi benzer puanlar vererek değerlendiriyorlarsa nesnellik kriteri sağlanmış olur. Örneğin bir işletmenin muhasebe biriminde çalışan personelin birinci ve ikinci yöneticisi tarafından verilen başarı değerlendirme puanları birbirine benzerse nesnellik sağlanmıştır. Puanlar arasında önemli ölçüde farklılık varsa değerlendirme sonuçları güvenilir olmaktan uzaktır.

GÜVENİLİRLİK VE HATA

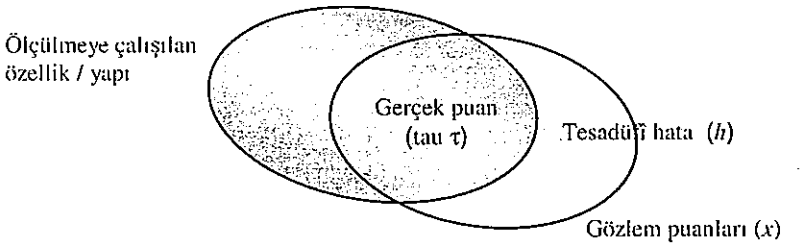
İster fiziksel, eğitsel, örgütsel veya isterse psikolojik nitelikte olsun bütün ölçeklere ve testlere ait veriler tam olarak *doğru* değildir. Söz konusu ölçüm verilerinde *hatalar* vardır. Bu hatalar; ölçüm aracının düzenleme biçiminden, uygulama prosedüründen, cevaplayıcıların kendilerinden veya verilerin kodlanmasından kaynaklanır. Hataların bir kısmı kontrol altına alınabilecek türden iken, diğerlerini kontrol altına almak zordur. Kontrol altına alınamayanlar *tesadüfî hata* olarak isimlendirilir. Güvenilirlik, bir anlamda kontrol altına alınamayan tesadüfî hataların ne ölçüde bulunduğunu saptama işlemidir. Tesadüfî hatalar azaltıldığı ölçüde, elde edilen gözlem verileri *güvenilir* olarak yorumlanır. Buna göre bir ölçüm işlemi Eşitlik 1-1'deki matematiksel modelle ifade edilir.

$$
\begin{aligned}
x &= g + h. && \text{(Lâtin harfleriyle gösterim biçimi)} \\
x &= \tau + \varepsilon. && \text{(Uluslar arası simgelerle gösterim biçimi)} \\
x &= \tau + \varepsilon_t + \varepsilon_s.
\end{aligned}
\qquad (1\text{-}1)
$$

x = Gözlem değeri.
g = Gerçek değer (τ).
h = Hata puanı (ε).
ε_t = Tesadüfî hata.
ε_s = Sistematik hata.

Gerçekleştirilen herhangi bir ölçüm işleminde x simgesi gözlem değerlerini gösterir. Gözlem değerleri *gerçek değer* ve *hata değerlerini* birlikte içerir (*bk.*, Şekil 1-3). Gerçek değer, hipotetik bir birimdir. Ana kütlede tam sayım/ölçüm yapılması halinde elde edilebilecek değerdir. Gerçek

değerin en yaygın tanımı $E(x)$ simgesiyle ifade edilen "beklenen değer"dir. Beklenen değer, sonsuz evrende yapılacak ölçümlerde x'in beklenen ortalama değeridir. Formüldeki tesadüfî hata ise, gözlem değerlerinden gerçek değer çıkarılarak bulunur ($h = x - g$).



Şekil 1-3. Klasik test kuramında gerçek puan ve tesadüfî hata.

Güvenilirlik, tek başına ölçeğin veya ölçüm aracının kendisiyle değil, *gözlem rakamlarıyla* ilgilidir.[31] Gözlem değerleri bir araştırma sonucunda (anket, gözlem, mülâkat, test veya deney olabilir) elde edilen rakamlardır. Gözlem değerlerinin varyansı[a] (değişkenliği), *gerçek değerlerin değişkenliğinin* yanında *hata paylarının değişkenliğini* de içerir (Lord ve Novick, 1968).[32] Buna göre ölçüm modelinin varyansı Eşitlik 1-2'deki gibi ifade edilir.

---

[a] Varyans (değişkenlik, oynama, hata): Varyans matematiksel bir terimdir ve bir serideki puanların dağılım aralığını veya puanların yayılışını gösterir. Bir testin veya ölçeğin varyansı, toplam puan üzerinden veya tek tek maddeler üzerinden hesaplanır. Varyans matematiksel olarak standart sapmanın karesidir (örnek kütle için s^2 ve ana kütle için ise σ^2 simgesi kullanılır). Toplam puan varyansı, birden fazla maddeye sahip bir test için veya aynı test birden fazla gruba uygulandığı zaman gruplar arasındaki puan değişimlerini incelemek için iyi bir ölçüdür. Maddenin varyans değeri ise maddeye gelen cevapların yayılışı hakkında bilgi verir. Maddelerin varyansı $\Sigma (x - m)^2 / n-1$ formülüyle hesaplanır. İki şıklı maddenin varyans değeri ise şu formülle hesaplanır: Varyans = $P * Q$. Formüldeki P test maddesini doğru yanıtlayanların oranını Q ise yanlış yanıtlayanların oranını gösterir. Maddenin varyansı p = ,50 olduğunda en yüksek değerini bulur. İki şıklı maddelerin varyansı aynı zamanda maddelerin güçlük derecesini gösterir. Bu maddelerin varyansı 0 ilâ ,25 arasında değişir. Bir maddenin varyansı büyüdükçe o maddenin testin/ölçeğin güvenilirliğine olan katkısı artar. Varyansı sıfıra yakın olan maddeler ise testten çıkarılır. Varyansın bir diğer anlamı, ana kütlede gözlenemeyen veya ölçülemeyen birimlerle ilgili hatadır. Varyans aynı zamanda örneklem hatasının karekökü dür.

$$\sigma_x^2 = \sigma_g^2 + \sigma_h^2 \ . \qquad (1\text{-}2)$$

σ_x^2 = Gözlem puanlarının varyansı.
σ_g^2 = Gerçek değerin varyansı.
σ_e^2 = Hata puanlarının varyansı.

Güvenilirlik, gerçek puanlardaki değişkenliğin gözlem puanlarındaki değişkenliğe bölünmesiyle bulunur. Teknik bir dille ifade etmek gerekirse, gerçek puan varyansının gözlem puanları varyansına bölünmesidir (*bk.*, Eşitlik 1-3).

$$r = \frac{\sigma_g^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_h^2}{\sigma_x^2}. \qquad (1\text{-}3)$$

Belirlenen bu formül hipotetik bir anlama sahiptir. Çünkü gerçek değerler ve bu değerlerin varyansı bilinmediğinden bu formülü çalıştırmak mümkün değildir. Güvenilirlik katsayısı bu nedenle korelasyon analizi, Cronbach alfa ve KR-20 gibi özel hesaplama yöntemleriyle tespit edilir.

Hata Kaynakları

Ölçümün içinde yer alan *hatalar* değişik nedenlerden kaynaklanır. Bilim adamları hata kaynaklarını genelde ikili guruplar içinde sınıflandırmışlardır. Bazı bilim adamları, *yöntem hatası* ve *özellik hatası* şeklindeki bir sınıflandırmayı uygun görürlerken diğerleri *sistematik hata* ve *tesadüfî hata* kavramlarını ön plana çıkarmışlardır. Bu sınıflandırmalarda, "hata" olgusu sadece farklı bakış açılarıyla değerlendirilmiştir. Biz de, araştırma sürecini göz önünde bulundurarak hata kaynaklarını dört grup altına inceleyebiliriz: (a) ölçüm aracından kaynaklanan hatalar, (b) uygulama işleminden kaynaklanan hatalar, (c) cevaplayıcıların kendilerinden kaynaklanan hatalar, (ç) verilerin kodlanmasından kaynaklanan hatalar (*bk.*, Tablo 1-1). Hangi sınıflandırma esas alınırsa alınsın sonuçta söz konusu hatalar verilerin güvenilirliğine gölge düşürür. Aşağıdaki bölümde farklı bakış açılarıyla hata kaynaklarına ilişkin değişik sınıflandırmalar ele alınmıştır.

Tablo 1-1. Hata Kaynakları

Ölçüm aracı hataları	Uygulama hataları	Cevaplayıcı hataları	Kodlama hataları
• Alan örneklemesi hataları • Maddelerin belirsiz olması • Yanlış kelimelerin kullanılması • Dereceleme hataları • Etiketleme hataları • Çift fikirli cümleler • Çift olumsuz cümleler	• Yetersiz bilgilendirme • Anketörler arasındaki farklılıklar • Yetersiz uygulama koşulları • Işık, ısı, gürültü • Kopya çekme • Farklı uygulamalar • Cevaplayıcıların uykusuz olmaları	• Bazı maddelere yanıt vermeme • Kendini gizleme • Dikkatsiz doldurma • İsteksiz doldurma • Yorgun olma • Duygusal davranma • Anlamama • Test kaygısı • Almış olduğu ilaçların etkisi	• Birden fazla şık işaretleme • Eksik şık bırakma • Yanlı olarak kodlama • Bilgisayara yanlış kodlama • Negatif maddeleri ters çevirmeme • Cevapsız bırakılan maddelere yanlışlıkla puan girme

Yöntem – özellik hatası. Herhangi bir ölçümdeki gözlem değeri, gerçek değer ve hata değerini birlikte içerir. Bunu aşağıdaki formülle ifade etmek mümkündür:

Gözlem değerleri = Gerçek değer + Hata (Yöntem hatası + Özellik hatası).

Yöntem hatası, araştırmacının uyguladığı metodun çok dikkatli bir şekilde tasarlanmamış olmasından kaynaklanır. Örneğin, anketin gürültülü bir ortamda uygulanması, çalışanların stresli oldukları zamana getirilmesi, anketin insanlar yolda yürürken ayak üstü doldurulmaya çalışılması, süper marketlerde tüketicilere yönelik olarak yapılan bir araştırmada kullanılan anketin sadece hafta sonu müşterilerinde uygulanması, üst düzey yöneticilere anketin posta ile gönderilmesi, fakat anketi kimin doldurduğunun tam olarak bilinememesi, anketin cevaplama süresinin 45 dakika gibi çok uzun olması, anketteki madde sayısının gereğinden fazla olması, ölçüm derecelerinin yanlış belirlenmesi, postayla gönderilen anketlerin sadece %25'inin

geri dönmesi, kolayda örnekleme yönteminin seçilmiş olması, anketin sadece tek bir yöntemle ve tek bir örnek kütle üzerinde uygulanması (ortak yöntem varyansı) gibi sorunlar bunlar arasındadır.

Özellik hatası, anket uygulanan bireylerin kişisel özelliklerinden kaynaklanır. Test alan veya kendilerine ölçek uygulanan kişilerin yorgun olmaları, alacakları test nedeniyle endişelenmeleri veya kaygı içinde olmaları, anket formunu isteksiz olarak doldurmaları, testi uygularken veya anketi doldururken arkadaşlarından yardım almaları, morallerinin düşük olması, ifadeleri yanlış okumaları veya yanlış anlamaları, kendilerine emrivaki yapılması nedeniyle anketi acele bir şekilde doldurmaları, stresli olmaları, anket doldururken aynı zamanda başka bir işle ilgilenmeleri, fiziksel olarak anketi uygun olmayan koşullarda doldurmaları, aşırı ölçüde uyarılmış/tetikte bulunmaları, kendilerine testin/anketin nasıl doldurulacağı konusunda gereğinden fazla bilgi verilmesi veya bazı kişilere bilgi verilirken diğerlerine verilmemesi, testte kişinin şansının yaver gitmesi özellik hatalarına neden olur. Cevaplayıcı çok uygun koşullarda dahi, *özellik hatası* yapabilir. Örneğin, yapılan bir araştırmada kişilere boyları ve kilolarıyla ilgili sorular yöneltildiğinde araştırmaya katılanlar boylarını olduğundan biraz daha yüksek ve kilolarını ise olduğundan düşük göstermişlerdir.[33] Bilim adamının, bu nedenle araştırma süreci içinde karşılaşabileceği muhtemel özellik hatalarını azaltmaya ve kontrol etmeye yönelik olarak bir ön çalışma yapmasında yarar vardır.

Sistematik hata – tesadüfî hata. Sistematik hata kavramını ön plana çıkaran bilim adamları bu terimle araştırma sürecinde; (a) yönteme ve uygulamaya ait yanlışlıkları, (b) sadece bir iki kişide değil, sistemli bir şekilde tüm grup çapında ortaya çıkan hataları ve (c) ölçeğin iç yapısının araştırma sonuçlarını etkilemesini kastetmişlerdir. Sistematik hata, teorik bir değerlendirmedir; bir test veya ölçek eğer amaçladığı alanın dışında başka bir şeyi daha ölçüyorsa ortaya çıkar. Örneğin *Liderlik Ölçeği* aynı zamanda *Zihinsel Sertliği* de ortaya çıkaracak maddelere sahipse sistematik hata içeriyor demektir. Sistematik hata metotla ilgili ise, *yöntem hatası* adını alır. Örnek kütledeki kişilerin yanlı olarak belirlenmesi sistematik hataya yol açar. Örneklemdeki kişiler hep belirli bir özelliğin etkisi altında kalarak yanıt vermişlerse bu da bir sistematik hatadır. Gözlemcilerden birinin sürekli olarak düşük puanlar vermesi, araştırmaya katılan cevaplayıcıların yaklaşık olarak tamamının olumlu veya olumsuz görüşler bildirmeleri sistematik hata örnekleridir. Sistematik hatada örneklem verilerinin ortalama-

sı gerçek değer ortalamasından farklıdır. Bu tür verilerde puanların dağılımından önce ortalamanın doğruluğu soruşturulabilir bir nitelik kazanmıştır. Bilim adamları, güvenilirliği etkileyen sistematik hatayı azaltmak için iki yönteme başvururlar: üçleme ve kalibrasyon.

Üçleme. Metodolojik çoğulculuğu simgeleyen bu yaklaşımda araştırmacı bilgi toplarken *üç farklı gözlem noktasından hareket etme yaklaşımını* benimser. Literatürde bu uygulama *üçleme* olarak isimlendirilmiştir.[a] Bilim adamı sistematik hatayı azaltmak için; bir taraftan anket sorularının dikkatli bir şekilde doldurulup doldurulmadığını incelemeli, diğer taraftan günlük yaşamdaki gözlemlerin bu verileri doğrulayıp doğrulamadığını araştırmalıdır. Yaklaşımın üçüncü ayağında ise, kapsamlı bir literatür taraması yapılır. Literatürdeki araştırma bulgularının anket sonuçlarını ve bilim adamının gözlemlerini teyit edip etmediği incelenir.

Kalibrasyon. Sistematik hatayı azaltmanın ikinci bir şekli, *ince ayar* (kalibrasyon) yöntemidir. Bu yaklaşımda bir maddeyi eleştiren gözlemcilere veya değerlendiricilere pilot araştırma sırasında ek bilgiler verilerek geri besleme yapılır. Geri beslemedeki amaç bilim adamının varsayımları, amacı konusunda değerlendiricileri tam olarak bilgilendirmesi ve değerlendirmenin bu bilgiler çerçevesinde doğru yapılmasını sağlamaktır. Ancak bu aşamada geniş ve kapsamlı tartışmalara girilmez, aksi halde ölçeğin cevaplandırılmasında tekrar yanlılık veya yönlendirme faktörü ortaya çıkar.

Kline'ye (1993) göre, sistematik hatalar psikolojide nispeten daha az önemsenmiştir. Çünkü bu tür hatalar bütün ölçümleri eşit derecede etkiler. Sistematik hatalar sadece bu tür hataları içermeyen test sonuçlarıyla karşılaştırma yapılmak istendiği zaman çeşitli güçlüklere neden olur.[34] Klasik test kuramında, test sonuçları üzerinde yeknesak bir etkiye sahip olan sistematik hata kaynakları güvenilmezlik kaynağı olarak görülmemiştir. Genelde bu tür hataların *gerçek puanın* içinde gizlendiği varsayılır.[35]

Tesadüfî hatalar ise, bilinmeyen, kontrol edilemeyen faktörlerden kaynaklanır ve ölçümün doğruluğunu daha ciddî bir şekilde etkiler. Test uygulanan ortamın fiziksel düzenleme biçimi, ışık, gürültü, anket uygulamasın-

[a] Üçleme (triangulation). Bu yaklaşım metodolojik çoğulculuğun yanında, fikir ve görüşlerin değerlendirilmesinde "yana" "karşı" ve "kararsız" olanları tespit etmek için de kullanılır. Bir ana kütlede yer alan kişilerin bir olguyla ilgili olarak "lehinde", "aleyhinde" ve "kararsız" olanların toplamı 100 olarak değerlendirilir.

da yeknesaklığın bulunmaması, anketi cevaplayan kişilerin içinde bulundukları ruh halleri, yorgunluk ve tahmin yürütme tesadüfî hataya neden olur. Özellik hatası da, kontrol edilmesinin çok zor olması nedeniyle *tesadüfî hata* olarak değerlendirilmiştir. Kişiler uygulanan ölçeği veya testi değerlendirirlerken bazen olduğundan daha yüksek ve bazen de olduğundan daha düşük puanları işaretlerler. Sonuçta bir örnek kütlede bu tür hatalar birbirini nötr hale getirir (*bk.* Tablo 1-2). Çünkü hataların sıfır aritmetik ortalama değerine ve normal dağılım eğrisine sahip olduğu varsayılır. Tesadüfî hatanın en önemli özelliği, verilerin değişkenliğini (varyansını) artırması fakat grubun ortalama puanını etkilememesidir.[36] Norland'a (1990) göre, "Geçerlilik analizi sistematik hatayı, güvenilirlik analizi ise tesadüfî hatayı kontrol eder."[37]

Araç – uygulama – cevaplayıcı – kodlama hataları. Bu sınıflandırma biçiminde araştırma süreci temel alınmıştır. Ölçüm aracından kaynaklanan hatalar daha önce de belirtildiği gibi ölçeğin/testin iç yapısından kaynaklanır. Maddelerin tutarlı olmaması, maddelerin birden fazla boyutu ölçüyor olması, ilgisiz maddeler içermesi ölçüm aracı hatasını gündeme getirir. Ayrıca birbiriyle karşılaştırılan ölçekler de hiçbir zaman birebir eşitliğe sahip değildir. Her zaman bu iki ölçek bazı farklı yönleri ölçüyor olabilir.

Uygulama biçiminden kaynaklanan hatalar yöntem hatalarıdır. Test alan kişilerin bir kısmına 15 dakika, başka bir kısmına 20 dakika süre verilmesi uygulama hatasıdır.

Tablo 1-2. Gerçek Puanlar ve Hata Puanları

Öğrenci no.	Toplam puan (x)	Gerçek puan (g)	Hata puanı (h)
1	48	45	+3
2	36	39	−3
3	42	40	−2
4	38	36	+2
5	36	37	−1
6	44	40	+4
7	32	31	+1
8	35	38	+3
9	43	45	+2
10	44	48	−4

Yine, test veren veya uygulayan kişilerin ön eğitimden geçirilmemeleri, ölçeği uygulayan anketçilerin bir kısmına eğitim verilmiş olması, fakat diğerlerinin herhangi bir eğitimden geçirilmemiş olmaları, anketlerin bir kısmının cevaplayıcıların kendileri, diğerlerinin anketçiler tarafından doldurulması, bir kısım anketin ise kimin tarafından doldurulduğunun tam olarak bilinmemesi uygulama hatalarını gösterir. Uygulama hataları büyük ölçüde kontrol edilebilir bir niteliğe sahiptir.

Cevaplayıcıların kendilerinden kaynaklanan hatalar; konsantre olamama, ilgi duymama, kendini gizleme, kendini tanımama, duygularının etkisinde kalma gibi nedenlere dayanır. Bu tür hatalar kısmen kontrol edilebilir. Ancak araştırmacı bu yönde yine de önemli bir çaba harcamalı ve bu tür hataları azaltacak önlemleri almalıdır.

Verilerin kodlanmasından kaynaklanan hatalar, cevaplayıcılar tarafından veya verileri bilgisayara giren araştırmacılar tarafından yapılır. Cevaplayıcıların dikkatsizlikle yanlış kutuyu işaretlemeleri, bir ifadeye yanıt vermemeleri, yanlış bir değerlendirmeyle yüksek puanları işaretlemeleri veya birden fazla şık işaretlemeleri kodlama hatası örnekleridir. Cevaplayıcılardan kaynaklanan kodlama hataları bir araştırmanın en önemli handikaplarından biridir. Araştırmacının yaptığı kodlama hataları ise rakamların bilgisayara yanlış girilmesi, negatif soruların tersine çevrilmemesi, ölçeğin niteliğine göre toplam veya ortalama puanlardan hareket edilmemesi, yanıtsız bırakılan ifadelerin (eksik yanıtlama) nasıl işleme alınacağının ve kukla değişkenlerin nasıl kodlanacağının bilinmemesi gibi faktörlerden kaynaklanır. Bilgisayara girme işlemi eğer optik okuma cihazlarıyla yapılmışsa araştırmacının kendisinden kaynaklanan kodlama hataları en alt düzeye iner. Herhangi bir ölçümün güvenilirliği gözlem değerlerinin gerçek değeri yansıtma oranına bağlıdır. Tekrarlanan ölçümlerde elde edilen değerler hep gerçek değerleri yansıtıyor olmalıdır. Teorik olarak bir ölçek, gerçek değerleri yansıttığı ölçüde güvenilirdir. Bu nedenle güvenilirlik, esas olarak ölçümdeki *hata oranını* azaltma çabasıdır.

$$Güvenilirlik\ Oranı = \frac{Gerçek\ Değer}{Gözlem\ Değeri\ (Gerçek\ Değer + Hata\ Değeri)}.$$

Bir araştırmacı, sınav öncesinde öğrencilerin yaşadıkları gerilimi belirlemeye yönelik olarak yaptığı bir araştırmada 10 ifadeden oluşan bir ölçek

kullanmış ve anket 12 öğrenciye uygulanmış olsun. Anket sonunda her bir öğrencinin toplam gerilim puanları belirlenmiş olacaktır. Bu puanlar gerçek değer ve hata değerinin her ikisini de içerir. Öğrenciler, ölçeğin iyi yapılandırılmamış olması nedeniyle veya yaşadıkları duyguları tam olarak aktaramamaları nedeniyle hatalı işaretleme yapmış olabilirler. Ancak bilim adamının bu tür hataları otomatik olarak saptaması ve ayıklaması mümkün değildir. Tesadüfî hatalar ölçüm aracının doğru bir şekilde yapılandırılması ve ölçüm işleminin kontrollü şartlarda yapılmasıyla azaltılabilir.

Belirli büyüklükteki bir örneklemde, hata puanlarının ortalaması sıfır olmak üzere verilerin normal dağılım özelliğine sahip olduğu varsayılır. Diğer bir deyişle, hata puanlarının ortalaması sıfır olduğundan eksi ve artı yöndeki hataların birbirini nötrleştireceği varsayımından hareket edilir. Bir örneklemde gerçek puanlar bilinmediğinden güvenirlilik matematiksel olarak hesaplanamaz, sadece tahmin edilebilir. Korelasyon katsayısı, Cronbach alfa, KR-20 veya KR-21 gibi değişik istatistiksel ve matematiksel yöntemler uygulanarak yapılan analizler sadece *güvenilirlik olasılığını* belirler.[38] Güvenilirlik katsayısının en önemli özelliği 0 ilâ 1 arasında değişmesi ve belirli bir oran üzerinden belirlenmesidir. Literatürde ,70 oranı genelde sınır değer olarak kabul edilmiştir. Bunun anlamı ölçüm sonuçlarının ,70 oranında *gerçek değeri* ve ,30 oranında *hata değerini* içerdiğidir.

Bir araştırmada örneklem hacmi büyüdükçe verilere ait dağılımın değişkenliği (varyansı) azalır, fakat örneklem büyüklüğü, dağılımın normale yaklaşacağı hakkında kesin bir garanti vermez. Verilerin güvenilirliği ve kalitesi örneklem hacmine değil, ölçümün doğru yapılmasına ve örneklemin temsil edicilik özelliğine bağlıdır.[39] Büyük, fakat temsil edicilik özelliği zayıf bir örneklemde çalışılmış ve dikkatsiz bir ölçüm yapılmışsa güvenilirlik sağlanamaz. Araştırmacı büyük örnek kütle hacimleriyle çalışıyorsa muhtemelen tesadüfî hatalar çok daha fazla ortaya çıkar.

KLASİK VE MODERN ÖLÇÜM KURAMLARINDA GÜVENİLİRLİK

Klasik Ölçüm Kuramı

Spearman tarafından önerilen klasik ölçüm kuramı, soyut kavramsal yapı modelleri üzerine kurulmuştur. Bu yaklaşımda, önce kavramsal yapılar ve daha sonra bu yapıları ölçecek değişkenler belirlenir. Bilim adamı, ölçümü bu değişkenlerden hareket ederek gerçekleştirir. Toplanan veriler, "gerçek

puan + hata" formülü çerçevesinde incelenir. Bu kuramda, ham puanlarda görülen gerçek puanlardan farklı olabilen tutarsızlıklar, "tesadüfî hata" olgusuyla açıklanır. Klasik test veya ölçüm kuramında güvenilirlik, bir ölçüm aracının çok iyi düşünülmüş, tasarlanmış ve standartlaştırılmış olduğunun kanıtı olarak önemlidir. Güvenilirlik, "yüksek ölçüde standartlaştırılmış testlerin ve ölçeklerin ortaya konması açısından" gereklidir.

Klasik ölçüm kuramının temelinde dört temel varsayım vardır. Birincisi, örnek kütleden elde edilen ölçüm değişkenine ait değerlerin ana kütlede normal dağılım özelliği gösterdiğidir. Bilim adamı örnek kütleden hareket ederek test sonuçlarını ana kütleye genellemek ister. İkincisi, ölçeği/testi oluşturan tüm maddelerin tau eşitliğine (gerçek puan değerine) sahip olduğudur. Üçüncüsü, her bir maddenin hata varyansının diğer maddelerin hata varyansından bağımsız olduğu ve dördüncüsü ise, maddelerin / ölçümlerin birbirine paralel olduğudur.[40] Paralellik bir testin içindeki maddeler arasında, farklı testler veya farklı formlar arasında sınanabilir. Maddelerin/ölçümlerin birbirlerine benzer olma açısından paralel olma özelliği sıralı olarak değişik düzeylerde gerçekleşir:[41]

1. Paralel.
2. Tau eşitliğine sahip.
3. Yaklaşık tau eşitliğine sahip.
4. Konjenerik (tek boyutlu).

Paralel olma özelliği. Paralel olma özelliğinde (Spearman 1904), farklı iki ölçüm sonucu elde edilen gözlem puanları ve "gerçek puanlar" birbirlerine eşittirler [$M(X_1) = M(X_2) = \tau_1 = \tau_2$]. Birinci maddeye ait gözlem puanı ikinci maddeye ait gözlem puanına ve onlar da gerçek puanlara eşittir. Aynı şekilde gözlem puanlarının hata varyansları da birbirine eşittir:

- Var (X_1) = Var (X_2).

Birinci maddenin varyansı ikinci maddenin varyansına eşittir. Maddelerin varyansları eşit olduğunda hata kovaryansları (ortak varyansları) sıfırdır.

- Ort. Var. $(h_1, h_2) = 0$.
- Var (h_1) = Var (h_2) .

Birinci maddenin hata varyansı ikinci maddenin hata varyansına eşittir. Klasik test kuramında hata öğesiyle ilgili olarak üç varsayımdan söz edilir:[42]

1. Hata bileşenlerinin aritmetik ortalaması sıfırdır ve bu nedenle gözlem puanlarının aritmetik ortalaması gerçek değerden sistematik olarak büyük ölçüde farklı çıkmayacaktır.
2. Ölçüm hataları da normal dağılım eğrisine uygundur.
3. Ölçüm hatalarıyla gerçek değerler arasında bir ilişki yoktur, bunlar birbirinden bağımsızdır.

Maddelerin ölçüm hatalarının varyanslarının birbirinden bağımsız olması, her bir maddenin içerdiği hata öğelerinin toplamının 0 olacağı varsayımına dayanır. Bunu aşağıdaki gibi örneklendirebiliriz:

$$\begin{aligned} X_1 &= G_1 + h_1 . \\ X_2 &= G_2 + h_2 . \\ X_3 &= G_3 + h_3 . \\ &\vdots \quad \vdots \\ \Sigma X &= \Sigma G + 0 . \end{aligned}$$

Günlük hayatta mükemmel paralelliğe sahip bir test oluşturmak gerçekleştirilmesi mümkün olmayan bir idealdir. Test kuramıyla ilgili literatürde yapılacak araştırmalar bilim adamlarının bu konuda önemli güçlüklerle karşılaştığını ortaya koyar.

Tau eşitliğine sahip olma özelliği. Herhangi bir ölçümde x_1 ve x_2 gözlem değerlerinin tau eşitliğine sahip olması; gerçek puanların eşit, fakat hata varyanslarının eşit olmadığı anlamına gelir (Eşitlik 1-4).

$$\tau_1 = \tau_2 \text{ fakat, } \sigma^2_{h_1} \neq \sigma^2_{h_2} . \qquad (1\text{-}4)$$

Paralel testlerde olduğu gibi tau eşitliğine sahip olan testlerde gözlem puanlarının kovaryansı gerçek puanların varyansına eşittir. Tau eşitliğine sahip testlerde hata varyanslarının eşit olmaması nedeniyle gözlem puanlarının varyansları da birbirine eşit değildir.

$$\sigma^2_{x_1} \neq \sigma^2_{x_2} . \qquad (1\text{-}5)$$

Tau eşitliğine sahip iki testin korelasyon katsayıları birbirine eşit değildir ve bu nedenle tau eşitliğine sahip iki testin güvenilirlik katsayıları farklı olabilir. Bir ölçekteki maddeler tau eşitliğine sahipse faktör analizi sonucunda faktör yükleri eşit çıkacaktır.[43]

Yaklaşık tau eşitliği. Yaklaşık tau eşitliği normal tau eşitliğine benzer; ancak *gerçek değer,* toplam puanlarda farklılaşarak ortaya *c* gibi sabit bir değer çıkmıştır. Bunun anlamı, gözlemlenen puanların ana kütleye ait ortalamalarının artık eşit olmadığıdır. Sonuçta gerçek puanların ortalamaları da eşit değildir (*bk.*, Eşitlik 1-6).

$$\tau_{1_p} = \tau_{2_p} + c. \qquad (1\text{-}6)$$

Gerçek puanlarda ilave sabit değerin bulunması nedeniyle gözlem puanlarının alındığı ana kütle ortalamaları ve gerçek puan ortalamaları eşit değildir. Yaklaşık tau eşitliğinde hata varyansları farklıdır. Alfa güvenilirliğinde maddelerin yaklaşık tau eşitliğine sahip olduğu varsayılır. Maddeler yaklaşık tau eşitliğine sahip değilse hesaplanan güvenilirlik katsayısı düşük değerliklidir.

Konjenerik ölçüm eşitliği. Jöroskog (1971) tarafından önerilen konjenerik ölçüm eşitliği tau ve yaklaşık tau eşitliğine benzer, ancak bu ölçümlerde gerçek puanlar doğrusal dönüştürme yöntemiyle farklılaştırılmıştır (*bk.*, Eşitlik 1-7).

$$\tau_{1_p} = a\tau_{2_p} + b. \qquad (1\text{-}7)$$

Konjenerik testlerde hata varyanslarının farklı olmasının yanı sıra, x_1 ölçümünün üçüncü bir değişkenle olan kovaryansı ile x_2 ölçümünün üçüncü bir değişkenle olan kovaryansı birbirine eşit değildir.

Klasik test kuramı; varsayımlarının zayıf olması, güvenilirlik ve geçerlilik rakamlarının ölçüm yapılan örnek kütleye bağımlı olması, maddelerin kalitesinin diğer maddelere bağımlı olması gibi nedenlerle eleştirilmiştir. Bu kuram, özellikle bir test düşük ve yüksek yetenekli kişilere uygulandığında *varyansların eşitliği* varsayımı nedeniyle zayıf olarak değerlendiril-

miştir. Test ve ölçeklerin güvenilirliği, örnekleme giren bireyler heterojen olduğunda daha yüksek, homojen olduğunda ise düşük çıkmaktadır.

Modern Ölçüm Kuramları

Modern ölçüm kuramları 1950'li yıllarda ortaya çıkmış, fakat daha çok 1970'li yıllardan sonra gelişme göstermiştir. Cronbach'ın geliştirmiş olduğu *genellenebilirlik* yaklaşımı ile *madde-yanıt kuramı* (MYK) modern ölçüm kuramları arasında en sık sözü edilenlerdir. Ancak bu konuda ilgili literatürü yakından takip etmek gerekmektedir. İstatistik ve psikometri dergilerinde matematik, cebir ve istatistik temelli olarak sürekli yeni modeller ve yaklaşımlar önerilmektedir.

Son yıllarda modern ölçüm kuramlarının gelişmiş olması klasik test kuramlarının (KTK) geçerliliğinin kalmadığı anlamına gelmez. Stage (1998) oldukça büyük örneklemlerde yaptığı araştırmalarda KTK ile MYK arasında kuramsal açıdan önemli farklılıklar bulunmakla birlikte sonuçların benzer çıktığını tespit etmiştir (aktaran Nelson).[44] Aynı bulguyu destekleyen başka araştırma bulguları da söz konusudur. Öyle anlaşılmaktadır ki, bilim adamı ölçümün güvenilirliğini klasik test kuramıyla inceledikten sonra belirlediği amaca göre ayrıca modern test kuramlarının yaklaşımlarından da yararlanabilecektir.

Genellenebilirlik kuramı. Cronbach ve arkadaşları (1972) tarafından geliştirilen ve modern ölçüm kuramı olarak isimlendirilen "genellenebilirlik" kuramında güvenilirlik daha geniş bir anlamda ele alınmıştır. Yapılan ölçümlerin "gerçek puana" ne ölçüde uygun olduğundan çok, ölçüm sonuçlarının *genelleme yapılacak evrene ne ölçüde uygun olduğu* konusu üzerinde durulur. Bu nedenle araştırmacılar, genelleme özelliğini kısıtlayan hata kaynaklarını araştırırlar. Bu kuramda, klasik test teorisinde olduğu gibi maddelerin paralel olduğu kabul edilmekle birlikte esas olarak hataların nereden kaynaklandığı ve hataların büyüklüğü tespit edilmeye çalışılır. Klasik test kuramında doğruluk ve yansızlık İngilizce "reliability" *güvenilirlik* sözcüğü ile ifade edilirken, genellenebilirlik kuramında doğruluğu tanımlamak için "generalizability" ve "dependability" kelimeleri tercih edilmiştir. İfadelendirmedeki nüans farkını Türkçede "genellenebilirlik" ve "dayanıklılık" terimleriyle karşılamayı uygun bulduk.

Bir ölçümde çok sayıda, belki yüzlerce ve hatta sonsuz denebilecek kadar potansiyel hata kaynağı vardır. Klasik test teorisinde hatalar iki grupta (tesadüfî ve sistematik olarak) sınırlandırılırken, genellenebilirlik

kuramında ikili bir sınıflandırmaya gidilmeden hataların araştırmacının yapacağı genelleme çalışmasına bağlı olarak genişleyebileceği belirtilmiştir. Ayrıca bu kuramda, saptanan tüm hataların bir şekilde tesadüfî veya sistematik hatayla bir ilintisi vardır. Klasik test kuramına göre ölçüm modelinin varyansı;

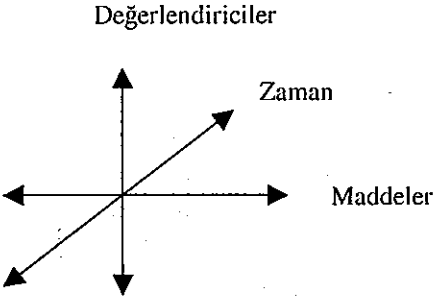
$$\sigma_x^2 = \sigma_g^2 + \sigma_h^2 \ , \qquad (1\text{-}8)$$

iken; genellenebilirlik kuramına göre ölçüm modelinin varyansı Eşitlik 1-9'daki gibi tanımlanmıştır.[45]

$$\sigma_x^2 = \sigma_e^2 + \sigma_{h1}^2 + \sigma_{h2}^2 + ... + \sigma_{hk}^2 \ . \qquad (1\text{-}9)$$

Klasik test kuramındaki "gerçek puan"ın yerini genellenebilirlik kuramında "evren" puanı almıştır (*bk.*, Şekil 1-4). Evren, araştırmacının genelleme yapmak isteği ana kütledir. Bir kişinin evren puanı, ölçüm sonuçlarının genellenebildiği tüm koşulların (değerlendiriciler, zaman, mekan) aritmetik ortalama puanıdır.

■ Evren puanı $(e) = \mu_{\text{değerlendiriciler, zaman, maddeler}}$.



Şekil 1-4. Evren puanını oluşturan bileşenler.

Kaynak. R. McCloy ve D. Putko, "Reliability Estimation [Güvenilirlik Tahmini]," <www.humrro.org/corpsite/download/ptc2002/ptcslides.ppt> (16.10.2002).

Ölçümde en önemli strateji, hataları mümkün olduğu kadar azaltmaktır. Aslında "hata" sözcüğü yanlış bir isimlendirmedir. Genellenebilirlik kura-

mında hata kavramı yerine, test/ölçüm durumunu etkileyen çeşitli "yönlerden" veya "yüzeylerden" söz etmek daha doğrudur. Modern ölçüm kuramında bu yönlerin veya yüzeylerin neler olduğu belirlenmeye çalışılır. Bunun için farklı *test koşullarında, farklı zamanlarda* ve *farklı araştırmacılar* tarafından yapılan ölçümlerde farklı yapıların ortaya çıkıp çıkmadığına bakılır. Genellenebilirlik kuramında araştırmacı şu sorulara cevap bulmaya çalışır:

1. Zaman içinde farklılıkların/değişkenliğin ortaya çıkma derecesi nedir?
2. Farklılıklar/değişkenlik ne ölçüde değerlendiricilerden kaynaklanmaktadır?
3. Farklılıklar/değişkenlik ne ölçüde uygulama koşullarından kaynaklanmaktadır?
4. Farklılıklar/değişkenlik ne ölçüde test maddelerinin kendisinden kaynaklanmaktadır?

Örneğin, *Stres Ölçeği* kişilere bir defasında normal çalışma zamanında ve daha sonra yoğun ve baskıcı iş koşullarında tekrar uygulanarak ne ölçüde farklılık ortaya çıktığı saptanmaya çalışılır. Genellenebilirlik kuramına göre bir testin güvenilirliği, testin/ölçeğin hangi koşullarda geliştirildiğine, uygulandığına ve yorumlandığına göre değişir (Cohen ve Sverdlik, 1999, aktaran Yazdani).[46] Genellenebilirlik kuramından yararlanmayı düşünen bir araştırmacı ölçüm uygulamasını bu kurama özgü belirli bir tasarım çerçevesinde gerçekleştirir. Basit araştırma tasarımları *tek yüzeyli* veya *iki yüzeyli* olarak düşünülür. Tek yüzeyli araştırma tasarımını örnek alacak olursak, *n* sayıda maddenin *k* sayıda kişiye uygulandığını düşünebiliriz. Bu çapraz tasarımda[a] maddelerin tümünün araştırmaya katılan kişilerin tamamına uygulandığı varsayılır ve bu tasarım *m* x *k* simgesi ile gösterilir. *İki yüzeyli* tasarımlarda ise yüzeylerin ögeleri arasında değişik uyuşma (veya yuvalanma) kombinasyonları söz konusudur.

Genellenebilirlik kuramında araştırmacı daha çok, sonuçların genellenebilirliği üzerinde durur, zaman içinde sistematik hataların ne ölçüde tekrar ettiğini araştırır. Genellenebilirlik kuramında farklıkları belirlemek için TEYVA Bileşenleri istatistik analiz tekniğinden yararlanılır. Modern

[a] Çapraz tasarım. Modele alınan değişkenlerle ilgili tüm şıkların veya birimlerin birbirileriyle ilişkili olmasıdır. Bir değişkendeki bazı şıklar diğer değişkendeki bütün şıklarla veya seçilmiş bir kaç şıkla ilişkili değildir. Bu ikincisi "yuvalanmış" tasarım olarak bilinir.

ölçüm kuramında güvenilirlik analizlerinin amacı, pratik nitelikteki sorulara yanıt bulmak ve testin kalitesi ve amaçlanan kullanımı hakkında belirli sorulara cevap vermektir. Araştırmacı bunun için kendisine şu soruyu sorar "Bu test/ölçek hangi amacı sağlamaya yönelik olarak güvenilirdir?"[47] Klasik test kuramından farklı olarak modern test kuramında güvenilirlik, pratik yararla ilişkilendirilmiştir.

Madde-yanıt kuramı. Son yıllarda geliştirilen modern ölçüm kuramlarından bir diğeri, tutum ölçekleri ve indekslerden çok standardize edilmiş başarı testleri ve skolastik testler için kullanılan *madde-yanıt kuramıdır* (MYK).[a] Bu yaklaşıma aynı zamanda *gizli özellikler kuramı*[b] da denir. Daha çok kişilik, yetenek, bilgi ve başarı testlerine ait maddelerin geliştirilmesinde kullanılır. Hollandalı istatistikçi ve matematikçi I.W. Molenaar MYK'nin gelişmesinde önemli ölçüde rol oynamıştır. Bu kuramda kişinin bir teste verdiği yanıtların *toplam puanı* değil, söz konusu test veya ölçeğin arka planındaki gizli yapılar ve modeller önemlidir. Gerçek puan modelinde, bir maddeye cevap veren kişilerin farklı yetenek düzeylerine sahip olmaları halinde o maddeyi nasıl işaretlemiş oldukları konusu üzerinde durulmazken MYK bu konuyu ele almıştır. MYK'de ölçek veya test değil, "madde"nin bizzat kendisi analiz yapılacak temel birim olarak görülür. Araştırmacı bu yaklaşımda bir maddeye gelen yanıtlardan hareket ederek cevapların öngörülen modele ne ölçüde uygun düştüğüne bakar. Öngörülen model, bir teste ait maddelere gelen yanıtlar ilk defa analiz edilirken Bayes yöntemine göre tahmin edilen değerlere ne ölçüde uygun olup olmadığına bakılarak belirlenir. Uygunsuzluklar yanıtlama hatası olarak değerlendirilir ve iki tür yanıtlama hatası vardır: maddelerin modele uygun düşmemesi (test puan hatası veya güvenilmezliği), kişi-model uyumsuzluğu (kişi hatası veya güvenilmezliği). Madde yanıt kuramında, aynı beceri düzeyine sahip kişiler, ait oldukları gruptan bağımsız olarak bir maddeye eşit oranda doğru yanıt vermişlerse söz konusu maddenin "yansız" olduğuna karar verilir.[48] Cevaplayıcıların %50'si söz konusu test maddesine "doğru" yanıt vermiş olmalıdırlar. Bu oran *eşik değeri* olarak kabul edilir. Test alan kişilerin etnik kökenleri, cinsiyetleri, sosyal statüleri ve ekonomik durumları farklı olabilir. Fakat bu kişiler eğer aynı yetenek ve beceri düzeyine sahiplerse maddeyi benzer oranlarda doğru olarak işaretlemiş olmalıdırlar. De-

---

[a] Item response theory.

[b] Gizli özellik kuramı (latent trait theory). Türkçeye bazı kaynaklarda *örtük özellikler kuramı* olarak çevrilmiştir.

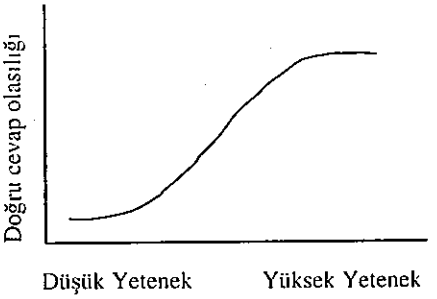
ğişik yetkinlik düzeylerinde maddelerin doğru yanıtlanma oranları da farklıdır. Bir gruptaki bireylerin yetkinlik düzeyleri z puanı cinsinden 0 ortalamaya ve 1 standart sapmaya sahiptir. Zayıf bir kişinin bir maddeye doğru cevap verme oranı %20 olasılığa ve –3,0 standart sapmaya sahip olabilir. Bir maddenin değişik gruplarda ne gibi bir davranış gösterdiğini ortaya çıkarmak için *diferansiyel madde fonksiyonu* (DMF)[a] analizi yapılır. İki tür *diferansiyel madde fonksiyonu* vardır: yeknesak fonksiyon ve yeknesak olmayan fonksiyon. *Yeknesak fonksiyonda* bir madde, yetenek/yetkinlik düzeyinden bağımsız olarak tüm gruplarda kişilere eşit oranda başarılı olma imkanı sağlar. Örneğin yetenek düzeyleri ne olursa olsun kadın ve erkeklerin bir maddede başarılı olma şansları eşit olabilir. Yeknesak olmayan fonksiyonda ise, bir maddenin gruplar ve yetenek düzeyleri açısından kişileri farklılaştırması söz konusudur. Gruplarla yetenek düzeyleri arasında bir etkileşim vardır, bu nedenle maddenin kişileri ayrıştırmasında *etkileşim etkisinden* söz ederiz. Örneğin bir maddeye düşük yetenekli erkekler daha fazla doğru yanıt verirken aynı maddeye yüksek yetenekli bayanlar daha fazla doğru yanıt vermiş olabilirler. Buna göre madde, kişileri hem cinsiyet hem de yetenek düzeyleri bazında farklılaştırmıştır.

Maddelere ilişkin parametre[b] tahminleri ana kütle parametrelerinden bağımsızdır. Madde-yanıt kuramı, klasik test kuramına göre daha güçlü varsayımlara sahiptir. Bu kuramda, maddelerin farklı kümelerde/gruplarda farklı sonuçlar verebileceğinden hareketle güvenilirlik, test/ölçek ile testi alan kişi arasındaki etkileşimin bir özelliği olarak değerlendirilir. Madde-yanıt kuramında *madde yanlılığı analizi* ile bir maddenin farklı gruplarda, örneğin erkeklerde ve bayanlarda veya üstün yetenekli ve düşük yetenekli kişilerde nasıl bir davranış göstereceği incelenir (*bk.*, Şekil 1-5).[49]

---

[a] Diferansiyel madde fonksiyonu – DMF (differential item functioning – DIF). Bir maddenin farklı gruplarda gösterdiği "davranışın" incelenmesi.

[b] Parametre. (1) Ana kütlenin sahip olduğu özellikler. Örneğin ana kütlenin aritmetik ortalaması, standart sapması varyansı gibi. (2) Bilgi. Bilgisayara girilen ve kullanıcıyı, değişkenleri, faktörleri tanımlayan veriler. (3) MYK'de test maddelerinin güçlük derecesi, ayırt edicilik değeri ve kişinin yetkinliğini gösteren teta değeri gibi teorik istatistiksel özellikler.



Şekil 1-5. Madde-yanıt kuramında yeteneklerle doğru cevaplar arasındaki ilişki.

Madde-yanıt kuramı, bir test sonucunun benzeri başka bir testte eş değer karşılığının bulunabileceğini öngörür. Bir kişinin bir testten elde edebileceği başarı, kişinin özellikleri tanımlanarak önceden tahmin edilebilir. Kişinin özellikleri doğrudan gözlenemediğinden bunlar gizli özellikler olarak tanımlanır. Madde-yanıt modelinde kişinin gözlenebilir test puanlarıyla gözlenemeyen özellikleri ve yetenekleri arasında bir ilişki olduğu varsayılır. Klasik yaklaşımda test oluştururken *gerçek puanın* tahminine dayalı olarak belirli sayıda doğru maddenin seçilmesi temel alınırken MYK'de test/ölçek maddeleri örnek kütlenin özelliklerine göre değiştirilebilir. Örneğin, zeki öğrencilere daha zor sorular içeren farklı bir test maddeleri uygulanıp sonuçta tüm kişiler aynı ölçek boyutunda sıralanabilir.[50] Madde-yanıt kuramı maddelerin özelliklerine ilişkin olarak "parametre" adı verilen dört temel bilgiyi kullanılır. Bunlar; bireysel yetkinlik parametresi, maddenin zorluk parametresi, maddenin ayırt edicilik parametresi ve maddenin yanıtlanmasında şans veya tahmin yürütme parametresidir.

Madde-yanıt kuramında bir dizi istatistiksel işlem sonucunda her bir test maddesi için "madde özellikleri eğrisi" (MÖE)[a] çizilir. MÖE grafiği normal olarak ham verileri işleyen SPSS ve Excell programlarıyla çizilmez, bu grafik matematiksel fonksiyonları işleyen özel matematiksel yazılımlarla üretilir.[b] MÖE grafiğine göre bir bireyin yetenek veya belirli bir özelliğe sahip olma derecesi arttıkça maddeye doğru bir şekilde cevap

[a] Madde Özellikleri Eğirisi (Item Characteristic Curve).

[b] Madde Özellikleri Eğrisi grafiğini çizmek için Perl, Winsteps, Biolog, SAS, DataDesk gibi yazılımlardan yararlanılabilir.

verme olasılığı da artar. Geliştirilen maddeler eğer bu özelliği sağlıyorsa bunlar güvenilir maddelerdir. Madde-yanıt kuramı klasik test kuramına göre şu avantajlara sahiptir:[51] Maddelerin parametre tahminleri, üzerinde araştırma yapılan örneklem grubundan bağımsızdır. Hesaplanan tahminler daha sonra başka örnek kütleler için de kullanılabilir. Ölçüm yapılan kişinin yeteneğine ilişkin tahminler, belirli madde ana kütlesinden seçilen maddelerden bağımsızdır. Bir ölçümde, yeteneğe ilişkin tahminin ne ölçüde duyarlı olduğu bilinebilir.

Madde yanıt kuramı modelleri üç grupta incelenir: tek parametreli Rasch modeli (MYK 1), iki parametreli Birnbaum modeli (MYK 2) ve üç parametreli Birnbaum modeli (MYK 3).

- MYK 1: Bu modelde maddelerin sadece *güçlüğü* göz önünde bulundurulur. Belirli bir yetenek düzeyendeki kişilerin doğru yanıt verme olasılığı araştırılır.

- MYK 2: Bu modelde maddelerin *güçlüğü* ve kişileri *farklılaştırma* durumu birlikte analiz edilir. Farklılaştırma, maddenin belirli bir yeteneğe sahip iki kişiyi ne ölçüde birbirinden ayırt ettiğidir.

- MYK 3: Bu modelde maddelerin *güçlüğü*, kişileri *farklılaştırma* durumu ve *şans/tesadüf* faktörü birlikte ele alınır. Şans faktörü, testi yanıtlayan kişinin bilerek değil, tahmin yürüterek cevaplama yapmış olmasıdır. Örneğin dört şıklı bir soruda ,25 şans faktörü söz konusudur.[52]

Madde yanıt kuramında, (a) çoktan seçmeli sorular, (b) *Doğru–Yanlış* şeklinde kodlanan ikili veriler ve (c) çok dereceli dereceleme ölçeği verileri analiz edilir. Madde yanıt kuramında bir ölçeğe fonksiyonu zayıf olan maddelerin yerine yapılacak analizlerden sonra yenilerinin konması işlemine *madde kalibrasyonu* adı verilir. Tek parametreli bir model için bir maddeyi en az 100 kişinin değerlendirmesi, iki parametreli bir testi 200-250 kişinin değerlendirmesi gerekirken üç parametreli bir model için söz konusu maddeyi en az 1000 kişinin alması gerektiği belirtilmiştir.[53] Ölçeği oluşturan kalibre edilmiş maddeler MYK parametre tahminlerini içeren istatistiklere sahiptir. İstatistiksel analizlerle güvenilirlikleri saptanmış olan maddeler "kalibre edilmiş maddeler havuzunda" saklanır.

Klasik test kuramında testin/ölçeğin tek bir güvenilirlik katsayısı vardır. MYK'de ise, bölgesel/yerel güvenilirlik değerleri söz konusudur. Diğer bir

deyişle, her bir noktadaki bilginin değeri önemlidir. [54] Madde yanıt kuramında ölçeğe, *en fazla bilgi veren* maddeler alınır.

ALINTI YAPILAN KAYNAKLAR

[1] F.N. Kerlinger, *Foundations of Behavioral Research*, (London:Holt, Rinehart and Winston, 1973), 455.

[2] R.K. Henson, "Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha [İç Tutarlılık Tahmin Değerlerinin Anlaşılması: Alfa Katsayısı Üzerine Kavramsal Bir İnceleme]," *Measurement and Evaluation in Counseling and Development*, Oct 2001, 177-189.

[3] A. Yu, "Reliability and Validity [Güvenilirlik ve Geçirlilik]," <http://seamonkey.ed.asu.edu/~alex/teaching/assessment/reliability.html> (02.09.2002).

[4] A. Marradi, "Reliability: A Dissenting View [Güvenilirlik: Karamsar Bir Görüş]," *Bulletin de Methodologie Sociologique*, (Eyl 1990), 56-71.

[5] T.A. Ackerman, "Testing and Measurment Issues [Test ve Ölçüm Sorunları]," <http://orme.uark.edu/aera2001rct.pdf.> (02.09.2002).

[6] M. Fichman, "Research Methods in Behavioral Sciences [Davranış Bilimlerinde Araştırma Yöntemleri]," <http:mario.gsia.cmu.edu/html-end/html/lects/reliability.pdf> (28.03.2003).

[7] R. Likert, "The Method of Constructing an Attitude Scale [Tutum Ölçeklerinin Oluşturulmasında Yöntem]," Der., . Fishbein, *Attitude Theory and Measurement*, (New York: John Wiley, 1967), 91.

[8] L.L. Thurstone, "Measurment of Social Attitudes [Sosyal Tutukların Ölçülmesi]," Der., M. Fishbein, *Attitude Theory and Measurement*, (New York: John Wiley, 1967), 23.

[9] L.J. Cronbach, *Essentials Psychological Testing* [Psikolojik Testlerin Esası], (New York: Harper and Row,1970), 154.

[10] R.A. Peterson, "A Meta-Analysis of Cronbach's Alpha [Cronbach Alfa İçin Meta Analizi]," *J. of Consumer Research*, 21(2), 381.

[11] L.M. Rudner, ve W.D. Schafer, "Reliability [Güvenilirlik]," <http://www.ed.gov/databases/ERIC_Digests/ed458213.html> (15.09.2002).

[12] Aynı.

[13] D. Dunworth, "Reliability and Validity [Güvenilrlik ve Geçerlilik]," <http://www.promissor.com/knowledge/askdrpsi/drcat20000918.asp> (21.12.2002).

[14] M.A. Syverson ve M. Barr, "How does the Learning Record Model Compare with Existing Methods of Measurement in Assessing Student Literacy Learning? [Öğrenme Kayıt Modelinin Öğrencilerin Başarılarını Değerlemede Mevcut Ölçüm Modelleriyle Karşılaştırılması]," <http://www.cwrl.utexas.edu/~syverson/olr/compare.html> (15.09.2002).

[15] D. Arkkelin, "Measurement in Research [Araştırmalarda Ölçüm]," <http://www.valpo.edu/home/faculty/darkkeli/syllabi/methods/p202outlines/ch8.htm> (25.09.2002).

[16] Toledo Universitiy, "A Brief History of Educational Measurement [Eğitimde Ölçme Uygulamalarının Kısa Bir Tarihi]," <http://homepages.utoledo.edu/CFOX2/history.htm> (30.11.2002).

[17] K. Vehkalahti, "Reliability of Measurement Scales [Ölçeklerin Güvenilirliği]," <http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/reliabil.pdf> (14.09.2002).

[18] Vehkalahti, "Reliability of Measurement."

[19] Vehkalahti, "Reliability of Measurement."

[20] Georg E. Matt, "Generalizability Theory [Genellenebilirlik Kuramı]," <http://www.psychology.sdsu.edu/faculty/matt/Pubs/GThtml/GTheory_GEMatt.html> (29.03.2003).

[21] A. Ferligoj ve A. Mreevar, "Assesment of Reliability [Güvenilirliğin Değerlendirilmesi]," <http://mrvar.fdv.uni-lj.si/pub/mz/mz15/socan.pdf.> (28.09.2002).

[22] Aynı.

[23] Aynı.

[24] Aynı.

[25] A. Yu, "Reliability and Validity [Güvenilirlik ve Geçerlilik]," <http://seamonkey.ed.asu.edu/~alex/teaching/assessment/reliability.html>

[26] D. Garson, "Scales and Standard Measures [Ölçekler ve Standart Ölçümler]." <http://www2.chass.ncsu.edu/garson/pa765/standard.htm> (15.09.2002).

[27] W.M. Rogers, N. Schmidt ve M. E. Mullins, "Correction for Unreliability of Multifactor Measures:Comparison of Alpha and Parallel Forms Approaches [Çok Faktörlü Ölçümlerde Güvenilirlik Düzeltmesi]," <http://io.psy.msu.edu/Schmitt/ormalpha.htm> (15.09.2002).

[28] D. Garson, "Scales and Standard Measures [Ölçekler ve Standart Ölçümler]," <http://www2.chass.ncsu.edu/garson/pa765/standard.htm> (07.09.2002).

[29] M. Stommel ve Diğerleri, "Confirmatory Factor Analysis as a Method to Asses Measurement Equivalencce [Ölçüm Eşdeğerliliğini Ölçme Yöntemi Olarak Teyit Edici Faktör Analizi]," <http://www.hsl.wisc.edu/ereserves_ content/nursing/fall2002/ N991/ N991_21F02.pdf>

[30] Willamette University, "Measurement: Reliability, Validity and Objectivity [Ölçüm: Güvenilirlik, Geçerlilik ve Nesnellik]," <http://www.willamette.edu/cla/exsci/356/356lec25.htm> (05.01.2003).

[31] B. Trochim, "How Stable and Consistent Is Your Instrument? [Ölçüm Aracımız Ne Ölçüde İstikrarlı ve Tutarlı]," <http://trochim.human.cornell.edu/tutorial/johnson/melody.htm> (27.07.2003).

[32] PennState, University Testing Services "Classical Test Theory Approach [Klasik Test Kuramı Yaklaşımı]," <http://www.uts.psu.edu/Classical_theory_frame.htm> (23.04.2004).

[33] C. Yu, "Reliability of Self-report Data [Cevaplayıcının Kendi Bildirimine Dayanan Verilerin Güvenilirliği9," <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/memory.html> (07.09.2002).

[34] P. Kline, *Handbook of Psychological Testing* [Psikolojik Test Elkitabı], (New York: Routledge, 1993), 29.

[35] Chung-Cheng University, "Reliability [Güvenilirlik]," <http://psy.ccu.edu.tw/testroom/Reliability.doc> (05.01.2003).

[36] W.M.K. Trochim, "Measurment Error [Ölçüm Hatası]," 2002,<http://trochim.human.cornell.edu/kb/measerr.htm> (08.09.2002).

[37] E. Van Tilburg Norland, "Controlling Error in Evaluation Instruments [Ölçüm Araçlarını Değerlerken Hatanın Kontrol Edilmesi]," <http://www.joe.org/joe/1990summer/tt2.html> (08.09.2002).

[38] B. Trochim, "Theory of Reliability [Güvenilirlik Kuramı]," <http://trochim.human.cornell.edu/kb/reliablt.htm> (07.09.2002).

[39] C. Yu, "Reliability of Self Report Data [Kişisel Anketerin Güvenilirliği]," <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/memory.html> (29.10.2002).

[40] MEİ, "Measurement Theories [Ölçüm Kuramları]," <http://www.measurementexperts.org/ClassicalTestTheory.htm> (13.10.2002).

[41] Penn State University, "Reliability [Güvenilirlik]," <http://www.personal.psu.edu/users/z/x/zxt105/HDFS526/CTT2.ppt> (23.12.2002).

[42] A. Pickering, "Classical Test Theory [Klasik Test Kuramı]," <http://homepages.gold.ac.uk/aphome/cttnotes.doc> (23.10.2002).

[43] MEİ, "Classical Test Theory [Klasik Test Kuramı]," <http://www.measurementexperts.org/ClassicalTestTheory.htm.> (09.12.2002).

[44] L.R. Nelson, "Some CTT and IRT Comments [KTK ve MYK Hakkında Bazı Yorumlar]," <http://www.lertap.curtin.edu.au/Documentation/IRTandCTTComments.htm> (24.10.2002).

[45] HumRRO, "Reliability."

[46] S. Yazdani, "Reliability [Güvenilirlik]," <http://www.atgci.org/medical%20education/reliability.ppt> (09.10.2002).

[47] Bärbel Knäuper, "Methods for Estimating Reliability [Güvenilirliği Tahmin Etmek İçin Yöntemler]," <http://www.psych.mcgill.ca/perpg/fac/knaeuper/psychtest/powerpoint/Lecture09_ReliabilityII.rtf> (13.10.2002).

[48] A. Yu, "True Score Model and Item Response Theory [Madde Yanıt Kuramı ve Gerçek Puan Modeli]."

<http://seamonkey.ed.asu.edu/~alex/teaching/WBI/measurement.html> (26.11.2002).

[49] M. Brannic, "Item Response Theory [Madde Yanıt Kuramı]," <http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm> (29.10.2002).

[50] Aynı.

[51] Alpine Media Corporation, "Psychometric Analysis [Psikometrik Analiz]," <http://www.alpinemedia.com/psy_certification_pat2.htm> (29.10.2002).

[52] Scrolla, Heriot-Watt University, "Item Analysis [Madde Analizi]," <http://www.scrolla.hw.ac.uk/focus/ia.html> (30.11.2002).

[53] Promissor, "Item Characteristic Curves [Madde Özellikleri Eğrisi]," <http://www.promissor.com/knowledge/askdrpsi/drcat20010323.asp> (30.11.2002).

[54] M. Brannic, "Item Response Theory [Madde Yanıt Kuramı]," t.y., <http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm> (24.10.2002).

ÖLÇÜM DÜZEYLERİ, ÖLÇÜM ARAÇLARI VE DERECELENDİRME

Her tür ölçme işleminde ve her tür ölçüm aracında güvenilirlik analizlerini yapmaya gerek yoktur. Klasik test kuramında güvenilirlik değerlendirmesi ağırlıklı olarak yansıtıcı ölçek niteliğindeki eşit aralıklı ölçek verileri için söz konusudur. Örneğin, klasik test kuramına göre maddeleştirilmiş ölçekler veya tek maddeli soru listeleri için güvenilirlik analizi yapılmaz. Aynı şekilde belirli nitelikteki oluşturucu ölçek maddeleri için de klasik teorideki güvenilirlik analizi yöntemlerini uygulamanın anlamı yoktur. Güvenilirlik değerlendirmesinde göz önünde bulundurulması gereken önemli bir diğer nokta ölçek dereceleridir. İkili veya çoklu ölçek derecelerinin güvenilirliği ne şekilde etkilediği araştırılmadan rasgele ölçek derecesi belirlenmemelidir. Bu tür sorunlara açıklık getirmek için bu bölümde klasik test kuramına göre ölçüm düzeyleri, ölçüm araçları ve derecelendirme konularının güvenilirlikle olan ilişkisi ele alınmıştır.

ÖLÇÜM DÜZEYLERİ VE GÜVENİLİRLİK

Araştırmacı, veri toplama aracının özelliğine göre değişik düzeylerde ölçüm yapabilir. Ölçüm düzeyleri bir psikofizikçi olan S.S. Stevens (1946) tarafından dört grup altında toplanmıştır: sınıflandırılmış ölçümler, puanları büyüklük sırasında değerlendirilen ölçümler, eşit aralıklı ölçümler ve oranlı ölçümler. Bu ölçümlerden elde edilen veriler ölçüm düzeyine uygun olarak aynı adla adlandırılır. Buna göre sınıflandırılmış, sıralı, eşit aralıklı veya oranlı ölçüm verilerinden söz ederiz.

Sınıflandırılmış Verilerde Güvenilirlik

Sınıflandırılmış (nominal) ölçümlerde maddeler *Evet-Hayır*, *Bayan-Erkek* veya *Başarılı-Başarısız* gibi ikili veya çok dereceli belirli kategorilere sahiptir. Bu şıklara gelen yanıtlar ana kütlede normal dağılım özelliği göstermez. Bilim adamı ölçüm maddelerini ve bu maddelere ait şıkları / cevaplama seçeneklerini herhangi bir sisteme bağlı olmadan serbestçe belir-

ler. Ölçüm yapılan maddeler iki gruptan birine girer. Maddeler ya birbirinden bağımsızdır veya en az üç dört tanesi bir araya getirilerek bu maddelerle bir ölçek oluşturulmuştur. Birbirinden bağımsız olan maddelerin cevap şıkları ise ikili veya çoklu olabilir. İkili veri yapılarında iki derece ve çoklu veri yapılarında ise ikiden fazla derece (şık) vardır.

■ Bağımsız nominal maddelerde cevap şıkları:

İkili veri yapısı	*Çoklu veri yapısı*
Cinsiyetiniz?	Yaşınız?
Bayan ..... (1)	20'den küçük ..... (1)
Erkek ..... (2)	20-25 .....(2)
	25'ten büyük.....(3)
Kurumunuzda kalite sistemi var mı?	Kalite sistemine inanıyor musunuz?
Evet ..... (1)	Evet, kesinlikle ..... (1)
Hayır.....(2)	Kısmen .....(2)
	Hayır .....(3)

Sınıflandırılmış verilerin önemli bir bölümü demografik sorularla ilgilidir. Bu bölümde önce demografik değişkenlerin güvenilirliği ele alınmış, daha sonra ölçek şeklinde düzenlenmeyen ve ölçek şeklinde düzenlenen nominal verilerin güvenilirliği konusu incelenmiştir.

Demografik değişkenler ve güvenilirlik. Demografik değişkenlere ait veriler, güvenilirliğin sadece belirli yönleriyle ilgilidir. Demografik değişkenlerde iç tutarlılık, yarıya bölme, alfa katsayısı gibi *ölçeklere özgü* yöntemler uygulanmaz. Demografik veriler bağımsız değişkenlerdir, iç tutarlılık güvenilirliği daha çok tutum ölçeği niteliğindeki bağımlı değişkenler için söz konusudur. Demografik değişkenlerin güvenilirliği, araştırmanın "yöntem ve özellik hatalarından arındırılmasıyla" ilgilidir. Duruma göre demografik veriler belirli bir örneklem büyüklüğünden elde edilmeli ve ana kütleyi temsil etme kabiliyetine sahip olmalıdır. Bazı araştırmacılar, demografik veriler için *güvenilirlik* sözcüğü yerine *veri kalitesi* ifadesini kullanma eğilimindedirler. Demografik veriler; etnik grup, cinsiyet, medenî hâl, yaş grubu, akademik branş, meslek, eğitim, deneyim, kıdem gibi değişkenlerden oluşur. Bu değişkenler gizli bir yapıyı değil; somut bir gerçeği ölçer ve bu nedenle ölçeklerde olduğu gibi paralel formlar, yarıya

bölme veya iç tutarlılık analizleri yapılmaz. Bu değişkenler sadece farklılıkları gösterir. Bir ölçüm sürecinde demografik değişkenler üç alanda bilim adamlarının araştırmalarına konu olur: nüfus sayımında, kuram geliştirme araştırmalarında ve demografik değişkenlere dayalı norm geliştirme çalışmalarında.

Nüfus sayımı ve seçim araştırmalarında katılımcıların demografik özelliklerini belirlemek önemlidir. Bu tür araştırmalarda verilerin kalitesi veya güvenilirliği bu verilerin ana kütle dinamikleriyle tutarlı olmasına bağlıdır. Demografik veriler sürekli değişmektedir, ancak örneklem seçilerek yapılan çalışmalarda bu değişikliklerin ne yönde olduğunu ekonometrik yöntemlerle güvenilir bir şekilde tahmin etmek mümkündür. Öte yandan tam sayım yapılmadığı durumlarda seçilen örnekleme ait demografik veriler belirli ölçüde hata içerir. Bu verilerin daha az hata içermesi ve sonuçta güvenilir olabilmesi için belirli örneklem büyüklüğüne sahip olması gerekir. Ana kütle dinamikleri uzun zaman diliminde değişebileceğinden kısa zaman diliminde (örneğin bir yıl gibi), toplanan veriler daha güvenilirdir. Araştırma örnek kütlesindeki oranların eğer ana kütleye genelleme yapılacaksa, ana kütledeki kadın ve erkeklerin dağılımına, katılımcıların ana kütledeki yaş dağılımına, işsizlerin dağılımına, boşanmaların dağılımına, suç oranlarının dağılımına, şeker hastalarının dağılımına gibi belirli kriterlere uygun düşmesi gerekir.

Kuram geliştirme araştırmalarında ise, esas amaç belirli bir hipotezi test etmek olduğundan bunun yanında katılımcıların demografik özellikleri de saptanmak istenir. Bu tür araştırmalarda demografik değişkenler ikincil bir öneme sahiptir. Demografik verilerin ana kütle oranlarını yansıtması gibi bir amaç peşinde koşulmaz. Çünkü söz konusu demografik değişkenlerin ana kütledeki dağılımına ilişkin herhangi bir tam sayım çalışması yapılmamıştır veya yapılması da imkansız olabilir. Eğer sonuçlar demografik değişkenler bazında genellenmek isteniyorsa o zaman bu değişkenlerin ana kütleyi temsil etmesi gerekir.

Demografik değişkenlere ait bilgiler doğrudan ilgili kişilerden toplanmışsa bu bilgilerin hatasız olduğu varsayılır.[1] Demografik değişkenler herhangi bir *özelliğin* daha iyi, daha kötü veya daha olumlu, daha olumsuz olma durumunu göstermez. Bu değişkenlerin şıklarından biri diğerinden daha doğru değildir. Demografik değişkenler bize çoğunlukla örneklemin yanlılığı hakkında bilgi verir. Araştırmanın amacı, okuyan ve okumayan tüm gençlere ulaşmak ve onların görüşlerini derlemek iken, anket çoğunlukla lise mezunu gençlere uygulanmışsa ölçüm değişkenlerinin sonuçları yanlı demektir. Araştırmaya katılanlara doğrudan yaşlarının ne olduğuna

ilişkin bir soru sorulmuşsa bu soruya bayan katılımcıların doğru yanıt vermemeleri halinde sistematik hata ortaya çıkacak ve verilerin güvenilirliği tehlikeye girecektir. Bazen demografik veriler ana kütleye genelleme yapmak için kullanılır. Özellikle alan araştırmalarında cinsiyet, yaş, medenî hâl, eğitim, ekonomik durum gibi faktörler açısından örnek kütlenin ana kütleyi tam olarak temsil etmesi gerekir. Örnek kütlenin demografik özellikler açısından temsil edicilik özelliği ortadan kalktığı durumda veriler ana kütleye genellenemez ve araştırmanın güvenilirliği sorunu gündeme gelir. Özellikle seçim araştırmalarında örneklemin demografik özelliği önem kazanır. Araştırmaya katılan kişilerin şehrin yerlisi veya misafir olması, yaş dağılımları, meslekleri, okur yazarlık durumu, gelir düzeyleri, kültürel özellikleri (şehre yeni göç etmiş olması, geniş aile yapısına sahip olması, dinî inanışları, yurt dışına çıkıp çıkmadıkları gibi faktörler) sonuçlar üzerinde büyük ölçüde etkili olur. Seçim araştırmalarının güvenilirliği katılımcıların demografik özellikler açısından yöresel toplumu temsil etme özelliğine bağlıdır.

Geliştirdiği psikometrik bir test için norm değerlerini oluşturmak isteyen bir araştırmacı ise, testin güvenilirliğini sağlam bir zemine dayandırmak için yaş, cinsiyet ve meslek gibi temel alınan gruplarda temsil edicilik ve örneklem büyüklüğü değerlerini dikkate almalıdır. Test normları, çoğunlukla temsili amaçlanan ilgili demografik gruplar için geçerlidir.

Ölçek şeklinde düzenlenmeyen maddelerde güvenilirlik. Bu uygulamada nominal ölçek maddeleri birbirlerinden bağımsızdır. Her bir madde başka bir kavramsal alanla ilgilidir. Maddeler demografik değişkenlerle veya bir konunun değişik yönleriyle birlikte ele alınarak analiz edilir. Ölçek şeklinde düzenlenmeyen maddeler ikili veya çoklu dereceye sahip olabilir. Bunun için söz konusu maddelerde veri yapısına uygun istatistiksel analiz yöntemi seçilir.

Ölçek şeklinde düzenlenmeyen sınıflandırılmış verilerin güvenilirliği, tekrar edilen ölçümlerdeki veya farklı kişiler tarafından yapılan değerlendirmelerdeki tutarlılığı belirlemeye yöneliktir. Sınıflandırılmış verilerin güvenilirliğini ölçmek için aşağıdaki teknikler kullanılır:

1. Oranlara dayalı güven aralığı.
2. Uyuşma yüzdesi.
3. Cohen kappa formülü.
4. Ki-kare istatistiksel analizleri.

Oranlara dayalı güven aralığı. Bu yaklaşım bir güvenilirlik analizi değildir, ancak elde edilen yüzde değerlerinin gerçekte hangi oranlar arasında dağılabileceğini göstererek rakamların daha doğru okunmasına imkan sağlar. Oranların güven aralığı, $Z_{\alpha/2} \times \sqrt{p.q/n}$ hata marjı formülü ile hesaplanır. Formülde p değeri ölçülmek istenen özelliğin oranını gösterir ve q ise $1-p$ anlamındadır. Örneğin, 85 kişi üzerinde yapılan bir araştırmada katılımcıların %65'i AB üyeliğine evet demiştir. Ana kütlede *Evet* yanıtını veren kişilerin %95 güven aralığında gerçek oranını bulmak için önce Eşitlik 2-1 ile hata marjı (HM)hesaplanır.

$$HM = 1{,}96 \times \sqrt{,65.,35/85}\ , \qquad (2\text{-}1)$$

$$HM = ,10\ .$$

Hata marjı bulunduktan sonra bu değer ölçülmek istenen özelliğin saptanan oranıyla toplanarak ve çıkarılarak ortalamanın güven aralığı (OGA) hesaplanmış olur.

$$OGA = ,65 + ,10 = ,75\ ,$$
$$OGA = ,65 - ,10 = ,55\ ,$$
$$OGA = ,75 - ,55.$$

Uyuşma yüzdesi. Gözlemcilerin yaptıkları değerlendirmelerin güvenilirliğini belirlemeye yöneliktir. Gözlemcilerin veya değerlendiricilerin uyuştukları madde sayısının toplam değerlendirme sayısına olan oranıdır. Bu yöntemin olumsuz yönü, tesadüfî uyuşmaları dikkate almamasıdır.

Cohen Kappa. Kappa istatistiği iki veya daha fazla gözlemci arasındaki uyuşmayı belirlemek için kullanılır. Kappa yönteminin uyuşma indeksinden farkı, şans faktörünün etkisini ortadan kaldırmasıdır. Kappa hesaplama yöntemi için "Güvenilirlik ve Korelasyon Analizleri" bölümünden daha fazla bilgi edinilebilir.

Ki-kare istatistik analizleri. İstatistiksel analiz programı SPSS'te ki-kare analizlerine dayalı olarak üç test değeri elde edilir: (a) Phi katsayısı, (b) kontenjan katsayısı, (c) Cramer *V* değeri. Phi katsayısı 2x2 şeklindeki veriler için uygunken, kontenjan katsayısı ve Cramer *V* değeri daha büyük

tablolar için uygundur. Farklı iki gözlemci, yarışmacılar için *Başarılı – Başarısız* şeklinde değerlendirmelerde bulunduklarında verilen puanlar arasındaki korelasyon/ilişki, phi değeri ile gösterilir. Diğer korelasyon katsayılarında olduğu gibi phi katsayısı 0 ilâ 1 arasında değişir. Öte yandan kontenjan katsayısı değerlerinin, oluşturulan tablonun büyüklüğünden etkilenmesi nedeniyle büyük tablolarda kullanılmaması önerilmiştir. Bir analizde 2x2 tablosu için örneğin ,74 katsayısı elde edilirken 5x5 tablosu için ,81 gibi bir değer çıkabilir.[2] Büyük tablolarda daha çok Cramer *V* değerleri kullanılır. Üç veya daha fazla dereceli ölçekler için kullanılan Cramer *V* değerinin uygulanabilmesi için iki ön koşul vardır. Bunlardan birincisi verilerin nominal nitelikte olması ve ikincisi ise verilerin tesadüfî olarak seçilen örneklemlerden elde edilmesidir. Cramer *V* değeri ayrıca ki-kare istatistik analizinin sınırlamalarına tâbidir. Verilerin dağılımı hakkında herhangi bir varsayımda bulunmadığından Cramer *V* nominal değişkenler arasındaki ilişkileri incelemek için iyi bir testtir.

Ölçek şeklinde düzenlenen ikili veri yapılarında güvenilirlik. Bazı veriler *Evet-Hayır*; *Doğru-Yanlış*; *Katılıyorum-Katılmıyorum* gibi iki şıklı bir niteliğe sahiptir ve bu şekildeki en az üç veya daha fazla madde bir araya getirilerek bu maddelerden bir ölçek oluşturulmuştur. Verileri ikili niteliğe sahip *ölçeklerde* güvenilirlik analizi için aşağıdaki hesaplama yöntemlerinden yararlanılır:

1. KR-20.
2. KR-21.
3. Madde-yanıt kuramı ve/veya Rasch formülü.
4. Loevinger *H* katsayısı.

Sayılan hesaplama yöntemlerinden ilk ikisi nispeten basit iken son iki yöntem daha karmaşık bir niteliğe sahiptir. Yaygın kullanılan istatistiksel analiz programlarında bu hesaplama yöntemleri bulunmaz. Örneğin, istatistiksel analiz programı SPSS'teki Reliability mönüsü altında KR-20 veya KR-21 için ayrı bir hesaplama seçeneği yoktur. Ancak, bu modüldeki *alfa güvenilirlik katsayısının* aynı zamanda iki şıklı değişkenler için de güvenilirlik hesaplamasını başarılı bir biçimde yaptığı bildirilmiştir. Bununla birlikte İnternet'te, KR-20 formülünü uygulayarak el ile yapılan bazı araştırmalarda aynı sonucu alamadıklarını bildiren yazarlar vardır. Bu nedenle SPSS'te iki şıklı değişkenler için güvenirlilik analizi yapılırken elde edilen sonuçların el ile yapılan hesaplama sonuçlarıyla doğrulanmasında yarar

vardır. Rasch formülü de SPSS'te bulunmamaktadır. Rasch'ın tek boyutluluk önermesini test etmek için iki testten yararlanılır: Martin-Löf testi (M-L testi) ve Madde Ayrıştırma Tekniği.[a] Bu tekniklerin uygulanması literatürde henüz belirli bir yaygınlığa kavuşmamıştır.[3] Bunun dışında güvenilirlik tahmini için *logaritmik benzerlik oranı*,[b] *logaritmik doğrusal-maksimum*[c] ve *en küçük kareler*[d] yöntemi gibi tekniklerden yararlanılabilir. Loevinger *H* katsayısının ise bu amaçla geliştirilmiş olan MSP isimli yazılımla hesaplandığı bildirilmiştir.

KR-20 formülü. Kuder-Richardson tarafından geliştirilen bu formülün uygulanabilmesi için veriler 0 ve 1 şeklinde kodlanmalıdır. KR-20 formülü aynı zamanda alfa güvenilirlik değeri olarak bilinmekle birlikte teknik anlamda bu tam doğru değildir. KR-20 formülü sadece iki şıklı değişkenlere uygulanabilirken alfa güvenilirlik formülü aynı zamanda çok dereceli değişkenler için de kullanılabilmektedir. KR-20 formülüyle ilgili temel varsayım, ölçüm maddelerinin sadece tek bir yapıyı ölçüyor olmasıdır. Bunun için maddelerin içerikleri benzer olmalıdır. Soruların niteliği birbirinden önemli ölçüde farklı olarak hazırlanan matematik ve Türkçe / Edebiyat gibi derslerin testlerinde bu güvenilirlik analizinin kullanılması doğru değildir.

KR-21 Formülü. KR-21 formülü özellikle eşit zorluğa sahip maddelerden oluşan ve sınıf ortamında uygulanan çoktan seçmeli testler için uygundur. Bu formülün uygulanabilmesi için testteki soruların zorluk derecelerinin eşit olması gerekir. Bu güvenilirlik hesaplama yönteminde bireysel olarak test/ölçek maddeleri dikkate alınmaz. KR-20 formülüne göre daha düşük değerlikli katsayılar elde edilmesine karşılık KR-21 basit ve güvenilir bir formüldür. KR-21 formülünün uygulanabilmesi için testteki madde sayısına (K), maddelerin aritmetik ortalamasına ve standart sapma (SS) değerlerine ihtiyaç vardır. KR-21 formülü, Eşitlik 2-2'de verildiği gibidir:

$$\text{KR-21} = (K * SS^2) - [\text{Ortalama} * (K - \text{Ortalama})] / (K-1) * SS^2 . \qquad (2\text{-}2)$$

---

[a] Splitter-item-technique.

[b] Log-likelihood ratio.

[c] Logit-linear maximum.

[d] Least-squares.

Amerikan Psikoloji Derneği, bir kişinin yetkinliği hakkında karar vermek için kullanılan testlerde KR_{20} veya KR_{21} katsayısının en az ,70 olması gerektiğini bildirmiştir. Testin KR_{20} güvenilirlik katsayısı ,80 ise "iyi", eğer ,90 ise "çok iyi" olarak değerlendirilir.[4] Nadir hallerde, negatif KR_{20} katsayılarıyla karşılaşıldığı bildirilmiştir. KR_{20} katsayısının negatif çıkması bu ölçümle ilgili varsayımların karşılanmadığı anlamına gelir.[5]

KR-20 veya KR-21 formülü testin genel olarak güvenilirliği hakkında bilgi verir, tek tek maddelerin güvenilirliği hakkında bir fikir vermez. Bir araştırmacı ölçek oluştururken iki şıklı (dereceli) bir ölçek oluşturmayı düşünüyorsa bu ölçeklerin sorunlu olduğunu unutmamalıdır. Örneğin *Doğru-Yanlış* şeklindeki veya tek bir doğrunun bulunduğu çoktan seçmeli sorularda KR-20 formülünün uygulanması halinde, elde edilen katsayı, güvenilirliğin tam bir göstergesi olmayabilir.[6] Bu tür ölçeklerde standart Pearson temelli yaklaşımlar yerine tetrakorik korelasyon matrislerinin kullanılması önerilmiştir. Tetrakorik korelasyon analizi, için veriler istatistiksel analiz programının çalışma sayfasına 1 ve 0 şeklinde kodlanır. Bu analizde iki şıklı değişkenlerin arka planında yatan gizli değişkenler arasındaki ilişki hesaplanır. Tetrakorik korelasyon[a] analizi SPSS'te bulunmamaktadır. Araştırmacıların bu amaçla hazırlanmış özel yazılımlardan yararlanmaları gerekir.

Rasch ve madde-yanıt kuramı modelleri. Rasch modeli, modern test yaklaşımlarından *gizli özellik kuramı*[b] çerçevesinde daha çok iki şıklı değişkenleri değerlendirmede kullanılan bir yaklaşımdır. Gizli özellik modellerinin bir diğer türü "madde-yanıt kuramı"dır. İkili değişkenler madde-yanıt kuramı çerçevesinde 1PL, 2PL ve 3PL modelleriyle de değerlendirilebilir.

Rasch modeli esasında iki dereceli değişkenler için kullanılır. Bununla birlikte yöntemin sahip olduğu özelliklerden ve nesnelliğinden hiçbir şey kaybetmeksizin çok kategorili maddeler için de uygulanabileceği belirtil-

---

[a] Tetrakorik korelasyon analizi, *Evet – Hayır* şeklinde yanıtlanan sorular arasındaki korelasyonu belirlemek için kullanılır. İlişki katsayısı normal korelasyon katsayısından daha yüksek çıkması nedeniyle Nunnally tarafından kullanılmaması tavsiye edilmekle birlikte istatistikçiler matematiksel model geliştirmek amacıyla bu tekniği kullanmaya devam etmektedirler. Bu tekniği kullanmak isteyen araştırmacılar, Statistica isimli yazılımın Reliability and Item Analysis mönüsünden yararlanabilirler.

[b] Gizli özellik modelleri (latent trait models).

sahiptir ve zorluk derecesi maddelere gelen olumlu yanıt oranlarına da yansır. Maddelere olumlu yanıt alma oranın artmasıyla birlikte gizli özelliğe ait değerin de arttığı düşünülür. Ölçeğin tek boyutluluğu veya türdeşliği *Loevinger H* katsayısı ile belirlenir. Elde edilen *H* istatistiği verilerdeki ölçekleme hatalarını dikkate alır. Mokken ölçeklerinin güvenilirlik analizi ve modele uygun olmayan maddelerin çıkarılması için MSP isimli yazılım geliştirilmiştir. Mokken'e (1971) göre, ölçeğin güvenilirliğini veya maddelerin "ölçek haline gelmiş olduğunu" (ölçeklendiğini) belirleyen kriterler aşağıdaki gibidir:[18]

- Loevinger *H* katsayıları.

,50 < *H*	: güçlü ölçek
,40 < *H* < ,50	: orta derecede güçlü ölçek
,30 < *H* < ,40	: zayıf ölçek
H < ,30	: oluşmamış ölçek

Mokken ölçeklerinde güvenilirliği hesaplamak için ayrıca *küme içi korelasyon analizi* yöntemine de başvurulabilir. Son yıllarda Mokken ölçekleriyle ilgili hesaplamaları yapabilmek için SPSS'e bir ek niteliğinde STAP isimli yazılım geliştirilmiştir (STatistical Appendix to SPSS).[a]

Sıralı Ölçek verilerinde Güvenilirlik

Sıralı ölçek verilerinde büyükten küçüğe veya küçükten büyüğe doğru anlamlı bir sıralama söz konusudur. Ancak veriler arasındaki mesafe tam olarak eşit değildir. Tutum ölçeklerinin madde dereceleri, gözlemcilerin üç veya daha fazla dereceli bir boyut üzerinde yaptıkları değerlendirmeler, kişilik testlerindeki maddelerin puanları, zekâ testlerindeki maddelerin puanları sıralı ölçek verisi niteliğindedir. Bu puanlara dayalı olarak güvenilirlik için gözlemciler arasındaki tutarlılık, maddeler arasındaki tutarlık ölçeğin tek boyutluluğu veya madde-toplam puan tutarlılığı araştırılır. Sıralı ölçek verileri normal dağılım özelliği göstermediğinden[b] klasik test kuramında sadece belirli istatistiksel teknikler uygulanır. Eğer modern test

---

[a] Bu konuda *bk.*, Niemoller, B., Van Schuur, W.H. ve Stokman, F.N., Stochastic Cumulative Scaling. STAP User's Manual Vol. 4, Technical Center, Faculty of Social Sciences, University of Amsterdam,1980.

[b] Belirli koşullarda bu verilerin de normal dağılım özelliğine sahip olduğu varsayılabilir. Bu konuda daha fazla bilgi için "Güvenilirlik ve Korelasyon Analizleri" bölümüne bakınız.

kuramı temel alınmışsa literatürde daha az başvurulan diğer korelasyon analizi yöntemlerine başvurulur ve bu teknikler aşağıdaki gibidir:

1) Klasik test kuramına göre.
   a) Spearman rho ($n > 20$ olan örneklemler için).
   b) Kruskal gamma.
   c) Kendall tau a, b, c ($n < 20$ olan örneklemler için).

2) Modern test kuramına göre.
   a) Polikorik korelasyon.
   b) Poliserial korelasyon.
   c) Teyit edici faktör analizi.

Bilim adamı sıralı ölçek verilerinin güvenilirliğini belirlerken örneklem büyüklüğünü, verilerdeki derece (şık) sayısını ve verilerin analizinde klasik veya modern yaklaşımlardan hangisinin temel alınacağını göz önünde bulundurmalıdır. Klasik test kuramına göre yapılan ölçümlerde kullanılan Spearman rho, Gamma ve Kendall testleri verilerdeki monotonluğu ölçer. Monotonluk, verilerin aynı yönde veya farklı yönde değişmesi anlamındadır. Aynı yönde değişme uyuşma veya güvenilirlik anlamına gelir.

Spearman rho. Örneklem büyüklüğü 20'den büyük olan çalışmalarda ikili veya çoklu veriler arasındaki ilişkiler Spearman rho (r_s) korelasyon analizi yöntemi ile test edilir. Örneğin, bir ölçeğin tek tek maddeleri arasındaki korelasyonlara bakılıyorsa 5 veya 7 dereceli ölçek maddeleri sıralı ölçek niteliğinde olduğundan Spearman rho korelasyon analizi uygulanır. Daha sonra korelasyon değerlerinin medyanı veya aritmetik ortalaması iç tutarlılığı gösteren güvenilirlik katsayısı olarak değerlendirilir.[19] Toplam puan ile maddeler arasındaki korelasyona bakılıyorsa yine Spearman korelasyon analizi tercih edilir.

Kruskal gamma. Bu testte, sıra büyüklüğüne sokularak eşleştirilmiş sıralı veriler arasındaki ilişkilerden hareket edilerek verilerin güvenilirliği araştırılır. Gamma değeri artı 1'e yakınlaştığı ölçüde iki değişken arasındaki korelasyonun, dolayısıyla güvenilirliğin yüksek olduğu sonucuna varılır.

Kendall tau. Kendall tau, ikili ve çok dereceli sıralı ölçek verileri arasındaki ilişkileri belirler. İki farklı gözlemcinin yaptığı değerlendirmeler puan büyüklüğü sırası içinde verilmişse veya ham puanlar daha sonra puan büyüklüğü sırasına sokulmuşsa Kendall tau analizine başvurulur.

Polikorik ve poliserial korelasyon. Polikorik korelasyon analizi, araştırmacının modern test kuramına göre çok dereceli ölçeklerde maddelerin gizli yapıyla ilgili olup olmadığını belirlemek için kullandığı bir tekniktir. Değişkenlerden birinin sıralı diğerinin ise eşit aralıklı olduğu durumda poliserial korelasyon analizi yöntemi uygulanır. Polikorik ve poliserial korelasyon analizleri PRELIS isimli istatistik yazılımda bulunur. Joreskog ve Sorbom tarafından geliştirilen *konjenerik*[a] *ölçüm modelinde* 5 veya 7 dereceli maddelerden oluşan ölçekler eşit aralıklı sayılmadığından bu tür ölçeklerin tek boyutluluğunu veya güvenilirliğini saptamak için polikorik korelasyon değerlerinden oluşan bir kovaryans matrisi oluşturulur ve bu matrise dayalı olarak *teyit edici faktör analizi* yöntemi uygulanır.

Eşit Aralıklı Ölçek Verilerinde Güvenilirlik

Eşit aralıklı veriler; Likert, Thurstone, Guttman, Bogardus ve *anlamsal farklılık* ölçeklerden elde edilen toplam puanlarla ilgilidir. Guttman ve Bogardus ölçeklerinde ifadeler arasındaki mesafenin tam olarak eşit olduğu söylenemese de *varsayımsal olarak eşit* olduğu kabul edilir.[20] Bunun yanında bir değişkenin şıkları dereceleme ölçeklerinde görüldüğü gibi, bir boyut üzerinde sıralanmamış da olsa şıkların aralarında belirli bir eşitlik varsa yine eşit aralıklı ölçek gibi değerlendirilir. Örneğin "günde kaç paket sigara içiyorsunuz?" sorusunun şıkları *İki paketten fazla, İki paket, Bir Buçuk Paket, Bir Paket, Yarım Paket* ve *Yarım Paketten Az* şeklinde belirlenmiş olabilir. Güvenilirlik analizi yapılmak istenen veriler eşit aralığa sahip beşli, yedili Likert ölçeği niteliğinde ise bu ölçeklerin toplam / ortalama puanlarına dayalı olarak aşağıdaki istatistik teknikler kullanılır:

1. Pearson korelasyon analizi.
2. Spearman korelasyon analizi.
3. Küme içi korelasyon analizi.
4. Varyans analizi.
5. Faktör analizi.

---

[a] Konjenerik ölçek (congeneric scale): Ortak genli tek boyutlu ölçek.

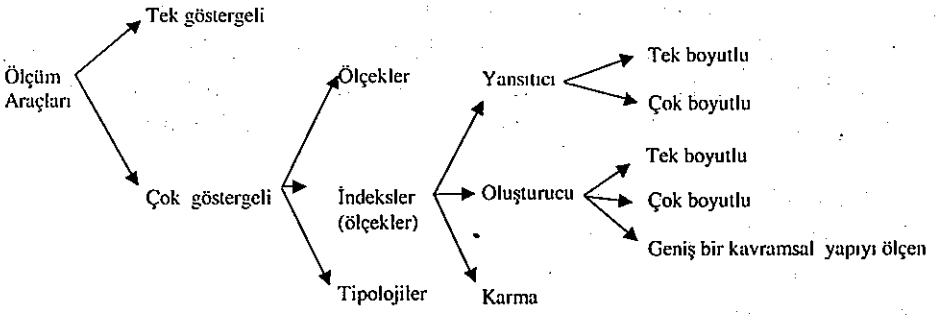
Eşit aralıklı ölçeklerde veriler normal dağılım özelliği göstermiyorsa Spearman *rho* yöntemi tercih edilir.

Oranlı Ölçek Verileri ve Güvenilirlik

Oranlı veriler; zekâ puanı, yaş, ücret, satış miktarı, üretim miktarı gibi gerçek verilerle ilgilidir. Bu verilerin güvenilirliği geçerli ölçüm araçlarının kullanılmasına, verilerin birinci elden toplanmasına, bilgilerin doğru verilmesine ve yinelenen ölçümlerde benzer sonuçlar elde edilmesine bağlıdır. Oranlı ölçek verilerinin güvenilirliğini saptamak için eşit aralıklı ölçek verilerinde kullanılan aynı istatistiksel analizlerden yararlanılır.

ÖLÇÜM ARAÇLARI VE GÜVENİLİRLİK

Güvenilirliği aynı zamanda ölçüm aracının türü ve niteliği açısından da değerlendirmek gerekir. Ölçüm araçları önce iki grupta ele alınır. Tek göstergeli ölçüm araçları ve çok göstergeli ölçüm araçları. Çok göstergeli ölçüm araçları birden fazla maddeden oluşan bileşik ölçümlere dayanır. Bileşik ölçüm araçları ise kendi içinde üç grupta sınıflandırılır. Bunlar; (a) ölçekler, (b) indeksler ve (c) tipolojilerdir (*bk.*, Şekil 2-2). Ele aldığımız bu sınıflandırmada, indeks türü ölçüm araçlarının ne olduğu ve nasıl geliştirilmesi gerektiği konusunda bilim adamları arasında tam bir mutabakat yoktur. Bu nedenle bilim adamı kullandığı ölçüm aracının hangi gruba girdiğini çok iyi bilmeli ve güvenilirlik analizlerini ölçüm aracının türüne uygun istatistiksel teknikleri kullanarak yapmaya çalışmalıdır.



Şekil 2-2. Ölçüm araçlarının sınıflandırılması.

Tek Göstergeli Ölçüm Araçları ve Güvenilirlik

Tek göstergeli ölçümler bir konuyla ilgili olarak değişik sayıda maddeden oluşan anketler, soru listeleri veya değerlendirmelerdir. Klasik test kuramında bu tür ölçümleri "ölçek" veya "indeks" olarak nitelendirmek doğru değildir. Trochim (2002) tek göstergeli ölçümleri, "yanıt ölçeği" (response scales) olarak isimlendirmiştir.[21] Burada *ölçek* sözcüğü genel terim olarak kullanılmıştır. Maddeleri birbirinden bağımsız olan *soru listesi* şeklindeki ölçümlerde kavramsal bir yapıdan söz edemeyiz. Çünkü, kavramsal bir yapının tek bir maddeyle ortaya çıkarılması mümkün değildir. Soru veya ifadeler birbirleriyle ilgili olabilir ancak bu ifadelerden genel bir indeks puanı veya tutum puanı hesaplanmaz. Çünkü maddeler alan örneklemesi[a] çerçevesinde belirlenmemiştir ve büyük ölçüde birbirinden bağımsızdır. Yanıt ölçeklerinin şıkları *Evet-Hayır* şeklinde ikili, duruma göre 3, 5 veya 7 dereceli olabilir. Bir maddenin, örneğin *Kuvvetle Kabul* şıkkından başlayıp *Şiddetle Ret* şıkkına kadar uzanan 5 dereceye sahip olması onun tek başına tutum ölçeği veya indeks olmasını gerektirmez. Belirli bir konudaki görüşleri derlemeyi amaçlayan bağımsız anket soruları, seçmen eğilimlerini saptamayı amaçlayan anketler, öğrencilerin eğitimden memnuniyetini araştıran ölçek şeklinde oluşturulmamış sorular tipik *yanıt ölçeği* örnekleridir. Araştırmacı bu tür ölçümlerde, arka planda var olduğu düşünülen gizli kavramsal yapıları, yetenek ve becerileri ortaya çıkarma gibi bir amaç peşinde değildir. Yorum ve değerlendirmeler doğrudan cevaplayıcıların yanıtlarına göre yapılır. Yanıt ölçeklerinde cevap şıkkı, cevap etiketleri veya derece sayısı için standart bir ölçü yoktur. Belirlenen derece sayısı ana kütle veya örnek kütledeki değişkenliği yeterli ölçüde ortaya çıkarmayabilir. Bu nedenle yanıt ölçekleri için birden fazla madde bulunmadığından "iç tutarlılık" güvenilirliği söz konusu değildir. Tek göstergeli ölçek türleri aşağıdaki gibidir:

1. *Çoktan seçmeli şıklara sahip, bağımsız maddelerden oluşan ölçekler* (Itemized category scales). "Maddeleştirilmiş ölçekler" adı da verilen bu uygulamada cevaplayıcılar belirli sayıdaki şıktan bir veya bir kaçını seçme durumundadırlar.

---

[a] Alan örneklemesi (domain sampling). Tek bir karakteristik veya tek bir özelliği temsil eden ölçüm alanı. Kapsamı sınırlı kavramsal bir yapıya ait bir ölçeğin faktör analizi sonucunda birden fazla faktör ortaya çıkarması alan örneklemesinin iyi yapılmadığı anlamına gelir.

2. *Karşılaştırmalı ölçekler* (Comparative scales). Bu ölçeklerde bir nesne veya kavram başka bir nesne veya kavramla karşılaştırılır. Sık kullanılan karşılaştırma ölçekleri çiftli karşılaştırma, tercih edilen şık-fiyat karşılaştırma ölçekleri ve birim-toplam-kazanç ölçekleridir (unity-sum-gain technique).

3. *İkili karşılaştırma ölçekleri* (Paired comparison scales). Ticarî markaların, programların veya uygulamaların ikili olarak karşılaştırılması ilkesine dayanır. (*ör.*, TV'de hangi programları seyretmeyi tercih edersiniz? Spor programlarını =1; haber programlarını =2) İkili karşılaştırma ölçeklerinde belirlenen madde sayısının faktöriyeli kadar karşılaştırma yapılır [Karşılaştırma çifti = $n\ (n - 1) / 2$]. Bu teknikte cevaplandırma süresini gereğinden fazla uzatmamak için karşılaştırma yapılacak nesne sayısı 10'dan az olmalıdır.

4. *Q sınıflandırma ölçekleri* (Q-sort scales). Stephenson tarafından 1935 yılında geliştirilen bu teknikte cevaplayıcılar veya uzmanlar araştırılan belirli özellikleri, sıfatları veya sayısı 140 kadar çıkan maddeleri normal dağılım özelliği gösteren bir dereceleme grubuna dağıtırlar. Değerlendirmede 10 veya daha fazla dereceleme grubunun bulunması gerekir. En olumludan en olumsuza doğru sıralandığında maddeler normal dağılım özelliğine sahip olmalıdır. Olumlu ve olumsuz maddelerin sayısı nispeten az, nötr nitelikteki maddelerin sayısı ise daha fazla çıkmalıdır.

5. *Tercih sıralamalı ölçekler* (Rank order scales veya Forced ranking scales). Cevaplayıcılar, belirlenen şıkların öncelik sırasını belirler.

6. *Sabit toplamlı ölçekler* (Constant sum scales). Toplamı 100 olacak şekilde şıklara belirli bir yüzde ağırlığı verilir.

7. *Resimli ölçekler* (Pictorial scales). Yüzler, merdiven resimleri vb. gibi grafik simgeler kullanılarak yapılan cevaplandırma ölçekleridir.

8. *Likert tipi maddeler.* Likert ölçeğine benzemekle birlikte toplam veya ortalama puanın temel alınmadığı ölçüm araçlarıdır.

Daha çok pazarlama ve yönetim-organizasyon araştırmalarında kullanılan yanıt ölçeklerinde seçilen maddelerin gizli bir yapıyı ortaya koyup koymadığı, ölçeğin tek boyutlu olup olmadığı, maddelerin ölçüm alanını temsil edip etmediği, alt boyutlara sahip olup olmadığı gibi konular araştırılmaz. Sadece varsayımsal olarak, tek bir maddenin alanı ölçtüğü düşün-

cesinden hareket edilir. Örneğin, "İdam cezasından yana mısınız?" sorusuna gelen yanıtların belirli bir örnek kütlenin görüşlerini yansıttığı iddia edilir, ancak bu konudaki iddialar hiçbir zaman güvenilir değildir. Bu nedenle *tek sorulu görüş bildirimine dayanan anketler/ölçekler* ve kamuoyu araştırmalarının sonuçlarını ihtiyatla karşılamak gerekir.

Yanıt ölçeklerinde nötr şıkları kullanma zorunluluğu yoktur. Belirlenen beş derecenin tamamı olumlu şıklar veya tamamı olumsuz şıklar şeklinde de belirlenebilir (*bk.*, Şekil 2-3). Ancak bu şekilde oluşturulan ölçeklere gelen cevaplar büyük ölçüde yanlıdır, gerçek durumu resmetmez.

	Vasat	Vasatın üstünde	İyi	Oldukça iyi	Çok iyi
Öğretim üyesinin derse hazırlıklı gelmesi	☐	☐	☐	☐	☐
Öğretim üyesinin ders sunum biçimi	☐	☐	☐	☐	☐
Öğretim üyesinin öğrencileri ders sırasında aktif hale getirmesi	☐	☐	☐	☐	☐
Öğretim üyesinin derse hakimiyeti	☐	☐	☐	☐	
nav notları	☐	☐	☐		
	☐	☐			

Şekil 2-3. Yanıt ölçeği şeklinde oluşturulmuş bir ölçüm aracı.

Tek göstergeli ölçüm aracı türlerinden bir diğeri, "Likert tipi maddeler" olarak isimlendirilen yaklaşımdır. Bu yaklaşımda Likert ölçeğine benzer şekilde ifadeler ve karşılarında 5'li veya 7'li ölçek dereceleri vardır. Fakat maddeler gizli yapıları ortaya çıkarmak üzere oluşturulmadığından birbirlerinden bağımsızdır. Bazı istatistikçiler Likert tipi maddelere verilen puanlardan hareket edip maddeler arasında *t*-testi veya diğer parametrik istatistikî analizleri uygulamaktan kaçınmamaktadırlar (Sisson ve Stocker, 1989, aktaran Classon ve Dormody).[22] Ancak tek tek maddelerin ana kütlede normal dağılım özelliği gösterip göstermediğini saptamak çok zordur. Bu maddelerin ana kütledeki dağılımı genelde çarpıktır ve ölçülmek istenen tutumu bütün yönleriyle kapsaması imkansızdır. Ayrıca cevaplayıcıla-

rın yanıtlarında tavan – taban etkisi[a] ortaya çıkması nedeniyle Likert tipi maddelerin güvenilirliği düşüktür. Yanıt ölçeklerine ait verilerin güvenilirliğini araştırmak isteyen bilim adamları benzer alt gruplarda aynı sonuçlara ulaşıp ulaşmadıklarına bakmalıdırlar. Tek göstergeli ölçümler için *iç tutarlılık* güvenilirliği yapılamamakla birlikte, araştırmacı gerektiğinde gözlemci içi tutarlılık, gözlemciler arası tutarlılık ve test-yeniden test yöntemlerinden yararlanabilir.

Tek göstergeli ölçüm araçlarından bir diğeri *madde-yanıt kuramı* çerçevesinde test edilen *ana kütle referanslı olmayan* bilgi ve beceri testleridir. Bu ölçüm araçlarında doğru yanıt sayısı belirli bir ana kütleye bağımlı olmadan belirlenir. Araştırmacı, ölçeğin veya testin değil, tek bir test maddesinin nasıl çalıştığı üzerinde odaklanır. Tek bir test maddesi kadınlarda, erkeklerde, işçilerde ve işverenlerde farklı bir şekilde çalışabilir. Bu tür uygulamalarda gerçek puan, bir dizi maddenin sonucuna göre belirlenmez. Araştırmacı her bir ölçüm puanına ait hata üzerinde odaklanır. Madde-yanıt kuramının temel alındığı başarı testlerinde bir maddenin yanlılığı, *madde özellikleri eğrisine* (MÖE) göre belirlenir. Eğer madde özellikleri eğrisi iki ana kütlede aynı ise madde yanlı değildir, farklı ise madde değişik ana kütlelerde farklı çalışıyor demektir. Bilgi ve beceri testlerinin analizinde sadece MYK esas alınmışsa iç tutarlılık analizleri, yarıya bölme, test-yeniden test yöntemi gibi uygulamalara baş vurulmaz.

İndeksler ve Güvenilirlik

İndeksler davranışları, kişilik özelliklerini, başarıyı veya sosyal ve örgütsel olguları ölçmeye yönelik olarak oluşturulan ölçüm araçlarıdır. Literatürde indekslerin ne olduğu ve indekslerin güvenilirlik analizlerinin nasıl yapılacağı konusunda bilim adamları çelişkili açıklamalar yapmışlardır. İndekslerin güvenilirlik analizleri konusunda net bir bilgi verilebilmesi için bu ölçüm araçlarının ne olduğunun çok iyi tanımlanması gerekir. ABD'li sos-

[a] Tavan-taban etkisi (floor or ceiling effect). Yetenek ve başarı testlerinde katılımcıların verdikleri yanıtların hep düşük puanlar üzerinde yoğunlaşması "taban etkisi", (çok zor sorular) yanıtlarının büyük çoğunluğunun yüksek puanlar üzerinde yoğunlaşması ise "tavan etkisi" (çok kolay sorular) olarak isimlendirilir. Güçlük düzeyi p = ,50 olan maddeler tavan-taban etkisi yaratmayan maddelerdir. Tavan-taban etkisi altına girmiş sayılmaması için maddelerin güçlük oranlarının ,30 ilâ ,70 arasında olması gerekir. Yüzde 30'un altındaki maddeler taban ve yüzde 70'in üstündeki maddeler ise tavan etkisi altındadır. Ancak testlerin, personel seçimi amacıyla kullanılması halinde *güçlük oranı* ile *seçim oranı* rakamlarının denkleştirilmesi gerekir. Eğer personel, katılımcıların en iyi üst %20'lik dilimi arasından seçilecekse güçlük oranının da buna göre düşünülmesi gerekir. En üst dilimde yer alan %20'lik kesim en zor bir maddeyi dahi ,20 civarında yanıtlamış olmalıdır. Kolay bir maddeyi yanıtlama oranı ise, %80'inin üzerine çıkmamalıdır.

yolog Babbie'nin indeks tanımlamasıyla Diamantopoulos ve Winklhofer'in indeks tanımlaması aynı değildir. Babbie'nin indeks kavramına getirdiği açıklamalar, ölçüm aracının belirli bir kavramsal yapıyla ilintili, tek boyutlu ve maddelerin paralellik özelliği göstermesini gerektirirken, Diamantopoulos ve Winklhofer'in indeks olarak tanımladıkları ölçüm aracında maddelerin paralellik özelliği göstermesi gerekmez. Diamantopoulos ve Winklhofer, indeksleri paralel yapılardan uzak ölçüm araçları olarak görmüşlerdir. Ölçüm çalışmaları için bir indeks geliştirmek isteyen bilim adamı şu soruların yanıtları konusunda net bir görüşe sahip olmalıdır.

1. Geliştirilen ölçüm aracındaki maddeler Guttman, Bogardus, Mokken ve Thurstone ölçeklerinde olduğu gibi hiyerarşik bir yapılanmaya ve tek boyutluluğa sahip olacak mıdır?
2. Ölçüm aracındaki maddelerden hareket edilerek bu maddelerin arka planındaki gizli bir yapı veya yapılar araştırılacak mıdır?
3. Ölçüm aracının oluşturulma nedeni sadece belirli bir boyutu ölçmek veya maddelerin hangi gruplar/bileşenler altında toplandığını görmek ve temel bileşenler arasındaki ilişkileri belirlemek midir?
4. Ölçüm aracının oluşturulma nedeni sadece belirli bir modeli test etmek midir? Modelin uyuşma durumunun kanıtlanmasıyla maddelerin güvenilirlik ve geçerliliklerinin de sağlanmış olduğu mu varsayılacaktır?

Birinci sorunun yanıtı "gerçek anlamda bir ölçek oluşturma" anlamına gelir. İkinci sorunun yanıtı "yansıtıcı bir ölçek" oluşturmaktır. Burada *gerçek* sözcüğünü kullanmamamızın nedeni *ölçek* sözcüğünün genel bir terim olarak kullanılması nedeniyledir. Teknik açıdan doğru bir şekilde ifadelendirirsek "yansıtıcı indeks" terimini kullanabiliriz. Üçüncü sorunun yanıtı ise, "oluşturucu bir ölçek" veya teknik dille ifade edilirse "oluşturucu bir indeks" meydana getirmektir. Dördüncü sorunun yanıtı ise teyit edici faktör analizi veya yapısal eşitlik modellerinden birini kullanmaktır.

"Ölçek" olarak isimlendirilen araçlarda maddelerin hiyerarşik veya kümülatif (yığışımlı) sıralanmasını sağlamaya yönelik çalışmalar güvenilirlik analizlerinin önemli bir ögesini oluştururken, *yansıtıcı indekslerde* hiyerarşik bir sıralanma söz konusu olmaksızın maddeler arasındaki iç tutarlılık önem kazanır. Ölçek kavramıyla daha çok Guttman, Mokken, Thurstone ve Bogardus türü ölçüm araçları kastedilir. Likert'in geliştirdiği ölçüm aracı ile *anlamsal farklılık ölçeği* maddelerin hiyerarşik sıralanma-

sına dayanmadığından ölçek değil, indekstir.[a] Literatürde sık kullanılan "Likert ölçeği" ifadesindeki *ölçek* sözcüğü genel bir terim olarak kullanılır. Günümüzde hiyerarşik yapılanmaya dayalı ölçekler büyük ölçüde kullanılmadığından, toplamalı ölçüm araçları olan indeksler, *ölçek* olarak adlandırılmaya başlanmıştır. Bu bölümde özel olarak *indeks* kavramını kullanırken "yansıtıcı indeks/ölçek" ve "oluşturucu indeks/ölçek" türlerinin her ikisini de kastediyoruz.

Chin'e göre (1998, aktaran Tutzauer) yansıtıcı indeksler arka plandaki gizli değişkeni ortaya çıkarmak için oluşturulmuş olan ölçüm araçlarıdır. Gizli değişkeni yansıttığı için kısaca "yansıtıcı ölçek" veya "yansıtıcı indeks" terimi kullanılır.[23] Sosyal bilimlerde yapılan çalışmaların büyük çoğunluğunda yansıtıcı ölçekler veya *sonuç göstergeleri* kullanılmıştır.[24] Yansıtıcı indekslerin çıkış noktası, mevcut literatür bilgileri veya belirli bir konudaki uzman görüşleridir. Oluşturucu indeks ise, ölçüm maddelerinin yeni ve farklı *bileşik bir değişkeni* oluşturduğu veya ortaya çıkardığı ölçeklere verilen addır. Oluşturucu indekslerin gizli değişkenle bir ilgisi yoktur. Literatürde bazen oluşturucu indekslerin de gizli değişkeni ortaya çıkardığından söz eden yazarlara, bilim adamlarına rastlanılır. Bu tür kullanımlar özensiz bir yaklaşımı (veya genel bir kullanımı) simgeler. Doğru kullanım biçimi oluşturucu göstergelerin *temel bileşen* veya *bileşenlerle* ilgili olduğudur. Yansıtıcı ölçeklerde *gizli yapılar* söz konusu iken, oluşturucu ölçeklerde *ortaya çıkan yapılardan* söz ederiz.

İndeksin tanımlanması. Maddeler arasında büyüklük/önem/ağırlık sıralaması olmaksızın geniş kapsamlı bir kavramsal yapıyı veya sınırları daraltılmış bir boyutu ölçmek amacıyla oluşturulan ve toplam/ortalama puanı alınarak ölçüm yapılan araçlar indeks olarak adlandırılır. İndekslerin temel iki özelliği; toplam/ortalama puanlara dayalı olması ve maddelerin hiyerarşik bir sıra içinde dizilmemesidir. İndeksler daha sonra içerdiği maddelerin özelliklerine göre farklı gruplar altında incelenebilir. En temel

[a] Ölçüm araçlarının sınıflandırılması konusunda bilim adamları arasında tam bir mutabakat yoktur. Ölçek/indeks sınıflamasının dışında bazı bilim adamları tutum ölçümünde kullanılan araçları (a) toplamalı ölçekler, (b) eşit aralıklı ölçekler ve (c) yığışımlı ölçekler şeklinde gruplandırmışlardır. Bir diğer sınıflandırma deterministik-probabilistik ölçek sınıflandırmasıdır. Ölçek sonucuyla yapılan tahminin belirgin veya belirleyici olması nedeniyle Guttman ölçekleri deterministik, yapılan tahminin belirli bir oranda geçerli olduğunu gösterdiği için Mokken ölçekleri probabilistik ölçek olarak nitelendirilmiştir. Literatürdeki ölçek-indeks-tipoloji ayrımı mevcut ölçüm araçlarının teknik alt yapısını daha iyi açıkladığından kitapta bu sınıflandırma biçimi tercih edilmiştir.

sınıflama, yansıtıcı indeks ve oluşturucu indeks biçiminde yapılan gruplandırmadır.

Yansıtıcı indeksler/ölçekler özellikle psikolojik, davranışsal, ekonomik ve siyasal içerikli ölçümlerde kullanılır. Bilim adamı arka plandaki gizli kavramsal bir yapıyı veya birden fazla gizli faktörü ortaya çıkarmak istediği zaman bu ölçüm araçlarından yararlanır. Yansıtıcı ölçeklerde maddelerin büyük ölçüde birbirleriyle tutarlı olması gerekir. Bu ölçekler esas itibariyle klasik test kuramının varsayımları çerçevesinde oluşturulur.

Oluşturucu indeksler/ölçekler gizli bir değişkeni ortaya çıkarma hedefi güdülmeksizin belirli bir kavramsal boyut çerçevesinde *evet – hayır* yanıtları temel alınarak veya Likert aracındaki gibi 5 veya 7 ölçek derecesi kullanılarak oluşturulan, tek bir toplam[a] veya ortalama puana göre değerlendirme yapılan ölçüm araçlarıdır. Oluşturucu indekslerde örtük yapılar araştırılmaz. Değerlendirme açısından bakarsak oluşturucu indekslerde değişik sayıda derecenin kullanıldığını görürüz. Oluşturucu indeksler iki şıklı, üç, beş veya yedi dereceli ölçümler temel alınarak oluşturulabilir. Oluşturucu indekslerin tek bir boyutu mu, birden fazla boyutu mu yoksa daha geniş *genel kavramsal bir yapıyı mı* ölçtüğü bilim adamları arasında tartışmalı olan bir konudur.

Literatürde indeks kavramıyla ilgili tartışmalar. Psikologlar, sosyologlar ve işletmeciler indeks kavramına farklı açıklamalar getirmişlerdir. Örneğin, psikologlar *indeks* anlamında *ölçek* terimini kullanarak maddelerin seçimi için *alan örneklemesi* yöntemine önem vermişlerdir. Psikologlar için önemli olan, maddelerin arka planındaki gizli yapılardır. Psikologlar çoğunlukla yansıtıcı ölçekleri veya ölçüm araçlarını geliştirmeye çalışmışlardır.

Sosyologlar ise indeks kavramıyla temel bileşenin veya bileşenlerin çıkarılmasını önemsemişlerdir. Bu yaklaşım "oluşturucu indeks/ölçek" geliştirmeyi gerektirir. Onlara göre indeksler esas olarak tek boyutludur. ABD'li sosyolog Babbie indeksi "bireysel vasıflara atanan puanların top-

---

[a] Genel uygulama olarak indekslerde toplam veya ortalama puanla çalışılır. Ancak araştırmacı toplam puanı 0 ilâ 10 veya 0 ilâ 100 arasındaki puan değerleriyle yeniden ölçeklendirerek de gösterebilir. Bunun için indeksin en küçük puanı belirli bir ham puandan çıkarılarak indeks baz puanı bulunur. Daha sonra elde edilen *indeks baz puanı* 10 veya 100 barem puanıyla çarpılarak, indeks baz puanının en küçük puanla toplandığı rakama bölünür. Örneğin, üç maddeden oluşan indeksin en küçük toplam puanı 3 olsun. Toplam puan 12'yi 10 üzerinden indeks puanlarına çevirmek için şu işlem yapılır: 12–3= 9; 9 x (10/13) = 6,9. Yeniden ölçekleme, özellikle farklı uzunluklarda ölçekler kullanılması ve ölçek puanları arasında karşılaştırma yapılmak istenmesi halinde uygulanır.

lanmasıyla elde edilen tek boyutlu ölçüm aracı" şeklinde tanımlamıştır.[25] Ona göre, indeks tek bir boyutu ölçen ve gözlem maddelerinin birbiriyle ilgili olduğu bir ölçüm aracıdır. İndeksteki bir dizi maddenin tek bir boyutu ölçmesi incelenen kavramın *geniş* veya *dar* bir biçimde tanımlanmasına bağlıdır. Eğer araştırılan kavram geniş bir içeriğe sahipse, belirlenen çok sayıda madde indeks standartlarını karşılayamaz ve bu şekilde oluşturulan ölçüm aracı indeks olarak kabul edilmez. Çünkü bu tür ölçüm araçlarında kavramın birden fazla yönü kuşatılmaya çalışılmıştır.[26] İndekslerde kavramın sadece tek bir yönü ele alınır ve bu yönle ilgili olarak seçilen maddelerin hepsi birbirleriyle yüksek ilişki katsayılarına sahiptir. Babbie, indekslerde maddelerin gizli bir yapıyı ortaya çıkarması veya belirli bir bileşene işaret etmesi konusuna değinmemiştir. "Abercrombie'ye göre indeks doğrudan ölçülemeyen bir şeyin göstergesidir. Neuman'a göre ise, belirli bir kavrama ait birbirinden ayrı birçok gözlem değişkenin toplanması veya kombine edilmesiyle oluşturulan ölçüm aracıdır (aktaran Kirsch)."[27] İndekslerin toplam puanları eşit aralıklı ölçek niteliğindedir, fakat maddelerine ait puanlar sıralı ölçek verisi olarak değerlendirilir. İndeksler, tutumların belirlenmesi veya belirli sonuçların tahmin edilmesi için kullanılır. Araştırmacılar indekslerle tutumları ölçerler veya belirli olayların veya olguların nedenlerini araştırırlar. Örneğin; suç işleme eğilimi, örgütsel bağlılık, geçimsizliğin nedenleri, bencilliğin nedenleri gibi.

Sosyologlar, "ölçek" kavramı ile; maddelerin hiyerarşik yapılandığı, güçlü maddenin seçilmesiyle ondan önceki bütün maddelerin işaretlenmiş kabul edildiği ölçüm araçlarını kastetmişlerdir. Onlara göre ölçekler daha çok arka plandaki kişisel özellikleri ölçen, gizli yapıları ortaya çıkarmayı amaçlayan ölçüm araçlarıdır. İndeksler ise "tek bir boyutu ölçen doğrusal bileşik ölçümlerdir." Oysa örtük yapılar, sadece hiyerarşik yapılı ölçeklerle değil, pekala doğrusal bileşik ölçüm aracı olan indekslerle de ortaya çıkarılabilir.

İşletmecilerin indeks kavramına ne şekilde yaklaştıklarını incelersek, tarihi süreç içinde ve değişik branşlarda farklı yaklaşımların olduğu görülür. Davranış bilimcileri yansıtıcı ve oluşturucu ölçeklerin (indekslerin) her ikisini de kullanma eğilimi içinde iken, pazarlamacıların, yönetim-organizasyon alanında çalışan bilim adamlarının, insan kaynakları ve üretim gibi branşlarda çalışanların daha çok oluşturucu ölçeklerle çalışmayı yeğledikleri görülür. Bununla birlikte bu eğilim son yıllarda ortaya çıkmıştır. İki binli yılların başına kadar gerek pazarlamacılar ve gerekse diğer işletmeciler ölçüm çalışmalarında çoğunlukla yansıtıcı ölçeklerle/göstergelerle çalışmışlardır. Ancak son yıllarda bu ölçek biçiminin somut/gerçek konuların ölçümünde yetersiz kalması nedeniyle oluşturucu

ölçeklere doğru bir yönelim ortaya çıkmıştır. İşletmeciler "oluşturucu ölçek" (veya indeks) kavramına sosyologlardan farklı bir anlam vermişler, oluşturucu ölçeği, "kavramsal yapıyı tam olarak kuşatan homojen değil, heterojen göstergelerden oluşan ölçüm aracı" olarak tanımlamışlardır. Bu yaklaşımda, tek bir *boyut* veya *boyutlar* yerine göstergelerin neden olduğu birinci düzey *bileşik değişkenler* ve duruma göre genel kavramsal yapıyı gösteren ikinci düzey *üst değişkenler* söz konusudur. Öyle anlaşılmaktadır ki, Babbie'nin bize tanıttığı "indeks" kavramı ne oluşturucu indeks, ne de yansıtıcı indekstir. Babbie'nin indeks tanımlamasını *temel bileşenlere işaret* eden basit yapılı bir "ölçek" olarak görebiliriz. Bu kitapta *indeks* terimini kullanırken Babbie'nin belirttiği anlamdan daha geniş manada ele alıyor, terimin yansıtıcı ve oluşturucu ölçeklerin her ikisini de kapsadığını düşünüyoruz.

Ölçek-indeks tartışmasında işletmecilerin toplamalı ölçek kavramıyla hiyerarşik yapılanmaya dayalı ölçek kavramı arasındaki ayırımı gözden kaçırdıkları görülmektedir. Babbie'nin hiyerarşik yapılanmaya dayalı olmayan ölçüm araçlarını indeks olarak isimlendirmesi doğrudur. Onun görüşlerine katılmadığımız nokta, indeksleri tek boyutlulukla sınırlandırmış olmasıdır. İndekslerin tek boyutluluğu ampirik gerçeklere uygun düşmemektedir. Teorik olarak "tek boyutluluk" fikrinden hareket edilse bile istatistikî analizler sonucunda bir indeksten (yansıtıcı veya oluşturucu olabilir) muhtemelen birden fazla faktör veya temel bileşen ortaya çıkar. Babbie'nin tanımlamasında eksik kalan bir diğer yön, indekslerin duruma göre gizli yapıları ortaya çıkarmak için veya temel bileşenleri belirlemek için de oluşturulabileceğidir.

Bazı işletmeciler, "oluşturucu ölçek" olgusunu indeks olarak tanıtmışlardır. Bize göre daha doğru bir tanımlama, oluşturucu ölçeğin, "indeksin bir türü" olduğudur. Oluşturucu ölçeği tek başına indeks olarak tanımlamak doğru değildir. Öte yandan "Yansıtıcı-Oluşturucu Ölçekler" başlığında da değinileceği gibi, bir ölçeğin yansıtıcı göstergeler ve oluşturucu göstergelerden oluşan *melez* veya *karma bir yapıya* sahip olması da söz konusu olabilir. Okuyucuların kafasını daha fazla karıştırmamak için şunu belirtmeliyiz ki, indeks kavramını iyi anlamak istiyorsak Babbie'nin belirttiği indeks kavramının anlamını genişletmeli, fakat işletmecilerin "oluşturucu ölçek" kavramını indeks terimiyle eşdeğerde değil, bu terimin bir alt grubu olarak ele almalıyız.

Oluşturucu indeks türleri. Oluşturucu indeksler, gizli bir yapıyı ölçmeyen ölçüm araçlarıdır. Literatürde oluşturucu indeksler / göstergeler kendi içinde ayrıca sınıflandırılmamıştır. Ancak farklı bilim adamlarının

konuya değişik biçimlerde yaklaşmaları nedeniyle bilim adamları arasında belirli bir mutabakat oluşuncaya kadar böyle bir sınıflandırma yapma ihtiyacı ortaya çıkmıştır. Oluşturucu indeksler üç farklı şekilde düzenlenebilir.

1. *Tek boyutlu oluşturucu indeksler.* Alan örneklemesi yönteminden yararlanarak neden olan göstergeler üzerinde temel bileşenler analizi yapmak suretiyle "tek bir boyut/bileşen" ortaya çıkarılmak istenir.
2. *Çok boyutlu oluşturucu indeksler.* Alan örneklemesi yönteminden yararlanarak neden olan (veya oluşturucu) göstergelerle birden fazla boyut veya bileşen ortaya çıkarılmak istenir.
3. *Geniş kavramsal yapıları ölçen oluşturucu indeksler.* Kavramsal alanın bütün yönlerinden madde alınarak belirlenen oluşturucu göstergelerle temel bileşenler saptanmaya ve temel bileşenler arasındaki ilişkiler test edilmeye çalışılır. Geniş kavramsal yapıları ölçen oluşturucu indekslerde oluşturucu ve yansıtıcı göstergeler karmaşık bir şekilde bir arada bulunabilir. Böyle bir durumda ölçekte hangi tür göstergelerin çoğunlukta bulunduğuna bakılır.

Literatürde böyle sınıflandırma olmadığından kendi geliştirdiğimiz bu yaklaşımı biraz daha ayrıntılı olarak ele alalım.

Tek boyutlu oluşturucu indeksler. Oluşturucu indeksler gizli yapıdan bağımsız olarak tek bir boyutu ölçmek veya ortaya çıkarmak üzere oluşturulabilir. Babbie'ye göre indeksler esasen tek boyutludur ve Likert ölçeklerinde olduğu gibi çok dereceli değil genellikle iki derecelidir. Bir indeksin puanı, *evet* yanıtlarının toplamı alınarak belirlenir. Babbie'nin tanımladığı indeks türlerinde ölçüm maddeleri ya soru işaretiyle biten cümleler şeklindedir veya maddeler görüş ve düşünceleri belirten düz cümleler şeklinde ifade edilmiştir. Babbie, indeks kavramının anlamını daraltmış ve indeksi sadece tek bir boyutun ölçüldüğü ölçüm aracı olarak düşünmüştür. Babbie'nin indeks tanımlaması, belirlenen göstergeler arasında ortak değişkenlik (kovaryans) olmaması nedeniyle aslında bir "oluşturucu" ölçektir. Çünkü bir maddeye verilen yanıt diğer maddeden tamamıyla bağımsızdır. Bir maddeye *Evet* yanıtı verilirken diğerine *Hayır* cevabı verilebilir. Maddelere verilen yanıtlar arasında tutarlılık bulunması gerekmez.

- Babbie'nin tanımladığı soru cümlelerinden oluşan indeks örneği.

1. Depreme karşı evinizde bir çanta hazırladınız mı?
2. Depreme karşı dolaplarınızı duvara tespit ettirdiniz mi?
3. Depreme karşı TV ve bilgisayarınızı tespit ettirdiniz mi?
4. Depreme karşı çocuklarınıza uygulama yaptırdınız mı?
5. Deprem için kolay ulaşabileceğiniz bir yerde düdük bulunduruyor musunuz?
6. Depreme karşı binanızı güçlendirdiniz mi?

- Babbie'nin tanımladığı düz cümlelerden oluşan indeks örneği.

1. Kendi kararlarıma çok güvenirim.
2. Eleştirilere açık değilimdir.
3. Diğerlerinin benim kadar deneyimli olmadıklarını düşünürüm.
4. Pişman olduğumu hatırlamam.
5. Katiyen vazgeçmem.
6. Esneklik, zayıflık işaretidir.

Babbie'nin indeks ölçüm araçlarını iki şıklı olarak değerlendirmesine karşılık bu ölçüm araçları gizli bir yapıyı ölçme amacından bağımsız olarak üç veya daha fazla dereceli olarak da düzenlenebilir. Ancak çok dereceli oluşturucu indekslerde, yansıtıcı indekslerde olduğu gibi madde derecelerinin "yaklaşık olarak eşit aralıklı" olması veya maddelerin birbirlerine paralel olması gibi bir zorunluluk yoktur.

- Tek boyutlu oluşturucu indekslerin özellikleri:

1. Alan örneklemesi yöntemine dayanır.
2. Sınırları dar, tek bir kavramsal yapı veya geniş bir kavramsal yapının belirli bir veçhesi ölçülmeye çalışılır.
3. Göstergelere yönelik olarak klasik test kuramına göre toplam puan-madde korelasyonu, iç tutarlılık, PLS modellemesi ve diğer güvenilirlik analizleri yapılır.

4. Temel bileşenler faktör analizi yöntemiyle tek bir boyut saptanmaya çalışılır.
5. Eğer birden fazla boyut ortaya çıkmışsa diğer boyutlara ilişkin maddeler ölçekten elenir veya çok faktörlü bir ölçek ortaya çıkar ki bu ölçekteki birincil boyutlar ikinci düzey faktörlerle temsil edilebilir. Böyle bir durumda muhtemelen birincil faktörler yansıtıcı göstergelerdir.

Çok boyutlu oluşturucu indeksler. Tek boyutlu indekslerin tersine bilim adamı bu ölçüm araçlarındaki göstergelerden hareket ederek keşfedici faktör analiziyle birden fazla boyut veya birden fazla bileşenin ortaya çıkmasını bekler. Göstergeler çoğunlukla çok dereceli ve birbiriyle ilintili maddeler şeklindedir. Temel bileşenler analizi sonucunda ortaya çıkan birden fazla bileşenin (faktörün) kendi aralarındaki korelasyon katsayıları eğer yüksek ise söz konusu bileşenler üst bir kavramsal yapıyı ölçüyordur. Korelasyon katsayıları düşük ise bileşenler birbirinden bağımsız olarak aynı kavramsal yapılarla ilgili olabilir. Bağımsız kavramsal yapılar alt ölçekler olarak nitelendirilir.

Geniş kapsamlı kavramsal yapıları ölçen oluşturucu indeksler. İndeksler, kavramsal alanın bütün yönlerinden madde alarak belirlenir. Bu nedenle de maddeler arasındaki korelasyon ve kovaryans katsayıları yüksek değildir. Bilim adamı *geniş kapsamlı kavramsal yapıları ölçen oluşturucu indekslerde* (GKKY-Oİ) güvenilirlik veya geçerlilik amacıyla maddeler arasındaki korelasyon katsayılarını dikkate almaz. GKKY-Oİ'de çoğunlukla tek bir kavramsal yapıya / somut bir ölçüm alanına işaret eden tek bir toplam / ortalama puan belirlenir. Bilim adamının amacı doğrusal bileşik değişkenleri ortaya çıkarmak da değildir. Bu uygulamada esas olarak oluşturucu göstergelerin etkisi tahmin edilmek istenir. Geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçek göstergeleri, faktör analizinden çok regresyon analizi yapmaya uygundur.

■ Çok boyutlu oluşturucu indekslerin özellikleri:

1. Maddeler alan örneklemesi yöntemiyle belirlenir.
2. Birden fazla bileşen tek bir kavramsal yapıyı veya çoklu yapıları ölçer.

3. Göstergeler birinci düzey bileşenleri ve birinci düzey bileşenler ise daha sonra ikinci düzey üst yapıları, faktörleri oluşturur.
4. Klasik test kuramına göre bileşenlere bağlı gösterge gruplarının her biri için ayrı olmak üzere iç tutarlılık ve diğer güvenilirlik analizleri yapılır.
5. Temel bileşenler analizi yöntemiyle optimum sayıda boyut / bileşen belirlenmeye çalışılır.

- Geniş kapsamlı kavramsal yapıları ölçen oluşturucu indeksler:

1. Alan örneklemesi yöntemine dayanmaz.
2. Sınırları geniş olan bir kavramsal yapının bütün yönleri ölçülmeye çalışılır.
3. Göstergelerde paralellik şartı aranmaz. Göstergeler anlam, içerik ve yönelim açısından birbirlerinden önemli ölçüde farklı olabilir.
4. Faktöriyel yapılar değil, kavramın temsil edilme derecesi önemlidir.
5. İç tutarlılık, madde-test korelasyonu ve yarıya bölme gibi güvenilirlik analizleri yapılmaz.
6. Göstergelerin ağırlığı ve model üzerinde odaklanılır.

Yukarıdaki açıklamalardan da anlaşılacağı gibi, oluşturucu indeksler, bir konuyu duruma göre oldukça geniş veya tam tersine daraltılmış bir çerçevede ele alacak bir biçimde oluşturulabilir. Bazı oluşturucu indeksler oldukça geniş bir "bileşen" yelpazesine sahiptir. Örneğin bir iş tatmini envanteri olan *İş Tanımlama İndeksi*'nde Ücret, İş Arkadaşları, Yönetim Uygulamaları, Haberleşme gibi birden fazla birinci düzey bileşen vardır. "İş tatmini" kavramsal yapısı ise ikinci düzey bileşen veya faktördür. Böyle olunca birinci düzeydeki her bir bileşen veya faktör ölçme özelliğine bağlı olarak tek boyutlu oluşturucu indeks (alt ölçek) veya yansıtıcı ölçek olarak değerlendirilir.

İndeksler ve güvenilirlik. İndekslerin güvenilirliğini saptamak için ne tür bir indeks olduğuna bakmak gerekir. Daha önce indeksleri iki temel sınıflama içinde ele almıştık: yansıtıcı indeksler ve oluşturucu indeksler.

Yansıtıcı indeksler ve güvenilirlik. Yansıtıcı indeksler gizli kavramsal yapıları ortaya çıkarmak amacıyla oluşturulduğundan klasik test kuramı çerçevesinde yapılan tüm güvenilirlik analizleri bu ölçekler için de geçerlidir. Ortak faktör analizi, Cronbach alfa, madde-toplam puan korelasyon katsayıları, test-yeniden test, paralel formlar gibi güvenilirlik analizleri, modelin güvenilirliği için teyit edici faktör analizi ve yapısal eşitlik modelleri yansıtıcı indekslerde uygulanabilecek temel güvenilirlik analizi yöntemleridir.

Oluşturucu indeksler ve güvenilirlik. Oluşturucu indekslerde yapılacak güvenilirlik analizleri literatürde tartışmalı bir konu olarak kalmıştır. Bazı bilim adamları bu tür ölçeklerde alfa iç tutarlılık analizlerini uygularken diğerleri bu uygulamaya karşı çıkmışlardır. Karmaşıklığın nedeni, oluşturucu ölçeklerin ne tür bir indeks olduğu konusundaki anlaşmazlıktan kaynaklanır. Çok düzeyli oluşturucu ölçeklerin içinde aynı zamanda yansıtıcı göstergelerden oluşan ve gizli yapıları ortaya çıkan faktörler de varsa böyle bir durumda birinci düzeyde iç tutarlılık güvenilirlik analizlerinin yapılması doğaldır. Ancak gerek birinci düzeydeki göstergeler ve gerekse ikinci düzeydeki faktörler birbirinden bağımsızsa bilim adamları böyle bir durumda ölçek belirli bir boyutu ölçüyor veya ortaya çıkarıyor olsa dahi iç tutarlılık analizleri yapılmasına karşı çıkmışlardır. Chin'e göre (1998) oluşturucu ölçeklerde PLS modelleme (kısmî en küçük kareler) tekniği uygulanmadan güvenilirlik analizi yapılamaz (aktaran Ashill ve Jobber, 2002).[28]

Çok boyutlu oluşturucu ölçeklerde ise, önce temel bileşenler analizi yöntemiyle temel boyutlar belirlenir ve daha sonra her bir boyut veya alt ölçek için önceki paragrafta ele alınan güvenilirlik analizleri yapılır. Araştırmacı bu bölümde isterse bu amaçla hazırlanmış özel yazılımlardan yararlanarak "bileşik indeks/ölçek güvenilirlik analizi" yöntemini uygulayabilir. Ancak, örneklem hacmi yetirence büyük değilse (100'den küçükse) ve veriler normal dağılım özelliği göstermiyorsa değişkenler arasında çoklu koşutluk özelliği yoksa yine PLS modellerinden yararlanılır.[29] Araştırmacı güçlü bir kuramsal temele sahip değilse ve geliştirdiği maddelerin birlikte değişme özelliği yoksa PLS modellerini kullanmayı düşünmelidir.

Geniş kavramsal yapıları ölçen oluşturucu indekslerde ise, maddeler arasında paralellik özelliği bulunmadığından Cronbach alfa, yarıya bölme, madde-toplam puan korelasyonu gibi güvenilirlik analizleri yapılmaz.[30] Bu tür ölçeklerde saptanan düşük iç tutarlılık geçerliliği etkilemez.

Oluşturucu indeksin geçerliliğinin sağlanmasıyla güvenilirliğinin de temin edilmiş olduğu düşünülür.

Likert ölçekleri ve güvenilirlik. Toplamalı tutum ölçeği grubunda değerlendirilen bu araç Rensist Likert tarafından geliştirilmiştir ve bu nedenle onun ismiyle anılır. Bir indeks olmasına karşın, daha çok "ölçek" sözcüğüyle birlikte kullanılır. Kavramsal yapıyı hassas bir şekilde ölçme özelliğinden uzak, gelişigüzel oluşturulmuş bir Likert ölçeğinin kullanılması halinde araştırmacı Tip II hatası yapar. Tip II hatası, karşılaştırmalı ölçümlerde mevcut olmayan sunî etkilerin/ilişkilerin sanki varmış gibi ortaya çıkarılması ve H_0 hipotezinin reddedilmesidir. Özensiz olarak oluşturulan Likert ölçeklerinde yüksek veya düşük puanlar gerçek durumu yansıtmaz.

Likert ölçeklerinin güvenilirliğini seçilen model ve ölçeğin oluşturulma biçimi çerçevesinde ele almak gerekir. Likert ölçekleri klasik veya modern test kuramı çerçevesinde güvenilirlik analizlerine tâbi tutulabilir. Likert ölçeklerinde güvenilirlik analizleri için araştırmacı öncelikle hangi modeli kullanacağına karar vermelidir. Klasik test kuramını temel alan araştırmacılar *yansıtıcı göstergelere* sahip bir ölçek oluşturmuşlarsa yarıya bölme, madde-toplam puan korelasyonu, alfa katsayısı, paralel formlar yöntemi, test-yeniden test yöntemleri ile keşfedici ortak faktör analizi yöntemlerinden yararlanabilirler. Tek veya çok boyutlu *oluşturucu göstergelere* sahip bir ölçek oluşturmuşlarsa bu kez temel bileşenler faktör analizi yöntemiyle birlikte iç tutarlılık analizlerini yapabilirler. Ancak göstergeleri birbirinden bağımsız olan *geniş bir kavramsal yapıyla* ilgili olarak oluşturucu bir Likert ölçeği tasarımına sahip iseler, bu bilim adamları güvenilirlik analizlerini bir kenara bırakmalı ve sadece geçerlilik üzerinde odaklanmalıdırlar. Modern yaklaşımlardan genellenebilirlik kuramını temel alan araştırmacılar ise *genellenebilirlik katsayısı* ve *güvenilirlik indeksi* formüllerini kullanabilirler. Literatürde Likert ölçeklerinin güvenilirliği çoğunlukla klasik test kuramı çerçevesinde ve "yansıtıcı ölçek" tasarımına uygun olarak yapılmıştır. Yayımlanan bilimsel makalelerin %90'ından fazlasında yansıtıcı ölçek tasarımı kullanılmıştır.

Araştırmacı Likert ölçeklerinin güvenilirliğini, yansıtıcı veya oluşturucu ölçek oluşturma süreciyle birlikte ele alıp değerlendirmelidir. Bu çerçevede ölçek maddelerinin kaç dereceli olarak oluşturulacağı, hangi ifadelerin seçileceği, ölçeğin uzunluğunun ne olacağı, ölçeğin dilinin uygun olması, ölçeğin dengeli veya dengesiz oluşturulması, tek boyutlu veya birden fazla boyutlu / faktörlü / bileşenli olma durumu incelemeye alınmalıdır.

Yansıtıcı göstergelere sahip Likert ölçekleri üzerinde yapılan ilk güvenilirlik analizleri ölçeğin iki yarısı arasındaki korelasyonun araştırılması üzerinde odaklanmıştır. Murphy ve Likert 1938'de ölçek kalitesi kavramını gözden geçirmişler, ölçeğin kalitesinin belirlenmesinde derece sayısı kadar ölçekteki madde sayısının da önemli olduğunu vurgulamışlar ve ölçekteki madde sayısı azalırken derece sayısının artırılmasıyla yarıya bölme güvenilirliğinin yükseldiğini bulmuşlardır.[31] Ölçek maddeleriyle ölçek dereceleri arasında denge kurulması daha sonra Bendig (1954), Komorita (1963), Komorita ve Graham (1965) gibi başka araştırmacılar tarafından da teyit edilmiştir (aktaran Munshi).[32]

Likert'te ölçek verilerinin dağılımı aşırı bir şekilde sağa veya sola çarpık ise maddeler ve dereceleri üzerinde yeniden çalışma yapmak gerekir. Dağılımın sağa veya sola çarpık olması yanıt yanlılığını gösterir. Araştırmacı yanıt yanlılığının önüne geçmek için ölçekteki maddelerin yarısını pozitif nitelikte ve diğer yarısını ise negatif nitelikteki maddelerden oluşturmalı ve ayrıca bu maddeleri ölçekte rasgele bir düzende sıralamalıdır. Klasik test kuramında işaretleme istikrarlılığı veya yeknesaklığı aritmetik ortalama değeri ile belirlenmeye çalışılırken modern test kuramlarından Rasch modelinde işaretleme yeknesaklığı veya düzensizliği "kişi-uyuşum indeks değeri" (pearson fit) ile belirlenir. Kişi-uyuşum indeks değeri, cevaplayıcıların ölçekteki maddelere ne denli tutarlı cevap verdiklerini belirler. Kişilerin cevapları *aşırı uyuşma* veya *yetersiz uyuşma* içinde olabilir. Klasik test kuramında ve Rasch modelinde, kişiler Likert ölçeğini eğer tuhaf veya eksantrik bir şekilde yanıtlamışlarsa bu kişilerin anketleri ölçümden çıkarılır.

Likert ölçeklerinde yanıtlayıcılar bazı ifadelere işaretleme yapmamışlarsa *atama yöntemine* başvurulur veya toplam puan yerine ortalama puanlardan hareket edilir. Likert ölçeklerinin güvenilir sonuçlar verebilmesi için maddelerin kökünde sadece tek bir yargı bulunmalı ve bir ifadeye farklı iki düşünce sıkıştırılmamalıdır.

Likert ölçeklerinde derece sayısı kadar önemli olan bir diğer nokta, derecelerin ifadelendirilmesi veya derecelerin etiketleme[a] biçimidir. Etiketler maddenin köküyle uyumlu ve anlamlı olmalıdır. Chang (1997), ölçek etiketlerinin farklı iki zamanda önemli ölçüde değiştirildiğinde sonuçların da değiştiğini ortaya koymuştur (aktaran Farmer).[33] Bazı yazarların ileri sürdüğü "araştırmacılar Likert ölçeklerinde dereceler için hangi etiketleri kullanacakları konusuyla çok fazla ilgilenmemelidirler" görüşünün gerçeği yansıtmadığı belirtilmiştir.[34] Araştırmacı ölçek maddelerinin köküne göre

[a] Etiket (label, response anchors). Ölçek dereceleri için yapılan tanımlamalar.

değişik etiketlerden yararlanabilir. Örneğin, bu etiketlerden bazıları Tablo 2-2'deki gibidir:

Tablo 2-2. Likert Tutum Ölçeğinde Kullanılabilecek Etiket Örnekleri

Miktar	*Seviye*	*Katılma*	*Derece*	*Puan*
Çok fazla	Çok yüksek	Kuvvetle katılıyorum	Çok iyi	5
Fazla	Yüksek	Katılıyorum	İyi	4
Gerektiği kadar	Aynı	Kararsızım	Vasat	3
Az	Düşük	Katılmıyorum	Kötü	2
Çok az	Çok düşük	Hiç katılmıyorum	Çok kötü	1

Etiketlerin tercihinde en çok tartışma konusu olan husus, etiketlerde nötr noktanın kullanılıp kullanılmayacağıdır. Bu konu, katılımcıların tarafsız olup olmayacaklarıyla ilgilidir. Nötr noktanın bulunmamasının ölçeğin güvenilirliğini etkilemesi konusunda farklı görüşler söz konusudur. Pazarlama araştırmalarında bilim adamları cevaplayıcıların kesin tercih yapmalarını istediklerinden nötr noktanın kullanılmaması eğilimine sahiptirler. Pazarlama araştırmalarında orta noktanın bulunmamasının ölçeğin geçerlilik ve güvenilirliğini etkilemeyeceği iddia edilmiştir.[35] Fakat öte yandan yapılan bazı araştırmalarda orta noktanın bulunmadığı durumda cevaplayıcıların daha fazla pozitif tutum gösterme eğilimi içinde oldukları görülmüştür. Ölçekte nötr noktanın bulunması halinde cevaplayıcıların %10 ilâ %20 kadarının nötr şıkkı işaretledikleri bulunmuştur. Ölçeğe eğer nötr şıkkı alınmazsa ölçüm hatasının artma ihtimali ortaya çıkar.[36] Sonuç olarak bu konuda, pek çok yazar derece sayısının belirlenmesinde ölçeğin içeriğinin, işlevinin ve ölçüm koşullarının etkili olacağını belirtmişlerdir (Cox, 1980; Friedman, Wilamowsky ve Friedman, 1981; Komorita, 1963; Matell ve Jacoby, 1971; Wildt ve Mazis, 1978; Garland'dan alınmışır).[37]

Likert ölçekleriyle ilgili olarak üzerinde düşünülmesi gereken bir diğer nokta ölçekte *Bilmiyorum*, *Benimle İlgili Değil*, *Görüşüm Yok* şıklarının yer alıp almayacağıdır. Ölçeğin kullanılması sırasında bazı cevaplayıcıların bu şıkları işaretlemesi uzak bir ihtimal değilse, ölçekte diğer derecelerden biraz uzakta altıncı veya sekizinci derece olarak bu şıklardan birine yer verilmelidir. *Kararsızım* şıkkı yerine *Bilmiyorum* veya *Görüşüm Yok* şık-

kını yazmak doğru değildir. Ancak cevaplayıcılar büyük ölçüde madde üzerinde düşünmekten kaçınarak bu şıkları işaretlemişlerse ölçeğin kullanılamaması gibi bir durum da söz konusu olabilir.[38]

Likert ölçeklerinin orijinalinde, cevaplayıcılar doldurdukları ölçek derecelerine numara verildiğini görmezlerken daha sonraki yıllarda ölçek derecelerine açık bir şekilde numara verilmeye başlanmıştır.[39] Ölçek dereceleri 1'den 5'e kadar veya 1'den 7'ye kadar numaralandırılabilir. Önceki yıllarda görülen –3, –2, –1, 0, +1, +2, +3 şeklindeki numaralandırma biçiminde cevaplayıcıların daha fazla pozitif sayılara yönelmeleri nedeniyle bu yaklaşım terk edilmiştir.[40]

Likert ölçeklerinin güvenilirliği cevaplayıcıların inançlarını, düşüncelerini, hislerini, tutumlarını veya davranışlarını dürüst ve samimi bir şekilde açıklamalarına bağlıdır. Bu konuda bilinçli bir gizleme söz konusu ise bu ölçeklere ait verilerin güvenilirliğinden söz edilemez. Likert ölçeklerinde güvenilirliği etkileyen bir diğer etken kişilerin gerçekten hissettikleri gibi değil, doğru olacağını düşündükleri gibi yanıt vermeleridir. Kişilerin ilgi duydukları konuda yüksek puan vermeleri ve sevmedikleri konuda ise düşük puan verme eğilimi içinde olmaları bu tekniğin güvenilirliği etkileyen bir diğer faktördür. Cevaplayıcıların "sosyal beğenilirlik" faktörü çerçevesinde cevap verme eğilimi içinde olmaları nedeniyle bu ölçüm aracının tek başına kullanılması halinde *ortak yöntem varyansı* sorunuyla karşılaşılır. Ortak yöntem varyansı, ölçümler arasındaki bulaşma etkisiyle korelasyon katsayılarının yüksek çıkmasıdır.

Anlamsal farklılık ölçekleri ve güvenilirlik. Bir başka indeks türü olan anlamsal farklılık ölçekleri C.E. Osgood tarafından 1950'li yıllarda geliştirilmiştir. Anlamsal farklılık ölçeği, yansıtıcı Likert ölçeğine benzer. Ancak Likert ölçeklerinden farklı olarak bu yöntemde ifadeler değil, sıfat çiftleri vardır. Sıfat çiftleri *Dostça/Düşmanca* örneğinde olduğu gibi iki kutuplu veya *Dostça/Dostça değil* örneğinde olduğu gibi tek kutuplu olarak düzenlenir. Bu ölçüm aracında derecelendirme boyutu sıfat çiftlerinin orta bölümünde yer almıştır. Yansıtıcı Likert ölçeklerinde olduğu gibi güvenilirlik için iç tutarlılık analizi, yarıya bölme yöntemi ve paralel formlar yöntemi ile test-yeniden test yöntemleri uygulanabilir. Osgood kendi yaklaşımında Likert ölçeklerinin yapısal düzenlemesine üç faktör ekleyerek farklı bir yapılanma ortaya çıkarmak istemiştir.

Değerlendirme faktörü. Bu faktörde iyi-kötü, taze-bayat, başarılı-başarısız gibi kelime çiftleriyle değerlendirme yapılır. Bu tür ölçekler daha çok başarı testleri için uygundur.

Potansiyeli açığa çıkarma faktörü. Bu faktörde ölçüm yapılan kişinin gelecekte gösterebileceği davranışları ve yetenekleri ortaya çıkarma önem kazanır. Güçlü-zayıf, kabiliyetli-kabiliyetsiz, büyük-küçük değerlendirmeleri bu faktörle ilgilidir.

Faal olma faktörü. Bu faktörde ölçüm yapılan kişinin hareket ve davranışları temel alınır. Kişinin aktif veya pasif olması, medenî cesareti, hızlı veya yavaş olması, çekingenliği bu faktör kapsamında değerlendirilir.

Uygulamada sıklıkla değerlendirme faktörü üzerinde durulmuş, potansiyeli açığa çıkarma ve aktivite faktörü daha az kullanılmıştır. Anlamsal farklılık ölçekleri için yansıtıcı Likert ölçeklerinde olduğu gibi öncelikle faktör analizi ve daha sonra iç tutarlılık analizleri yapılır. Bu ölçeklerde belirli bir alanla ilgili olarak tek boyutluluğu ortaya çıkarmak için sıfat çiftlerinin en az üç maddeden oluşması gerekir.

Stapel ölçekleri ve güvenilirlik. Stapel ölçekleri tek kutuplu ifadelerden oluşan ve daha çok pazarlama araştırmalarında kullanılan ölçüm araçlarıdır. Çift kutuplu ifade geliştirme güçlüğünün çekildiği durumlarda Stapel ölçeklerinden yararlanılır ve bu özelliğiyle anlamsal farklılık ölçeklerini ikame eden bir araçtır. Ölçeğin oluşturulması ve uygulanması anlamsal farklılık ölçeğine göre daha kolaydır. Bu ölçekte marka, ürün veya hizmetle ilgili olarak değerlendirilmesi istenen ifadeler/kavramlar belirlenir ve her bir ifade için +3 ilâ –3 veya +5 ilâ –5 arasında bir dereceleme ölçeği belirlenir. Artı 3, *Çok iyi tanımlıyor* ve –3 ise *Çok zayıf tanımlıyor* ifadesi ile açıklanır. Ölçekte sıfır nötr noktası bulunmaz, tam ortada bulunan nötr noktasının yerine değerlendirilmesi istenen ifade veya kavram yazılır. Stapel ölçeklerinin güvenilirliği oluşturulma biçimine göre değerlendirilir. Eğer yansıtıcı ölçek niteliğinde oluşturulmuşsa Likert ölçekleri için uygulanan güvenilirlik analizlerinin tümü Stapel ölçekleri için de geçerlidir. Oluşturucu ölçek niteliğinde oluşturulmuşsa tek boyutlu, çok boyutlu veya boyutsuz olma durumu araştırılır. Ölçekteki her bir madde diğerinden bağımsızsa maddeler arasında iç tutarlılık analizi yapmaya gerek yoktur. Ölçeğin geçerliliğinin sağlanmasıyla güvenilirliğinin de sağlandığı varsayımından hareket edilir.

Ölçekler ve Güvenilirlik

Literatürde sık kullanılan ölçekler beş grup altında toplanır: Thurstone ölçekleri, Bogardus ölçekleri, Guttman ölçekleri, Rasch ve Mokken ölçek-

leri. Söz konusu tutum ölçeklerinin türüne göre yapılan güvenilirlik analizleri de farklılaşır. Aşağıdaki bölümde kısaca bu tür ölçeklerde uygulanabilecek güvenilirlik analizlerine değinilmiştir.

Thurstone ölçekleri. Thurstone ölçekleri 1929'da, bir kişinin belirli bir konuda sahip olabileceği değişik tutumların *odak noktasını* tespit etmek üzere geliştirilmiştir.[a] Ölçeklerin geliştirilmesi iki yönteme göre yapılır. *Çiftli karşılaştırma* adı verilen birinci yöntemde belirli sayıda hakeme maddeler ikili gruplar halinde verilerek karşılaştırma yapmaları istenir. Hakemler sonunda hangi çiftlerin birlikte bulunmasının daha uygun olacağına karar verirler. *Eşit görünen aralıklar* adı verilen ikinci yöntemde ise tutum nesnesiyle ilgili maddeler 100 – 200 kadar uzmana değerletilir. Uzmanlar, ifadeleri 11 dereceli bir ölçek üzerinde değerlendirerek her bir ifadeye bir puan verirler. Daha sonra her bir ifade için uzmanların vermiş oldukları puanların aritmetik ortalama veya medyan değerleri saptanır. Hakemlerin bir maddeye önemli ölçüde farklı puanlar vermeleri halinde söz konusu maddeler ölçekten çıkarılır. Bu yöntemde maddelerin tek bir boyutu ölçme güvenilirliği için "uzmanlar arası uyuşma" ve "farklılaştırma indeksi" hesaplamaları yapılır. Maddelerin güvenilirliği, belirlenen ölçüm boyutunda eşit olarak dağılımına bağlıdır. Ancak gerçek hayatta bu koşulu tam olarak sağlamak oldukça güçtür.

Thurstone 1925 ilâ 1932 yılları arasında iyi yapılandırılmış ölçeklerin hangi özellikleri taşıması gerektiğine ilişkin 24 makale yazmıştır. Bu makalelerde iyi yapılandırılmış bir ölçeğin şu özellikleri taşıması gerektiğini belirtmiştir: tek boyutluluk, doğrusallık, soyutlama, değişmezlik, ölçüm nesnesinden etkilenmeme ve kişilerin puanlarından bağımsız olma.[41] Ölçümün doğrusallığı elde edilen verilerle aritmetik işlemler yapılabilmesidir. Ölçüm, araştırma yapılan örneklemden bağımsız olmalı ve ölçüm aracında işaretleme yapılmamış eksik veri bulunmamalıdır. Güvenilir nihaî ölçek maddeleri homojen ve tek bir boyutu ölçen bir özelliğe sahip olmalıdır.

Thurstone ölçeklerinde *Doğru-Yanlış* veya *Kabul Ediyorum-Kabul Etmiyorum* şeklinde sadece iki şık vardır. Bunlara bazen *Kararsızım* şıkkının eklendiği de görülür. Thurstone ölçekleri için dereceleme şıklarının artması veya azalması güvenilirlik analizlerine konu olamamıştır. Bu ölçeklerde daha çok uygun madde sayısının ne olması gerektiği konularında araştır-

[a] Geliştirildiği yıllarda dahi çok az kullanılan Thurstone ölçeklerine günümüzdeki bilim adamları fazla bir ilgi göstermemekte ve bu nedenle kullanılmamaktadır. Yöntemin ders kitaplarında yer almasının nedeni, öğrencileri pedagojik yönden yetiştirmeye yöneliktir.

malar yapılmıştır. Thurstone ölçekleri için 15 ilâ 30 arasındaki madde sayısının yeterli olacağı belirtilmiştir. On beşten az madde sayısı güvenilirliği azaltırken 30'dan fazla madde sayısı cevaplayıcıları isteksiz hale getirebilir.[42] Thurstone ölçeklerinde bir kişinin puanı, sadece en yüksek tartı değerine sahip maddenin puanı temel alınarak veya *Evet* yanıtı verilen maddelerin tartı puanlarının ortalaması alınarak belirlenir.

Thurstone ölçekleri için alfa, yarıya bölme güvenilirliği ve toplam puan – madde korelasyonu yöntemleri maddelerin tartı puanlarının farklı olması nedeniyle uygun değildir. Ölçek derecelerinin ikili olması, bu ölçeklerde KR-20 veya KR-21 güvenilirlik analizlerinin uygulanabileceği anlamına da gelmez. Thurstone ölçeklerinde ancak paralel formlar güvenilirliği ve test-yeniden test güvenilirliği yöntemleri uygulanabilir. Thurstone ölçekleri için temel güvenilirlik analizi, değerlendirme yaptırılan 100 kadar hakem arasındaki uyuşma yüzdesidir. Günümüzde Thurstone ölçeklerinin yerini maddelerin güvenilirliğini test eden Rasch ölçekleri almıştır.

Guttman ölçekleri. Guttman ölçekleri 1940'lı yıllarda Louis Guttman tarafından geliştirilmiştir. Guttman, 1944 yılında Likert ölçeklerindeki ham puanın özgün/örnek/standart bir cevaplama biçimi olmadığı sürece belirsiz kalacağını iddia etmiştir.[43] Ona göre, ham puanlar iyi bir ölçü olarak değerlendirilmez. Ham puanlar örneklem açısından yanlıdır ve gerçeği tam olarak göstermez. Yine ona göre bir ölçeğin ölçek sayılabilmesi için uç bir ifadeyi kabul eden bir kişinin, görüşlerinde o kadar keskin olmayan başka bir kişiye göre, uç ifadeden daha az keskin olan diğer ifadelerin hepsini kabul ettiği varsayılmalıdır. Guttman, "Biz, bir dizi maddeyi eğer bu maddeler kişileri keskinlik açısından ayırt edebiliyorsa ölçek olarak kabul ederiz"[44] görüşünü ortaya atmıştır.

Guttman yönteminde hem cevaplayıcılar hem de indeks maddeleri derecelendirilir. Katılımcılardan gelen cevaplar "scalogram" adı verilen bir tablo içinde sınıflandırılır. Bu tabloda her bir sütun bir kişiye denk gelir ve sütunlar azalan ham puanlar çerçevesinde sağa doğru uzanır. Tablodaki her bir sırada ise maddeler sıralanmıştır. Sıralar da ham puanların aşağıya doğru giderek azaldığı bir düzende ölçeklendirilir. (Bazı uygulamalarda kişilerle maddeler yer değiştirmiş olabilir.) Guttman'ın yapmak istediği, başarılı kişi ve maddelerin sol üst köşedeki hücrelerde yer alması başarısız kişi ve maddelerin ise sağ alt köşede bulunmasıdır. Böylece başarılılar ile başarısızlar arasında tam bir ayrım ortaya çıkacaktır. Guttman ölçeklerinin güvenilirliğinde "ölçeklenme" kavramı üzerinde durulmuştur. Ölçeğin *tek boyutluluğunu* belirlemek için *yeniden üretilebilirlik* (test-yeniden test uygulamalarında aynı puanların elde edilebilmesi) kriteri temel alınır. Sa-

dece bu kriteri karşılayan maddeler ölçeğe alınır. Eğer bir ölçek tek boyutlu ise, bir kişi diğerinden daha olumlu tutuma sahipse her bir ifadeye diğer kişilere göre eşit şekilde veya diğerlerinden daha olumlu bir şekilde yanıt vermelidir.

Duruma göre 5 ilâ 10 maddeden oluşan[a] Guttman ölçeğinde ilk ifade, *daha zayıf iken* son ifade *daha güçlü* tutumu veya eğilimi yansıtır. Guttman ölçeklerindeki maddelerin *yeniden üretilebilirlik* katsayısını hesaplamak için ölçek en az 100 kadar kişi üzerinde sınanır. Doğru tahmin yüzdesi yeniden üretilebilirlik katsayısı olarak isimlendirilir. Ölçeğin tek boyutluluğu için en az ,90 yeniden üretilebilirlik katsayısı aranır. Ancak üretilebilirlik katsayısının yüksek olması ölçeğin aynı zamanda geçerli olduğu anlamına gelmez. Yeniden üretilebilirlik katsayısı Eşitlik 2-4'teki formülle hesaplanır:

$$\textit{Yeniden üretilebilirlik} = \frac{1 - \textit{Hata sayısı}}{\textit{Yanıt sayısı}} . \qquad (2\text{-}4)$$

Aynı formülü değişik bir şekilde de ifade edebiliriz.

$$\textit{Yeniden üretilebilirlik} = \frac{\textit{Doğru tahmin sayısı}}{\textit{Toplam tahmin sayısı}} . \qquad (2\text{-}5)$$

Formülün simgesel olarak gösterimi ise, Eşitlik 2-6'daki gibidir:

$$YÜK = 1 - \frac{h}{kn}. \qquad (2\text{-}6)$$

$YÜK$ = Yeniden üretilebilirlik katsayısı.
h = Hata sayısı.
k = Madde sayısı.
n = Vaka sayısı.

[a] Ölçeğin geliştirilmesinde 10-12 kadar madde olması önerilirken, ölçek geliştirildikten sonra nihai ankette 4–6 maddenin yeterli olacağı belirtilmiştir. Pilot araştırma sırasında 20-30 kadar cevaplayıcının bulunması yeterli görülürken esas araştırmada bu sayı 100'den az olmamalıdır.

Guttman, ölçeklerinde maddeler *ortak geçişkenliğe* sahiptir (conjoint transitivity)[a], toplam puandaki belirsizlik ancak ortak geçişkenlik özelliği ile ortadan kaldırılabilir. Ölçeğin, anket uygulanan bireyleri tek bir boyut üzerinde farklılaştırıp farklılaştırmadığı *skalogram* analizi ile saptanır (*bk.*, Tablo 2-3). SPSS'te Guttman ölçeklerinin güvenilirlik analizlerini yapmak için bir hesaplama yöntemi yoktur.

Tablo 2-3. İdeal Bir Skalogram Tablosu (Tek Boyutlu ve Ölçeklenmiş Bir Tasarım)

Madde no				Kişi
2	4	1	3	sayısı
X	X	X	X	4
—	X	X	X	3
—	—	X	X	2
—	—	—	X	1
1	2	3	4	

X = Kabul
— = Ret

■ Nihai ölçekte maddelerin sıralanış biçimi aşağıdaki gibidir:

2 numaralı madde → puanı 1
4 numaralı madde → puanı 2
1 numaralı madde → puanı 3
3 numaralı madde → puanı 4

Az sayıda madden oluşan ölçeklerde maddelerin skologram değerlerini tespit etmek kolaydır, fakat madde sayısı arttığı zaman yığışımlı ölçek değerini bulmak için *Boolean analizi* gibi istatistiksel analiz tekniklerinden yararlanılır.

Rasch ölçekleri. Bu yöntemde ham puanlardan hareket edilerek eşit aralıklı bir ölçek oluşturulur. Rasch ölçeklerinde madde-yanıt kuramındaki

[a] Ortak geçişkenlik (conjoint transitivity). Matematiksel hesaplama yöntemiyle ilgili bir kavramdır. Ölçekteki her bir madde kişilerin yeteneklerini / görüşlerini belirli bir sıra içinde farklılaştırıyorsa bu maddeler *monoton türdeşliğe* sahiptir denilir. Eğer, eş anlı olarak kişiler de bu maddeleri doğal güçlük sırası içinde derecelendiriyorlarsa *çift yönlü monotonluk* söz konusudur ve Guttman bu durumu *ortak geçişkenlik* terimiyle ifade etmiştir.

tek parametreli lojistik hesaplama modeli kullanır. Bu ölçekte maddeler nispî güçlük derecelerine göre hiyerarşik bir sıraya sokularak tek boyutlu kavramsal yapıya uygun olmayan maddeler elenir. Sonuçta maddeler tek bir yeteneği veya tek bir özelliği ölçüyor olmalıdır.

Rasch ölçekleri daha önce "Sınıflandırılmış Verilerde Güvenilirlik" başlığında ele alınmıştır. Bu konuda okuyuculara ilgili başlığa başvurmalarını öneririz.

Bogardus ölçekleri. Bogardus ölçekleri bireylerin kendileriyle farklı sosyal kesimler arasındaki duygusal veya tutumsal mesafeyi ölçmek amacıyla geliştirilmiştir. Sosyal kesimler; etnik gruplar, farklı ülkelerin insanları, farklı dinî inanışlara sahip kişiler, farklı meslek grupları veya sosyal statü grupları olabilir. Ölçek, bireylerin farklı gruptan olan kişilerle gerçekleştirmeyi düşünebilecekleri ilişkilerin yoğunluğunu ve iletişimin sıklığını dikkate alarak ağırlıklandırılmıştır. Ölçeğin toplam puanı, en yoğun ilişkiyi tanımlayan madde puanına göre belirlenir. Bu ölçeklerin güvenilirliği ile ilgili olarak en fazla eleştiri yapılan konu, ölçek maddeleri arasındaki mesafenin metrik olarak tam eşit olmamasıdır. Puanlamada maddeler arasındaki mesafe eşit aralıklı ölçek gibi derecelendirilir, fakat bunu tam olarak ölçecek ve belirleyecek bir ölçüm aracına sahip olduğumuz söylenemez.

Mokken ölçekleri ve güvenilirlik. Mokken ölçekleri daha önce "Sınıflandırılmış Verilerde Güvenilirlik" başlığı altında ele alınmıştır. Bu konuda okuyuculara ilgili başlığa başvurmalarını öneririz.

Yansıtıcı – Oluşturucu Ölçekler ve Güvenilirlik

Ölçekler/indeksler ya *gizli bir yapıyı* ortaya çıkarmaya yönelik olarak veya belirli *kavramsal yapılara işaret etmek üzere* oluşturulurlar. Gizli bir yapıyı ortaya çıkarmaya yönelik olarak oluşturulan ölçeklere "yansıtıcı ölçekler" denir. Yansıtıcı ölçekler, gözlem değişkenlerinde değişkenliğe neden olan arka plandaki örtük kavramsal yapıları ortaya çıkarmayı hedefler. Bu ölçüm araçlarında pek çok gösterge tek bir kavramsal boyuttan, gizli faktörden etkilenir. Gizli faktörler "neden" olan bağımsız değişkenlerdir. Yansıtıcı göstergelerden oluşan ölçüm araçları, büyük ölçüde klasik test kuramına dayanır (Lord ve Novick, 1968, aktaran Diamantopoulos ve Winklhofer) ve bu ölçeklerde özellikle alan örneklemesi modeli temel alınır.[45]

Ancak pek çok durumda göstergeler etkilenen değil etkileyen ve oluşturan değişkenlerdir. Bazen birden fazla gösterge bir araya gelerek yeni bir

kavramsal yapının ortaya çıkmasına neden olur.[46] Örneğin; eğitim düzeyi, malî durum, oturulan semt, evin mülkiyeti ve araba sahibi olmak gibi birden fazla gösterge bir araya gelerek "sosyoekonomik düzey" (SED) adını verdiğimiz yeni bir kavramsal yapının ortaya çıkmasını sağlar. Bu tür ölçüm araçlarına *oluşturucu ölçek* adı verilir. Oluşturucu ölçeklerde her bir gösterge bir bütünün bağımsız bir parçasıdır. Oluşturucu göstergeler tanımlaması, ilk kez Blalock (1964) tarafından *bir yaratma veya değişimin nedeni olan maddeler* anlamında kullanılmıştır. Oluşturucu veya yansıtıcı olma özelliği, bir ölçekteki/indeksteki maddelerin etkileyen veya etkilenen olma durumunu belirler. Gizli yapılar[a] ise, duruma göre nedensel veya sonuç değişkenleridir.

Oluşturucu ölçek ve indeks ölçüm araçlarının niteliği konusunda bilim adamları arasında tam bir mutabakat bulunmamaktadır. Daha önce "İndeksler ve Güvenilirlik" başlığı altında oluşturucu ölçekleri/indeksleri üç grupta sınıflandırmıştık: tek boyutlu oluşturucu indeksler, çok boyutlu indeksler ve geniş kapsamlı kavramsal yapıyı ölçen oluşturucu indeksler (GKKY-Oİ). Bu bölümde daha çok geniş kapsamlı kavramsal yapıyı ölçen oluşturucu indeksler üzerinde durulmuştur.

Geniş kapsamlı kavramsal yapıları ölçen oluşturucu indekslerinin özellikleri. GKKY-Oİ, sahip oldukları belirli özelliklerle yansıtıcı ölçeklerden kolaylıkla ayırt edilebilir. Daimantopoulos ve Winklhofer (2003) bu özellikleri aşağıdaki gibi sıralamışlardır:[47]

1. Yansıtıcı ölçeklerde maddeler birbirlerinin yerine ikame edilebilir ve gerektiğinde çıkarılabilirken GKKY-Oİ'de bir maddenin çıkarılması yapının bir parçasının eksik olması anlamına gelir.
2. GKKY-Oİ'deki oluşturucu göstergeler arasındaki korelasyon katsayıları modelin tanımlanması açısından herhangi bir anlam ifade etmez. Bu nedenle göstergelerin geçerliliğini ampirik olarak sınamak çok zordur ve ölçek bu açıdan problemli olarak görülmüştür.

---

[a] Aslında oluşturucu ölçekler için "gizili yapı" kavramını kullanmak doğru değildir. Daha doğru ifadelendirme, "bileşik değişken" veya "temel bileşen" şeklinde olabilir. Ölçüm aracından birden fazla "yapı" ortaya çıkmışsa o zaman kavram çoğul ekiyle ifade edilir: bileşik değişkenler veya temel bileşenler. Ancak literatürde özellikle özensiz kullanıcıların sık aralıklarla *gizli yapı* kavramını kullandıklarını görüyoruz. Okuyucu kavramlar arasındaki anlam farklılıklarını göz önünde bulundurmalıdır.

3. Göstergeler arasındaki korelasyonların negatif veya pozitif olması, katsayıların yüksek veya düşük olması önemli değildir. Bu bulgulara rağmen ölçek anlamlı bir şekilde belirli bir yapıya veya yapılara işaret ediyor olabilir.
4. Geniş kapsamlı kavramsal yapıları ölçen oluşturucu göstergelerde hata terimi yoktur. Oluşturucu ölçeklerde hata terimi yerine *karışıklık* (veya "artık varyans") terimi kullanılır ve terim Grek harflerinden 14'üncüsü olan "zi" (ξ) harfi veya simgesiyle gösterilir. Maddelerin *karışıklık* terimleri birbirleriyle ilişkisizdir.
5. En az üç göstergesi bulunmayan bir model, GKKY-Oİ olarak tanımlanamaz. Modele ilişkin olarak tahmin yürütülebilmesi için geniş modellerle çalışılmalı ve gizli değişkene ilişkin sonuçları içeren belirli sayıda gösterge bulunmalıdır.
6. Geniş modellerde dahi sorunlarla karşılaşılabilir. Artık varyansın ortaya çıkabilmesi için üst düzey gizli değişkenden diğer alt düzeydeki gizli değişkenlere yönelik en az iki rota tanımlaması yapılmış olması gerekir. Alt düzeydeki değişken de sonuç göstergelerini içeriyor olmalıdır.

Sıralanan bu özellikler dikkatle incelenirse, klasik test kuramındaki güvenilirlik ve geçerlilik analizlerinin GKKY-Oİ için uygun olmadığı görülür. Oluşturucu ölçeklerdeki karışıklık terimi (distrubance term) maddelerin ölçülmek istenen kavramsal yapıyı tam olarak ortaya çıkarıp çıkaramayacağıyla ilgilidir. Modeli veya kavramsal yapıyı ölçmek üzere oluşturulan ölçeğin maddeleri eksiksiz ise karışıklık terimi, $\xi = 0$'dır. Karışıklık veya rahatsızlık terimi; kuramdaki yetersizliği, verilerdeki yetersizliği, modelin basit olması nedeniyle ölçümdeki yetersizliği veya maddeler arasındaki regresyon doğrusunun tam doğrusal bir niteliğe sahip olmadığını gösterir. Bir ölçümde *kavramsal yapı*, *maddeler* ve *karışıklık terimi* arasındaki ilişkiler Eşitlik 2-7'deki formülle ifade edilmiştir:

■ Oluşturucu ölçeklerde *karışıklık* (veya rahatsızlık) terimi.

$$\text{Kavramsal yapı } A = \gamma_{1\,(\text{madde } 1)} + \gamma_{1\,(\text{madde } 2)} + \cdots \gamma_{n\,(\text{madde } n)} + \xi \,. \qquad (2\text{-}7)$$

Formülde γ simgesi birinci maddenin kavramsal yapıya yaptığı katkıyı gösterir. GKKY-Oİ göstergelerinin ölçüm kalitesini değerlendirmede kul-

lanılacak hesaplama yöntem ve teknikleri literatürde henüz çok yenidir ve bilim adamları arasında bu konuda tam bir mutabakat yoktur.

GKKY-Oİ geliştirme. Daimantopoulos ve Winklhofer (2003) geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçek geliştirmede dört temel soruna işaret etmiştir: kavramsal alanı belirleme, göstergeleri belirleme, göstergelerin çoklu doğrusallık özelliğini araştırma ve göstergelerin dış geçerliliğini belirleme.[48]

GKKY-Oİ'lerde araştırılan kavramsal alanın yansıtıcı ölçeklere göre çok daha belirsiz olduğu bildirilmiştir. Yansıtıcı ölçeklerde gizli faktörler ölçüm alanını net bir şekilde ortaya koyarken GKKY-Oİ'lerde temel bileşen daha geniş ve bir ölçüde belirsizdir. GKKY-Oİ'lerde, ölçüm aracına girecek maddeleri belirlemek için göstergelerin tamamına ulaşmak gerekir. Yansıtıcı örneklerde alan örneklemesi yöntemi kullanılırken oluşturucu ölçeklerde kavramsal alanın geniş olması nedeniyle tüm sınırları temsil edecek şekilde madde seçilmesi gerekir. Tek bir madde dahi olsa alanın tüm temel ögeleri ölçekte temsil edilmiş olmalıdır. GKKY-Oİ'lerde bilim adamının karşısına çıkacak önemli bir sorun, çoklu doğrusallık veya maddelerin koşutluk özelliğidir. Oluşturucu ölçüm modeli çoklu regresyon analizine dayandığından model örneklem büyüklüğünden ve maddelerin koşutluk özelliğinden etkilenir. Maddeler aşırı ölçüde koşutluk özelliğine sahipse, bir taraftan herhangi bir göstergenin bileşik değişken üzerindeki etkisini saptamak güçleşecek ve diğer taraftan ise ölçüm özelliği zayıflamış maddelerin "artık göstergeler" olarak değerlendirilip ölçekten çıkarılması gerekecektir. Dördüncü sorun, dış geçerlilikti. Dış geçerlilik, GKKY-Oİ maddelerinin "dış değişken" adı verilen başka bir değişken veya başka değişkenlerle karşılaştırılması esasına dayanır. Bu karşılaştırmada anlamlı korelasyon katsayısına sahip olan maddeler ölçeğe alınır. Ancak bu karşılaştırma sırasında bir maddenin korelasyonu düşük çıkmışsa bu maddenin ölçekten çıkarılması gerekir. Öte yandan çıkarma işlemine başvurulması halinde yapının değişmesi tehlikesi söz konusu olacağından çıkarma işleminde yapının değişip değişmediğinin, daralma durumunun ortaya çıkıp çıkmadığının ayrıca araştırılması gerekir. Geçerlilik çalışmalarında "*dış değişken* neye göre belirlenecektir?" sorusunun yanıtı net değildir. Daimantopoulos ve Winklhofer bu konuda ölçülmek istenen yapının özünü tanımlayan ve tek bir göstergeden oluşan "global madde" yaklaşımının kullanılmasını önermişlerdir.[49] GKKY-Oİ maddelerinin çok sayıda olması bu ölçeğin uygulanabilirliğini azaltacağından bir şekilde madde ayıklama yöntemine de başvurulması gerekmektedir. Literatürde bunun için üzerinde anlaşmaya varılmış bir yöntem bulunmamaktadır.

Bilim adamı, GKKY-Oİ maddelerinin sayısının tespiti için temel bileşenler analizini veya kısmî en küçük kareler yöntemini kullanabilir. Göstergelerin geçerliliğini saptamak için kullanılabilecek bir diğer yaklaşım *çoklu göstergeler ve çoklu nedenler – ÇGÇN* (multiple indicators and multiple causes – MIMIC) modelini sınamaktır.[50]

GKKY-Oİ'lerin kullanım alanları. Yansıtıcı ölçekler psikoloji ve davranış bilimleri alanında kullanılırken oluşturucu ölçekler, sosyoloji, yönetim organizasyon, pazarlama, insan kaynakları ve siyaset bilimi gibi alanlarda yaygınlık kazanmıştır. Geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçekler sosyoekonomik durumun saptanmasında olduğu gibi kişisel özelliklerin belirlenmesi amacıyla kullanılabilmesine karşılık, bilim adamları bu yönteme daha fazla örgütsel ve sosyal nitelikteki kavramsal yapıların ölçümü amacıyla başvurmuşlardır. Bu ölçeklerde çalışma grupları, firmalar, örgütler inceleme alanı olarak görülmüştür. Bilim adamı, geliştirdiği ölçüm aracının yansıtıcı mı yoksa tek boyutlu mu, çok boyutlu mu veya geniş kavramsal yapıları ölçen oluşturucu bir ölçek mi olduğu konusunda net bir fikre sahip olmalı ve araştırma raporunda bunu açık bir şekilde ifade etmelidir. Bilimsel dergi editörleri ve hakemler gizli yapıların ölçülmek istenmesi halinde oluşturucu ölçeklerden çok yansıtıcı ölçekleri kullanan yazarların çalışmalarını yayımlama eğilimi içindedirler.[51]

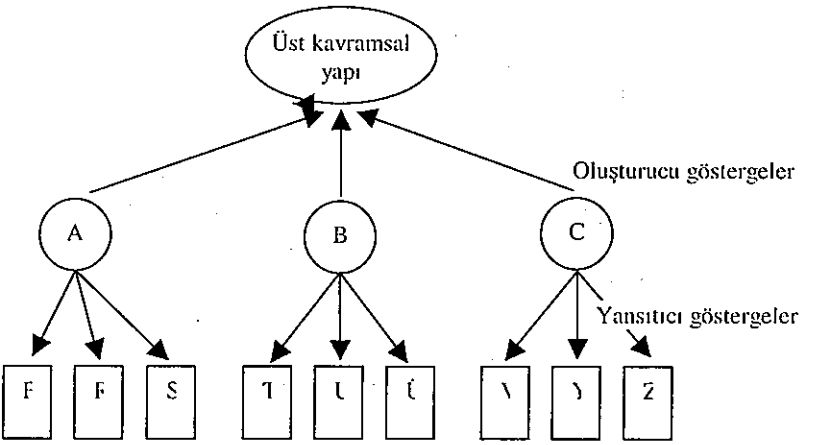
Ölçeğin oluşturucu - yansıtıcı olma özelliğinin saptanması. Belirli durumlarda ölçek veya indeks geliştirilirken söz konusu ölçeğin yansıtıcı mı yoksa oluşturucu ölçek mi olacağı; eğer oluşturucu ölçek ise tek boyutlu, çok boyutlu veya geniş kapsamlı kavramsal bir yapıyı mı ölçeceği önceden belirlenmelidir. Ancak bazı durumlarda göstergeler arasındaki ilişkilerin doğrusallığını hemen veya kolaylıkla tespit etmek mümkün olmayabilir. Edwards ve Bagozzi'nin (2000) belirttiği gibi bir yapı ile bir dizi gösterge arasındaki ilişkilerin doğrusallığını belirlemek bazen oldukça zordur (aktaran, Siguaw).[52] Geliştirilen ölçekle ilgili olarak, literatürde yapılan incelemeler kavramsal yapının yansıtıcı nitelikteki ölçüm araçları kullanılarak ölçüldüğünü ortaya koymuşsa bilim adamı bu verileri temel alır. Böyle bir veri bulunamamışsa maddeler arasındaki korelasyonlar ve faktör analizi sonuçları incelenerek ölçeğin yansıtıcı / oluşturucu olup olmadığına karar verilir. Bollen (1989) yansıtıcı ve oluşturucu ölçek tercihinin bilim adamının nedensellik önceliğini göstergeler ve gizli değişkenlerden hangisine verdiğine göre belirleneceğini belirtmiştir (aktaran Diamantopoulos ve Winklhofer, 2003).[53] Kişilik ve değişik tutumların ölçülmesi gibi konularda geliştirilen göstergeler arka plandaki bir özelliği ölçmeye çalıştığın-

dan genelde yansıtıcı bir özelliğe sahiptir. Göstergeler eğer belirli bir kavramsal yapıyı ortaya çıkaracak açıklayıcı maddeler olarak görülüyorsa oluşturucu indeks geliştirme yöntemine başvurulur.

Yansıtıcı ölçek geliştirme çalışmalarında *ortak varyansı* ortaya çıkararak maddeler arasındaki korelasyonlara büyük ölçüde önem verilirken oluşturucu indeks geliştirme çalışmalarında soyut varyans açıklanmaya çalışılır. Yansıtıcı ölçek geliştirme, ortak faktör analizi ve / veya güvenilirlik analizleri yöntemiyle başlarken *oluşturucu indeks* geliştirme çalışmalarında ise, öncelikle ölçeğin türü araştırılır. Tek boyutlu ve çok boyutlu ölçeklerde temel bileşenler analizi; geniş kapsamlı kavramsal yapıları ölçen oluşturucu indekslerde ise koşutluk özelliğiyle birlikte kısmî en küçük kareler yöntemi uygulanır. Geniş kapsamlı kavramsal yapıları ölçen oluşturucu indekslerde koşutluk özelliği varsa temel bileşenlerin veya ayırt edici yapıların ortaya çıkması zorlaşır.

Çok düzeyli oluşturucu ölçekler. Bilim adamı, *oluşturucu ölçek* geliştirirken ya *tek bir temel bileşeni* ortaya çıkarmayı hedeflemiştir veya *birden fazla temel bileşeni* ortaya çıkarmak istiyordur. İkincisinde, çok düzeyli bağımsız oluşturucu yapıların varlığından söz ederiz. Çok düzeyli, bağımsız *oluşturucu yapılara* sahip ölçeklerin anlamı, ölçmeye çalıştığımız kavramsal bir yapının kendi içinde değişik sayıda alt ölçeklerden oluşabileceğidir. Üst düzeyde kavramsal yapıyı ortaya çıkaran boyutların (veya bileşenlerin) birbirleriyle yüksek derecede korelasyon içinde olmaları gerekmez. Örneğin, iş tatminini ölçmeye çalışan bir bilim adamı iki yaklaşımdan birini seçebilir. Birinci yaklaşımda sadece genel iş tatmini üzerinde odaklanarak iş tatminin ögelerini bir kenarda tutabilir. Böyle bir durumda tek bir oluşturucu veya yansıtıcı yapı ortaya çıkacaktır. İkinci yaklaşımı tercih etmişse; *yönetim biçimi, iş arkadaşları, ücret, çalışma koşulları* gibi birden fazla *oluşturucu alt ölçek* oluşturma durumunda kalacak ve bu alt ölçeklerin hepsi birlikte *iş tatmini* kavramsal yapısının ortaya çıkmasına neden olacaktır. Daha dikkatli bir şekilde incelenirse bu tür ölçek geliştirme çalışmaları aynı zamanda yansıtıcı ve oluşturucu ölçeklerin her ikisini de içeriyor olabilir. Örneğin *ücret* kavramsal yapısıyla bu yapıyı ölçen maddeler arasındaki ilişkiler *yansıtıcı* nitelikte iken, kavramsal yapıların bir araya gelerek *iş tatmini* isimli üst kavramsal yapıyı ortaya çıkarması *oluşturucu yapı* örneğidir. Bilim adamı, ölçeklerin oluşturucu / yansıtıcı yapısına karar verirken kavramsal yapılarla maddeler arasındaki ilişkileri ve maddelerin bağımlılıklarını göz önünde bulundurmalıdır. Örneğin, ücret kalemine ilişkin maddeler arasındaki bağımlılık çok yüksek iken Ücret, Yönetim Biçimi, Çalışma Koşulları, İş Arkadaşları gibi yapılar arasındaki

ilişkiler bir ölçüde birbirinden bağımsızdır. Yapılar arasındaki korelasyon katsayıları çok yüksek çıkmayabilir. Ancak bu bağımsız yapılar bir araya gelerek ayrı bir üst yapı oluşturduğuna göre alt testlerin toplam/ortalama puanları *oluşturucu gösterge* olarak değerlendirilir.



Şekil 2-4. Yansıtıcı – oluşturucu yapılar arasındaki ilişkiler.

Şekil 2-4'de A, B ve C testlerinin bileşik puanları oluşturucu göstergelerdir. Üç farklı bir test bir araya gelerek yeni ve farklı bir *bileşik değişken* ortaya çıkarmıştır.

Yansıtıcı ve oluşturucu ölçüm araçları ve madde sayısı. Yansıtıcı ölçeklerde daha "cimri" bir yapı (daha az sayıda madde içermesi) söz konusu iken geniş kapsamlı kavramsal yapıları ölçen oluşturucu indekslerde madde sayısı görece daha fazladır. Maddelerin birbirleriyle yüksek korelasyon katsayılarına sahip olma gibi bir zorunluluğun bulunmaması nedeniyle oluşturucu ölçeklerde madde elemesine, madde düşürülmesi yöntemine başvurulmaz.

Geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçekler ve güvenilirlik. Oluşturucu maddelerde gizli bir yapının ortaya çıkarılması söz konusu olmadığından oluşturucu ölçekler, diğer testlerde olduğu gibi iç tutarlılık, yarıya bölme ve paralel formlar yöntemlerinin uygulanmasına

müsait değildir (Bollen 1984; Bollen ve Lennox 1991, aktaran Chin).[54] İç tutarlılık ve tek boyutluluk gibi yöntemler oluşturucu ölçeklerde ölçüm modelinin kalitesini yargılamak için kullanılamaz. Geniş kapsamlı kavramsal yapıları ölçen oluşturucu bir ölçeğin maddeleri arasındaki korelasyonlar negatif, pozitif işaretli ve hatta sıfır korelasyon katsayısına sahip olduğu halde *geçerli* olabilir. Çünkü bu ölçeklerde bileşik değişken, *neden* değil, bir *sonuçtur*. Bu tür ölçüm araçlarında korelasyon ve iç tutarlılık analizleri geçersizdir (Mathieson ve d. 1996, aktaran Chwelos ve Benbasat).[55] Bilim adamı, geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçeklerde faktör yüklerini değil, göstergelerin ağırlıklarını incelemeli ve ağırlıkların ne ölçüde 1,0'e yaklaştığına bakmalıdır. Madde ağırlıkları, standart regresyon analizindeki beta katsayılarıdır. Araştırmacı maddeleri karşılaştırarak göreceli olarak değerlerini belirlemelidir. Yansıtıcı yapılara göre oluşturucu yapılarda maddelerin ağırlıkları daha düşüktür.[56] Geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçeklerde, maddelerden çok modelin ve kuramsal yapıların güvenilirlik ve geçerliliği araştırılır. Bunun için, geçerlilik ve güvenilirliği eş zamanlı olarak analiz eden PLS yazımlının kullanılması önerilmiştir. Bilim adamı geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçeklerde güvenilirlik analizi yapmayı düşünüyorsa modelin güvenilirliği için önce PLS analizini yapmalı ve daha sonra iç tutarlılık, faktör yükü ağırlığı ve çapraz yüklerin ne olduğunu araştırmalıdır (Chin, 1998, aktaran Ashill ve Jobber).[57]

Geniş kapsamlı kavramsal yapıları ölçen oluşturucu ölçekler ve istatistiksel analiz yazılımları. Yansıtıcı yapılar LISREL, AMOS, EQS gibi yapısal eşitlik modellerini test eden yazılımlarla analiz edilirken geniş kapsamlı kavramsal yapıları ölçen oluşturucu yapılar için PLS Graph 3.0 ve SAS gibi yazılımların kullanılması önerilmiştir. PLS Graph yazılımı hem yansıtıcı ve hem de oluşturucu yapıları analiz etme özelliğine sahiptir.[58] PLS Graph yazılımı *kısmî en küçük kareler* (KEKK) yöntemini kullanır. Yöntem, bireysel maddelerin ağırlıklarını optimize ederek (olabilecek en iyi hale getirerek) modelin içerdiği bağımlı değişkenin varyansını maksimize etmeye çalışır. Bu analiz sonucunda elde edilen R^2 değeri ve yapılar arasındaki anlamlı ilişkiler belirlenen modelin ne ölçüde iyi çalıştığını gösterir.[59]

Geniş kapsamlı kavramsal yapıları ölçen oluşturucu yapılarda göstergelere ait nispeten *düşük ağırlık değerleriyle* karşılaşılmışsa, bu durum hemen ölçüm modelinin "zayıf" çıktığı şeklinde yorumlanmamalıdır. Bilim adamı modeli aynı zamanda yansıtıcı ölçek özelliğiyle de değerlendirerek son kararını ondan sonra vermelidir. PLS yazılımının, göstergelerde önem-

li olmayan rotalara ve faktörlere düşük ağırlık değerleri verdiği bildirilmiştir. KEKK yöntemi temel bileşenler analizine benzer, bu analizin tekrarlamalı hesaplamalara dayanan bir bileşimidir. Yazılıma öncelikle *dış yapılar* ve daha sonra *iç yapılar* tanıtılır. Dış yapılar, *neden olan* (bağımsız) değişkenler (antecedents[a]) olgusuyla ilintilidir. İç yapılar ise, sonuçlardır ve bağımlı değişkeni tanımlar. PLS Graph yazılımında KEKK yönteminin uygulanabilmesi için "karmaşık yapılı oluşturucu ölçeklerde" gösterge sayısının en az 10 katı ve "tek bir iç yapıyı" ortaya çıkarmayı amaçlayan diğer oluşturucu ölçeklerde ise gösterge sayısının en az üç katı kadar katılımcıya veya örnek kütle büyüklüğüne ulaşılması gerekir (Barclay ve d., 1995, aktaran Bontis).[60]

Tipolojiler ve Güvenilirlik

Cevaplayıcıların belirli bir konunun alt dilimleriyle ilgili olarak sahip oldukları görüşlerin karşılaştırmalı olarak saptanmasına imkan sağlayan ölçüm araçlarıdır. Araştırmacı, tipolojilerde[b] nominal ölçek verisi niteliğindeki iki değişken ve boyutun ortak etkisini birlikte görmek ister. Ölçekler ve indeksler çoğunlukla tek boyutlu olarak oluşturulurken, tipolojiler çok boyutludur. Bir tipolojide iki veya daha fazla alt boyut arasındaki etkileşimler araştırılır. Basit bir tipolojide iki değişken ve her birinde nominal ölçek verisi niteliğinde iki alt seçenek vardır ve böylece dört gözlü bir pencere ortaya çıkar (*bk.*, Tablo 2-4). Penceredeki her bir göze "tip" adı verilir.

Tablo 2-4. Kişilik Tiplerinin Temel Alındığı Bir Tipoloji Örneği

Kişilik Tipleri	Duygusal	Rasyonel
İçedönük	A	B
Dışadönük	C	D

[a] Neden olan değişkenler veya göstergelerdir. Oluşturucu ölçeklerde *göstergelerin* her biri dış yapıları tanımlar. Temel bileşenler ise iç yapılardır. İç yapılar göstergelerin fonksiyonudur. Yansıtıcı ölçeklerde ise, durum tam tersinedir. Bu ölçeklerde neden olan değişkenler, gizli faktörlerdir. Gizli faktörler bağımsız değişkenlerdir ve dış yapıları tanımlar. Bağımlı değişkenler ise göstergeler veya değişik ölçüm sonuçlarıdır ve iç yapıları tanımlar. Burada göstergeler dış yapıların fonksiyonudur.

[b] İngilizce *tipoloji* kavramını tam olarak Türkçe ifade etmek istersek "belirli tip sınıflarına sokma" anlamında *tipleştirme* sözcüğünü kullanabiliriz.

Şekil 2-5'te de görüldüğü gibi, tipolojideki veriler nominal veya sıralı ölçek verisi niteliğindedir ve daha çok bağımsız değişken olarak kullanılır.[61] Tipoloji verileriyle bağımsız değişken olarak ki-kare analizi ve bağımlı değişken olarak ise iki yönlü varyans analizi (İYVA) tekniği uygulanır. Örneğin, bir öğrencinin öğretmeninin verdiği ödevlerle ilgili görüşleri şu şekilde saptanabilir: (a) Bilgisayar ve TV izleme ödevlerinin her ikisini de seviyor. (b) Bilgisayar ve TV izleme ödevlerinin her ikisini de sevmiyor, (c) Bilgisayar ödevini seviyor, fakat TV izleme ödevini sevmiyor, (ç) Bilgisayar ödevini sevmiyor fakat TV izleme ödevini seviyor. Tipolojiler için klasik test kuramının güvenilirlik analizlerini yapmak zordur. Bu tür ölçümlerde ancak test-yeniden test yöntemiyle cevaplayıcıların tutarlı yanıt verip vermedikleri araştırılabilir.

DERECELENDİRME GÜVENİLİRLİĞİ

Derecelendirme güvenilirliği, 2 dereceden başlayıp 15 ve hatta bazı yazarlara göre 20 dereceye kadar uzanan ölçüm noktalarıyla ilgilidir. Literatürdeki tartışma daha çok Likert ölçeklerinin çevresinde yoğunlaşmıştır.

Çift Sayı İle Biten Dereceleme Ölçekleri

Pazarlama, yönetim-organizasyon, muhasebe gibi belirli bilim alanlarında araştırmacılar çift sayı ile biten dereceleme ölçeklerini tercih etme eğilimi içindedirler. Çift sayı ile biten dereceleme ölçeklerinde ölçeğin son noktası 2, 4, 6, 8 veya 10 gibi bir değerdir. Bu tür dereceleme ölçeklerinin güvenilirliği konusundaki bulgular çelişkilidir. Yapılan araştırmalarda güvenilirliği etkileme açısından çift sayı ile biten dereceleme ölçeklerinin lehinde ve aleyhinde görüş bildiren bilim adamları vardır. Çift rakamla biten dereceleme ölçeklerinde, iki serili (iki dereceli) yaklaşımları diğer çift rakamlı dereceleme ölçeklerinden ayrı tutmak gerekir. Bu tür ölçekler başarı veya başarısızlığı belirleyen ölçüm araçlarıdır ve genellikle başka bir şekilde de kodlanmaz. Sorun, Likert tipi ölçeklerde 4, 6 veya 8 gibi çift rakamlı ölçek derecelerinin kullanılıp kullanılamayacağıyla ilgilidir. Pazarlama gibi belirli bilim dallarında çift sayılı ölçek derecelerinin kullanılması literatürde yaygınlık kazanmıştır.

Tek Sayı İle Biten Dereceleme Ölçekleri

Araştırmacılar Likert türü bir ölçek geliştirmeye karar verdikleri zaman bu ölçekte her bir ifadenin kaç dereceli bir boyut üzerinde değerlendirileceğini de belirlemek zorundadırlar. Literatürdeki kaynaklar incelendiğinde Likert ölçeklerinin genelde 3 dereceden başlayıp 11 dereceye kadar uzanan tek sayılı ölçekler olduğu görülür. Ölçeğin kaç dereceli olarak düzenleneceğine anket uygulanacak kişilerin yaş düzeylerine, okur yazarlık durumuna, eğitim durumlarına, motivasyonlarına ve kişilerin zihinsel kapasitelerine bakılarak karar verilir.[62] Likert (1932) yazmış olduğu orijinal makalesinde derece sayısının o kadar önemli olmadığını, bir ölçek eğer beş dereceli olarak yapılandırılmışsa ortadaki üç rakamının *nötr nokta* olarak ayrılmasının gerektiğini belirtmiştir. Daha sonraki yıllarda yapılan araştırmalarda da güvenilirlik katsayısı büyüklüğünün ölçekteki derecelerden bağımsız olduğu ifade edilmiştir. Ancak son zamanlarda yapılan araştırmalarda ölçüm boyutundaki derece sayısının artmasıyla birlikte cevapların varyansının arttığı ve derece sayısının bir ölçeğin geçerlilik ve güvenilirliğini etkilediği ortaya çıkmıştır.[63] Bir dereceleme ölçeğinin varyansı, ölçek dereceleri 5'ten az olduğunda en alt düzeye düşmekte (ölçek ayırt edici özelliğini kaybetmekte) ve 7'nin üstüne çıktığında ise varyansta önemli bir değişiklik gözlenmemektedir.[64] Cicchetti, Showalter ve Tyrer (1985) gözlemciler arası güvenilirliğin 7 dereceye kadar arttığını bu dereceden sonra ise her hangi bir artış gözlenmediğini ifade etmişlerdir (aktaran Hoon).[65] Araştırmalar ölçekteki derece sayısı 9'un üzerine çıktığında güvenilirliğin düştüğünü ortaya koymuştur (Bendig, 1953; Preston ve Colman, 2000, aktaran Hoon).[66] Bu konuda yapılan diğer araştırmalarda da maksimum güvenilirliğin 5 veya 7 dereceli ölçekle elde edilebileceği sonucuna varılmıştır. Buna göre tek sayı ile biten dereceleme ölçeklerinin çift sayı ile biten dereceleme ölçeklerinden daha güvenilir olduğunu söyleyebiliriz.

ALINTI YAPILAN KAYNAKLAR

[1] H.M. Jr. ve A.B. Blalock, *Methodology in Social Research* [Sosyal Araştırmalarda Yöntem], (New York: MacGraw Hill, 1971), 20.

[2] L.A. Becker, "Crosstabs [Çapraz Tablolar]," <http://www.uccs.edu/~lbecker/ctabs1.htm> (13.10.2002).

[3] N. Verhelst, "Testing the Unidimensionality Assumption of Rasch Model [Rasch Modelinin Varsayımları Altında Tekboyutluluğu Ölçme]," <http://www.ppm.ipn.uni-kiel.de/mpr/issue15/art2/verhelst.pdf> (13.10.2002).

[4] BCSP, "Examinations [Sınavlar]," <http://www.bcsp.org/bcsp_newsletters/2_99Newsletter/examinations.html> (07.10.2002).

[5] Ryerson, "Exam Analysis [Sınav Analizi]," t.y., <http://www.ccs.ryerson.ca/fac_staff/index.cfm?cblockID=235> (07.10.2002).

[6] University of Northern Colorado, "Exam Scoring [Sınav Puanlaması]," 03 Eyl 2002, <http://www.unco.edu/it/security/scoring.htm> (07.10.2002).

[7] A.V. Alphen, R. Halfens, A. Hasman ve T. Lmbos, "Likert or Rasch [Likert mi Rasch mı?" <http://www.rasch.org/rmt/rmt82d.htm> (16.10.2002).

[8] "Rasch Analysis of Beta Test [Beta Testinin Rasch Analizi]," <http://www.eskimo.com/~miyaguch/iqtest/rasch.html> (28.09.2002).

[9] WINSTEPS, "Reliability of Measures [Ölçümlerin Güvenilirliği]," <http://www.winsteps.com/winman/reliability.htm> (22.05.2004).

[10] Aynı.

[11] Institute for Objective Measurement, "Relating Cronbach and Rasch Reliabilities [Cronbach ve Rasch Güvenilirlikleri Üzerine]," <http://www.rasch.org/rmt/rmt132i.htm> (13.10.2002).

[12] Institute for Objective Measurement, "KR-20 or Rasch Reliability: Which Tells the "Truth"? [KR-20 mi Yoksa Rasch Güvenilirliği mi: Hangisi Gerçeği Söylüyor?]," <http://www.rasch.org/rmt/rmt113l.htm> (13.10.2002).

[13] "What Advantages Does Item Response Theory (IRT) Item-pattern Scoring Offer ?[Madde Yanıt Kuramı Ne Tür Avantajlar Sunuyor?]," <http://www.google.com.tr/search?q=cache:q_wbfh8Ru2UJ:www.ctb.com/static/about_assessment/popup_faq8.jsp+reliability+%22standard+error+of+measurement%22&hl=tr&ie=UTF-8&inlang=tr> (05.08.2003).

[14] R.K, Hambleton, "Hambleton's 9 Theses [Hambleton'un 9 Önermesi]," <http://www.rasch.org/rmt/rmt62d.htm> (22.05.2004).

[15] M. McAlpinev, "Item Analysis [Madde Analizi]," <http://www.scrolla.hw.ac.uk/symp/mcalpine.ppt< (22.05.2004).

[16] Aynı.

[17] D. Garson, "Scales and Standard Measures [Ölçekler ve Standart Ölçümler]," <http://www2.chass.ncsu.edu/garson/pa765/standard.htm> (16.10.2002).

[18] R.J. Mokken, "Psychometric Quality of Scales [Ölçeklerin Psikometrik Kalitesi]," <http://homeusers.brutele.be/nova-inrct/Nederl/stress/Psychometric%20Quality%20of%20Scales%20Extended%20version.doc.> (16.10.2002).

[19] "Scales and Standart Measures [Ölçekler ve Standart Ölçümler]," <http://www2.chass.ncsu.edu/garson/pa765/standard.htm> (13.10.2002).

[20] Babbie, "Indexes, Scales, and Typologies [İndeksler, Ölçekler ve Tipolojiler]," <http://www2.chass.ncsu.edu/Judge/PS371/Lecture6b_part2.htm> (24.09.2002).

[21] W.M.K Trochim, "General Issues in Scaling [Ölçeklemede Genel Sorunlar]," <http://trochim.human.cornell.edu/kb/scalgen.htm> (30.11.2002).

[22] D.L. Classon ve T.J. Dormody, "Analyzing Data Measured by Induvidual Likert Type Data [Likert Tipi Maddelerin Analizi]," <http://pubs.aged.tamu.edu/jae/pdf/Vol35/35-04-31.pdf> (30.11.2002).

[23] C. Tutzauer, "Strategic Information Systems Planning Success [Başarının Planlanmasında Stratejik Bilişim Sistemleri]," <http://www.google.com.tr/search?q=cache:uAV3IKr4w5YJ:www.acsu.buffalo.edu/~tutzauer/SISP.pdf+formative++reflective+retest+reliability+indicators&hl=tr&ie=UTF-8&inlang=tr> (17.07.2003).

[24] K.A Bollen, "Indicator: Methodology [Göstergeler: Yöntembilim]," <www.unc.edu/~bollen/indicatorencyc.pdf> (13.07.2003).

[25] D.W. Kirsch, "Indices, Scales and Typologies [İndeksler Ölçekler ve Tipolojiler], <

[26] G.D. Israel, "Analyzing Survey Data [Alan Araştırması Verilerinin Analizi]," <http://www.google.com.tr/search?q=cache:4DLL92Rp_g4J:edis.ifas.ufl.edu/BODY_PD007+index+babbie+scale&hl=tr&ie=UTF-8&inlang=tr> (15.07.2003).

[27] D.W. Kirsch," Indices, Scales and Typologies [İndeksler, Ölçekler ve Tipolojiler]," <http://comp.uark.edu/~yangwang/soci5013/kirsch.pdf> (22.05.2004).

[28] N.J Ashill ve D. Jobber, "An Empirical Investigation of the Factors Affecting the Scope of Information Needed in a MkIS [MkIS'de İhtiyaç Duyulan Bilgi Alanını Etkileyen Faktörlerin Araştırılması]," 2002, <http://www.bradford.ac.uk/acad/management/external/pdf/workingpapers/Booklet_02=14.pdf> (23.04.2004).

[29] W.W. Chin, "Partial Least Squares For Researchers: An overview and presentation of recent advances using the PLS approach [Araştırmacılar İçin Kısmi En Küçük Kareler Yöntemi]," <http://disc-nt.cba.uh.edu/chin/icis2000plstalk.pdf> (25.04.2004).

[30] Cheryl Burke Jarvis, Scott B. Mackenzie ve Philip M. Podsakoff, "A Critical Review of Construct Indicators and Measurement [Yapısal Göstergeler ve Ölçümlerinin Eleştirel Bir değerlemesi]," <http://www.bauer.uh.edu/mark/papers/Jarvis_JCR_2003.pdf> (23.04.2004).

[31] J. Munshi, "A Method for Constructing Likert Scales [Likert Ölçeklerinin Oluşturulmasında Yöntem]," <http://munshi.sonoma.edu/working/LIKERT.HTML> (30.11.2002).

[32] Aynı.

[33] W.L. Farmer ve Diğerleri, "Latent Trait Theory Analysis of Changes in Item Response Anchors [Gizli Özellik Kuramının Analizinde Madde Yanıt Çıpalarındaki Değişiklikler]," <http://www.cami.jccbi.gov/AAM-400A/Abstracts/2001/FULL%20TEXT/0104.pdf.> (30.11.2002).

[34] Aynı.

[35] R. Garland, "The Mid-Point on a Rating Scale: Is it Desirable? [Dereceleme Ölçeklerinde Orta Nokta: Arzu Edilir mi?]," <http://marketing-bulletin.massey.ac.nz/article2/research3b.asp> (01.12.2002).

[36] Usability Professionals, "Usability Testing Methods: Subjective Measures Part Ii - Measuring Attitudes And Opinions [Test Yöntemlerinin Kullanılabilirliği]," <http://www.upassoc.org/html/1999_archive/usability_testing_methods.html> (01.12.2002).

[37] Garland, "The Mid-Point."

[38] RMS, "Dont Know [Bilmiyorum]," <http://oassis.gcal.ac.uk/teaching/rms/misc/dknow.html> (01.12.2002).

[39] Usability Professionals, "Usability Testing."

[40] Aynı.

[41] V.B. Lober, "Fundamental Measurement for Psychology [Psikoloji İçin TTemel Ölçümler]," <http://www.rasch.org/memo64.htm> (18.10.2002).

[42] "Attitude Scales [Tutum Ölçekleri]," <http://www.chssc.salford.ac.uk/healthSci/distres/resmeth/attitude.htm> (30.11.2002).

[43] W.B. Lober, "Fundamental Measurement For Psychology [Psikoloji İçin Temel Ölçümler]," <http://209.41.24.153/memo64.htm> (30.11.2002).

[44] "Attitude Scales."

[45] A. Diamantopoulos ve H. Winklhofer, "Index Construction with Formative Indicators: An Alternative to Scale Development [Oluşturucu Göstergelerle İndeks Oluşturma: Ölçek Oluşturmaya Alternatif]," <www.lboro.ac.uk/departments/bs/research/1999-41.doc> (11.07.2003).

[46] J.A. Siguaw, "Formative vs. Reflective Indicators in Measure Development: Does the Choice of Indicators Matter [Ölçek Geliştirmede Oluşturucu Ölçeklere Karşı Yansıtıcı Göstergeler: Herhangi Birisinin Seçilmesi Durumu Değiştirir mi?]," Ithaca, New York, 2003.

[47] Diamantopoulos ve Winklhofer, "Index Construction."

[48] Diamantopoulos ve Winklhofer, "Index Construction."

[49] Aynı.

[50] Aynı.

[51] Academy of Marketing Science Review, "Submission Guidelines [Yazım Rehberi]," <http://www.google.com.tr/search?q=cache:mgWP0y-qDFMJ:www.amsreview.org/submit.htm+formative+scales+reflective&hl=tr&ie=UTF-8&inlang=tr> (09.07.2003).

[52] Siguaw, "Formative vs. Reflective."

[53] Diamantopoulos ve Winklhofer, "Index Construction."

[54] W.W. Chin, "Issues and Opinion on Structural Equation [Yapısal Eşitlik Modellerinde Sorunlar ve Görüşler]," Modeling<http://www.misq.org/archivist/vol/no22/issue1/vol22n1comntry.html> (07.07.2003).

[55] P. Chwelos ve I. Benbasat, "Empirical Test of an EDI Adoption Model [EDI Uyum Modelinin Ampirik Olarak Testi]," <http://ebusiness.commerce.ubc.ca/internal/UBCBEBR2000-003.pdf.> (11.07.2003).

[56] Aynı.

[57] N.J. Ashill ve D. Jobber, "An Empirical Investigation of the Factors Affecting the Scope of Information Needed in a MkIS [MkIS İçin İhtiyaç Duyulan Bilgileri Etkileyecek Faktörlerin Ampirik Olarak Araştırılması]," <http://www.google.com.tr/search?q=cache:78O-3bFRyWcJ:www.bradford.ac.uk/acad/management/external/pdf/workingpapers/Booklet_02%3D14.pdf+pls+formative+items&hl=tr&ie=UTF-8&inlang=tr> (14.07.2003).

[58] W. Ulaga ve A. Eggert, "Relationship Value in Business Markets: Development of a Measurement Scale [İş Piyasalarında İlişki Değeri: Ölçüm Skalası Geliştirme]," <http://www.google.com.tr/search?q=cache:8xsTblwZwDkJ:www.mcse.external.xerox.com/isbm/dscgi/ds.py/Get/File-228/2-2003.pdf+formative+scales+reflective&hl=tr&ie=UTF-8&inlang=tr> (09.07.2003).

[59] N. Bontis, "Intellectual Capital and Business Performance in Malaysian Industries [Malezya Endüstrisinde Entelektüel Sermaye ve İşletmelerin Başarısı]," <http://www.business.mcmaster.ca/mktg/nbontis/ic/publications/JIC1-1Bontis.pdf> (11.07.2003).

[60] Bontis, "Intellectual Capital."

[61] Bu konuda daha fazla bilgi için *bk.*, <http://www2.chass.ncsu.edu/Judge/PS371/Lecture6b_part2.htm> (18.10.2002).

[62] H. Bernal, S. Wooley ve J.J. Schensul, "The Challenge of Using Likert Type Scales [Likert Tipi Ölçeklerin Kullanılmasındaki Güçlükler]," <http://www.hsl.wisc.edu/ereserves_content/nursing/fall2002/N991/N991_03F02.pdf> (30.11.2002).

[63] Peterson, 383.

[64] B Stennet, "Qpinion Survey Ratig Scales [Kanaat Araştırmalarında Ölçek Dereceleri]," <http://www.assessmentplus.com/articles/opinion_survey_rating_scales.pdf> (07.09.2002).

[65] K.K. Hoon, "An Analysis of Optimum Number of Response [Optimum Yanıt Sayısı Analizi]," <http://www.kams.org/journal/m1-05.htm> (01.12.2002).

[66] Aynı.

GÜVENİLİRLİK ANALİZİ YÖNTEMLERİ, GÜVENİLİRLİK İNDEKSİ VE GÜVENİLİRLİK KATSAYILARI

Klasik test kuramı kapsamındaki güvenilirlik analizi yöntemleri tarihsel süreç içinde sürekli olarak gelişme göstermiştir. Öte yandan her geçen gün bilim adamları yeni analiz yöntemleri önermeye devam etmektedirler. Bu bölümde ele alınan teknikler literatürde belirli bir yaygınlığa kavuşmuş olanlardır. Bölümde klasik test kuramına göre önce güvenilirlik analizi yöntemleri ele alınmış, daha sonra güvenilirlik indeksi ve güvenilirlik katsayılarının anlamları ve türleri üzerinde durulmuştur.

GÜVENİLİRLİK ANALİZİ YÖNTEMLERİ

Güvenirlilik analizleri iki grupta değerlendirilir: Form sayısı ve seans sayısı açısından. Araştırmacı güvenilirlik analizleri için tek bir form veya birden fazla form kullanabilir. Ayrıca güvenilirlik analizini tek bir seansta veya birden fazla seansta gerçekleştirmek isteyebilir. Uygulama biçimine göre yapılacak güvenilirlik analizleri de değişir. Psikometriciler kolay hatırlanması nedeniyle güvenilirlik analizlerini basit bir şekilde dört grupta ele almışlardır:

1. İç tutarlılık güvenilirliği.
2. Test-yeniden test güvenilirliği.
3. Paralel formlar güvenilirliği.
4. Gözlemciler arası güvenilirlik.

Bu dört maddeyi, seans ve form sayısı faktörünü dikkate alarak tablolaştırdığımızda, çok daha geniş ve kapsamlı bir liste elde ederiz (*bk.*, Tablo 3-1). Araştırmacı, söz konusu tablodaki sınıflandırmayı temel alarak çalışmasının niteliğine uygun bir güvenilirlik analizi planı geliştirmelidir.

Tablo 3-1. Klasik Test Kuramında Güvenilirlik Analizi Yöntemleri

Seans sayısı	Form sayısı: Tek	Form sayısı: İki
Tek	İç tutarlılık güvenilirliği Cronbach alfa Yarıya bölme Spearman-Brown Guttman formülü Rulon formülü Korelasyon analizi Lambda 4 Theta Omega Kuder – Richardson 20 Kuder – Richardson 21 Faktör analizi Hoyt varyans analizi Gözlemciler arası güvenilirlik Uyuşma indeksi Cohen Kappa Korelasyon (r, rho, tau) Phi Tekrarlamalı TYVA Regresyon analizi Genellenebilirlik Güvenilirliği Genellenebilirlik katsayısı Dayanıklılık indeksi	Paralel formlar Spearman-Brown Korelasyon analizi Guttman formülü Varyans analizi
İki	Test-yeniden test Pearson korelasyon analizi Küme içi korelasyon (n<15) Kappa (kategorik verilerde) Varyans analizi Tekrarlanabilirlik katsayısı t-Testi Gözlemci içi test[a] Uyuşma indeksi Korelasyon (r, rho, tau)	Paralel formlar Korelasyon analizi Varyans analizi

[a] Gözlemci içi test. Tek bir gözlemcinin farklı iki zamanda gözlem veya değerlendirme yapması.

İç Tutarlılık Analizleri

İç tutarlılık güvenilirliğinde, tek bir ölçüm aracı kullanılarak ve tek bir seansta ölçüm yapılarak maddelerin belirli bir kavramsal yapıyı tutarlı bir şekilde ölçüp ölçmediği araştırılır. Hogan, Benjamin ve Brezinski (2000) tarafından yapılan bir araştırmada ABD'de APA tarafından yayımlanan *Directory of Unpublished Experimental Mental Measures* [Yayımlanmamış Ampirik Zihinsel Testler Yıllığı] isimli kütükte yer alan ölçüm araçlarının %75'inde sadece iç tutarlılık güvenilirliğinin verildiği bildirilmiştir (aktaran Henson, 2001).[1] Bir test veya ölçek için iç tutarlılık güvenilirliğini yapmak gereklidir, fakat yeterli değildir. Ölçüm çalışmasının niteliğine göre aynı zamanda diğer güvenilirlik analizleri de yapılmalıdır. Örneğin, güvenilirlik analizi yapılması düşünülen ölçüm aracı iş hayatında kullanılacak bir bilişsel test ise aynı zamanda test-yeniden test güvenilirliği; ölçüm aracı bir ölçek ise paralel formlar güvenilirliği de hesaplatılmalıdır. İç tutarlılık güvenilirliğini belirlemek için değişik yöntemlerden yararlanılabilir. Araştırmacı yaptığı bilimsel çalışmanın tez, bilimsel makale veya endüstride kullanılmak üzere yeni geliştirilen bir test olmasına göre değişik düzeydeki analiz ve yöntemlerden yararlanır.

Güvenilir test ve ölçekler, maddeleri arasındaki iç tutarlılığı yüksek olan araçlardır. Ancak, Cattell (1966, 1978) belirli bir dereceden yüksek olan iç tutarlılığın ölçeğin veya testin geçerliliğini düşürebileceği tezini ileri sürmüştür. Bu olgu, "gaz yapmış şişkin mideye" benzetilebilir. Cattell'e göre bir testte / ölçekte çok sayıda madde varsa ve bu maddeler büyük ölçüde birbirine benziyorsa ölçek şişkin gibi gözükür. Fakat aslında bu maddelerin hepsi ya sadece çok spesifik bir alanı ölçüyordur veya bu maddelerden pek azı ölçülen alanla yüksek derecede ilişkilidir. İngilizcede *bloated specific* olarak adlandırılan bu olguyu Türkçede "şişkin özgünlük" olarak kavramlaştırabiliriz.

Bir ölçeğin çok sayıda madde ile "dolu" gibi gözükmesi belirlenen maddelerin alanı çok iyi kapsadığı anlamına gelmez. Alanın iyi kapsanma-

dığı durumda ise testin / ölçeğin geçerliliği zayıflar.[a] Bu tür ölçeklerde maddelerin hepsi birbirine benzer, maddelerde aynı olgu değişik bir şekilde yeniden ifade edilmiş gibidir. "Türk filmlerini severim." ile "Benim favorim Türk filmleridir." ifadeleri arasında önemli bir farklılık yoktur. Son yıllarda geliştirilen testlerde ve ölçeklerde ölçülmeye çalışılan kavramsal yapılar şişkin özgünlük olgusunu anımsatacak şekilde çok sayıda madde ile çok dar sınırlarda ele alınmaya başlanmıştır. Araştırmacı kuramsal bir çerçeveye dayanmadan belirli bir takım tahminlerde bulunmak için bir test / ölçek oluşturmuşsa şişkin özgünlük bir sakınca oluşturmaz. Ancak kişilik gibi oldukça geniş kavramsal içeriğe sahip yapılarda *şişkin özgünlük* özelliği, ölçüm aracını olumsuz yönde etkiler.[2] Bu tür ölçeklerde çok sayıda madde olsa bile bu maddeler kavramsal alanı yeterli genişlikte kapsamadığı için güçsüzdür. Bir araştırmacı kişiliğin bütün alanlarını kapsayacak şekilde değişik nitelikte çok sayıda maddeden oluşan bir test geliştirdiği zaman, maddeler arasındaki korelasyon ve maddelerin toplam puanla olan korelasyon değerleri düşük çıkar (,10 ilâ ,20 arasında değişebilir). Bu nedenle özellikle kişilik testlerinde ve genel amaçlı tutum ölçeklerinde her bir faktörle ilgili olarak yeteri kadar (10–20 civarında) madde belirlemek gerekir.[3] Öte yandan kişilik ve genel tutum ölçeklerinin dışında, özgün bir konuyu ölçmeyi amaçlayan test ve ölçeklerde ise maddelerin yeniden ifade edilmiş gibi gözükmesinde sakınca yoktur. Bu tür ölçeklerde maddelerin geniş değil, dar kapsamlı olması önemlidir. Ray'a (2002) göre, maddelerin alanın "genişliğini" kapsama derecesi tek başına esas amaç olmadığı gibi buna önceden de karar verilemez. Maddelerin alanı kapsama genişliği deneysel olarak yapılacak analizler sonucunda belli olur.[4]

Araştırmacı iç tutarlılık analizleri için değişik hesaplama ve istatistiksel analiz yöntemlerinden yararlanabilir. Aşağıdaki bölümde bu istatistikî teknikler üzerinde durulmuştur.

---

[a] Maddeler ile madde gruplarının analizi psikometrisyenlerle psikologlar arasında uzun yıllar tartışma konusu olmuştur. Psikologlar *madde analizini* ön plana çıkarırlarken psikometrisyenler dört veya beş maddeden oluşan *madde gruplarının* analizine önem vermek gerektiğini belirtmişlerdir. Zeka testlerinde, kişilik testlerinde toplam puan ile maddeler arasındaki korelasyon çoğunlukla ,10 veya ,20 gibi düşük rakamlar halinde seyreder. Toplam puanla yüksek korelasyona sahip madde sayısı bu tür ölçeklerde iki üç maddeyi geçmez. Bu nedenle Cattell dört veya beş maddeden oluşan madde gruplarını temel analiz birimi olarak görmüştür. Bu konuda daha fazla bilgi için *bk.* C. Brand, "Factor Analysis [Faktör Analizi]," <http://www. cycad.com/cgi-bin/Brand/quotes/qfa.html> (18.10.2002).

Maddeler arası korelasyon katsayılarının ortalaması. Bu analizde ölçeğin/testin toplam puanları hesaplamaya katılmaz. Sadece maddeler arasında korelasyon analizi yapılır ve değişkenlerin korelasyon katsayılarının ortalaması alınır. Bu analiz, test/ölçek maddelerinin ne ölçüde birbirleriyle ilişkili olduğu hakkında bilgi verir. İstatistikî analiz programı SPSS'te korelasyon analizi yapmak için Correlate, Bivariate düğmeleri seçili hale getirilir. Normal dağılım özelliği göstermeyen sıralı ve eşit aralıklı ölçeklerde Spearman korelasyon analizi uygulanır. Maddeler arasındaki korelasyonu görmenin ikinci bir yolu Reliability mönüsü altında korelasyon matrisi düğmesini seçili hale getirmektir (*bk.*, Tablo 3-2).

Maddeler arası korelasyon analizinde iki değişken arasındaki ilişki negatif değerli olarak gözüküyorsa maddelerin aralarında ters bir ilişki var demektir. Bu maddelerden biri veya duruma göre her ikisi de ölçekten çıkarılabilir. Bunun için her bir maddenin diğer maddelerle olan ilişkisine bakılarak karar verilir. Bu konuda bir diğer yöntem maddelerin toplam puanlarla olan korelasyonuna bakmaktır. Maddeler arası korelasyon katsayılarının ortalaması alınırken negatif işaretli maddelerin korelasyon katsayıları ortalamaya katılmaz. Aritmetik ortalama, bu maddeler ölçekten çıkarıldıktan sonra hesaplanır.

Tablo 3-2. Maddeler Arası Korelasyon Analizi

	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*
D1								
D2	,91							
D3	,77	,75						
D4	,49	,87	,44					
D5	,71	,62	,92	,61				
D6	,60	,45	,68	,81	,75			
D7	,44	,50	,31	,77	,70	,77		
D8	,67	,42	,72	,79	,64	,72	,72	

Tablo 3-2'de görülen korelasyon rakamları toplanıp aritmetik ortalaması alındığında ,66 değerinin elde edildiği görülür. Bu rakam tek başına güvenilirlik katsayısı olarak değerlendirilebilir. Ancak uygulamada daha çok alfa güvenilirlik değeri kullanıldığından, eğer alfa değeri verilmişse bu rakamı ayrıca vermeye gerek yoktur. Çünkü alfa değeri ile karşılaştırıl-

dığında maddeler arası korelasyon değerlerinin ortalaması daha düşük çıkar. Alfa değeri hesaplatılmış olsaydı bu değer belki de ,76 çıkacaktı. Alfa değeri (düşük değerlikli güvenilirlik katsayısı olarak adlandırılsa da) muhtemel bütün yarılar arasındaki ilişkileri dikkate aldığından korelasyon ortalamalarından daha güçlüdür.

Madde – toplam puan korelasyonu katsayılarının ortalaması. Tutum ölçeklerinde (indekslerde) beş veya yedi dereceli ölçeğin; bilgi ve başarı testlerinde ise çift rakamlı değerlerden de oluşabilen ölçeğin/testin toplam puanlarıyla her bir maddeye ait puanların korelasyonunun alınmasıdır. Bu işlem *madde güvenilirliği* olarak adlandırılır. Madde-toplam puan korelasyon katsayılarının ortalaması *testin güvenilirliğini* verir (*bk.*, Tablo 3-3). Uygulamada, kimi yazarlar bu işlemi okurlarına *madde analizi* kavramıyla tanıtmışlardır. Ancak madde analizi daha geniş bir kavramdır. Madde analizi, testin tamamının ve testteki her bir maddenin kalitesini tanımlayan bir dizi hesaplama yöntemi ve hesaplama süreçlerine verilen bir isimdir. Madde analizi üç kuramsal çerçevede yapılır: klasik test kuramına göre, madde-yanıt kuramına göre ve Rasch ölçüm yöntemine göre. Madde-yanıt kuramı ve Rasch analizinin her ikisinde de maddelerin arka planındaki gizli özellikler araştırılır.

Tablo 3-3. Madde – Toplam Puan Korelasyon Analizi

	D1	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*
D1								
D2	,91							
D3	,77	,75						
D4	,49	,87	,44					
D5	,71	,62	,92	,61				
D6	,60	,45	,68	,81	,75			
D7	,44	,50	,31	,77	,70	,77		
D8	,67	,42	,72	,79	,64	,72	,72	
Top.	,75	,67	,62	,70	,71	,74	,70	,73
r =	,70							

Klasik test kuramında madde analizi yapılmasının amacı; maddelerin aritmetik ortalama ve standart sapmalarını tespit etmek, bilgi ve yetenek testleri için madde güçlük analizini[a] yapmak, başarılı olanlarla başarısızları ortaya koyan farklılaştırma (veya ayırt etme) analizini[b], farklılaştırma katsayısını[c] belirlemek, çeldirici analizini[d] yapmak ve madde-toplam puan korelasyonu ile güvenilirliği hesaplamaktır.

---

[a] Madde güçlük analizi (Item difficulty analysis): Klasik test kuramında bir testteki doğru yanıt sayısının toplam madde sayısına bölünmesidir. Madde güçlük analizi, *p* simgesiyle gösterilir. Maddelerin güçlüğü soruların ve onları yanıtlayan kişilerin özelliklerinden etkilenir. İyi bir bilgi testinde maddelerin büyük çoğunluğunun *p* oranlarının ,30 ilâ ,80 arasında olması gerektiği belirtilmiştir (*bk.*, J. Kehoe, "Basic Item Analysis [Temel Madde Analizi]," <http://ericae. net/pare/getvn.asp?v=4&n=10.> (16.10. 2002). Ancak bu ölçüler genel bir kuralı yansıtmaz. Maddelerin güçlük oranları testin uygulanış amacına, test uygulanan kitlenin yetenek düzeyine ve başarılarına göre değişir. Bazı bilgi sınavlarında *p* oranının ,50 ilâ ,70 arasında olması gerektiği belirtilmiştir. Testin *p* oranı arttıkça farklılaştırma indeksi düşeceğinden güçlük katsayıları *D* indeks değerine göre belirlenmelidir. Ayrıca maddelerin *p* oranları cevap şıklarının ikili, üçlü, dörtlü ve beşli olmasına göre de değişir. Doğru ve yanlış şeklinde düzenlenen iki şıklı cevaplarda ideal *p* oranı ,85 olarak belirlenirken; beş şıklı test sorularında bu oranın ,60 veya ,70 olması gerektiği belirtilmiştir (*bk.*, *Lord, F.M.* "The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties," *Psychometrika*, 1952, 18, 181-194. <http://www. was hington. edu/oea/item.htm> (16.10. 2002). Madde güçlük oranını belirlemede bir diğer yaklaşım minimum ve maksimum değerlerin ortalamasını almaktır (*ör.*, ,30 + ,80/2=,55). Modern ölçüm kuramlarından madde-yanıt kuramında ise madde güçlük analizi, kişinin yetkinliği ve maddenin güçlük oranının kesiştiği noktada belirlenir. Testi alan kişinin bir maddeyi doğru olarak yanıtlama olasılığının en az %50 olması gerekir.

[b] Ayırma analizi (Discrimination Index): Kendilerine test uygulanan bir grupta testin gerçekten başarılı olan kişilerle başarısız olanları ayırt etme gücüdür ve *ayırma indeksi* kavramıyla tanımlanır. Ayırma indeksinin simgesi *D* harfidir. Normal dağılım özelliği gösteren büyük bir grupta kişilerin toplam puanları yüksekten küçüğe doğru sıraya dizilir ve Truman Kelley kuralı gereğince ilk %27'lik dilim başarılılar ve son %27'lik dilim ise başarısızlar grubu olarak alınır. Farklılaştırma indeksinde başarılılar grubundaki doğru yanıt sayısı başarısızlar grubundaki doğru yanıt sayısından çıkarılır ve iki gruptan içinde daha fazla üye bunan grubun sayısına bölünür. Analiz sonucunda *D*'nin büyük çıkması, bir maddenin başarılı grup lehine kişileri iyi ayırdettiğini gösterir. Ayırma indeks değerinin ,30 ilâ ,50 (bazı yazarlara göre ise ,20 ilâ ,40) arasında olması önerilmiştir (Oosterhof, 2001). *Bk.*, <http:// www.flaguide.org/cat/multiple choicetest/multi ple_ choice_test4.html> (23.04.2001).

[c] Ayırma katsayısı (Discrimination coefficients): Nokta – iki serili korelasyon katsayısı ile iki serili korelasyon katsayıları hesaplama yöntemlerine dayanır. Hesaplama kolaylığı olması nedeniyle pek çok kişi ayırma indeksi yerine ayırma katsayıları ile çalışmayı yeğler.

[d] Çeldirici analizi (Distractors analysis): Bir testteki yanlış şıkların işaretlenme biçimlerinin incelenmesidir. Her bir yanlış şıkkın yararlı olup olmadığını belirlemek için farklılaştırma indeksi veya farklılaştırma katsayısı kullanılır. Eğer bir şık testi alan kişilerin hiç birisi tarafından tercih edilmemişse veya bir çeldirici şıkkını testi alan kişilerin %5'inden daha azı tercih etmişse bu şıkta değişiklik yapılır veya bu şık bütünüyle testten çıkarılır.

Gizli özellik kuramlarında ise testin toplam veya ortalama puanından daha önemli olan arka planda bulunan gizli faktörlerdir. Güvenilirlik, testi alan adayın durumu ile testin zorluk veya kolaylığının ortak etkileşimine bağlıdır. Madde-yanıt kuramında madde analizi üç parametreden etkilenir: güçlük, farklılaştırma ve şans faktörü. Güvenilirlikte verilerin bu üç parametreye ne ölçüde uygun düştüğüne bakılır. Bir parametreli Rasch analizinde sadece tek bir parametre, "güçlük/yetkinlik" veya "konum" değişkeni dikkate alınır. Rasch analizinde testin farklılaştırma özelliğinin yeknesak olduğu varsayılır ve şans faktörünün de var olduğu ilkesinden hareket edilir.

Görüldüğü gibi güvenilirlik analizi, madde analizine ilişkin değişik hesaplama işlemlerinden sadece birisidir. Madde-toplam puan korelasyonunu bu nedenle "madde analizi" olarak isimlendirmek doğru değildir. Ancak testin bütün olarak güvenilirliği, madde analiz sürecinde ele alınan diğer faktörlerden de etkilenir.

Madde-toplam puan korelasyonunda toplam puanlar *eşit aralıklı* veri niteliğinde, madde puanları ise *sıralı veri* olarak değerlendirildiğinden Spearman sıra korelasyonu analizi uygulanır.

Spector'e göre (1992), madde-toplam puan korelasyonu analizinin yapılabilmesi için 100 ilâ 200 arasında cevaplayıcının olması gerekir (aktaran Gaddy).[5] Başka kaynaklarda ise madde sayısının en az beş katı kadar cevaplayıcının bulunması öngörülmüştür. Bu örneklem büyüklükleri sadece esas araştırma için değil, aynı zamanda pilot araştırma için de düşünülmelidir. Madde-toplam puan korelasyon katsayısı eğer ,30'un altındaysa (Örneklem büyüklüğüne bağlıdır. Dört yüz veya daha fazla katılımcının bulunduğu büyük örneklemlerde ,20 gibi daha düşük korelasyon katsayıları da kabul edilebilir.) bu maddelerde ciddi bir sorun var demektir.[6] Bu maddeler ölçekten düşürülebilir. Ancak maddeyi düşürmek ilk alternatif olarak düşünülmemelidir. Bu tür maddeler düşürüldüğünde ölçekteki çeşitliliğin azalması ve dolayısıyla maddelerin çok dar bir alana sıkışması ihtimali vardır. Daha önce belirtildiği gibi bu durum, tipik bir *şişkin özgünlük* halidir. Bu nedenle çıkarmak yerine öncelikle maddenin içeriğinde değişiklik yapılıp yapılamayacağı araştırılmalıdır. Ayrıca, düşürülen bir maddenin alfa değeri üzerindeki etkisi incelenmelidir. Eğer maddenin alfa değerini yükseltme etkisi çok az ise, ölçekten çıkarma yerine maddede değişiklik yapma yoluna başvurmak daha doğru olur.

Geliştirilen ölçekte eğer alt boyutlar/faktörler varsa veya ölçek, *alt ölçeklerden* meydana gelen bir batarya şeklinde ise madde analizi alt ölçeğin

toplam puanlarıyla bu ölçeğe ait maddeler arasında yapılır. Bir ölçek tek boyutlu (faktörlü) veya çok boyutlu (faktörlü) olabilir. Çok faktörlü ölçeklerde genel toplam puanla maddeler arasındaki korelasyona bakıldığında bir çok maddenin korelasyon katsayısı düşük görünür. Sadece tek boyutlu ölçeklerde madde–toplam puan korelasyonu yüksektir. Öte yandan madde–toplam puan korelasyonu negatif çıkmışsa bu maddelerin ayırt etme özellikleri düşük demektir. Bu maddeler zayıf bir şekilde ifade edilmiş, yanlış kodlanmış, tersine çevrilmemiş veya kasıtlı bir şekilde cevaplandırılmış olabilir. Negatif işaretli maddelerin ölçekten çıkarılması gerekir.

Madde-toplam puan korelasyonunda, toplam puana kendisiyle ilişki kurulan madde de dahil edildiğinden korelasyon katsayısı bir ölçüde şişkin çıkar. Bunu önlemek için kendisiyle ilişki kurulan maddenin puanı toplam puandan çıkarıldıktan sonra korelasyon analizi yapılır. Bu şekilde elde edilen değer *düzeltilmiş madde–toplam puan korelasyon katsayısı* olarak isimlendirilir.[7]

Madde-toplam puan korelasyonunun değişik bir şekli madde-medyan değeri korelasyonudur. Literatürde daha az görülen bu uygulamada maddelerin medyan değeriyle her bir madde arasında korelasyon analizi yapılır. Hesaplama sonucunda elde edilen ,30'luk korelasyon katsayısı güvenilirlik açısından "iyi" bir değer olarak yorumlanır.

İki şıklı değerlere ait korelasyon analizi. İstatistikte *biserial korelasyon analizi* olarak isimlendirilen bu yöntemde, sürekli veri niteliğindeki toplam puan değerleri ile maddeleri 1-2 veya 0-1 şeklinde kodlanan puan değerleri karşılaştırılır. Katılımcılar sorulara *Doğru* veya *Yanlış* şeklinde cevap vermişlerse testin güvenilirliği *biserial korelasyon* analizi ile yapılır. İstatistiksel analiz programı SPSS'te *biserial korelasyon* analizi Bivariate düğmesi altında Pearson şıkkı seçilerek hesaplanır. Bir maddenin toplam puanla negatif yönde ilişkili olup olmadığını görmek için ise *point-biserial korelasyon* analizine başvurulur. Bu analiz, Reliability mönüsünde Interclass correlation coefficient düğmesi seçili hale getirilerek yapılır.

Cronbach alfa değeri. Cronbach alfa; Likert türü toplamalı ölçeklerde, anlamsal farklılık ölçeklerinde, Stapel ölçeklerinde toplam veya ortalama puana dayanan diğer psikometrik testlerde ve bileşik maddelerden oluşan indeks türü ölçüm araçlarında maddelerin birbirleriyle *tutarlı olup olmadığı-*

nı ve maddelerin *hipotetik bir değişkeni*[a] *ölçüp ölçmediğini* belirler. Çok sayıda maddeden oluşan ölçeklerde iç tutarlılığı ölçen alfa katsayısı ilk olarak 1937 yılında Kuder-Richardson tarafından ikili veri yapılarının güvenilirliğini belirlemek için geliştirilmiştir. Daha sonra 1945 yılında Louis Guttman tarafından esaslı bir şekilde değiştirilmiş ve Guttman yazdığı makalesinde alfa katsayısını "L3" olarak kodlamıştır. Guttman'dan altı yıl sonra 1951'de bu hesaplama yöntemi Cronbach tarafından yeniden ele alınarak elde edilen katsayı *alfa* (α) olarak isimlendirilmiştir.[b] Bu nedenle McDonald (1999), yöntemin daha doğru bir şekilde G-C alfa yöntemi olarak isimlendirilmesini önermiştir.[8] Fakat, terimin günümüzdeki yaygın kullanım biçimi *Cronbach alfa* şeklindedir.

İç tutarlılık – tek boyutluluk ilişkisi. Esas olarak bir "güvenilirlik indeks değeri" olan Cronbach alfa, test veya ölçeğin içerdiği maddelerin birbiriyle ne ölçüde tutarlı olduğu ve arka planda bulunan gizli, hipotetik değişkeni ne derece temsil ettiği hakkında bilgi verir. Literatürde alfa değerinin gizli değişkene ilişkin olarak tek bir boyutu mu yoksa birden fazla boyutu mu temsil ettiği konusunda çeşitli tartışmalar vardır. Bazılarına göre, "Rulon yöntemine göre hesaplanan ve muhtemel bütün yarılar arasındaki korelasyon katsayılarının ortalamasını gösteren"[9] alfa değeri, faktör analizi sonucunda hesaplanan birinci faktörün doymuşluk değeridir. Bu yorum, tek bir faktöre işaret eder. Bununla birlikte bilim adamlarının büyük çoğunluğuna göre alfa her zaman tek boyutluluğu göstermez. Alfa değerinin esas işlevi iç tutarlılığı saptamasıdır. Alfa, maddelerin belirsiz olmadığına işaret eder.

Alfa iç tutarlılık hesaplaması yapılırken ölçek/test maddelerinin aritmetik ortalama değerlerinin ve varyanslarının eşit olup olmadığına bakılmaz, sadece eşit olduğu varsayılır. Alfa değerinde, iç tutarlılıkla tek boyutluluk olgularını ayrı ayrı değerlendirmek gerekir. Bir test, iç tutarlılık değeri yüksek olduğu halde birden fazla boyuta sahip olabilir. Türdeşlik ise, tek boyutluluk anlamına gelir. Bir testte birden fazla boyut/faktör varsa o ölçüm aracı türdeş değildir. İç tutarlılık, sadece ölçeğe ait maddeler arasındaki korelasyon ve kovaryans katsayılarının yüksek olduğuna işaret eder.

---

[a] Literatürde yazarların "değişken" ve "faktör" kavramlarını birbirinin yerine kullanabildikleri görülmektedir. Faktörler de bir değişkendir, ancak faktör anlamındaki değişkenle madde anlamındaki değişkenler birbirlerinden farklıdır. Burada okuyucuların bu tür tanımlamalara aşinalık kazanması için faktör yerine değişken sözcüğü tercih edilmiştir.

[b] Cronbach'ın 1951'de yazmış olduğu makale, *Social Science Citation Index* verilerine göre 2200'den fazla makalede referans olarak gösterilmiştir.

Cronbach alfa değeriyle ilgili varsayımlar. Cronbach alfa değeri klasik test teorisine dayanır ve bu nedenle bu değerle ilgili varsayımlar aynı zamanda klasik test teorisinin varsayımlarıdır. Klasik test teorisine göre güvenilirlilik, gerçek değerlerle gözlemlenen değerler arasındaki ilişkinin yüksek olmasıdır ve tesadüfî hata düşük kaldığı oranda bu ilişki yüksek çıkar. Alfa değeri hesaplanırken klasik test teorisine ait varsayımların göz önünde bulundurulmadığı durumda araştırmacı alfa katsayısını gereğinden yüksek veya gereğinden düşük değerlendirebilir. Alfa değerinin hesaplanmasında aşağıdaki varsayımlar temel alınır.

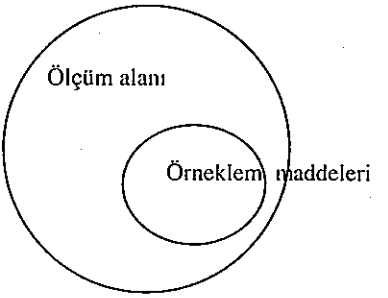
Alan örneklemesine dayanması. Alfa değeri, ölçülmek istenen özellikle ilgili maddelerin sınırları belirli bir kavramsal alandan örnekleme yöntemiyle seçildiğini varsayar (*bk.*, Şekil 3-1). Örneklem maddelerinin seçildiği alan, duruma göre birden fazla boyut/faktör içerebilir. Cronbach alfa, tek göstergeli/maddeli ölçekler için uygun değildir. Çünkü tek göstergeli ölçekler alanı temsil etme gibi bir iddiayla oluşturulmaz. Tek göstergeli ölçeklerde her bir madde farklı bir alana ait olabilir. Bunun yanında sınırlı sayıda maddeden oluşan ve sadece tek boyutluluğu kabul eden Bogardus[a], Guttman[b] ve Thurstone ölçeklerinde de alfa iç tutarlılık analizi yapılmaz. Bu ölçüm araçlarında katı bir şekilde sadece "tek boyutluluk" ve

---

[a] Amerikalı sosyolog Emory Bogardus (1882-1973) tarafından geliştirilen sosyal mesafe ölçeği, kişilerin değişik uluslardan başka bireylerle sosyal ilişkilere girme istekliliğini belirlemek için kullanılır. Bogardus 1932'de sosyal ilişkiler açısından insanlar arasındaki mesafeyi belirlemek için yedi soru belirlemiştir. Bu sorularda katılımcıların araştırılan ülke insanlarıyla ilgili olarak 1. evlenirim, 2. yakın arkadaşlığını kabul ederim, 3. kendisiyle aynı büroda çalışabilirim, 4. yakın komşum olmasını kabul ederim, 5. selam verdiğim uzak bir komşum olmasını kabul ederim, 6. aynı mahallede yaşayabilirim, 7. ancak aynı şehirde yaşayabilirim ifadelerini seçmeleri istenir. Sonunda her ülke insanlarıyla ilgili olarak 1 ilâ 7 arasında bir puan elde edilir. Bir tür indeks olan Bogardus ölçeklerinde iç tutarlılık güvenilirliği aranmaz. Ancak test – yeniden test güvenilirliği ile cevaplayıcıların tutarlı yanıt verip vermedikleri araştırılabilir. Bogardus ölçeklerinde ifadelerin sosyal mesafeyi mi yoksa coğrafi uzaklık anlamında geometrik uzaklığı mı gösterdiği konusunda araştırmalar yapılmış ve ölçek ifadeleri geometrik bir temele oturtulmaya çalışılmıştır. Bu konuda daha fazla bilgi için *bk.*, P.J. Ethington, "The Entellectual Construction of "Social Distance" Toward a Recovary of Georg Simmel's Social Geometry [Sosyal Mesafenin Entelektüel Olarak Oluşturulması ve Georg Simmel'in Sosal Geometri Kuramı]," <http://www.cybergeo. presse.fr/essoct/texte/socdis.htm> (17.10.2002).

[b] Guttman ölçeğinde ifadeler içerdiği güce göre sıralanmıştır. "Guttman yarıya bölme güvenilirliği" guttman ölçekleriyle ilgili değildir. Bu tür ölçeklerin güvenilirliği için maddeler arası tutarlılığı belirleyen alfa katsayısı değil, Guttman *yeniden üretilebilirlik katsayısı* kullanılır. Çünkü Guttman ölçeğinde maddeler arasında en önemsizinden önemliye doğru uzanan hiyerarşik bir sıralanma vardır.

"ölçeklenme" olgusu vurgulandığından alandaki dağılımı tam olarak temsil etme gibi bir kaygıyla hareket edilmez.



Şekil 3-1. Maddelerin alandan örnekleme yöntemiyle seçilmesi.

Tek bir boyutu ölçüyor olması. Alfa değeri tek bir özellik veya karakteristiği, tek bir faktörü ölçmeye yönelik olarak belirlendiği zaman daha güçlüdür. Birden fazla boyutlu ölçeklerde testin tamamı için hesaplanan alfa güvenilirlik katsayıları düşüktür veya yüksek çıkmış olsa bile olsa bu katsayılar düşük değerliliğe sahiptir. Birden fazla boyutlu ölçeklerde, her bir boyut bir alt test gibi düşünülüp alfa değeri ayrıca hesaplatılmalıdır. İstenirse testin tamamına ait güvenilirlik, ya alt testlerin güvenilirlik rakamlarının medyan değeri alınarak veya Lisrel istatistikî analiz programındaki "bileşik güvenilirlik" değeri temel alınarak belirlenebilir.

Maddelerin örnekleme bağımlı olması. Alfa değeri, ölçeğin kullanıldığı her bir farklı örneklem için yeniden hesaplanmalıdır. Bir araştırmacının daha önceden geçerlilik ve güvenilirlik analizi yapılmış ve yüksek çıkmış bir ölçüm aracını temin edip tekrar uygulaması, fakat alfa güvenilirlik analizi yapmaması doğru değildir. Çünkü alfa değeri gruplar arasında önemli ölçüde farklılıklar gösterir. Pilot araştırma sırasında hesaplanan alfa katsayısı ölçeği iyileştirmeye yöneliktir. Araştırma sonuçlarının güvenilirliği "esas araştırma" sırasında elde edilen alfa katsayılarının yüksekliğine bağlıdır.

Dereceleme ölçeği. Ölçekte/testte bütün maddeler için aynı yanıt biçimi veya dereceleme ölçeği kullanılmalıdır. Maddelerin bir kısmında ikili ve diğer kısmında üçlü veya beşli dereceleme ölçeği kullanılamaz.

Maddelerin paralel olması. Geliştirilen maddeler, içerik olarak araştırılan özellikle ilgili olguların değişik şekillerde yeniden ifade edilmiş biçimlerinden oluşur. Maddeler arasındaki ilişkiler teorik olarak; "paralel", "tau eşitliğine sahip", "yaklaşık tau eşitliğine sahip" veya "konjenerik" olarak değerlendirilir. Maddelerin paralel olması, tüm maddelerin "hata varyanslarının ve bunun yanında gerçek puan varyanslarının birbirine eşit olması" anlamına gelir. Ölçekteki her bir madde diğer maddeler kadar gizli özelliği iyi bir şekilde ölçme özelliğine sahiptir. Maddeler gizli özelliği ölçme açısından mükemmel olunca her bir maddedeki hata payları da birbirine eşittir ($e_1 = e_2 = e_3 = e_4 = e_5$). Paralel maddelere sahip ölçeklerde, her bir maddenin puanıyla gerçek puan arasında tam bir korelasyon vardır. Gerçek puanla tam bir korelasyon varsa o zaman maddeler arasındaki korelasyonlar da aynı çıkar. Bu olgu sadece ölçek maddelerinin varyansları birbirine eşit olduğu zaman gerçekleşebilir. Maddelerin *tau eşitliğine* sahip olması "maddelerin her birinin belirli oranda gerçek puanı içermesi, fakat hata varyanslarının farklı olması" anlamına gelir. Maddelerin *yaklaşık tau eşitliğine sahip olması* "maddelerin aynı güvenilirlik (korelasyon) katsayılarına sahip olması fakat aritmetik ortalama ve standart sapmalarının ise farklı olması" anlamına gelir. Konjenerik ilişkide ise maddeler arasında ortak genden kaynaklanan bir ilişki olduğu vurgulanır. Ancak her bir madde söz konusu ortak geni (özelliği) değişik ölçülerde içerir. Her bir madde gizli özellikten değişik derecelerde etkilenir. Maddelerin gerçek puan varyanslarının eşit olması gerekmez. Konjenerik ilişkilerde "eşit hata varyansı" ve "eşit gerçek puan ortalaması" gibi koşulların sağlanma durumu araştırılmaz. Bununla birlikte her bir madde eğer gerçek puanla daha güçlü bir şekilde ilişkili ise ölçek o denli güvenilirdir. Uygulamada, ölçeklerde paralellik ve gerçek tau eşitliği nadiren sağlanabilir (Cortina 1993, aktaran Yu).[10] Maddeler en azından yaklaşık tau eşitliğine sahipse maddeler arasındaki korelasyon katsayılarının ortalamaları alfa değerine eşit çıkar ($\rho_{\lambda}^2 = \alpha$). Ölçekteki maddeler konjenerik nitelikte ise bu tür ölçeklerin güvenilirliği LISREL ve benzeri programlarda bulunan *teyit edici faktör analizi* yöntemi kullanılarak belirlenir. "Heterojen maddelerden oluşan ölçeklerde" veya maddelerinin her biri farklı bir yapıyı ölçen "oluşturucu indekslerde" maddeler paralel olmadığından Cronbach alfa değerinin hesaplatılması anlamlı değildir. Bu tür ölçeklerde PLS-PC ve PLS-Graph yazılımlarıyla Kısmî En Küçük Kareler – KEKK (Partial Least Sequares –

PLS) istatistik analiz yöntemi, test-yeniden test veya paralel formlar yöntemi uygulanır.

Negatif ifadeler. Alfa değerinin hesaplanabilmesi için negatif içerikli maddeler tersine çevrilerek puanlanır. Negatif içerikli olmamakla birlikte bir madde negatif korelasyon katsayısına sahipse bu maddenin anlamı gözden geçirilir. Cümlenin olumsuz yüklemle bitmiş olmasına değil, anlamının negatif olup olmadığına bakılır.

Varyansların eşitliği. Ölçekteki maddelerin gerçek değerlerinin ve varyanslarının (her ne kadar böyle olmasa da) eşit olduğu varsayımından hareket edilir.

Verilerin dağılım biçimi. Alfa değeri sağa veya sola çarpık verilerden büyük ölçüde etkilenir. Bu tür verilerde alfa değeri düşük çıkar. Bu nedenle verilerin normal veya normale yakın bir dağılım eğrisine sahip olması gerekir. Cronbach'ın kendisi, alfa değerini güçlendirmek için gözlem değerlerinden ayrık, uç değerlerin çıkarılmasını ve hesaplamanın buna göre yeniden yapılmasını önermiştir. Böyle bir durumda güvenilirlik verilerin tamamı için değil, bir bölümü için araştırılan yapıyı temsil eden bir değer olarak ortaya çıkar.

Toplam puanların varyansı. Alfa, ölçeğin toplam puan varyansından etkilenir. Toplam puanların varyansı yüksekse alfa değeri de yüksek çıkar. Homojen örneklemlerde herkes benzer görüşlere sahip olacaklarından toplam puanlarının varyansı düşük olur ve sonuçta güvenilirlik rakamı da düşük çıkar.

Ölçüm hatası. Alfa değeri, maddelere ait ölçüm hatası ortalamalarının sıfır olduğu varsayımına dayanır. Diğer bir deyişle maddelerin ölçüm hataları tesadüfîdir ve bu hatalar birbiriyle ilişkili değildir. Bu varsayımın karşılanmadığı durumda (uygulamada bunu sağlamak tam olarak mümkün olmasa da) alfa değeri olduğundan daha yüksek çıkar.

Tau eşitliği. Maddelerin "yaklaşık tau eşitliğine" sahip oldukları varsayılır. Örneğin, ölçüm yapılan iki maddede araştırılan gerçek değerin (kavramsal yapının) her ikisinde de sabit bir oranda mevcut bulunduğu varsayılır. Diğer bir deyişle maddeler gizli değişkeni eşit ölçüde içeriyor

olmalıdır. Cronbach alfa için bu varsayım karşılanamamışsa hesaplanan değer olduğundan daha düşük çıkar. Gerçek hayatta bu koşulun tam olarak sağlanamadığını bildiğimizden alfa değeri, *düşük değerlikli bir katsayı* olarak adlandırılır.[11]

Maddelerin yanıtlanma oranı. Anketteki maddelerin %85'ine tam olarak yanıt verilmişse eksik maddelere atama yöntemi uygulandıktan sonra alfa değeri hesaplatılmalıdır. Uygulamada cevapsız bırakılan maddelerin yerine ölçeğin / faktörün aritmetik ortalaması, regresyon değeri karşılığı veya ölçeğin medyan değeri ikame edilir.

Norm referanslı testler için uygun olması. Alfa değeri norm referanslı testler için uygundur. *Teşhis* amaçlı ölçümlerde veya *başararak geçme* durumunu saptamaya yönelik olarak uygulanan kriter referanslı testlerde bu hesaplama yapılmaz.

Hız testleri. Alfa hesaplaması hız testleri için uygun değildir. Bu testlerde cevaplayıcıların önemli bir bölümü tüm şıklara cevap verme konusunda güçlükle karşılaştığından alfa güvenilirliği yerine paralel formlar güvenilirliği araştırılır.

Bilgisayar uyarlı testler. Alfa hesaplaması bilgisayarlardaki programlar aracılığıyla kişilerin yeteneklerine uygun olarak test sorusu gönderen ölçüm türleri için uygun değildir. Uyarlı testler bilgisayar ortamında değil, kağıt-kalem testi şeklinde de uygulansa durum değişmez. Bu tür testlerde MYK modellerinden yararlanılır.

Cronbach alfa katsayısına alternatif diğer güvenilirlik katsayıları. Literatürde Cronbach alfa değerinin yeterince güçlü olmaması nedeniyle alternatif güvenilirlik katsayılarının geliştirilmesi yoluna başvurulmuştur. Bu çerçevede sık kullanılan iki yöntem *bileşik güvenilirlik katsayısı* ve *çıkarılan ortalama varyans* değeridir.

Bileşik güvenilirlik. Cronbach alfa değerine alternatif olarak kullanılabilecek bir diğer yöntem, *bileşik güvenilirlik* (BG) katsayısıdır. Alan araştırmalarında kullanılan ölçüm araçlarının güvenilirliği bazen *bileşik güvenilirlik* katsayısıyla kanıtlanmaya çalışılır (Werts ve diğerleri 1974, aktaran Esteves).[12] Werts, Linn ve Jöreskog (1974) ise bu yaklaşımı "gizli

değişken güvenilirliği" (latent variable reliability) olarak adlandırmışlardır.[13]

Werts ve arkadaşlarına göre, bileşik güvenilirlik "arka plandaki özelliğe atfedilebilecek ölçüm varyansının oranıdır. Diğer bir deyişle özellik varyansının "özellik ve hata varyansının toplamına" olan oranıdır."[14] Bileşik güvenilirlik katsayısı ,50'nin üzerinde çıkmışsa bu ölçekler güvenilir kabul edilir. Nunnually'e göre ise BG katsayısı en az ,60 olmalıdır ve literatürde eşik değer olarak daha çok bu rakam kabul edilmiştir. Werts ve arkadaşlarının belirlediği bileşik güvenilirlik katsayısı, Teyit Edici Faktör Modeli ve Yapısal Eşitlik Modeli çerçevesinde özel istatistiksel yazılımlar kullanılarak hesaplanabilir. Bu amaçla en çok LISREL yazılımı kullanılmıştır. Lisrel yazılımında teyit edici faktör analizi, belirli bir faktör için öngörülen modelin iç tutarlılığını gösterir. Ancak literatürde bileşik güvenilirlik analizi çok faktörlü ölçekler için değil yine tek faktörlü veya tek boyutlu ölçekler için kullanılmıştır. Bileşik güvenilirlik katsayısının alfa katsayısına göre güvenilirliği daha gerçekçi bir şekilde gösterdiği ifade edilmiştir. Bileşik güvenilirlik katsayısında, maddelerin farklı faktör yüklerine ve hata varyanslarına sahip oldukları varsayılır. Oysa alfa değerinde faktör yükü değerlerinin ve hata varyanslarının eşit olduğu düşüncesinden hareket edilmiştir.[15] Werts ve arkadaşlarının geliştirdiği BG formülü Eşitlik 3-1'deki gibidir.[16]

$$\rho_c = \frac{(\Sigma \lambda_i)^2 \text{ var } X}{(\Sigma \lambda_i)^2 \text{ var } X + \Sigma \text{ var } (ei)} \quad . \qquad (3\text{-}1)$$

ρ_c = Bileşik güvenilirlik.
λ_i = X değişkenin Yapısal Eşitlik Modeli ile belirlenen faktör yükü.
e_i = X değişkeni için Yapısal Eşitlik Modeli ile belirlenen hata varyansı.

Eğer ölçek tek boyutlu ise bileşik güvenilirlik katsayısı ile alfa katsayısı birbirine yakın çıkar. Ölçek çok boyutlu ise alfa değeri ile BG katsayısı önemli ölçüde farklı çıkabilir.

Çıkarılan ortalama varyans. Alfa katsayısına alternatif olarak güvenilirliği belirlemenin bir diğer yöntemi "çıkarılan ortalama varyans" (average variance extracted) değerini kullanmaktır. Bu yaklaşım da istatis-

tik yazılım LISREL kullanılarak hesaplanır. Çıkarılan ortalama varyans değeri, göstergelerde arka plandaki gizli yapı tarafından temsil eden ortalama değişkenliği gösterir. Fornell ve Larcker çıkarılan ortalama varyans değerinin güvenilir sayılabilmesi için en az ,50 olması gerektiğini bildirmişlerdir (aktaran N. Schen).[17]

Çok boyutlu ölçekler ve Cronbach alfa. Cronbach alfa, çok boyutlu ölçeklerde yeterince güçlü olmayan bir iç tutarlılık ölçüsüdür. Çok sayıda yazar, alfa değerinin sadece tek boyutlu ölçekler için iyi bir güvenilirlik katsayısı olduğunu belirtmiştir. Geliştirilen ölçeklerin çoğunda birden fazla alt boyut olduğundan bu tür ölçeklerde alfa değerinin ölçeğin tamamının güvenilirliğini göstermek üzere kullanılması doğru değildir. Çünkü, çok boyutlu ölçeklerde alfa değeri görece daha düşük çıkar. Düşük çıkması bir kural değildir. Bazen çok faktörlü olmasına rağmen alfa değeri yüksek çıkabilir. Alfa değeri ölçeğin tek boyutlu olduğunu belirlemez, sadece maddelerin birbirleriyle tutarlı olup olmadığı hakkında bilgi verir. Örneğin, maddeler arasıdaki korelasyon katsayıları düşük olduğu halde (yüksek iç tutarlılık) alfa değeri yüksek çıkabilir. Cronbach ilk ifadelerinde (1947, 1951, aktaran Schmitt) çok boyutlu ölçeklere ait güvenilirliğin sadece aynı faktör yapılarını ölçen paralel formlarla tahmin edilebileceğini belirtmiştir.[18]

Bazı bilim adamları, çok boyutlu ölçeklerde düşük çıkan alfa değerini yükseltmek için Spearman-Brown *zayıflığı düzeltme* formülünü kullanırlar. Ancak bu yaklaşım doğru değildir. Araştırmacı düşük çıkan değeri zayıflığı düzeltme formülü ile yükseltmek isterken bu kez güvenilirliği olduğundan daha yüksek gösterme gibi bir durumla karşılaşır.[19]

Çok boyutlu ölçeklerde alfa değerini dört temel öge içinde ele almak gerekir: Ölçekteki faktör sayısı, her bir faktörün kapsadığı madde sayısı, faktör içindeki maddeler arasındaki korelasyon katsayıları ve faktörlerle maddeler arasındaki korelasyon katsayıları. Çok boyutlu ölçeğin oluşturulma biçimi, alt faktörlerin her birinin ayrı bağımsız bir boyut olarak ortaya çıkmasına neden olabilir. Literatürde, faktörlerin araştırılan kavramsal bir yapının ögeleri mi olduğu yoksa bağımsız boyutlar/yapılar olarak mı değerlendirilmesi gerektiği tartışmalı bir konudur. Bu nedenle birden fazla faktör içeren ölçeklerde güvenilirlik değerleri her bir faktör için sanki ayrı birer alt test imiş gibi ayrı ayrı verilmelidir.

Bazı psikometriciler ölçüm aracı çok boyutlu ise, bu tür ölçeklerin güvenilirliğinin, yarıya bölme yönteminde olduğu gibi, her bir faktöre ait

maddelerin yaklaşık olarak eşit olmasına bağlı olduğunu belirtmişlerdir.[20] Bu bilim adamlarına göre, faktörlere ait maddelerin sayısı arasında büyük farklılıklar bulunmamalıdır.

Kullanılan ölçekte birden fazla faktör varsa araştırmacı böyle bir durumda çok boyutlu ölçeklerin güvenilirliğini ölçen formüllerin yanı sıra, faktör bazında alfa katsayısı ile birlikte test-yeniden test, paralel formlar güvenilirlik yöntemlerinden birini daha uygulamayı denemelidir.[21] Çok boyutlu ölçeklerde kompozit alfa değeri, farklı faktörlere dayanan maddelerin gerçek varyanslarının eşit olmaması nedeniyle düşük çıkar. Raykov (1997) *yapısal eşitlik modelini* kullanarak çok boyutlu ölçeklerde ne kadar bir güvenilirlik aşınması ortaya çıktığını belirlemeye yönelik cebirsel bir hesaplama yöntemi geliştirmiştir (aktaran Ferligoj ve Mrvar).[22] Fakat önemli olan bu aşınmanın büyüklüğünü tespit etmek değil, çok boyutlu ölçekler için doğru rakamı verecek bir güvenilirlik rakamı saptamaktır. Çok boyutlu ölçeklerde güvenilirlik analizleri için kompozit alfa değeri yerine karmaşık matematiksel formüllere dayanan değişik hesaplama yöntemleri önerilmiştir. Psikometriyle ilgili güvenilirlik ölçümlerinin bu alanı gelişme aşamasındadır ve matematikçiler ile psikometriciler henüz alfa değerinde olduğu gibi basit ve kolay anlaşılan bir formül üzerinde belirli bir mutabakata ulaşamamışlardır. Aşağıdaki bölümde literatürde henüz gelişme aşamasında olan bu yaklaşımlara değinilmiştir.

Yapısal eşitlik modeli. Bazı araştırmacılar çok boyutlu ölçeklerin güvenilirliği için, *yapısal eşitlik modelini* (YEM) ve bu modelin içindeki uyuşma (goodness-of-fit) analizinin kullanılmasını önermişlerdir.[23] Yapısal eşitlik modeli, gizli faktörler arasındaki nedensellik ilişkilerini araştıran bir tekniktir. Yapısal eşitlik modeli, eş anlı olarak gözlem değişkenlerinin kendi aralarındaki ilişkileri, gözlem değişkenlerinin gizli değişkenlerle olan ilişkilerini ve gizli değişkenler arasındaki ilişkileri ölçer. YEM'deki gizli değişkenler de faktör analizindeki faktör ağırlıkları gibi faktör yüklerine sahiptirler. Bu faktör yükleri veya ağırlıkları gizli değişkenlerin güvenilirliğini belirlemek için kullanılır.[24] Gizli değişken yüklerinin en az .70 olması gerektiği belirtilmiştir.

Tarkkonen yaklaşımı. Çok boyutlu ölçekler için önerilen bir diğer yaklaşım Lauri Tarkkonen'in (1987) önerdiği ve son yıllarda Vehkalahti tarafından daha da geliştirilen yaklaşımdır. Ortak faktör analizin genelleş-

tirilmesine dayanan bu yaklaşımın *psikometrik alfa değeri, gerçek değer modeli* gibi pek çok modeli kapsadığı ileri sürülmüştür.[25]

Konjenerik analizi. Çok boyutlu ölçekler için Karl Jöreskog'un konjenerik analizi yöntemi uygulanabilir. Bu yöntemde hesaplama varsayımları çok katı değildir. Konjenerik analizinde maddelere ait gerçek puanların veya maddelerin hata varyanslarının eşit olduğu gibi bir varsayımda bulunulmaz. Jöreskog'un (1971) konjenerik ölçüm analizi, (KÖA) ölçek maddelerinin farklı gerçek puan varyans değerlerine sahip olduğu düşünüldüğünde uygulanabilecek bir yöntemdir. Konjenerik ölçümlerde, maddeler gerçek değerleri yüksek korelasyona sahip çiftler olarak ele alınır. Jöreskog bunun için varyans analizi kapsamındaki *maksimum benzerlik tahmin değerlerini* kullanmıştır.[26] Konjenerik analiz modeli, hâlâ test temelli bir yaklaşıma sahiptir ve elde edilecek çözüm kullanılan örnekleme ve göstergelerin yapıyla doğrusal ilişki içinde bulunmasına bağlıdır.[27]

En büyük alt sınır güvenilirliği. *En büyük alt sınır güvenilirliği* (greatest lower bound reliability), tek bir yöntem olmaktan çok, birkaç hesaplama yöntemini içeren teorik bir kavramdır. Bu kavram Jackson ve Agunwamba (1977) tarafından önerilmiştir (aktaran Sočan).[28] Onlara göre en büyük alt sınır güvenilirlik değeri hataların kovaryans matrisinin belirlenmesiyle tespit edilebilir. Bu matrisi belirlemek için değişik yazarlar tarafından farklı yaklaşımlar önerilmiştir. Son olarak Berge ve Kiers (1991) *minimum rank faktör analizi* adını verdikleri hesaplama yöntemini geliştirmişlerdir (aktaran Ferligoj ve Mrvar).[29] Güvenilirliği çok daha doğru bir şekilde ölçmesine karşılık, karmaşık hesaplama yöntemleri nedeniyle bu prosedürün uygulanması yaygınlık kazanmamıştır. Bu yaklaşımın bir diğer yetersizliği modelin ana kütle kovaryans matrisi hakkında bilgi sahibi olduğumuzu varsayıyor olmasıdır. Yazarlar küçük ölçeklerde büyük ölçüde yanlılığın ortaya çıktığını bu nedenle 1000 veya daha büyük örnek kütle büyüklükleriyle çalışılmasını önermişlerdir.

Ortak faktör modeli. Bir ölçekte birden fazla gizli değişken (faktör) varsa araştırmacı *ortak faktör modelinin* varsayımlarından yararlanabilir. Ortak faktör modelinde ölçekteki tüm maddeler veya bir test bataryasındaki tüm testler, ölçülmek istenen yapıyı temsil eden faktörle yüklüdür. Bir ölçekte/testte üç tür varyans aranılabilir: ortak faktör varyansı, grup faktörü varyansı ve spesifik varyans. Grup faktörü varyansı sadece belirli

maddelerde ortak olan özelliklerdir. Spesifik faktör varyansı ise her bir maddeye özgüdür. Ortak faktör modeli özellikle "genel yetenek" ölçümlerinde sık kullanılır. Bir kişinin ortak faktör puanı (g), batarya içindeki alt testlere ait standardize edilmiş puan değerlerinin ortalamaları alınarak bulunur.[30] Ancak bu puan, diğer alt testlerin içerdiği hata varyanslarından değişik ölçülerde etkilenmiş durumdadır. Genel ölçek veya test bataryasının güvenilirliğini hesaplamak için, her bir alt ölçeğin veya testin standardize edilmiş bileşik puanlarının kendi arasında ve bu puanlarla g puanları arasında Spearman veya Pearson korelasyon analizi yapılır. Hesaplama bu yönüyle madde-toplam puan korelasyonuna benzer. Ancak maddelerin yerini faktörler, toplam puanın yerini ise g faktörü almıştır. Ölçüm aracı çok boyutlu bir ölçek ise, içerdiği boyutların her biri alt bir ölçek gibi değerlendirilir ve ölçekteki madde sayılarının da yaklaşık olarak eşit olması arzulanır. Buna göre alt ölçekler genel faktörü ve grup faktörünü eşit ölçüde içeriyor varsayılır. Ölçeğin/testin tamamına ait güvenilirlik, alt ölçekler sanki paralel formlar imiş gibi değerlendirilip korelasyon katsayısı ile tespit edilir.

Bileşik puanların güvenilirliği. Bileşik puanların güvenilirliğinde[a] bataryadaki / ölçekteki faktör veya test sayısı, ortalama test güvenilirliği ve testler arasındaki korelasyon katsayılarının ortalaması dikkate alınır. Bileşik puanların güvenilirliği Eşitlik 3-2'deki formülle hesaplanır.[31]

$$r_{bp} = 1 - \frac{k - \left(k\bar{r}_{ii}\right)}{k + \left(k^2 - k\right)r_{ij}} . \qquad (3\text{-}2)$$

r_{bp} = Bileşik puanların güvenilirliği.
k = Bataryadaki test sayısı veya bir ölçekteki alt ölçek/faktör sayısı.
$\bar{r}_{ii}$ = Ortalama test güvenilirliği.
r_{ij} = Test/faktör puanları (toplam veya ortalama değerleri) arasındaki korelasyon katsayılarının ortalaması.

[a] Okuyucu "bileşik güvenilirlik katsayısı" kavramıyla "bileşik puanların güvenilirliği" kavramlarını birbirine karıştırmamalıdır. İlki bu amaçla geliştirilmiş özel bir terim iken ikincisi tek bir güvenilirlik rakamı elde etmeye yönelik bir hesaplama biçimidir.

Madde-toplam puan korelasyonları. Bu yöntemde, ölçeğin eğer birden fazla özellik içerdiğinden kuşkulanıyorsa veya literatürdeki bulgulardan veya önceki çalışmalardan kavramsal yapının birden fazla faktörden oluştuğu tahmin ediliyor veya biliniyorsa madde-toplam puan korelasyonu yöntemine başvurulur. Söz konusu korelasyon analizleri sonucunda çoklu özellik-çoklu madde (multitrait/multiitem) korelasyon matrisi oluşturulur. Madde-toplam puan korelasyonunda düzeltilmiş korelasyon katsayıları eğer ,40'ın üzerinde ise bu maddelerin söz konusu faktörü ölçtüğüne karar verilir. Bu ölçü aynı zamanda bir iç tutarlılık kriteridir. Matriste maddelerin ayırt edicilik özelliği, ölçülen faktör ile maddeler arasındaki korelasyon yüksek iken diğer faktörlere ait maddeler arasındaki korelasyon düşük ise açık bir şekilde ortaya çıkmış olur. Söz konusu korelasyonlar arasındaki farklılığın anlamlılığı ise, Steiger'in (1980) bağımlı korelasyonlar için *t*-testi formülü ile hesaplanır.[32] Çoklu özellik – çoklu madde matrisini el ile hesaplamak ve oluşturmak zor olduğundan bu konuda SAS isimli istatistik analiz yazılımının MULTI isimli makrosundan yararlanılabileceği bildirilmiştir.

Madde-yanıt kuramında çok boyutluluk. Araştırmacı güvenilirlik analizlerini madde-yanıt kuramı çerçevesinde yapıyorsa çok boyutlu ölçeklerin güvenilirliği için, *çok boyutlu madde-yanıt kuramına*[a] (ÇBMYK) ait modellerden yararlanabilir. Bu tür modellerde test maddelerinin hiyerarşi modeline göre kalibre edilmesi gerekir.[33]

Psikometrik test bataryaları ve Cronbach alfa. Çok boyutlu ölçeklerin güvenilirliğinde olduğu gibi birden fazla testten oluşun test bataryalarının toplam puan değeri için de aynı durum söz konusudur. Özellikle personel seçim testlerinde, öğrencilerin okullara yerleştirmelerinde kullanılan testlerde birden fazla puan vardır ve genellikle bu puanlar bazen basit bir şekilde bazen de belirli katsayılarla çarpılmak suretiyle toplanarak (bileşik hale getirilerek) tek bir puan olarak ifade edilir. Tek bir puan haline getirme işleminde testlerdeki soru sayısı ve testlerin önemi göz önünde bulundurularak ağırlıklandırma yapılabilmektedir. Ancak bu ağırlıklandırmanın bileşik puanların güvenilirliğini arttırmada önemli bir işlevi olmadığı bulunmuştur.[34] Farklı testlerin ham puanlarının toplanarak tek bir puan haline getirilmesiyle her bir testin kendi içindeki alana ait varyansları veya puan değişkenliklerini görmek mümkün olmaz. Bu nedenle farklı testlerden elde

[a] Multidimensional IRT testing methods (MIRT).

edilen ham puanlar yerine standardize edilmiş puanların toplanması gerekir.[35] Öte yandan standardize edilmiş bile olsa testlerin ağırlıklarını eşit saymak maddelerin içeriği açısından doğru olmayabilir. Örneğin, üniversite giriş sınavından bir cebir probleminin çözümü, öğrencinin iki dakika gibi bir süre kullanmasını gerektirirken bir dilbilgisi sorusu belki de en fazla 15 saniyesini alacaktır. Böyle olunca testlerin gerektirdiği zihinsel çaba, hata değişkenliği ve cevaplama süreleri farklı olacaktır. Ham puanlar standartlaştırılmış bile olsa ağırlık verilmeden toplanması bileşik puanların gerçeği yansıtmaması olgusuyla sonuçlanabilir. Güvenilirliği arttırmanın yollarından biri, daha zor olan testlere belirli bir ağırlık vermektir. Rudner (2000), bileşik puanların güvenilirliğini arttırmaya yönelik bir formül önermiştir. Bu formüle göre bileşik puan güvenilirliği artarken bileşik geçerlilik azalabilmektedir.[36] Bu nedenle araştırmacı bileşik puanlarla çalışırken geçerlilik ve güvenilirlikten hangisini ön plana çıkaracağını veya ikisi arasında nasıl bir denge kuracağını belirlemelidir.

Çok boyutlu ölçek ve testlerde maddelere ağırlık verilmesi. Çok boyutlu ölçeklerde ağırlık kullanılmadığı durumda alfa güvenilirlik katsayısı alt sınırdaki güvenilirlik değerlerini veren *ihtiyatlı* bir rakamdır. Ölçekteki maddelere ağırlık verildiğinde, alfa katsayısı yerine bir faktör analizi türü olan *temel bileşenler analizli* sonucu elde edilen *theta katsayısının* kullanılmasının daha doğru olacağı belirtilmiştir.[37] Theta katsayısı alfa katsayısından daha büyüktür. Maddelere ağırlık verildiğinde theta değeri *maksimum alfa katsayısı* olarak değerlendirilir.

Alfa katsayısı - alfa güvenilirlik indeksi. Alfa, istatistiksel bir test değil, matematiksel hesaplamalara dayanan güvenilirlik indeks değeridir. Bazı yazarlar bu nedenle *katsayı* teriminin kullanılmasına karşı çıkmışlardır. Çünkü katsayı korelasyon analizi gibi istatistiksel işlemler sonucunda elde edilir. Bununla birlikte kimi yazarlar da alfa indeks değeri için *alfa puanı* teriminin kullanılmamasını dile getirmişler ve katsayı kelimesini önermişlerdir. Cronbach'ın kendisi *katsayı* terimini kullanmıştır. Burada hatırda tutulması gereken nokta, katsayı kelimesi kullanılsa da bunun istatistiksel bir hesaplamaya dayanmadığıdır. Güvenilirlik indeks değeri olarak isimlendirilmesi, arka planda yatan gizli kavramsal yapıdaki değişkenlik hakkında bilgi vermesi nedeniyledir. Bazı matematiksel ve istatistiksel yazılımlarda alfa katsayısının kare kökü "alfa indeks değeri" olarak tanımlanmıştır. Bu değer, gözlem puanlarıyla evren puanları arasındaki korelasyonun karesine işaret eder.

Alfa değeri ve pilot araştırma sonuçları. Güvenilirlik değerlendirmesi esas araştırma sonuçlarına dayalı olarak yapılır. Bilim adamının ölçek geliştirme çalışmaları sırasında yaptığı pilot araştırma sonuçlarında alfa değeri yüksek veya düşük çıkabilir, ancak her iki sonuç da yanıltıcıdır. Pilot araştırma uygulaması ölçekteki önemli hataları gidermeye ve test maddelerini kalibre etmeye yöneliktir. Pilot araştırma sırasında kendilerine test veya ölçek uygulanan kişiler ölçümü gerçek koşullarında uygulamadıklarından bazen güvenilirlik değeri oldukça düşük çıkar. Ölçüm aracının güvenilirliğini sadece pilot araştırma sonuçlarına bakarak değerlendirmek prematüre bir karardır.[38] Bilim adamı alfa değerine ilişkin sonuçları raporlarken pilot araştırma ve esas araştırma sonuçlarının her ikisini de vermelidir. Ancak araştırmacının bu sonuçları Fisher *z'* puanlarını kullanarak birleştirmesine gerek yoktur ve bu yaklaşım doğru da olmaz. Pilot araştırma örneklemi ile asıl uygulama örneklemi farklı ise alfa değerlerinin benzer veya farklı çıkmasının bir anlamı yoktur.

Alfa katsayısının standart hatası. Cortina (1993, aktaran Schmitt) bir ölçeğin çok faktörlü olma ihtimalinin bulunması veya ölçüm değerlerinin büyük ölçüde örneklem hatasından etkileniyor olması nedeniyle maddeler arasındaki korelasyon katsayılarının benzerlik yerine geniş bir dağılım gösterebileceğini ifade etmiştir. Cortina, hesaplama sonucunda maddeler arasındaki korelasyon katsayıları önemli ölçüde değişkenlik gösteriyorsa alfa değeri raporlanırken *alfa kesinlik değeri* veya *alfa değerinin standart hatasının* da verilmesinin doğru olacağını belirtmiştir. Bu istatistik maddeler arasındaki korelasyonların dağılımı hakkında bilgi verir. Maddeler arasındaki korelasyonlar sıfır ise indeks, 0 değerini verir. Korelasyonların dağılımı büyük ölçüde farklı ise daha yüksek bir indeks değeri elde edilir. Feldt, Woodruff ve Salih (1987) alfa katsayısının standart hatasını hesaplamaya yönelik bir formül önermişlerdir. Araştırmacılar eğer alfa değerinin doğruluğunu değerlendirmek istiyorlarsa örneklem hatasıyla ilgili olan *Feldt indeks* değerini hesaplatabilirler (aktaran Schmitt).[39]

Alfa değerinin yanlış kullanılması. Az sayıda madde içeren ölçeklerde alfa tek boyutluluğa işaret edebilir, fakat hesaplanma amacı tek boyutluluğu ortaya çıkarmak değildir. Alfa, Spearman'ın *g* faktörünü göstermez. Çünkü bir ölçekte birbirinden bağımsız iki faktör bulunabilir.

Alfa güvenilirlik değerinin büyüklüğü. Nunnelly'e (1998) göre, alfa güvenilirlik değeri ,70'den büyük olmalıdır. George ve Mallery'e (2003) göre ise alfa değerinin > ,9 olması mükemmel; ,8 –,9 arasında olması iyi; ,7 – ,8 arasında olması kabul edilebilir; ,6 – ,7 arasında olması kuşkulu; ,5 – ,6 arasında olması zayıf ve ,5'in altında olması ise kabul edilemez olarak değerlendirilir (Aktaran Gliem, 2003).[40] Ancak bu tanımlamalar geneldir. Son yıllarda araştırmacılar güvenilirlik katsayılarının test ve ölçeğin niteliğine göre değişebileceğini belirtmişlerdir. Örneğin zeka testleri gibi bilişsel testlerde alfa güvenilirlik katsayısının en az ,80; psikolojik kavramsal yapıları ortaya çıkarmayı amaçlayan ölçeklerde, yetenek ve becerileri ölçen testlerde ise en az ,70 olması gerektiği belirtilmiştir.[41]

Alfa indeks değeri, madde çıkarılarak, yeni maddeler ilave edilerek veya mevcut maddelerde değişiklik yapılarak artırılabilir. Bunun için ölçeğin tamamına ait güvenilirlik katsayıları hesaplandıktan sonra her bir değişkenin korelasyon katsayılarına ve alfa değerlerine bakılır. Korelasyon katsayısı ve alfa değeri düşük olanlar ölçekten çıkarıldıktan veya değişiklik yapıldıktan sonra hesaplatma yeniden yapılır ve alfa değerindeki değişiklik gözlemlenir. Muhtemelen bu kez alfa değeri daha yüksek çıkacaktır. Bir madde inceleme dışı bırakıldığında alfa katsayısı önemli ölçüde artıyorsa kural olarak bu madde ölçek dışında bırakılır. Ölçekte önceden belirlenmiş olan madde sayısından daha önemli olan asgarî güvenilirlik seviyesini tutturmaktır.

Alfa analizi sonucunda ölçek büyük ölçüde güvenilir çıkmamışsa maddeler bütünüyle yeniden gözden geçirilir ve gerekli değişiklikler yapılır. Araştırmacı ilke olarak ,50 ve daha düşük çıkan alfa değerlerini yorumlamaya çalışmamalı, böyle bir durumda güvenilirliği artıracak çalışmalar üzerinde odaklanmalıdır.

Alfa değeri ve örneklem. Araştırmada kullanılan ölçek bir kişilik testi ise, psikolojik veya psikiyatrik teşhis koyma amaçlı bir çalışma ise Cronbach alfa değerinin toplumun değişik kesimleri için ayrı ayrı hesaplatılması gerekir. Psikolojik ve psikiyatrik nitelikteki ölçeklerin normal bireyler, rahatsızlık şikayetiyle başvuran bireyler için; diğer kişilik ve ilgi envanterlerinin de öğrenciler, yetişkinler geneli ve belirli bir meslekte çalışan özel gruplar için ayrı ayrı hesaplatılması doğru olur.

Güçlü alfa değeri. Cronbach alfa indeks değerinin yeterince güçlü olmaması nedeniyle pek çok bilim adamı bu katsayının güvenilirliğin gös-

tergesi olarak ileri sürülmesine itiraz etmiştir. Son yıllarda matematikçiler alfa katsayısını güçlendirmeye yönelik çalışmalar yapmışlardır. Bu çalışmaların sonunda *güçlü alfa katsayısı*[a] olarak isimlendirilen yeni yaklaşımlar ortaya çıkmıştır. Bu yaklaşımların temel felsefesi, birkaç maddede görülebilecek farklı kaynaklara ait küçük veri değişkenliklerine veya gözlem hatalarına karşı çok duyarlı olan alfa değerini maddeler bazında değil, veri kütlesi bazında daha duyarlı hale getirmektir. Güçlü alfa formülü üzerinde ilk kez Wilcox (1992) çalışmıştır. Wilcox, varyans ve kovaryansın[b] tahmininde ayrık değerleri içermeyen dizinin orta bölümde yer alan değerlere ait varyans ve kovaryansları hesaplamaya dahil etmiştir. Christmann ve Alest ise (2002), ayrık gözlemlerden veya ölçüm sapmalarından etkilenmeyecek bir alfa formülü üzerinde çalışmışlardır. Bu bilim adamları varyans ve kovaryansın hesaplanması için *kovaryans matrisinin* kullanılmasını önermişler ve kovaryans matrisinin S-Plus, R, EQS ve SAS gibi istatistik analiz programlarıyla kolayca hesaplanabileceğini belirtmişlerdir.[42]

Alfa güvenilirlik değerinin artırılması. Alfa güvenilirlik değeri yeterince yüksek çıkmamışsa araştırmacı birkaç yöntemden yararlanabilir. Bunlardan birincisi *kalibrasyon* yöntemi, ikincisi *üçleme* yöntemi ve üçüncüsü ise *zayıflığı yenme* yöntemidir.

Kalibrasyon; ölçekteki maddelerin gözden geçirilerek incelenmesi, bu inceleme sonrasında yanlış anlam içermesi sebebiyle tutarsız işaretlemeye neden olan, ayrık puan içeren, katılımcıları ayrıştırmayan maddelerin saptanması ve bu maddelerde gerekli iyileştirme çalışmalarının yapılmasıdır. Kalibrasyon araştırmacının tek başına kendisi tarafından veya birkaç kişiden oluşan bir uzmanlar grubu tarafından yapılır. Uzmanlar, sorunlu maddeleri okuyarak birlikte tartışırlar ve bu maddeleri yeniden ifadelendirirler. Madde iyileştirme oturumlarına *kalibrasyon toplantıları* adı verilir. Kalibrasyon toplantıları ölçeğin geliştirilme aşamasında pilot araştırmadan önce veya sonra yapılır. Bu toplantılarda sonucun ne çıkması gerektiği konusu üzerinde durulmaz. Amaç, maddelerin birbirleriyle tutarlı olmasını

---

[a] Güçlülük (Robustness). Genel hatlarıyla güçlü olma ve istenen sonucu gerçeğe yakın olarak verme. İstatiksel analizlerde bir test, önceden belirlenmiş dayandığı varsayımları tam olarak doğrulmamasına rağmen gerçeğe yakın sonuçlar veriyorsa bu testin güçlü olduğundan söz edilir.

[b] Ortak varyans (kovaryans). İki değişkene ait sapmaların ne ölçüde birbirine denk veya yakın olduğunun ölçüsüdür. kov (x, y) = top. $[(x - x_{ort})\ (y - y_{ort})]$. Denklik yüksek olduğunda formülden en yüksek değer elde edilir. *Bk.*, <http://www2.chass.ncsu.edu/garson/pa765/correl.htm>

sağlamaktır. Bazı bilim adamları ölçeğin pilot araştırmasını kalibrasyon toplantılarıyla gerçekleştirirler. Ölçek önce beş kişilik bir gruba uygulanarak analitik bir değerlendirmeye tâbi tutulur. Değerlendirme sonunda gerekli değişiklikler yapıldıktan sonra duruma göre ikinci veya üçüncü bir kalibrasyon toplantısı daha yapılır. Kalibrasyon toplantılarına katılan kişiler *kalibrasyon örneklemi* olarak isimlendirilir. Kalibrasyon toplantısının bir diğer şekli, personel başarı değerlendirmesi yapacak yöneticilere, sportif etkinliklere puan verecek gözlemci hakemlere ve sınav kağıtlarını değerlendiren öğretmenlere değerlendirme ve puan verme yöntemleriyle ilgili belirli yeteneklerin kazandırılmasıdır. Bu tür kalibrasyon toplantıları senede birkaç kez tekrarlanır. Öte yandan Rasch modelinde kalibrasyon, test maddelerinin belirli yetenek düzeylerine uygun güçlük / zorluk düzeyinin saptanması çalışmalarına verilen addır.

Üçleme, birden fazla ölçüm yönetimini kullanarak sistematik hatayı (yöntem hatasını) azaltmaktır. Araştırmacı anket yöntemiyle topladığı verileri inceleyerek bu verilere ait sonuçların; (a) gözlem değerleriyle ve (b) önceki araştırma bulgularıyla tutarlı olup olmadığını inceler. Eğer tutarsızlıklar saptamışsa bunların araştırmada kullanılan yöntemin hangi ögelerinden kaynaklandığını bulmaya ve yönteme ilişkin bu olumsuz etkenleri ortadan kaldırmaya çalışır.

Zayıflığı yenme yöntemi, araştırmacıya doğrudan düşük alfa güvenilirlik katsayısını iyileştirme fırsatı sağlamaz. Fakat değişik madde sayılarıyla çalışılması halinde alfa güvenilirlik katsayısının ne olabileceği hakkında bir fikir verir.

Lambda 4. Alfa değerinin gerçek güvenilirliği tam olarak göstermemesi ve düşük kalması nedeniyle Monte Karlo teknikleri kullanılarak alfa değerini daha güçlü yapacak teknikler üzerinde çalışılmıştır. Bu konuda H. G. Osburn (2000), yaptığı çalışmalar sonucunda *maksimum lambda 4* değerinin güvenilirliği tam olarak gösterdiğini iddia etmiştir (aktaran Wuensch).[43]

Lambda 4 [lam'da] (λ_4), aynen alfa katsayısı gibi hesaplanır. Bu yöntemde ölçeğin muhtemel bütün yarıya bölme güvenilirlikleri hesaplanır, fakat yarıya bölme çiftlerinden sadece bir tanesi, güvenilirlik katsayısı en yüksek olanı dikkate alınır. Alfa katsayısının hesaplanması için $2n$ sayıda maddeden oluşan herhangi bir testte ,$5(2n)!/(n!)2$ kadar yarıya bölme olasılığı söz konusudur. Maksimum lambda 4'ü elde etmek için mümkün bütün yarılar için λ_4 hesaplanır ve en yüksek λ_4 değeri, güvenilirlik katsayısı

olarak seçilir. Ölçekteki madde sayısı 4-5 gibi küçük bir rakam olduğunda bu yöntemi uygulamak kolaydır. Ancak 10 maddeli bir ölçekte 126 muhtemel yarıya bölme olasılığı olacağından her biri için lambda $_4$ değerini hesaplamak imkansız hale gelir. Çok sayıda maddeden oluşan ölçeklerde λ_4 değerini hesaplamak için bu amaçla yazılmış programlardan yararlanılır.[a] Henüz yaygın kullanıma sahip olmayan bu yöntemde λ_4 katsayısının hesaplanma biçimi için okuyucular ilgili literatüre başvurmalıdırlar.

Theta. D.J. Armor (1974) tarafından türetilen theta [te'ta] güvenilirliği, alfa güvenilirliğine benzer.[b] Theta güvenilirliği, faktör analizi sonucu ortaya çıkan birinci faktördeki iç tutarlılığı ölçmek üzere geliştirilmiştir. Theta katsayısı Eşitlik 3-6'daki formül aracılığıyla hesaplanır (aktaran Yafee):[44]

$$\Theta = \left(\frac{k}{k-1}\right)\left(1 - \frac{1}{\lambda_1}\right). \qquad (3\text{-}6)$$

k = Ölçekteki madde sayısı.
λ_1 = En yüksek özdeğer (eigenvalue).

İstatistiksel analiz programı SPSS'in mevcut sürümleri theta değerini doğrudan hesaplamamaktadır. Ancak faktör analizi çıktılarından hareket edilerek theta değeri hesaplanabilir. Vernon Greene ve Edward Carmines (1979) Cronbach alfanın değişik bir şekli olan theta'nın maksimum alfa değerine eşit olduğunu bildirmiştir (aktaran Vehkalahti, 2004).[45] Faktör analizi sonucunda hesaplanan theta'nın sınır değeri ,5 olarak belirlenmiştir. Hesaplama sonucunda theta değeri ,5'in üzerinde olan faktörler alı konmakta ,5'in altında olan faktörler ise ret edilmektedir.[46] Theta değeri, alfa değeri gibi yorumlanır.

Omega. Heise ve Bohrnstedt (1970) tarafından geliştirilen bu yaklaşım çok boyutlu ölçekler için önerilmiştir. Bir test veya ölçekteki maddeler için faktör analizi yöntemi uygulanarak analiz sonucunda maddelerin ortak

---

[a] Bu programlardan biri Karl L. Wuensch tarafından yazılan Lambda4.sas isimli yazılımdır.

[b] Theta güvenilirliği konusunda daha fazla bilgi edinmek için *bk.*, David J. Armor, "Theta Reliability and Factor Scaling [Theta Güvenilirliği ve Faktör Ölçeklemesi]," Der., Herbert L. Costner, *Sociological Methodology*, San Francisco: Jossey-Bass, 1974.

varyansının belirlenmesidir. Omega (ω), maddelerin toplam varyansına etki eden ortak varyansın oranını belirler. Alfada olduğu gibi Omega katsayısı da 0 ilâ 1 arasında değişir. Omega ile alfa arasındaki ilişkiler Kent Smith (1974) tarafından incelenmiştir. Smith, bilim adamının tek bir ortak faktör modelini araştırması halinde Omega'nın en iyi çözüm olacağını belirtmiştir (aktaran Vehkalahti).[47] Omega'nın literatürdeki diğer güvenilirlik katsayılarıyla olan ilişkisi Tablo 3-4'teki gibidir:[48]

Tablo 3-4. Omega Değerinin Diğer Güvenilirlik Katsayılarıyla İlişkisi

Model	*Kovaryans matrisi*	*Korelasyon matrisi*
Tek bir ortak faktör modeli	$\omega > \alpha$	–
Tau eşitliğine sahip maddeler	$\omega = \alpha$	–
Paralel maddeler	$\omega = \alpha$	$\omega = \alpha$
Tam paralel maddeler	$\omega = \alpha$	$\omega = \alpha$

Omega değeri, değişik şekillerde yorumlanabilir: (a) Omega, maddeleri temsil eden faktör *F* ile, madde puanlarının toplamı olan *Y* arasındaki korelasyon değerinin karesidir. (b) Omega *Y* ve *Y`* simgeleriyle gösterilen aynı ortalama ve aynı varyansa sahip, ayrıca ortaklaşa olarak homojen olma özelliğini üzerlerinde taşıyan ve tek faktör modeline tam olarak uyan iki test puanı arasındaki korelasyondur. (c) Omega bir testteki toplam (veya ortalama) puan ile sonsuz sayıda madde içeren homojen ölçüm alanındaki maddelerin ortalamaları arasındaki korelasyonunun karesidir.

Kuder-Richardson homojenlik analizi. Kuder ve arkadaşı Richardson 1937'de norm temelli[a] bir testin güvenilirliğini belirlemek için bu yöntemi geliştirmişlerdir. Bu yöntem tek bir formun uygulanmasıyla güvenilirliği ölçmektedir. Yarıya bölme yöntemine benzemekte ve testi farklı yarılardan

[a] Norm temelli test: Başarı veya başarısızlık sınırını belirlemeye yönelik olarak geliştirilen ve daha çok ana kütleyi temsil eden örneklem verilerini temel alan ölçümlerdir. Okullarda öğrencilerin ders başarılarını ölçmek için kullanılan testler ile iş hayatında personelin başarılarını ölçmek için kullanılan testler bu gruba girer. Norm grupları değişik ölçütlere göre belirlenir: ulusal normlar, bölgesel normlar, kültürel normlar, gelişme normları, meslek normları, yaş, cinsiyet, etnik köken ve eğitim normları en sık kullanılanlardır.

bölerek bu yarılar arasındaki korelasyonu araştırmaktadır. Kuder–Richardson 20 ve 21 formülleri iki şıktan oluşan bilgi testleri ile başarı ve başarısızlığı gösteren psikometrik testlerdeki türdeşlik güvenilirliğini hesaplar.

Kuder–Richardson 20. Kuder-Richardson 20 formülünde test maddelerinin eşit zorlukta olmadığı varsayımı söz konusudur. Maddeler/sorular iki şıklı/dereceli ölçek değerlerine sahipse KR-20 formülü uygulanır (*bk.*, Tablo 3-5). Kuder-Richardson formülünün uygulanabilmesi için ölçüm aracı iki paralel bölüme ayrılır. Paralel iki bölümün puanları arasında belirlenen formüle göre korelasyon analizi yapılır. Bu yöntem özellikle tek bir yapıyı ölçtüğü zaman uygundur.

■ Kuder-Richardson 20 formülü:

$$\rho_{KR20} = \frac{k}{k-1}\left[1 - \frac{\sum pq}{S^2}\right]. \qquad (3\text{-}7)$$

p = Doğru yanıt oranı. $q = 1 - p$.
k = Testteki madde sayısı.
Σpq = Bir maddeye doğru yanıt veren cevaplayıcıların yüzdesi ile yanlış yanıt veren cevaplayıcıların yüzde çarpımlarının toplamı.
S^2 = Toplam puanların standart sapmasının karesi (varyans).

Cronbach alfa, iki şıklı değişkenler (doğru / yanlış) için uygulandığında Kuder-Richardson 20 formülüne eş değerdedir. Cronbach alfa katsayısı hesaplanmışsa ayrıca Kuder-Richardson 20 formülünü uygulamaya gerek yoktur. Test maddelerinin, sınav sorularının zorluk dereceleri farklıysa KR-20 formülü uygulanır. Kuder-Richardson 20 formülü için Σpq değerleri Tablo 3-6 çerçevesinde oluşturulabilir.

Tablo 3-5. Kuder – Richardson 20 Formülü İçin Veri Tablosu

Kişiler	Maddeler										
	1	2	3	4	5	6	7	8	9	10	Toplam
1	1	0	0	1	1	1	0	1	1	1	7
2	1	1	0	1	1	0	1	1	1	1	8
3	0	0	1	1	1	0	0	1	0	1	5
4	0	1	0	1	0	1	1	1	1	1	7
5	0	0	0	0	0	0	0	1	0	0	1
6	1	0	0	0	0	1	1	1	0	0	4
7	1	0	0	0	1	0	0	0	1	1	4
8	1	0	1	0	0	1	1	0	1	1	6
9	1	1	1	1	1	0	0	1	1	1	8
10	1	1	1	0	0	1	1	1	1	1	8
p	,80	,40	,40	,50	,50	,50	,50	,80	,70	,80	
1-p	,20	,60	,60	,50	,50	,50	,50	,20	,30	,20	
pq	,16	,24	,24	,25	,25	,25	,25	,16	,21	,16	2,17
S											2,29
S^2											5,28

Tablo 3-6. Kuder-Richardson 20 Formülünde Başarılı ve Başarısız cevapların Dağılım Oranları

Madde no	Doğru yanıtların oranı (p)	Yanlış yanıtların oranı (q)	pq
1	p_1	q_1	p_1q_1
2	p_2	q_2	p_2q_2
:	:	:	.
k	p_k	q_k	p_kq_k
-	-	-	Σpq

Tablodaki değerler formüldeki yerine konarak gerekli hesaplamalar yapılır.

$$r_{KR20} = \frac{10}{10-1}\left[1-\frac{2,17}{5,28}\right], \qquad (3\text{-}8)$$

$$r_{KR20} = \frac{10}{9} \cdot \frac{3,11}{5,28}, \tag{3-9}$$

$$r_{KR20} = ,65. \tag{3-10}$$

KR-20 formülü, az sayıda maddeden oluşan (10 –15 madde gibi) bilgi testleri için uygulanmışsa ,50 gibi düşük bir değer dahi güvenilir kabul edilir.[49] Fakat 50'den fazla madde içeren bir testin KR-20 güvenilirlik katsayısı muhtemelen ,80'in üzerinde çıkacaktır. Testteki madde sayısı arttıkça güvenilirlik değeri de artar. Bu nedenle bilgi testlerinde herhangi bir konuyla veya bölümle ilgili olarak en az 10 test maddesinin sorulması önerilir. Öğrencilerin üniversite *giriş sınavları* gibi yaşamlarını etkileyecek testlerde maddelerin KR-20 güvenilirlik katsayısı ,80'den daha düşük olmamalıdır. Aynı şekilde personel seçimlerinde kullanılacak BYB testlerinde de asgari ,80 güvenilirlik katsayısına ulaşılmalıdır.

KR-20 formülünün hız testlerinde uygulanmaması önerilmiştir. Bu testlerde belirlenen süre içinde yanıtlanan madde sayıları farklı olacağından sonuçlar yanlış çıkabilir. Hız testleri için test-yeniden test güvenilirlik analizi yönteminin uygulanması daha doğrudur.[50]

KR-21 formülü. KR-21, çoktan seçmeli sorular ve ölçekler için kullanılır. Çoktan seçmeli sorularda sadece tek bir doğru yanıt varsa doğru yanıtlar 1 ve yanlış yanıtlar da 0 şeklinde kodlanır. KR-21 formülü, test maddelerinin yaklaşık olarak *eşit zorlukta* olduğu varsayımı altında kullanılır. Bu şart sağlanamadığı zaman formül, güvenilirlik değerini olması gerekenden daha düşük hesaplar. Bu nedenle giderek zorlaşan *progresif matris* testi maddelerinin güvenilirliğini ölçmek için uygun değildir.

- Kuder – Richardson 21 formülü.

$$\rho_{KR21} = \frac{k}{k-1}\left[1 - \frac{\bar{x}\left(1 - \frac{\bar{x}}{k}\right)}{S^2}\right]. \tag{3-11}$$

k= Testteki madde sayısı.

x = Grup puanlarının aritmetik ortalaması.
S^2 = Puanların standart sapmasının karesi (varyans).

Yarıya bölme yöntemi. Yarıya bölme, testin iç tutarlılığını ölçmekte yararlanılan bir diğer yöntemdir. Likert ölçeklerinde, psikometrik testlerde, öğrencilerin bilgilerini ölçmeye yönelik olarak oluşturulan başarı testlerinde bu yönteme başvurulabilir. Yarıya bölme yönteminde aşağıdaki prosedüre uyulur.

Test veya ölçek cevaplayıcılara uygulandıktan sonra belirli bir sisteme bağlı olarak iki eşit yarıya bölünür. Bu işlemde maddeler; (a) rasgele, (b) tek – çift sıralaması içinde, (c) birinci yarı – ikinci yarı şeklinde, (ç) her bir yarı alt boyutları / faktörleri eşit ölçüde içerecek şekilde veya (d) kolaylık ve zorluk açısından maddeler her iki yarıda dengeli olacak şekilde bölünür. Ölçüm aracı bir test ise ve test maddeleri de kolaydan zora doğru sıralanmışsa birinci yarı ve ikinci yarı yerine tek – çift maddeler şeklinde bölme uygulamasına gidilir. Ölçüm aracı bir ölçek ise, maddelerin zorluğu söz konusu olmayacağından bu kez maddelere gelen yanıtların dağılım özelliği veya ölçeğin faktöriyel yapısı dikkate alınır. Varyansı yüksek olanlar *kolay karar verilemeyen* veya kolay işaretleme yapılamayan maddeler olarak değerlendirilir.[a] Bu tür maddelerin iki yarıda eşit ölçüde bulunmasına dikkat edilir. Çünkü bu yöntemde her iki yarının *yaklaşık olarak tau eşitliğine* sahip olduğu ve her bir yarıdaki maddelerin alandaki maddeleri temsil ettiği varsayılır.

İkinci aşamada her bir yarının toplam puanları bulunur (x ve x'). İki yarı arasındaki korelasyon analizi, toplam puanlara göre veya kullanılan ölçüm aracında belirli değişkenlere cevap verilmemişse aritmetik ortalama değerlerine göre yapılır ($\rho_{xx'}$). Elde edilen korelasyon katsayısı, *yarıya bölme güvenilirliği katsayısı* olarak isimlendirilir. Yarıya bölmede tek-çift rakam uygulamasına gidilmişse bu kez korelasyon katsayısı, *tek-çift güvenilirliği katsayısı* olarak adlandırılır. İki yarı arasındaki ilişkiler yazılım kullanılarak araştırılacaksa, korelasyon analizi dışında her bir yarı için ayrıca diğer istatistikî analizler de yapılır. Bu istatistikî analizlerde aşağıdaki konular araştırılır:

[a] Likert ölçeklerinde maddelerin yetenek açısından zorluğu veya kolaylığı söz konusu değildir. Onun yerine cevap şıklarının dağılımında *tercih etme* faktörü dikkate alınır. Bazı maddelerde çok kolay tercih yapılırken, diğerlerinde tercih yapmak daha zor olabilir.

1. Kişilerin testin her iki yarısında aynı toplam puanlara sahip olup olmadıkları.
2. Kişilerin testin her iki yarısında yüzdelik dilimlerinin aynı olup olmadığı.
3. Kişilerin testin her iki yarısında aynı z puanlarına sahip olup olmadıkları.

Araştırmacı yarıya bölme yöntemini uygulayabileceğini düşünerek ölçekteki madde sayısını mümkün olduğunca çift rakamlı olarak belirlemelidir. Tek rakamlı olarak belirlenmişse bu tür ölçekler için geliştirilmiş düzelteme formülü uygulanır.

Test maddeleri kolaydan zora doğru sıralanmışsa yarıya bölme işlemi testin tek rakamlı ve çift rakamlı maddeleri seçilerek yapılır. Her iki yarıdaki maddelerin homojen olduğu varsayılır ve her iki bölümün toplam puanları arasında Pearson korelasyon analizi yapılır. Ancak elde edilen sonuç iki yarı arasındaki güvenilirlik katsayısıdır. Bu rakamın testin / ölçeğin bütününü kapsaması için Spearman-Brown formülü kullanılarak düzeltme yapılması gerekir.[a] Düzeltme formülü Eşitlik 3-12'deki gibidir:[51]

$$\rho^2_{xx'} = r_{sb} = \frac{k \cdot r_{xx'}}{1+(k-1)r_{xx'}}. \qquad (3\text{-}12)$$

r_{sb} = Yarıya bölme güvenilirlik katsayısı.
k = Testin kaç katı olduğu (yarıya bölmede 2).
$r_{xx'}$ = Testin/ölçeğin iki yarısı arasındaki korelasyon katsayısı.

Spearman – Brown formülünün kullanılması sonucunda güvenilirlik katsayısı daha yüksek çıkar, ancak çıkan bu değer bir projeksiyondur. İstatistiksel analiz programı SPSS'te yarıya bölme yöntemi her iki yarıdaki test maddelerinin sayılarının ve zorluk derecelerinin eşit olduğu varsayımı altında yapılır. Eğer testin iki yarısı arasındaki madde sayısı eşit değilse Eşitlik 3-13'teki Spearman – Brown formülü kullanılır.[52]

---

[a] Spearman – Brown formülünün düşük korelasyon rakamlarını yükseltmek için kullanılması doğru değildir. Değişik kaynaklarda farklı sayıdaki ölçek maddelerinin güvenilirlik rakamlarını nasıl etkilediğini göstermek için yapılan uygulamalar eğitim amaçlıdır.

$$R = 2\,[S^2x - (S^2y_1 + S^2y_2)]\ 1/\ S^2x\ . \qquad (3\text{-}13)$$

Formülde S^2y_1 ve S^2y_2 testin/ölçeğin iki yarısına ait puanların varyansı ve S^2x ise testin bütün maddelerine ilişkin varyans değeridir. Spearman – Brown formülünün güvenilirlik katsayısına dayanan bir diğer hesaplama şekli Eşitlik 3-14'teki gibidir:

$$r_{sb} = 2r_{xy} / (1 + r_{xy})\ . \qquad (3\text{-}14)$$

r_{xy} = Birinci yarı ile ikinci yarı arasındaki korelasyon katsayısı.

Simge olarak kullanılan küçük r testin iki yarısı arasındaki korelasyonu gösterirken, büyük R_{sb} testin tamamına ait güvenilirlik katsayısına işaret eder. Testin tamamına ait güvenirlik, bazı formüllerde $r_{S\text{-}B}$ simgesiyle gösterilmiştir.

Yarıya bölme işlemi çok farklı şekillerde yapılabileceğinden çift-tek rakamlı maddelerin korelasyonu ile birinci-ikinci yarı veya ilk-son çeyrek ve ortadaki %50'lik dilime ait maddeler arasındaki korelasyon katsayıları farklı çıkabilir. Araştırmacı birden fazla yarıya bölme yöntemi uygulamışsa elde ettiği farklı güvenilirlik katsayılarının ortalamasını temel alabilir.

Yarıya bölme güvenilirliği ile alfa güvenilirlik rakamlarının birbirine benzer çıktığı bildirilmiş; farkın sadece virgülden sonra üçüncü hane rakamlarında görüldüğü ve bunun da önemsiz olduğu ifade edilmiştir.[53]

Yarıya bölme yöntemi zamana karşı yapılan ve hızın önemli olduğu bilgiyi ölçme testlerinde ve hız ögesi içeren diğer psikometrik testlerde uygulanmaz. Çünkü hız testlerinde[a] insanların testin birinci yarısında kullandıkları zaman ile ikinci yarısı için kullandıkları zaman dengeli olmadığından testin iki yarısı arasındaki korelasyon rakamları büyük ölçüde hata varyansı içerir.

[a] Hız testi, kişilerin başarılarının bütünüyle testin tamamlanma süresine göre belirlendiği nispeten kolay sorulardan oluşmuş ölçüm araçlardır. Testin tamamlama süresi çok kısa olduğundan hiçbir kişi testi zamanıda tamamlayamaz. Güç testlerinde, maddelerin %90'ının ne kadarlık bir süre içinde cevaplandığı temel alınırken hız testlerinde bu oran daha düşüktür.

Yarıya bölme güvenilirlik katsayısı en az ,70 olmalıdır.[54] Yarıya bölme güvenilirlik katsayısının büyüklüğü testin uzunluğu ile yakından ilgilidir. Az sayıda maddeden oluşan ölçeklerde (örneğin 8 madde gibi) sağlıklı sonuçlar alınmaz. Kline (1993) göre bir test/ölçek 10 maddeden daha az ise muhtemelen güvenilir değildir. Testte en az 10 madde bulunmalıdır.[55]

Homojen test maddelerinden oluşan bir test/ölçek için alfa iç tutarlılık analizi yapılmışsa ayrıca yarıya bölme hesaplaması yapmaya gerek yoktur. Çünkü alfa, tüm muhtemel yarıya bölme uygulamalarının bileşkesidir. Ancak araştırmacı olguyu aynı zamanda değişik iç tutarlılık analizi sonuçlarıyla okuyucularına yansıtmak istiyorsa raporunda yarıya bölme güvenilirlik analizi sonuçlarını da verebilir.

Yarıya bölme güvenilirliğini istatistiksel analiz programı SPSS'te yapmak isteyen araştırmacılar bunun için Scale, Reliability, Split-Half (Model) mönülerini kullanabilirler.

Yarıya bölme yöntemi ve Rulon formülü. Yarıya bölme konusunda günümüzde daha az uygulanan bir diğer yöntem Rulon'un (1939) güvenilirlik katsayısıdır. Bu yöntemde test önce iki yarıya bölünür. Yarı test puanları arasındaki farklılık bulunur. Daha sonra fark puanlarının varyansı ve akabinde toplam puanların varyansı bulunarak fark puanlarının varyansı toplam puanların varyansına bölünür (*bk.*, Tablo 3-7). Dizideki verilerde eğer uç değerler varsa hesaplama iki kez yapılır. Birincisinde bu değerler hesaplamaya alınırken ikincisinde çıkarılır. Buna göre güvenilirlik katsayısı Eşitlik 3-15'teki formülle hesaplanır.

$$\rho_{xx'} = 1 - \frac{\sigma^2_{fark}}{\sigma^2_x} \qquad \sigma^2_{fark} = \frac{\Sigma(x-\mu)^2}{N} \tag{3-15}$$

Formülde $\rho_{xx'}$ iki yarı arasındaki korelasyonu, σ^2_{fark} farklılık puanların varyansını, σ^2_x toplam puanların varyansını gösterir.

Tablo 3-7. Rulon Formülünün Hesaplaması

Katılımcı	Maddeler						1. yarı	2. yarı	Fark	Toplam
	1	2	3	4	5	6				
1	4	3	2	2	3	3	9	8	1	17
2	5	4	3	4	4	3	12	11	1	23
3	2	3	2	3	3	4	7	10	3	17
4	5	5	4	2	5	4	14	11	3	25
5	2	3	2	2	1	2	7	5	2	12
Toplam									10	94
Ortalama									2	31,33
Farklılık puanlarının varyansı									0,08	
Toplam puanların varyansı										21,76

$$\sigma^2 = 1 - \frac{0,8}{21,7} = 1 - 0,03 = 0,97 \; . \qquad (3\text{-}16)$$

Guttman ve yarıya bölme formülü. Guttman formülünün hesaplanabilmesi için maddeler tek ve çift olmak üzere iki gruba ayrılır. Guttman formülünde iki varsayım söz konusudur: (a) Birinci yarının güvenilirliği ile ikinci yarının güvenilirliği eşit değildir. (b) Birinci yarının varyansı ile ikinci yarının varyansı da eşit değildir. Her iki yarıdaki güvenilirlik ve varyansların eşit olmaması nedeniyle Guttman güvenilirlik katsayısı daha küçük çıkar.[56] Bu varsayımlara dayalı olarak hesaplama için Eşitlik 3-17'deki formül kullanılır.

$$_rG = 2\,(S2_t - S2_t1 - S2_t2) \,/\, S2_t \; . \qquad (3\text{-}17)$$

S^2_t = Ölçeğin toplam puanlarının varyansı.
S^2_{t1} = Ölçeğin birinci yarısına ait puanların varyansı.
S^2_{t2} = Ölçeğin ikinci yarısına ait puanların varyansı.

Spearman – Brown kehanet formülü. Spearman – Brown (S-B) formülü testteki madde sayısının aynı ölçüde (bir kat daha) artmasıyla birlikte güvenilirliğin ne olacağını kestirmek için kullanılır. Ancak artan bölümdeki maddelerin önceki maddelere paralel olması gerekir. Diğer bir deyişle testteki alt boyutlar da aynı kavramsal yapıyı ölçüyor olmalıdır. Bu varsayımın karşılanmadığı durumda S-B formülü uygulanmaz.

Loevinger H. Cronbach alfa yöntemine karşı başka bir yaklaşım J.A. Loevinger (1944) tarafından geliştirilen *türdeşlik indeksi*dir. Loevinger *H* katsayısı hiyerarşik madde yapısına sahip Mokken ölçeklerinde kullanılır. Mokken ölçekleri de Guttman ölçekleri gibi hiyerarşik bir yapıya sahiptir. Fakat Mokken ölçekleri deterministik değil probabilistik sonuçlar verir ve bu nedenle ölçeğin güvenilirliği "yeniden üretilebilirlik" katsayısı ile değil "Loevinger *H* katsayısı" ile belirlenir. Türdeşliği veya *ölçeklenme* olgusunu tanımlayan Loevinger'in *H* katsayısı[a] klasik güvenilirlik katsayısı olan rho'ya alfa değerinden daha yakın çıkar.[57] Bazı araştırmacıların yüksek güvenilirlik değeri verdiği şeklinde eleştiriler getirdikleri Loevinger'in *H* değerini hesaplamak için bu amaçla yazılmış istatistiksel analiz programlarından yararlanılabilir.[b]

---

[a] Loevinger'in *H* katsayısında, *H* harfi türdeşlik anlamına gelen İngilizce *homogeneity* kelimesinin kısaltmasıdır.

[b] Bu programlardan biri MSP (Mokken Scaling for Polychotomies) nonparametrik madde yanıt teorisi ölçekleme programıdır. W. Molenaar, P. Debets, K. Sijtsma ve B.T. Hemker tarafından yazılan bu program esas olarak Mokken ölçeklerinin analizi için geliştirilmiştir. Yazılımın 100 kadar çok dereceli ölçeği maksimum 10 sıralı ölçek kategorisi içinde ele alıp analiz yapabildiği bildirilmiştir. Mokken ölçeğinde Rasch ölçeklerinin tersine katılımcıların yanıtladıkları puanlar doğrudan programa girilebilmektedir. Ölçeğin geliştiricileri, cevaplayıcıların işaretlemeleriyle ortaya koydukları yanıt modellerinin mükemmel Guttman ölçeği koşullarını ihlal etmediğini ileri sürmüşlerdir. Araştırmacı pratik nedenlerle yanıt modellerindeki yanlışlanabilir işaretlemeleri analizden çıkarabilir veya analize dahil edebilir. Ölçeğin geliştiricileri iki dereceli ölçeklere göre çok dereceli ölçeklerde toplam puandan hareket etmenin gizli yapı parametresini, (θ) ortaya çıkarma konusunda zayıf kalacağını, ancak yine de cevaplar gizli özellik çerçevesinde işaretlenmişse ciddi bir sorunla karşılaşılmayacağını belirtmişlerdir. Bu konuda daha fazla bilgi için *bk.*, "Appendix 1 [Ek-1]" <http://polmeth.wustl.edu/pa/vanschu ur.web.doc> (24.05.2004). Yazılım, ProGAMMA isimli şirket tarafından pazarlanmaktadır. İlgi duyan okurlar İnternetten ücretsiz olarak yüklenebilen gösteri sürümünü inceleyebilirler.

Loevinger'in H katsayısı geçişkenlik[a] (transitivity) ölçüsünü verir. Bu yaklaşımda her bir maddenin doğru cevaplama oranı (p) ile türdeşlik katsayıları (H) hesaplanır. Loevinger'in H katsayısı üç düzeyde elde edilir:

1. H simgesi ile bütün ölçeğin güvenilirliği hesaplanır.
2. H_i ile, i maddesinin güvenilirliği veya yeniden üretilebilirlik katsayısı hesaplanır.
3. H_{ij} ise, i ve j maddelerinden oluşan madde çiftinin güvenilirliğini verir. Araştırmacı isterse güvenilirlik analizlerinde madde çiftlerinin güvenilirliğinden hareket edebilir.

H(i)'nin ölçekte bir madde olarak kalabilmesi için değerinin ,30'dan yüksek olması gerekir. Aynı şekilde bir ölçeğin kabul edilebilir bir ölçüm aracı olması için H değeri yine ,30'dan yüksek olmalıdır.[58] Mokken ve Lewis (1982) H değeri 0,50'den büyük ise aracın güçlü bir ölçek olduğunu, 0,40 ilâ 0,50 arasındaki bir değere sahip aracın orta derecede ölçeklenmiş olduğunu, 0,30 ilâ 0,40 arasındaki değerin ise zayıf bir ölçeklenme olgusunu gösterdiğini belirtmişlerdir.[59] Loevinger H katsayısı, Eşitlik 3-18 ve Eşitlik 3-19 formülleri çerçevesinde hesaplanır:[60]

$$H_{ij} = \frac{Cov(X_i, X_j)}{Cov_{\max}(X_i, X_j)}, \quad H_j = \frac{\sum_{i \neq j} Cov(X_i, X_j)}{\sum_{i \neq j} Cov_{\max}(X_i, X_j)}, \qquad (3\text{-}18)$$

$$H = \frac{\sum \sum_{i \neq j} Cov(X_i, X_j)}{\sum \sum_{i \neq j} Cov_{\max}(X_i, X_j)}. \qquad (3\text{-}19)$$

[a] Geçişkenlik. Yanıtlayıcılar "güç" bir soruya eğer doğru (veya olumlu) yanıt vermişlerse ondan önce gelen daha "kolay" soruların hepsine yine olumlu yanıt vermiş olmalıdırlar. Bu koşul sağlanmışsa maddeler "ölçeklenmiştir" ve "geçişkenlik sağlanmıştır" denir. Geçişkenlik matematiksel olarak maddelerin birbirini kapsama derecesidir. Eğer $X \geq Y$ ve $Y \geq Z$ ise o zaman, $X \geq Z$ olur. Geçişkenlik özelliği ile maddelerin dereceleri yerine kendileri sıralı ölçek niteliğini kazanır.

Loevinger *H* katsayısı başlangıçta iki dereceli maddeler için geliştirilmişken daha sonraları hesaplama formülleri çok dereceli maddelerin de güvenilirliği hesaplayacak şekilde genişletilmiştir.

Faktör analizi. İç tutarlılık analizinin bir diğer şekli, keşfedici faktör analizi sonucunda hesaplanan faktör ağırlıklarından (yüklerinden) yararlanmaktır. Maddeler eğer tek bir faktörü ölçüyorsa söz konusu faktörü temsil etme ağırlığı ,40 ilâ ,80 arasında değişebilir. Örneğin, bir ölçekte faktör analizi sonucunda iki bağımsız faktör ortaya çıkmışsa maddelerin yaklaşık yarısı birinci faktörde en az ,40 faktör yükü ağırlığına sahip olmalı ve aynı maddelerin ikinci faktördeki ağırlıkları ise sıfıra yakın veya negatif değerler olmalıdır. Öte yandan, diğer maddeler birinci faktörle ilgili olmamalı büyük ölçüde ikinci faktöre ait ağırlık değerlerine veya faktör yüklerine sahip bulunmalıdır.[61]

Hoyt alfa katsayısı. İlgili literatürde daha az rastlanılan ve C. Hoyt (1941) tarafından varyans analizi yöntemine dayalı olarak geliştirilen güvenilirlik analizi yöntemi, maddelerin iç tutarlılığını ölçer. Hoyt alfa değeri Cronbach alfa ile aynı anlamdadır. Hoyt yöntemi, KR-20 ve KR-21 alfa katsayısı tekniklerinin yanında üçüncü bir alfa katsayısı verir. Hoyt yönteminin daha çok bilgisayar programlamasını gerektiren kapsamlı test uygulamalarında kullanıldığı bildirilmiştir.[62] Hoyt yöntemine göre alfa değerini hesaplamak için tek yönlü varyans analizinden yararlanılır.

Varyans analizinde önce grup içi ve daha sonra gruplar arasındaki değerlerin kareler ortalaması bulunur. Elde edilen bu değerler Eşitlik 3-2'deki formüle uygun olarak yerleştirilir ve daha sonra 1'den çıkarılır. Böylece alfa değeri hesaplanmış olur.[a]

[a] Hoyt varyans analizi konusunda daha fazla bilgi için *bk.*, Hoyt, C. (1941). Test Reliability Estimated by Analysis of Variance [Varyans Analizi Yöntemiyle Güvenilirlik Tahmini]," *Psychometrika*, 6, 153-160.

$$\alpha = 1 - \frac{KO_{içindeki}}{KO_{arasındaki}}, \quad veya \quad \alpha = \frac{KO_{arasındaki} - KO_{içindeki}}{KO_{arasındaki}}.$$

$$\alpha = \frac{MS_{Between} - MS_{Within\ (error)}}{MS_{Between}}. \qquad (3\text{-}20)$$

$KO_{içindeki}$ = Grup içindeki değerlerin kareler ortalaması. Literatürde aynı zamanda "hata karelerinin ortalaması olarak da bilinir (MS_{error}).

$KO_{arasındaki}$ = Gruplar (kişiler) arasındaki değerlerin kareler ortalaması.

İçindeki kareler ortalamasını, 10 kişiye uygulanan dört maddeli ve beş dereceli bir ölçek üzerinde örnek vererek açıklayabiliriz. İçindeki kareler ortalaması, dört maddenin her biri için ayrı ayrı hesaplanırken arasındaki kareler ortalaması için kişiler temel alınır. Veriler istatistiksel analiz yazılımı SPSS'e *yığışımlı veri matrisi* olarak iki sütun halinde girilir. Bağımlı değişken olarak birinci sütuna kişilerin beş dereceli ölçek üzerinde maddelere verdikleri puanlar, ikinci sütuna ise bağımsız değişken olarak 1, 2, 3 ... *n* şeklinde kişi numaraları girilir. Çözümleme sonucunda Tablo 3-8'deki varyans analizi çıktısı elde edilir.

Tablo 3-8. Hoyt Alfa Değeri İçin Varyans Analizi Çıktısı

	SS (sum of squares) Kareler Ortalaması	*df* (Degrees of freedom) Serbestlik derecesi)	*MS* (mean Squarı) Kareler ortalaması	(F ratio) F oranı
Kişiler arası karşılaştırma (arasındaki)				
Hata (kişinin kendi içindeki)				
Toplam				

Hoyt varyans analizinde araştırmacı F ve p olasılık değerlerini raporlayabileceği gibi esas olarak bu tablodaki MS değerlerini kullanarak Hoyt güvenilirliğini belirler ve bu rakamı raporlar.

İç tutarlılığı ölçmenin kolay yöntemi. Hopkins ve Stanley (1981) iç tutarlılığı daha kolay yönden hesaplamaya imkan sağlayan bir yöntem geliştirmişlerdir. Bu yöntemde ölçümün standart hatası ölçekteki madde sayısına bölünür. Eğer oran %5'ten daha küçük çıkmışsa test veya ölçeğin yüksek iç tutarlılığa sahip olduğu söylenir.

Genellenebilirlik kuramı. Test puanlarının tutarlılığını ölçmek için kullanılabilecek bir diğer yöntem *genellenebilirlik* (*G*) yaklaşımıdır. Cronbach ve arkadaşları tarafından 1972 yılında geliştirilen bu yaklaşımda test/ölçek sonuçlarının ne ölçüde genellenebileceği ve test sonuçlarının ne ölçüde sistematik hata kaynaklarından etkilenebilir olduğu anlaşılmaya çalışılır. Amaç sadece sistematik hata kaynaklarını belirlemek değil aynı zamanda bu hata kaynaklarını kontrol ederek daha sonra verilecek olan pratik kararlarda bu hata kaynaklarını göz önünde bulundurmaktır. Genellenebilirlik kuramında hata kaynakları "yön" veya "yüzey" terimiyle ifade edilir. Bir ölçüm işlemi üç boyutlu bir küpe benzetilebilirse küpün her bir yüzeyinde yer alabilecek alt değişkenler hataları simgeler. Bir test veya ölçeğin güvenilirliği puanların hangi koşullarda geçerli olduğunun saptanmasına bağlıdır. Bunun için araştırmacı aşağıdaki aşamalardan geçerek güvenilirlik analizini yapar.

Birinci aşmada, hata faktörlerinin kaç "yüzeyde" araştırılacağı belirlenir: (a) değerlendiriciler arasındaki hatalar, (b) maddeler arasındaki hatalar, (c) farklı zamanlarda yapılan ölçümlerden kaynaklanan hatalar, (ç) uygulama şartlarının farklı olmasından kaynaklanan hatalar gibi. Basit ölçüm tasarımlarında hata, sadece tek bir yüzeyde araştırılır.

İkinci aşamada, analiz tasarımı belirlenir. Analiz tasarımında önce hangi *tesadüfi*[a] veya *sabit* değişkenlerin analize alınacağı saptanır. Daha sonra geliştirilen modele göre veriler arasındaki ilişkilerin *çapraz tasarım* veya *yuvalanmış tasarımdan* hangisine uygun olduğuna bakılır. Çapraz tasarımda bir yüzeydeki birimler diğer yüzeydeki tüm birimlerle ilgili ise bu yola başvurulur. Örneğin; *değerlendiriciler* x *kişiler* x *maddeler* (d x k x m) tüm düzeylerde hepsi birbiriyle ilgili olarak birlikte ele alınabilir. Yuvalanmış

[a] Tesadüfi değişken: Verileri tesadüfi gözlem sonuçlarına dayanan değişkenler. Sabit değişken: Verileri araştırmacının tasarımına göre düzenlenen değişkenler.

tasarımda ise bir yüzeydeki birimler diğer yüzeydeki birimlerden hepsiyle değil, sadece bazı birimlerle ilgilidir. Örneğin, *değerlendiriciler* x *iş*. tasarımında değerlendiriciler yüzeyi koşullar yüzeyinin sadece iş düzeyi ile ilişkilendirilmiştir (*d:i*).

Üçüncü aşamada araştırmacı, kişiler arasındaki karşılaştırmaları ya kriter bir puanı veya belirli bir gruba ait norm değerini temel alarak yapar. Kriter değer sabit bir rakamdır. Bir gruba ait puanlar arasındaki karşılaştırmalarda ortaya çıkan değişkenlik, *nispî hata* olarak isimlendirilir. Kriter değere göre yapılan karşılaştırmalardaki değişkenlik ise *mutlak hata* olarak isimlendirilir.[63]

Bilim adamı, tespit ettiği potansiyel hata kaynakları ile ilgili olarak her bir yüzeydeki varyansın büyüklüğünü saptamak için *varyans bileşenleri analizi* yöntemini uygular (Bu konuda daha fazla bilgi için *bk.*, "Varyans Analizi ve Güvenilirlik"). Böylece, birden fazla hata kaynağı tek bir analiz ile test edilmiş olur. Güvenilirlik nispî bir kavramdır ve büyük ölçüde genelleme yapılacak *evren* ile ilgilidir. Her bir evrende hata kaynakları da değişeceğinden değişik ana kütlelerde genellemeye dayanan güvenilirlik analizlerini yeniden yapmak gerekir. Genellenebilirlik kuramında ölçülmeye çalışılan özelliğin değişmediği gibi gizli bir varsayım söz konusudur. Genellenebilirlik katsayısı diğer güvenilirlik katsayıları gibi yorumlanır. Bu yöntemde her bir yüzeydeki varyansı belirlemek için TYVA bileşenleri çıktısındaki *beklenen ortalamaların karesi* sütunundaki değerler dikkate alınır.

İstikrarlılık Analizleri

Bir testin veya ölçeğin istikrarlılığı farklı zamanlarda yapılan ölçüm sonuçlarının benzer çıkması ile belli olur. İstikrarlılık test-yeniden test yöntemi ile belirlenir. Test-yeniden test yönteminin her tür test ve ölçeklerde uygulanması zordur. Daha çok standardizasyonu yapılmak istenen test ve ölçeklerde uygulanır. Bilimsel bir kuramı test etmek amacıyla geliştirilen test ve ölçeklerde bu yöntemin uygulanması zaman ve maliyet öğesi açısından çoğunlukla mümkün değildir. Testlerin standardizasyonu amacıyla uygulandığında ise norm grubunu oluşturan kişilerin tamamına veya ana kütlenin bütününe değil, örnekleme yapılarak seçilen bölümüne uygulanır. Öte yandan test-yeniden test yöntemi güvenilirliği belirleyen tek ölçüt olarak kullanılmamalıdır.

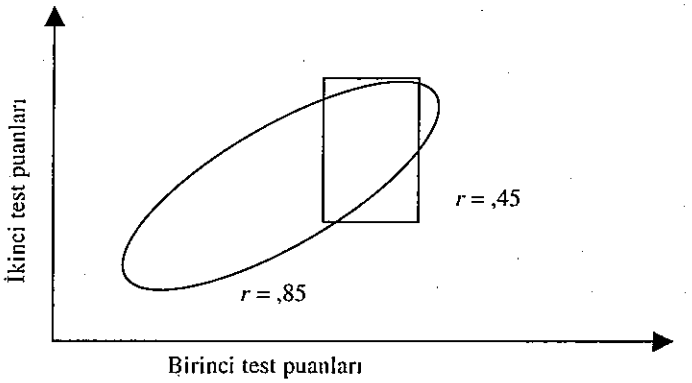
Zaman aralığı. Test-yeniden test yöntemi, kullanılan ölçeğin/testin duruma göre uzun veya kısa zaman aralıklarında yeniden sınanmasıdır. Test, aynı veya benzer özellikteki kişilere farklı iki zaman diliminde uygulanır. Test-yeniden test yönteminin uzun zaman aralığında uygulanabilmesi için ölçülen özelliğin oldukça istikrarlı bir yapıya sahip olması gerekir. Bu nedenle her türlü test/ölçek için standart zaman dilimi uygulanmaz. Bireye ait kişilik özellikleri (trait) oldukça kalıcı iken tutumlar (görüşler, davranışlar) değişkendir. Bir araştırmacının *Avrupa Birliğine Karşı Tutumlar Ölçeği* için üç aylık bir süreden sonra test-yeniden test yöntemini uygulaması halinde güvenilirlik rakamları düşük çıkabilir. Test-yeniden test yönteminde zaman dilimi *kısa zaman* veya *uzun zaman* dilimleri şeklinde belirlenir.

Kısa zaman aralığında yapılan ölçümler. Belirli nitelikteki testlerin güvenilirliği kısa zaman aralığında yapılan ölçümlerle belirlenir. Örneğin parmak becerisini küçük pimleri kullanarak ölçen bir testte test-yeniden test uygulaması iki dakika gibi bir zaman aralığından sonra yapılabilir. Bu tür testler bir günde birkaç defa tekrarlanır. Fiziksel dayanıklılık, alet kullanma, güç ölçümü gibi testlerde en az 7 günlük bir sürenin geçmesi ve bu süre içinde kişinin gücünü yeniden toplaması beklenir. Öte yandan süre çok uzun olursa bu kez olgunlaşma, öğrenme, çalışma gibi diğer faktörlerin ölçüm sonuçlarını etkileme tehlikesi ortaya çıkar. Fiziksel performans ölçümlerinin güvenilirliğini test etmek için birinci ölçümde araştırmaya katılan kişilerin hepsine yeniden test uygulaması yapmaya gerek yoktur. Önceki ölçüme katılan kişilerin %25 ilâ %50'sine yeniden test uygulanması güvenilirlik katsayısını hesaplamak için yeterli olabilir.[64] Ancak bu uygulama genel bir kuralı belirtmez. Tutum ölçeklerinde ve diğer ölçümlerde yeniden test uygulamasına önceki ölçüme katılan kişilerden özellikli olan sadece bir bölümünün alınması (tesadüfî örnekleme yöntemine başvurulmaması) test-yeniden test güvenilirlik katsayısının ranj kısıtlaması nedeniyle düşük çıkmasına neden olur (*bk.* Şekil 3-2).

Uzun zaman aralığında yapılan ölçümler. Kişilik, zeka ve yetenek gibi testlerde güvenilirliği daha doğru bir şekilde saptayabilmek için aradan iki üç ay gibi bir zaman geçmesi gerekir. Kısa zaman aralıklarında ölçüm yapılmış ve ölçüm sonuçlarının güvenilirliği de yüksek çıkmışsa bu rakamlar sağlıklı değildir. Ölçüm konusuna göre uygun süre belirlenmelidir. Kanaat ve görüş belirleme araştırmaları için bir yıl çok uzun bir süre iken fizyolo-

jik bir ölçüm veya kişilik ölçümleri için uygun olarak değerlendirilebilir. Kişilik testinde ölçeğin tutarlılığı ilk ölçümden altı ay sonra ikinci kez ve bir yıl sonra üçüncü kez tekrar sınanır. Kline (1993), psikolojik testlerde test-yeniden test ölçümleri için arada en az üç aylık bir sürenin bulunması gerektiğini belirtmiştir.[65] Test-yeniden test yöntemi üç ay veya altı ay gibi oldukça uzun bir süre sonunda tekrarlanmış ve güvenilir sonuçlar elde edilmişse bu rakamlar güvenilirlik katsayısı olarak değil, *tutarlılık katsayısı* olarak isimlendirilir. İstikrarlılık tek bir testin puanlarından hareket edilerek veya paralel testleri farklı zamanlarda uygulayarak belirlenir.

İkiden fazla zaman diliminde yapılan ölçümler. Test-yeniden test uygulamasında bilim adamı pilot araştırma yaptığı örnek kütlede ikiden fazla ölçüm yapabilir. İlk ölçümü iki günlük bir aradan sonra, ikinci ölçümü bir haftalık bir aradan sonra ve üçüncü ölçümü ise bir aylık bir aradan sonra yapabilir. Bu uygulamanın sonunda korelasyon rakamları karşılaştırılarak önemli ölçüde değişiklik ortaya çıkıp çıkmadığına bakılır.



Şekil 3-2. Test-yeniden test uygulamasının örneklemin sadece belirli bir bölümünde sınanması (ranj kısıtlaması sorunu).

Kaynak. "Reliability [Güvenilirlik]," t.y., <http://www.onid.orst.edu/~uttlb/psy470/reliability.pdf> (28.12.2002).

İkiden fazla zaman diliminde yapılan ölçüm sonuçları %95 güven aralığında ve ±1,96 standart sapma aralığında kalıyorsa sonuçlar arasında önemli bir farklılık olmadığına karar verilir. Test-yeniden test karşılaştır-

maları sadece korelasyon rakamlarına bakılarak değil, aynı zamanda tanımlayıcı istatistikî analiz sonuçları incelenerek de değerlendirmeye alınır. Bu çerçevede;

1. birinci, ikinci ve üçüncü test sonuçlarının ortalamaları ve standart sapmaları incelenir,
2. katılımcıların her üç testte benzer yüzdelik dilimlerde yer alıp almadıklarına bakılır,
3. katılımcıların her üç testte benzer puanlar alıp almadıkları incelenir,
4. katılımcıların her üç testte benzer z veya T puanlarına sahip olup olmadıklarına bakılır ve,
5. Fisher z'puanlarından hareket edilerek birleşik güvenilirlik katsayılarının anlamlı olup olmadığı incelenir.

Böylece normal dağılım eğrisinde, kişilerin yerlerinde önemli bir değişikliğin ortaya çıkma durumu araştırılır. Kişilerin yerlerinde önemli bir değişiklik yoksa testin güvenilir olduğundan söz edilir. Test-yeniden test güvenilirlik analizine karar vermek için belirli koşulların oluşması veya oluşturulması gerekir. İki test uygulanacağının katılımcılara önceden bildirilmesi, zaman planlaması yapılması, ikinci toplantı veya uygulama tarihinin belirlenmesi ve örneklem büyüklüğünün saptanması bunlar arasındadır. Araştırmacı "yapmış olmak için" bu yönteme başvurmamalıdır.

Testin paralel gruplarda uygulanması. Testin farklı iki zaman diliminde yapılması konusunda güçlük varsa bilim adamı test-yeniden test uygulaması yerine geçmek üzere *paralel grup* uygulaması yöntemine başvurabilir. Ancak bu uygulamada ölçülmek istenen özellik açısından grupların birbirine tam paralel olması gerekir. Örneğin, öğrencilerin *zihinsel sertliklerini* belirlemeyi amaçlayan bir test, aynı sınıfın A ve B şubelerindeki öğrencilere birlikte uygulanabilir. Bu sınıflarda cinsiyet ve yaş dağılımı ile diğer kişilik özellikleri açısından özel bir kümelenme olmadığı sürece sonuçlar test-yeniden test güvenilirlik analizlerine tâbi tutulabilir.

Yetenek ve beceri testleri de aynı ana kütleden seçilmiş paralel nitelikteki farklı örneklemlerde uygulanabilir. Örneğin, bir sayısal yetenek testi en az beş farklı örneklemde uygulanmış ve güvenilirlik katsayıları ,80 ilâ ,90 arasında çıkmışsa testin istikrarlılık güvenilirliğine sahip olduğu söylenir. Test, en az üç örneklemde uygulanmış ve güvenilirlik katsayıları orta-

lama ,60 ilâ ,70 arasında iken bir örneklemin korelasyon katsayısı düşük çıkmışsa bu testi ihtiyatla uygulamak gerekir. Farklı üç örneklemde uygulanan test güvenilirlik katsayıları ,60'nin altına düşmüşse bu testler güvenilir olarak kabul edilemez.

Paralel formların farklı zaman dilimlerinde uygulanması. Araştırmacı güvenilirlik analizleri için paralel formlar yönteminden yararlanmayı düşünmüşse t_1 zamanında kendi geliştirdiği formu ve daha sonra t_2 zamanında cevaplayıcılara alternatif formu vererek iki forma ait toplam değerler arasında korelasyon analizi yöntemini uygular. Paralel formların farklı iki zamanda sınanması test-yeniden test uygulamasına benzer.

Gözlemci içi güvenilirlik. Gözlemci içi güvenilirlik, bir tür "test-yeniden test" uygulamasıdır. Bu yöntemde eğer birden fazla gözlemci bulunamamışsa aynı gözlemciye birden fazla zamanda değerlendirme yaptırılarak ölçüm sonuçlarının tutarlılığı araştırılır.

Test-yeniden test uygulamasının yapılma amacının belirlenmesi. Test-yeniden test uygulaması belirli koşullarda gerçekleştirilir. Bilim adamı her tür test ve ölçek için test-yeniden test yöntemini uygulamak zorunda değildir. Yeniden test uygulaması, güvenilirlik analizinden *bağımsız* olarak aşağıdaki koşullarda gerçekleştirilir:

Birinci ölçüm sağlıklı yapılamamışsa. Yeniden test, eğer birinci ölçümde cevaplayıcılar yeterince ilgili davranmamışlarsa, ilgileri dağılmışsa, bazı değişkenler kontrol altına alınmamışsa uygulanır. Araştırmacı, sadece beklediği puanları alamadığı için yeniden test uygulamasına girişmemelidir.

Araştırmacı iki test arasında bir kazanç elde etmek istiyorsa. Araştırmacı yeniden test uygulamasını eğer iki test arasında bir takım uygulamalar, müdahaleler yapmışsa ve bu müdahalelerin sonucunu görmek istiyorsa uygulamalıdır.

İlk ölçüm başka birisi tarafından yapılmışsa. Araştırmacı başka birisi tarafından yapılan ölçümlerin doğruluğunu ve güvenilirliğini test etmek için yeniden test uygulaması içine girebilir.

Literatürdeki uygulamalar. Literatürde benzeri kavramsal yapıların güvenilirlik analizlerinde test-yeniden test yöntemi uygulanmışsa bu yöntem tercih edilebilir. Böyle bir durumda yeniden test verme açısından ortalama sürenin ne olduğuna ilişkin literatür araştırması yapılmasında yarar vardır.

Testin içeriği. Yeniden test uygulamasına testin içeriğine bakarak karar vermek daha uygundur. Psikolojik yapıları / özellikleri ortaya çıkarmayı hedefleyen bir test / ölçek kısa zaman aralığında uygulanmamalıdır. Çünkü aradan fazla bir zaman geçmediği sürece test sonuçları benzer çıkacaktır.

Klinik karar verme durumunda. Araştırmacı test / ölçek sonuçlarına bakarak klinik kararlar verme durumunda ise, yeniden test uygulamasını düşünebilir.

Yeniden test uygulama olanağının bulunması. Hastanelerde, marketlerde ve okullarda çoğunlukla yeniden test yapma olanağı vardır. Ancak, bir çok araştırmada bilim adamları yeniden test yapma olanağı bulamazlar. Cevaplayıcılar çok dağınık yerlerdedirler ve yeniden test yapmanın maliyeti çok yüksektir. Bu gibi durumlarda yeniden test yapma uygulamasından vazgeçilir.

Araştırmacı test-yeniden test yönetimini uygulamışsa araştırma raporunda aradan geçen sürenin uzunluğunu bildirmeli ve süreyi nasıl tespit ettiği hakkında bilgi vermelidir. Test-yeniden test değerleri çocuklarla yaşlılarda ve yetişkinlerde farklı çıkar. Yeniden test değerlerinde çocuklarda yetişme, yaşlılarda ise kayıtsızlık etkisi görülür.

Test-yeniden test yöntemi, okullarda öğrencilere uygulanan çoktan seçmeli sorular için uygun değildir. Bu tür uygulamalarda öğrenme etkisi büyük olduğundan elde edilen sonuçlar testin güvenilirliği hakkında fazla bir bilgi vermez. Daha uzun bir süre içinde uygulandığında ise öğrenme ve olgunlaşma etkisi nedeniyle yine sonuçlardaki hata varyansının artma ihtimali vardır.

Test-yeniden test yöntemi tek bir birey üzerinde uygulandığı zaman elde edilen değerlerin güvenilirlik ölçüsü olarak *değişim katsayısı* (V_k)[a] kullanılır (*bk.*, Eşitlik 3-21). Değişim katsayısı, bir dizi ölçüm sonunda elde edilen değerlere ait standart sapmanın aynı değerlerin aritmetik orta-

[a] Coefficient of variation.

lamasına olan oranıdır ve çıkan sonuç yüzde cinsinden ifade edilir. Örneğin bir atletin 100 metre koşuya ilişkin farklı zamanlarda ve yerlerde yapılan zaman ölçümleri arasındaki fark değişim katsayısı ile gösterilir. Kesinlik ifade etmesi için değişim katsayısının küçük bir yüzde değerine sahip olması gerekir.

■ Değişim katsayısı formülü.

$$Degisim\ katsayisi = \frac{\text{standart sapma}}{\text{aritmetik ortalama}}.100. \qquad (3\text{-}21)$$

Değişim katsayısının gözlemci içi değerlendirmelerde ölçüm aracının değil, ölçüm yapılan kişinin istikrarlılığını gösterdiği belirtilmiştir. Ölçümlerde değişim katsayısının %6-7 düzeyinde çıkması yüksek güvenilirlik / kesinlik ve % 12,5'in üzerinde çıkması ise düşük güvenilirlik / kesinlik derecesi olarak yorumlanır.[66]

Test-yeniden test tanımlayıcı istatistiksel analiz sonuçları. Test-yeniden test uygulamalarında korelasyon analizi sonuçlarından önce merkezî dağılım ölçüleri, standart sapma, varyans, çarpıklık ve basıklık katsayıları verilerek her iki ölçümdeki sonuçlar karşılaştırmalı olarak ele alınır.

Test-yeniden test korelasyon analizi sonuçları. Test-yeniden test korelasyon analizi daha çok Pearson ve Spearman teknikleri kullanılarak yapılır. Ancak bu yöntemlerin 15 veya daha küçük örneklem hacimlerinde güvenilirlik katsayısını olduğundan daha büyük gösterdiği belirtilmiştir. Bu nedenle küçük örneklemlerde test-yeniden test korelasyonu için *küme içi korelasyon analizi* yöntemini (intraclass correlation analysis) uygulamak daha doğrudur.[67]

Test-yeniden test korelasyon katsayısı. Test-yeniden test korelasyon katsayısı en az ,80 olmalıdır.[68] Bazı bilim adamları ,70 güvenilirlik katsayısının da yeterli olabileceğini belirtmişlerdir. Uzun zaman diliminde sınanan test-yeniden test korelasyon katsayısı, araya giren zaman faktörü nedeniyle yarıya bölme yöntemine göre daha düşük korelasyon katsayıları verecektir. Ayrıca test-yeniden test korelasyon katsayıları paralel formlar güvenilirlik katsayısına göre daha düşük değerliklidir. Bilim adamı korelasyon katsayısının karesini alarak iki test arasındaki ortak varyansı göre-

bilir. Böyle bir durumda r^2 simgesinden yararlanılır. Güvenilirlik pek çok yazar tarafından ortak varyans olarak tanımlanmıştır.

Test-yeniden test uygulamasının dezavantajları. Ölçümde test-yeniden test uygulaması her zaman daha iyi sonuçlar vermez. Bu uygulamanın bazı olumsuz etkileri söz konusu olduğundan bilim adamının bu tür olumsuzlukları göz önünde bulundurması gerekir. Bu olumsuzluklar aşağıdaki gibidir:

Ezberleme / öğrenme etkisi. Test veya ölçeği alan kişiler eğer aradan belirli bir zaman geçmemişse test sorularını ezberleyebilirler. Hatta eğer kendilerine tekrar bir test uygulanacağı bildirilmişse böyle bir durumda teste hazırlık yapabilirler veya görüşlerini gözden geçirebilirler. Bu uygulama sonuçta güvenilirlik katsayısını etkiler.

Uygulama etkisi. Test veya ölçeği alan kişiler ilk uygulamayla birlikte testin nasıl uygulanacağı, hangi aşamalardan geçileceği, yanıtların nasıl verileceği konusunda bilgi sahibi olurlar. Bir sonraki uygulama, öncekine göre daha kolay gerçekleşir. İkinci uygulamada puanların daha yüksek çıkması testin istikrarlılığını değil, uygulamaya aşina olunduğunu gösteriyor olabilir.

Zaman aralığı. Zaman aralığının kısa olması halinde ezberleme ve öğrenme etkisi ortaya çıkar, uzun olması halinde ise gelişme etkisi ile karşılaşılır. Uzun zaman diliminde fiziksel, zihinsel ve duygusal olgunlaşma etkisi ortaya çıkar.

Zorluk düzeyi. Eğer, test maddeleri çok kolaysa veya çok zorsa böyle bir durumda test-yeniden test güvenilirliği yüksek çıkar. Kolay olması halinde katılımcılar testi her uyguladıklarında kolay bir şekilde çözeceklerdir. İdeal bir güvenilirlik katsayısı elde edebilmek için testin zorluk düzeyi ,50 civarında olmalıdır.

Katılımcılar. Katılımcıların niteliği test-yeniden test uygulamalarını etkiler. Test eğer klinik bulgular gösteren bir grupta uygulanmışsa ikinci kez yine aynı grupta denenmelidir. Kaldı ki bu denemede dahi katılımcıların klinik özellikleri nedeniyle aynı sonuçlar alınmayabileceğinden düşük güvenilirlik rakamlarıyla karşılaşılabilir. Oysa normal özellikler gösteren katılımcılarda güvenilirlik rakamları daha yüksek çıkar.

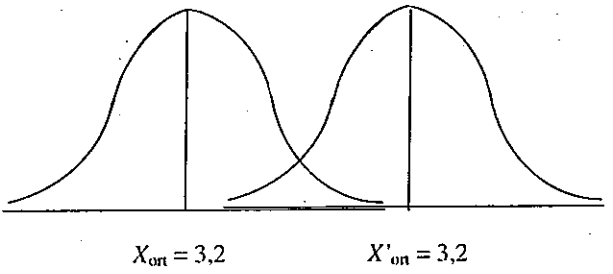
Örneklem büyüklüğü. Test-yeniden test uygulamalarında korelasyon katsayısı kadar örneklemin standart hatası da önemlidir. Örneklem hacmi büyüdükçe standart hata küçülür. Bu nedenle test-yeniden test uygulamalarının her birinde örneklem hacmi en az yüzer kişiden oluşmalıdır.[69] Ancak örneklem büyüklüğü tek başına yeterli değildir, örneklemin ana kütleyi temsil edecek şekilde seçilmiş olmasına da dikkat etmek gerekir. Literatürde güvenilirlik analizi içeren yayımlanmış makaleler üzerinde yapılan bir araştırmada bu yazıların %59'unda örneklem büyüklüklerinin 100'den az olduğu bulunmuştur. Korelasyona dayalı güvenilirlik analizlerinin öznellikten kurtulması için en az 400 kişi üzerinde sınanması önerilmiş ve geçerlilik analizlerinde bu rakamın daha da büyük olması gerektiği belirtilmiştir.[70]

Eş Değerlilik Analizleri

Bir test veya ölçümün eş değerlilik güvenilirliği iki şekilde saptanır. Bunlardan birincisi alternatif formlar, ikincisi ise gözlemciler arasındaki tutarlılık yöntemidir.

Alternatif formlar güvenilirliği. Güvenilirlik, kullanılan alternatif formlar arasında yüksek derecede ilişki bulunmasıyla belirlenir. Alternatif form ifadesi genel bir terimdir. Bu terim; *paralel formlar*, *eş değer formlar* ve *karşılaştırılabilir formlar* anlamında kullanılır.[71]

Paralel form; herhangi bir ana kütlede yapılan ölçümlerde formların / testlerin; (a) içeriklerinin farklı olması, (b) madde sayılarının aynı olması, (c) aritmetik ortalama değerlerinin aynı olması, (ç) standart sapma ve varyans değerlerinin eşit olması, (d) maddelerin güçlük derecelerinin eşit olması ve (e) bu formlar başka bir ölçüm aracıyla karşılaştırıldığında her bir testin o ölçüm aracıyla benzer korelasyon katsayılarına sahip olması anlamına gelir (*bk.*, Şekil 3-3).



Şekil 3-3. Paralel formlarda iki ölçüme ait gözlemlenen test puanlarının aritmetik ortalama ve varyans değerlerinin eşit olması.

Paralel formlardaki maddelerin ifadelendirme biçimleri, zorluk dereceleri ve ayırt etme güçleri birbirine benzer olmalıdır. Paralel testteki bir madde asıl testteki benzeri maddeyle doğrudan ilişkili olmalı ve benzer bir temayı ele almalıdır. Paralel formlar yöntemi, daha çok standardize edilmiş psikometrik testler için uygulanır. Bu tür testlerde çoğunlukla bir ölçüm aracının iki farklı versiyonu aynı zaman diliminde birden geliştirilir. Bazen de araştırmacılar daha önceden geliştirilmiş bulunan bir testi taklit ederek o testin yeni versiyonlarını oluştururlar. Tutum ölçekleri için de paralel formlar yöntemi kullanılabilir ancak söz konusu tutum ölçeklerinin her ikisinde de alt faktörlerin sayısı ile faktörlerin içerdiği madde sayıları eşit olmalıdır. Koşulların zorluğu nedeniyle tutum ölçeklerinin içerdiği yapılar bakımından birebir örtüşmesi daha zordur. Tutum ölçekleri için *eş değer formlar yöntemi* daha uygundur.

Eş değer formlar, paralel formlardaki gibi istatistiksel sonuçlar açısından yakın bir benzerliğe sahip değildir. Formlardaki madde sayıları da aynı değildir. Fakat, ham puanlardaki farklılıklar belirli bir dağılım aralığında kalıyorsa veya önceden belirlenmiş norm değer aralıkları içine giriyorsa söz konusu farklılıklar önemsenmez. Bu farklılıklara tolerans gösterilir.

Karşılaştırılabilir formlarda ise, ölçüm araçları içerik olarak birbirlerine oldukça benzer gözükmesine karşılık bu konuda istatistik sonuçları ortaya koyacak herhangi bir araştırma henüz yapılmamıştır.[72] Karşılaştırılabilir formlarda da madde sayıları eşit değildir.

Araştırmacı geliştirdiği ölçeğin alternatif formlara karşı güvenilirliğini de test etmek istiyorsa bunun için değişik yöntemlere başvurabilir. Birincisi daha araştırmanın başlangıç aşamasında aynı kavramsal yapıyı ölçen başka bir test veya ölçek bulmasıdır. Paralel form içerik ve zorluk açısından geliştirilen formun veya testin büyük ölçüde aynısı olmalıdır. Bilim adamı karşılaştırma yapma amacıyla alternatif form/madde bulmak için değişik kaynaklardan yararlanabilir ve bunlar aşağıdaki gibidir:

1. Yüksek lisans ve doktora tezlerinde kullanılan formlar.
2. Üniversitelerdeki bilim adamlarının kendi bilimsel araştırmalarını yapmak ve makalelerini yazmak için kullanmış oldukları formlar.
3. Özel bilimsel araştırma merkezlerinin geliştirmiş oldukları formlar.
4. İnternet kaynaklarından sağlanan formlar.
5. Kurumsal bilgi ve madde bankalarından sağlanan formlar.

Doktora tezi hazırlayan araştırmacıların, doktora sonrası araştırma yapan bilim adamlarının ve standart test geliştirmek isteyen araştırma kuruluşları yetkililerinin daha titiz bir çalışma için yabancı ülkelerdeki test merkezlerinin kaynaklarını araştırmalarında yarar vardır. Literatürde sık başvurulan yabancı kaynaklardan bazıları aşağıdaki gibidir:

1. Educational Testing Service.
2. Buros Mental Measurement Yearbook.
3. Tests in Print.
4. Score Index.
5. APA, Directory of Unpublished Experimental Mental Measures.
6. Handbook of Tests and Measurement in Education and the Social Sciences.
7. Measures for Psychological Assessment.
8. Annual Review of Psychology.
9. PsycLit.

Gerçek anlamda birbirinin aynı olan iki paralel test formunda ölçümün standart hataları da (*ÖSH*) birbirine eşittir. Ancak günlük yaşamda birebir benzerlik veya aynılık tam olarak sağlanamayacağından formlar belirli ölçüde tesadüfî hata ögelerini içerir. Bu nedenle elde edilen güvenilirlik katsayısı gerçek durumu olduğundan biraz daha yüksek yansıtıyor olabi-

lir.[73] Testlerde özellikle zorluğun eşitlenmesi konusunda güçlüklerle karşılaşılabilir. Paralel formlar uygulamasında, her iki test birbirinden bağımsız olarak birlikte veya birkaç gün arayla uygulanarak korelasyon analizi sonuçlarına bakılır.

İkinci yöntem, test/ölçek geliştirme sırasında oluşturulan madde havuzundan birbirine benzer nitelikte iki ayrı ölçek/test geliştirmek ve bu ölçekleri birlikte uygulamaktır. Test geliştiren profesyonel kuruluşlar ve araştırmacılar aynı testin A ve B formlarını birlikte oluştururlar. Bazen, özellikle bilgi testlerinde ikiden fazla formun dahi oluşturulduğu görülür. Araştırmacı böylelikle bir taraftan güvenilirliği test etmeye yönelik ikinci bir forma sahip olurken, diğer taraftan kopya çekme olgusuna karşı kullanabileceği, telafi sınavı yapabileceği veya sonuçlara itiraz gelmesi halinde uygulayabileceği ikinci bir aracı daha kullanma şansını elde eder. Ancak bu tür paralel formlardaki maddelerin hepsi tek bir testin/ölçeğin içine yerleştirilmez. Katılımcılara iki ayrı test halinde verilir. Geliştirilen iki ölçekteki madde sayısı aynı ise paralel form olarak nitelendirilir. Madde sayısı farklı olarak belirlenmişse *eş değer formlar* olarak işlem yapılır. Örneğin, asıl anketteki madde sayısı 200–300 civarında iken bu maddelerin kısaltılmış başka bir sürümü cevaplayıcılara *eş değer form* olarak ayrıca uygulanabilir. Ancak bu tür uygulamalarda kısa form veya test etmek için geliştirilmiş bulunan form, asıl formu ikame edecek bir ölçüm aracı olarak düşünülmemelidir. Bu formlar cevaplayıcıların benzer şıkları işaretleyip işaretlemediklerini görmek için kullanılır.

Alternatif formlar farklı zaman dilimlerinde de uygulanabilir. Araştırmacı eğer bu tür bir tasarımdan hareket ediyorsa *eşitlik* ve *istikrarlılığı* birlikte ölçmek istiyor ve görmek istiyordur. Alternatif formlar arasındaki korelasyon değerleri belirli bir zaman geçtikten sonra dahi önemli ölçüde değişmiyorsa ölçek/test hem güvenilir ve hem de istikrarlıdır. Tam tersine önemli ölçüde farklılık varsa, hata varyansı ölçeklerin içeriğinin farklı olmasından ve aradaki zamandan kaynaklanıyordur. Paralel formları farklı zamanlarda uygulama prosedürü şu şekilde yapılır. Önce t_1 zamanında katılıcılara *A Formu* verilir. Daha sonra ardan belirli bir süre geçince bu kez t_2 zamanında *B Formu* verilir ve iki formun toplam puanları arasındaki korelasyona bakılır. Elde edilen değer "eş değerlilik katsayısı" olarak isimlendirilir.

Alternatif formlar eğer tek bir zaman diliminde ve tek bir seans halinde uygulanıyorsa testi alacak kişiler iki gruba bölünür. Transfer ve uygulama etkisini ortadan kaldırmak için birinci gruba önce *X* testi ve daha sonra *Y*

testi uygulanırken ikinci gruba önce *Y* testi daha sonra *X* testi verilir.[74] Bu uygulamaya ise *dengeleme* adı verilir. Böylece testi erken bitirme, testin son sorularına yanıt vermeme ve yorgunluk gibi faktörlerin etkisi bir ölçüde giderilmiş olur.

Alternatif form oluşturmanın güçlükleri. Alternatif form oluşturmanın bir takım güçlükleri vardır. Bu zorluklardan birincisi eşit formlar güvenilirlik katsayısının tutucu olması ve güvenilirliği olduğundan düşük göstermesidir. İkincisi, alternatif test veya ölçek oluşturmanın veya bulmanın oldukça güç olmasıdır. Üçüncüsü ise, alternatif formları katılımcılara uygulama güçlüğüyle karşılaşılmasıdır. Bilim adamı, söz konusu zorlukları yenmek için zaman planlaması yapmalı ve belirlediği formların birbirine alternatif olma özelliğini tam olarak sağlamaya çalışmalıdır.

Alternatif form oluşturmada matematik modellerden yararlanma. Son yıllarda alternatif veya paralel test formları geliştirmek için değişik matematik modeller geliştirilmeye başlanmıştır. Bunlardan ikisi minimizasyon ve maksimizasyon yaklaşımlarıdır. 1989 yılında Adema ve Van der Linden güvenilirliği maksimize edecek modeller üzerinde çalışmışlardır.[75] Minimizasyon modellerinde belirli kısıtlayıcılar altında minimum madde sayısına sahip paralel bir test oluşturma amaçlanır. Kısıtlayıcılar; test uzunluğu tanımıyla, testler arasında geçişme olmaması koşuluyla, güvenilirlik katsayısıyla, doğru işaretlenen puanların aritmetik ortalamasıyla ve kategori konusuyla ilgilidir. Buna göre her test maddesi bir kere seçilmeli ve paralel formlarda az sayıda ve eşit miktarda madde olmalıdır. Maksimizasyon modellerinde ise, – test uzunluğu testi oluşturan araştırmacı tarafından belirlenmek üzere – maksimum test güvenilirliğine sahip bir test oluşturma amacı güdülür. Bu yaklaşım testlerin güvenilirlik katsayılarının eşit olacağı gibi bir düşünceyi garanti altına almamakta, fakat paralel test uzunluklarının eşit olacağını belirlemektedir. Minimizasyon modeli için geçerli olan kısıtlayıcılar maksimizasyon modeli için de geçerlidir.[76] Bilim adamları minimizasyon ve maksimizasyon modellerini kullanmak isteyen araştırmacılar için matematiksel algoritmalar geliştirmişlerdir. Konuya ilgi duyan araştırmacılar, bu amaçla geliştirilmiş bulanan bu algoritmalardan ve ayrıca söz konusu algoritmaları çalıştıran matematik yazılımlarından yararlanabilirler.

Alternatif formlar güvenilirlik katsayısı. Alternatif formlarda güvenilirlik katsayısı ,90 olmalıdır. Alternatif formlarla yapılan değişik güvenilirlik analizi çalışmalarında korelasyon katsayılarının ,85 ilâ ,95 arasında değiştiği gözlenmiştir. Güvenilirlik katsayısı ,85'in altına düştüğünde formlar karşılaştırılabilirlik özelliğini kaybeder.[77]

Alternatif formlar arasındaki ilişkilerin hipotez testi ile sınanması. Araştırmacı isterse, kullandığı alternatif formlara ait maddelerin aynı yapıyı ölçüp ölmediği hipotezini test etmek için *yapısal eşitlik modeli* istatistiksel analiz yönteminden yararlanabilir.[78]

Madde-yanıt kuramında paralel formlar oluşturma. Araştırmacı ölçüm aracının güvenilirliğini madde-yanıt kuramı ile test ediyorsa bu kurama uygun olarak paralel formlar oluşturabilir. Bunun için tek bir form geliştirirken uyguladığı prosedürü takip eder. Her bir madde için aynı parametreler temel alınarak *madde bilgi fonksiyonlarının* toplamından oluşan *test bilgi fonksiyonu* (TBF) ve *test özellikleri eğrisi* (TÖE) değerlerini kullanır.[79]

Alternatif formlar yönteminin dezavantajları. Alternatif formlar yöntemi güvenilirliği ölçme konusunda belirli kolaylıklar sağlamasının yanında bazı sakıncaları da beraberinde getirir. Alternatif test formları aynı zaman diliminde veya yakın zaman dilimlerinde kişilere verildiğinde öğrenme ve olgunlaşma etkisiyle karşılaşmak kaçınılmazdır. Alternatif form uygulamasının sonuçları önceki form uygulamasının sonuçlarından etkilenir. Buna "bulaşma" etkisi denir. İkinci uygulama belirli ölçüde lekelenmiş, saflığını kaybetmiştir. Bulaşma etkisini ortadan kaldırmak için derece sayısı, ifadelerin düzenlenmesi ve ifadelerin içeriği açısından farklı olan ölçek veya testler geliştirmek gerekir.

Gözlemciler arası tutarlılık. Eş değerlilik yöntemlerinden ikincisi, gözlemciler arası tutarlılıktır. Gözlemciler önceden belirlenmiş bir puanlama sistemine bağlı olarak belirli bir olguyu bağımsız bir şekilde değerlendirebilirler. Bu değerlendirmelerde verilen puanların birbirine benzer olması, söz konusu puanların güvenilir olduğunu gösterir. Gözlemci puanlarının eş değerliliği konusu kapsamlı bir içeriğe sahip olması nedeniyle bir sonra başlıkta ayrıntılı bir şekilde ele alınmıştır.

Gözlemciler Arası Tutarlılık

Gözlemciler değişik nedenlerle değerlendirme yapabilirler. Örneğin doktorlar hastalarını, yöneticiler personelinin başarılarını, antrenörler futbolcularını, jüri üyeleri artistik patinaj yarışmasına katılan sporcuları değerlendirirler. Gözlemciler arasındaki değerlendirmenin güvenilirliği, ölçüm rakamları arasındaki tutarlılığa bağlıdır. Gözlemciler arası güvenilirlik "farklı gözlemciler tarafından yapılan değerlendirmelerin puan ortalamalarındaki sapma derecesidir (Tinsley ve Weiss, 1975, aktaran Humpro)."[80] Bu yöntem literatürde değişik isimlerle anılmıştır. Değerlendiriciler arası tutarlılık, hakemler arası tutarlılık, kayıtçılar veya puanlayıcılar arası tutarlılık deyimleri bunlar arasındadır. Bu yöntemin geçerlilik analizlerinde kullanılan, hakem veya uzmanların test içeriğinin ölçülmek istenen amaca uygun olup olmadığını belirledikleri değerlendirmelerle karıştırılmaması gerekir. Literatürde, gözlemciler arasındaki değerlendirmelerde *güvenilirlik* kavramıyla *uyuşma* kavramlarının birbirinin yerine kullanılıp kullanılamayacağı konusu tartışılmıştır. Güvenilirlik, korelasyon katsayısına dayanırken uyuşma, matematiksel bir hesaplamadır. Düşük varyans değerleri ve tek bir değerlendirme hedefi (ölçüm nesnesi) koşullarında *uyuşma* kavramının kullanılması daha uygun olur.

Gözlemciler, belirli bir olguyla ilgili olarak puan veren veya değerlendirme yapan kişilerdir. Gözlemciler, yapılan ölçümlerde benzer puanları vermişlerse sonuçlar güvenilir demektir. Ölçüm aracı kullanılarak yapılan değerlendirmelerde gözlemciler arasındaki uyuşmanın en az ,80 düzeyinde olması istenir. Ölçüm sadece gözlemde bulunmak suretiyle yapılmışsa bu kez değerlendiriciler arasındaki uyuşmanın ,70 olmasının yeterli olacağı belirtilmiştir (Szymanski ve Linkowski, 1995, aktaran VRI, 2002).[81] Ölçüm yapan gözlemcilerin tutarlı puanlar vermeleri için test verme veya ölçüm yapma özellikleri konusunda kendilerine video çekimleri izletilir veya ölçüm işlemi tekrar tekrar yaptırılarak deneyim kazanmaları sağlanır. Gözlemciler arası tutarlılık ancak ,80 düzeyine geldiği zaman kendilerinden bağımsız gözlemci/değerlendirmeci olarak yararlanılır.[82] Araştırma *uzun dönemli* bir niteliğe sahipse farklı gözlemciler arasındaki ölçümler değişik zamanlarda tekrar edilerek güvenilirliğin devam edip etmediği sınanmalıdır. Ölçüm verileri sürekli veya kesikli bir niteliğe sahip olabilir. Araştırmacı, verilerin tutarlılığını ölçmek için rakamların niteliğine uygun hesaplama yöntemlerinden yararlanır.

Kesikli verilerde gözlemciler arası tutarlılık. Ölçüm aracı kesikli verilerden (nominal veya sıralı veriler) oluşuyorsa, gözlemciler arası *uyuşma yüzdesi* güvenilirlik değeri olarak alınır. *Gözlemciler arası uyuşma*, "farklı gözlemcilerin değerlendirilen kişiler hakkında tam olarak aynı yargılarda bulunmalarıdır."[83] Değerlendiricilerin uyuştukları madde sayısı toplam madde sayısına bölünerek *uyuşma oranı* bulunur. Örneğin, bir hastanın durumunu bir hafta süre ile gece ve gündüz vardiyasında takip eden iki hemşire hastanın durumunu *İyileşiyor*, *Değişiklik Yok* ve *Kötüleşiyor* şıkları altında değerlendirsinler. Bu değerlendirmede hemşireler ,75 oranında *İyileşiyor* değerlendirmesini yapmışlarsa ölçüm sonuçları güvenilirdir. Uyuşma yüzdesi gözlemcilerin sayısı ile negatif yönde ilişkilidir. Gözlemci sayısı arttıkça uyuşma yüzdesi azalır. Uyuşma yüzdesi bir istatistik değil, indeks değeridir ve bu hesaplama değerlendirmede şans faktörünün etkisini dikkate almaz.

Kesikli verilerde gözlemciler arasındaki tutarlılığı hesaplamanın ikinci yöntemi *Cohen kappa* yöntemidir. Bu uygulamada birinci test uygulayıcısının verdiği testte geçme ve başarılı olma sayısı ile ikinci uygulayıcıda geçme ve başarılı olma sayısı karşılaştırılır. (Kappa formülünün hesaplanması için "Korelasyon Analizi ve Güvenilirlik" bölümüne bakınız.)

Kesikli verilerde gözlemciler arasındaki tutarlılığı hesaplamanın üçüncü yöntemi *phi katsayısıdır*. Bu istatistiksel yöntemde iki gözlemciye ait değerlerin birbirinden bağımsız olduğu varsayılır. Matriste herhangi bir hücredeki beklenen değer 5'ten küçük ise *Fisher kesin testi* (Fisher exact test) veya diğer bir adlandırmayla *Yates düzeltme faktörünün* kullanılması gerekir.

Dördüncü yöntem, Kendall uyuşma katsayısıdır. Kendall *W* uyuşma katsayısı aynı zamanda *Wilcoxon bağımlı örneklem*[a] *işaret testi* (*W*) olarak da bilinir. Bu yöntemde satırlarda değerlendiriciler ve sütunlarda değerlendirilen maddeler gösterilir. Hesaplanan *W* katsayısı 0 ilâ 1 arasında değişir.

Sürekli verilerde gözlemciler arası tutarlılık. Ölçüm aracı sürekli verilerden oluşuyorsa bu kez iki gözlemciye ait puanlar arasında korelasyon analizi yapılır. Örneğin Davranış Bilimleri Ana Bilim Dalı yüksek lisans öğrencilerinin derslere ek kaynaklardan yararlanarak hazırlıklı gelme durumu dört maddeden oluşan 5 dereceli bir ölçek üzerinde (*Büyük Ölçüde Hazırlanmış, Kısmen Hazırlık Yapmış, Belli Değil, Hazırlık Yap-*

[a] Aynı test, farklı iki gözlemci tarafından değerlendirilmekle birlikte ölçümlerin birbiriyle ilgili olması nedeniyle "bağımlı örneklem" grubunda değerlendirilir.

mamış, Bütünüyle İlgisiz) iki öğretim üyesi tarafından takip edilmiş olsun. Sömestr sonunda öğrencilerin toplam puanları arasında korelasyon analizi yapılarak gözlemciler arasındaki tutarlılığa bakılır. Korelasyon katsayısı yüksekse ölçüm sonuçlarının (toplam puanların) güvenilir olduğuna karar verilir. Küçük örneklemlerde ($n < 15$) Pearson r korelasyon katsayısının[a] şişmesi nedeniyle bu test yerine Safrit (1976) değerlendiriciler arası güvenilirliği tespit etmek için *küme içi korelasyon* analizinin kullanılmasını önermiştir (aktaran Humpro).[84] Pearson r gözlemcilerin ortalama değerleri hakkında herhangi bir varsayımda bulunmadığından farklılık bulunup bulunmadığı hipotezini test etmek için t-testinden yararlanılabilir. t-Testi sonucunda elde edilen anlamlılık düzeyi, iki değerlendirici arasında anlamlı bir farklılık bulunmadığını gösterir.

Sürekli verilerde gözlemciler arasındaki güvenilirliği hesaplamanın bir diğer yöntemi *Kendall tau* değeri ile çalışmaktır. Parametrik veriler eğer normal dağılım özelliği göstermiyorsa gözlemciler arasındaki tutarlılığı belirlemek için uygun bir korelasyon ölçüsüdür. Veriler normal dağılım özelliği göstermiyor ve $n>20$ ise bu kez Spearman rho kullanılır.

Uygulama alanları. İş yerlerinde personelin başarı değerlendirme ölçümleri, personel seçimi için "değerleme merkezi" yöntemiyle yapılan mülakat ölçümleri, psikometrik yetenek ve beceri testleri, hastanelerde hastalarla ilgili ölçümler (tansiyon, nabız, ateş, kan sayımları gibi), klasik sınav kağıtlarının değerlendirilmesi, sportif karşılaştırmalarda sporculara puan verilmesi birden fazla bağımsız gözlemci tarafından yapılır. Ölçüm sürecinde farklı gözlemcilerden, pratisyenlerden veya teknisyenlerden yararlanılıyorsa bu kişilerin ölçüm uygulamasını hep aynı şekilde, standart bir uygulamaya sahip olarak yapıp yapmadıkları önem kazanır. İki farklı kişinin yaptığı ölçüm sırasında aradan geçen zaman ve ölçüm yapılan kişilerde ortaya çıkan değişikler sonuçları etkileyebilir.

Gözlemci etkisi. Güvenilirlik değerleri, gözlemcilerin özelliklerinden etkilenir. Gözlemcilerin değerlendirme yaptıkları kişilerle ilgili bir özelliği ön plana çıkarmaları *hâle etkisi* olarak isimlendirilir. Hâle etkisinde kalınarak yapılan bir değerleme güvenilirlik rakamlarını düşürür. Gözlemcileri etkileyen ikinci faktör *basma kalıp yargılardır*. Gözlemcilerin belirli bir gruba yönelik olarak sahip oldukları önyargılar yapılan değerlendirmeyi

[a] Pearson r kümeler arası korelasyon analizi olarak tanımlanır. Pearson analizinde değişken verileri hem metrik değildir, hem de verilerde varyans paylaşımından söz edilemez.

etkiler. Bir diğer faktör *algılama farklılıklarıdır*. Değerlendiricilerin önceki deneyimleri, sahip oldukları bilgiler yapılan değerlendirmeleri etkileyebilir. Gözlemci etkisini azaltmak için eğitim, egzersiz, puanlama yönergeleri ve ölçüm aracında revizyona gidilebilir. Gözlemcilere değerlendirme yaptırılmadan önce test/ölçek veya ölçümü yapılacak etkinlik hakkında ayrıntılı bilgiler verilmeli ve gözlemcinin kafasına takılan sorular açıklığa kavuşturulmalıdır. Gözlemci ölçüm biçimine veya aracına aşina olmalı bu konuda her hangi bir tereddüt geçirmemelidir. Değerlendirme sırasında belirli aşamalar varsa değerlendirici bu aşamaları çok iyi öğrenmeli ve test yönergelerine uygun biçimde davranış göstermelidir. Gözlemci, yapılan ölçümün mantığını, arka plandaki felsefeyi çok iyi anlamalı ve olabildiğince objektif bir biçimde değerlendirme yapmalıdır. Gözlemcilerin dikkatsiz olmaları testi gelişi güzel bir biçimde uygulamaları, sistematik veya tesadüfî hatanın artmasına neden olur.

Analiz Yöntemlerinin Karşılaştırılması

Her bir güvenilirlik analizinin kendisine göre avantaj ve dezavantajları vardır. Bilim adamı, öncelikle yaptığı ölçümün ve kullandığı ölçüm aracının niteliğine uygun güvenilirlik analizlerinin hangileri olduğunu araştırmalıdır. Ölçüm *gözleme dayanıyorsa* en iyi yöntem gözlemciler arası güvenilirlik analizlerinin uygulanmasıdır. Fakat bunun için birden fazla gözlemci olması gerekir. Eğer birden fazla gözlemci bulunamıyorsa böyle bir durumda tek bir gözlemcinin farklı durumlarda yaptığı gözlem sonuçları değerlendirilir.[85] Bu uygulamaya *gözlemci içi değerlendirme* adı verilir.

Araştırma kağıt-kalem testi veya ölçek gibi bir *ölçüm aracı* kullanılarak yapılıyorsa *keşfedici faktör analizi* yöntemiyle alt boyutlar ortaya çıkarıldıktan sonra güvenilirlik analizi için ilk etapta alfa güvenilirlik katsayısı düşünülmelidir. Alfa katsayısı iç tutarlılık güvenilirliğinin üst sınırını oluşturur. Alfa değeri düşük çıkmışsa ya test çok kısadır veya maddeler çok az ortak bir öze sahiptir. Böyle bir durumda alternatif formlar gibi diğer güvenilirlik analizlerini yapmaya gerek yoktur. Çünkü bu analizler yapıldığında güvenilirlik rakamları daha da düşük çıkar. Alfa değeri eğer yüksek çıkmışsa alternatif formlardan elde edilen korelasyon katsayıları da yüksek çıkar.[86] Nunnelly (1978) tutum ölçekleri için en az iki güvenilirlik analizinin yapılmasını önermiştir. Bunlar alternatif test formları ve alfa güvenilirlik katsayısıdır (aktaran HFS).[87] Araştırmacı üçüncü bir güvenilirlik analizi olarak yarıya bölme güvenilirliğini düşünebilir. Fakat yarıya bölme yöntemi yerine alfa değerini kullanmak daha anlamlıdır. Yarıya bölme yönte-

minin haklı görülebileceği tek yer maddelerin üçlü veya daha fazla dereceler halinde puanlandırılması durumudur. Yarıya bölme yöntemi için uygun sayılabilecek bir diğer durum araştırmacının kısa zaman diliminde alternatif form bulamamasıdır. Böyle bir durumda testin iki yarısı katılımcılara sanki alternatif form imiş gibi düzenlenerek verilir.[88]

Alternatif formlar yönteminde belirli bir kavramsal yapıyı ölçmeye yönelik olarak aynı zaman diliminde iki form birden geliştirilir. Fakat tek bir ölçek yerine iki ölçeği birden geliştirmeye çalışmak zordur. Onun yerine önceden geliştirilmiş alternatif başka bir form bulma yoluna gitmek daha doğru bir yaklaşım olur. Matematik testi, Türkçe testi gibi bilgiyi ölçmeye yönelik ölçümlerde ise, paralel formlar geliştirmek nispeten daha kolaydır.

Güvenilirliği belirlemek için gerektiğinde alternatif formlar yerine test-yeniden test yöntemi uygulanabilir. Test-yeniden test yöntemi, alet kullanılarak yapılan fiziksel performans testleri ile, kontrol grubunun kullanıldığı deneysel ve yarı deneysel araştırmalar için uygundur. Deneysel araştırmalarda ön test ve son testin her ikisinde de ölçümler ayrıca kontrol grubunda da yapılır.[89] Bununla birlikte, test-yeniden test yönteminin ciddî yetersizlikleri söz konusudur. Bu yöntem "alan örneklemesi" modelini dikkate almadığından sadece toplam puanlar arasındaki korelasyona dayanır. Bir testin alfa değeri düşük, fakat test-yeniden test güvenilirlik değeri yüksekse bu durum testin yüksek güvenilirliğe sahip olduğu şeklinde yorumlanmamalıdır. Güvenilirlik için öncelikle test maddelerinin iç tutarlılığa sahip olması gerekir. Maddeler arasındaki korelasyon yüksek değilse testin/ölçeğin toplam puanının herhangi bir özelliği veya vasfı ölçtüğünü söylemenin bir anlamı yoktur. Eğer test-yeniden test korelasyon değeri düşük ise alternatif formlar korelasyon değeri de düşük çıkacaktır. Bir test farklı iki zamanda uygulandığında korelasyon katsayıları yüksek değilse diğer güvenilirlik kanıtlarını aramaya gerek yoktur. [90]

Güvenilirlik analizlerinin her biri farklı bir değer verir. Alfa katsayısı ve paralel formlar yöntemiyle gözlemciler arası değerlendirme ve test-yeniden test yönteminden daha yüksek sonuçlar elde edilir.

İNDEKS VE KATSAYILAR

Literatürde farklı anlamlara sahip olan *güvenilirlik indeksi* ile *güvenilirlik katsayısı* terimlerinin birbirlerinin yerine kullanıldığı görülür. Okuyucuların

yanlış kullanmalarını önlemek amacıyla bu kavramların anlamları üzerinde kısaca durmakta yarar vardır.

Güvenilirlik indeksi. Kuramsal anlamda güvenilirlik indeksi, gerçek puanla (evren puanlarıyla) gözlem puanları arasındaki korelasyona işaret eder.[a] Bu, hesaplanması mümkün olmayan kuramsal bir değerdir.

Uygulamada ise, korelasyon analizlerine dayanan güvenilirlik katsayısının karekökü *güvenilirlik indeksi* olarak isimlendirilir (*bk.*, Eşitlik 3-22).[91] Aynı şekilde Cronbach alfa katsayısının karekökü de güvenilirlik indeksidir. Güvenilirlik indeksi, *k* sayıda madde korelasyonları arasındaki hata içermeyen gerçek puan oranını yansıtır.[92] Güvenilirlik indeksi, güvenilirlik katsayısından daha büyük bir değer verir, çünkü test puanları evrensel alandaki puanlarla paralel formlar testine tâbi tutulmuş gibi düşünülür (*bk.*, Tablo 3-9). Alandaki puanların kavramsal yapıyı temsil etme özelliği yüksek olduğundan indeks değeri daha büyük çıkar (*bk.*, Şekil 3-4).

$$\text{Güvenilirlik indeksi} = \rho_{tx} = \sqrt{\rho_{xx'}} \quad . \tag{3-22}$$

ρ_{tx} = Güvenilirlik indeksi.
x ve x' = Paralel form gözlem değerleri.
$\rho_{xx'}$ = Paralel formlar korelasyon katsayısı (rho).

Tablo 3-9. Güvenilirlik Katsayısı ve Güvenilirlik İndeksi

Belirlilik katsayısı = Korelasyon katsayısının karesi (testteki gerçek puan varyansının oranı)	Güvenilirlik katsayısı = Korelasyon katsayısı	Güvenilirlik indeksi = Güvenilirlik katsayısının karekökü, Gözlem puanlarıyla evren puanları arasındaki korelasyonun karesi
,24	,49	,70
,41	,64	,80
,65	,81	,90

[a] Bazı kaynaklarda "gözlenen puanlarla gerçek puanlar arasındaki korelasyonun karesi" olarak tanımlanmıştır (*bk.*, Yaşar Baykul, *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*, (Ankara: ÖSYM, 2000, s.143.).

İstatistiksel olarak, muhtemel bütün yarılar arasındaki korelasyon katsayılarının ortalaması yine güvenilirlik indeksidir. Bazı araştırmacılara göre *katsayı* istatistiksel analizlerle ilgilidir, *indeks* ise matematiksel hesaplamalara dayanır. Güvenilirlik indeksi kavramı, ayrıca *ölçümün standart hatası* anlamında kullanılır. Okuyucu indeks kelimesini kitapta kullanıldığı yere ve kullanım amacına göre anlamlandırmalıdır.

Güvenilirlik katsayısı. Güvenilirlik katsayısı, iki ölçüm arasındaki korelasyon değerine dayanır. Bu kavram değişik yazarlar tarafından zaman zaman alfa katsayısı, belirlilik katsayısı, genellenebilirlik katsayısı ve gözlemciler arası tutarlılık katsayısı ile aynı anlamda kullanılmıştır. İstatistiksel anlamda ise paralel iki ölçüm değişkeni arasındaki ilişkiyi gösterir.[93] İki ölçüm birbirine paralel ise verilerin kendi aralarındaki değişimler doğrusal bir niteliğe sahiptir ve bu nedenle elde edilen korelasyon değeri aynı zamanda güvenilirlik katsayısıdır.[a] Güvenilirlik katsayısı, paralel ölçümlerin niteliğine ve uygulanan hesaplama formülüne göre farklı isimlerle adlandırılır. Bu isimlendirmeler aşağıdaki gibidir:

1. İstikrarlılık katsayısı (t_1 ve t_2 zamanında yapılan ölçümler).
2. Eş değerlilik katsayısı (alternatif formlar).
3. Alfa katsayısı (muhtemel bütün yarılar arasındaki tutarlılık).
4. KR-20 iç tutarlılık katsayısı (Maddeler arasındaki tutarlılık).
5. KR-21 iç tutarlılık katsayısı (Maddeler arasındaki tutarlılık).
6. Yarıya bölme güvenilirlik katsayısı (iki yarı arasındaki güvenilirlik).
7. Spearman-Brown güvenilirlik katsayısı.

Belirlilik katsayısı. Güvenilirlik katsayısıyla ilgili bir diğer kavram, *belirlilik katsayısı*dır. Belirlilik katsayısı, korelasyon katsayısının karesidir. İki ölçüm arasındaki ortak varyansı açıklar. Literatürde bu katsayıya aynı zamanda *değişkenlik* veya *varyans* denir. Bazı bilim adamlarına göre ise, kore-

[a] J.L. Melia, ölçümler arasında paralellik varsa korelasyon katsayısı sıfıra yakın bile çıksa bunun güvenilirlik katsayısı olarak isimlendirileceğini, ancak ölçümler arasında parallellik yoksa korelasyon değeri 1 bile olsa bunun güvenilirlik katsayısı olarak isimlendirmenin doğru olmayacağını ifade etmiştir. Bununla birlikte ona göre, güvenilirlik katsayısında ölçümler arasındaki değişkenliğin tam bir doğrusallık gösterdiğini saptamak çok zordur. Güvenilirlik katsayısında "doğrusallık" fonksiyonu ihmal edilir (*bk.*,<http://www.uv.es/~meliajl/ Research/Description_books.htm> (17.12.2002).

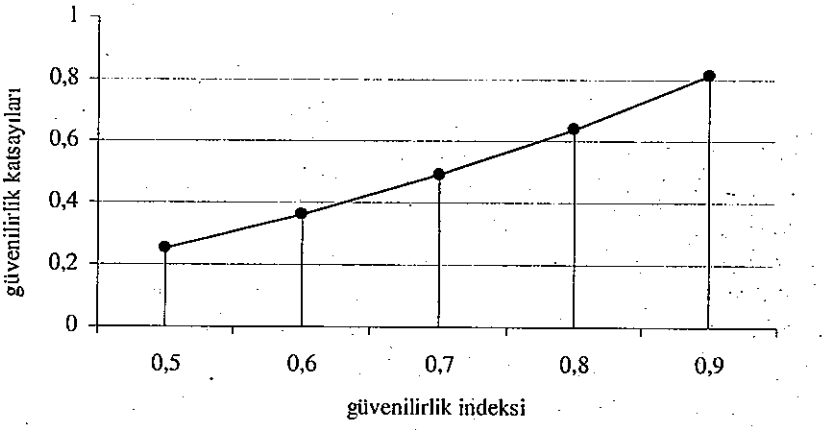
lasyon katsayısı "güvenilirlik katsayısı" değil, *güvenilirlik indeksidir.* Bu değeri katsayıya dönüştürmek için karesini almak gerekir (Gronlund ve Linn, 1990, aktaran Dawson).[94] Kopeikin'e göre (2000) ise güvenilirlik katsayılarının hepsinde korelasyon katsayılarının kareleri alınır.[95] Bu yazarlar "güvenilirlik katsayısı" terimini, "belirlilik katsayısı" anlamında kullanmışlardır.[a] Ancak literatürdeki genel uygulama korelasyon katsayısının aynı zamanda güvenilirlik katsayısı olarak değerlendirilmesi yönündedir.

Belirlilik katsayısı iki ölçüme ait ortak değişkenlik alanını belirler. Eğer iki ölçümden biri "neden" olan bağımsız değişken ise; belirlilik katsayısı, bağımsız değişkenin (X) veya bağımsız değişkenlerin (X_1, X_1, ... X_n) bağımlı değişken (Y) üzerindeki etkisini belirler. Korelasyon analizi sonucunda örneğin ,80 gibi bir değer elde edilmişse r^2 = ,64 çıkacaktır ve bu rakam, "iki paralel formun ,65 oranında ortak varyansa sahip olduğu" şeklinde yorumlanır. Testlerden birinin toplam puanı bilinirse diğer testten toplam kaç puan alınabileceği ,65 oranında tahmin edilebilir demektir.

Kimi istatistikçiler, *belirlilik katsayısını* iki değişken yerine ikiden fazla değişken arasındaki korelasyon değeri için kullanmayı yeğlemişlerdir. Bir ölçekte ikiden fazla madde arasındaki korelasyon katsayılarının ortalamaları alınmışsa güvenilirlik katsayısı bu değerin karesi alınarak bulunur. Bu anlamda güvenilirlik katsayısı, ikiden fazla ölçüm arasındaki ortak varyansı gösterir. Bazı istatistik kaynaklarında bu durumu belirlemek için *çoklu korelasyon katsayısı* terimi kullanılmıştır. Çoklu korelasyon katsayısı R simgesi ile gösterilir.

---

[a] Kopeikin makalesinde, Pedhazur ve Schmelkin'den yaptığı alıntı ile aslında Cronbach alfa değerinin de teorik olarak karesi alınmış bir korelasyon katsayısı olduğunu ileri sürmüştür. Her nekadar alfa simgesinde "karesi" simgesi (^{2}) gösterilmiyorsa da bu değerin kavramsal alana ait maddeler havuzundan halen oluşturulmaya çalışılan test maddelerine benzer maddeleri içeren rastgele seçilmiş iki teste ait puanlar arasındaki korelasyonun karesini gösterdiğini ifade etmiştir.



Şekil 3-4. Güvenilirlik indeksi ile güvenilirlik katsayıları arasındaki ilişkiler.

Güvenilirlik Katsayılarının Büyüklükleri

Cronbach alfa veya diğer güvenilirlik katsayılarının ne olması gerektiği konusu bilim disiplinleri ve araştırma alanlarına göre farklılık gösterir. Bu konuda literatürde değişik bilim adamları çok sayıda ve farklı katsayı önerilerinde bulunmuşlardır. Araştırmacı bu katsayı önerilerini kendi araştırmasının özelliklerini göz önünde bulundurarak değerlendirmelidir.

Sağlık ölçümlerinde. Güvenilirlik rakamları insanların yaşamları hakkında karar vermek için (tedaviye karar vermek, hastaneye yatırmak, ameliyat etmek gibi) kullanılıyorsa alfa katsayısı için en alt düzey ,90 olmalıdır.

Okul testlerinde. Okul testleri iki grupta değerlendirilir. Birincisi okul yerleştirme testleri ve ikincisi ise ders başarı testleridir. Kişilerin okullara yerleştirilmelerinde kullanılacak testlerin güvenilirlik katsayılarının yüksek olması gerekir. Bu tür testler için ,90 ve hatta ,95 güvenilirlik katsayılarıyla çalışılması önerilmiştir.[96] Ders başarı testlerinde ise güvenilirlik katsayıları ,50 veya ,60 düzeyine kadar düşebilir.

Personel seçimi amacıyla yapılan psikometrik ölçümlerde. personel seçimlerinde kullanılacak testlerde ,90 – ,95 güvenilirlik katsayıları aranır. Ancak bu tür uygulamalarda test veya ölçekler istihdam etmek, hayatî kararları vermek için tek başına kullanılmamalıdır.

Bilimsel araştırmalarda. Yüksek güvenilirlik katsayısı kuralı, sadece araştırma yapmak ve belli bir ölçeği geliştirmek için kullanıldığı zaman bir miktar esnetilebilir.[97] Öte yandan sosyal bilimlerde alt düzey ,70 olarak kabul edilmiştir.

Grupların karşılaştırılmasında. Ölçümler bireylerin değil, grupların karşılaştırılması amacıyla kullanılacaksa daha düşük güvenilirlik katsayılarıyla çalışılabilir. Grup karşılaştırmalarında güvenilirlik katsayısının ,70+ olması gerektiği belirtilmiştir.[98]

Bireysel ölçümlerde. Bireysel ölçümler; kişilik, zeka, ilgiler, yetenek ve beceriler gibi konuları kapsar. Kişilere bu konuda test veya ölçüm sonuçlarına bağlı olarak geri besleme yapılacaksa güvenilirlik katsayılarının yüksek olması gerekir. Bilim adamları bu tür ölçümlerde en az ,85+ veya ,90+ güvenilirlik katsayısı ararlar.

Standart testlerde. Yetenek ve başarı[a] olmak üzere iki sınıfta toplanan standart testler, belirli bir norm grubu veya kriter grubu temel alınarak testin ortalama değerleri, dağılım aralığı, varyansı ve standart sapması saptanmış olan ölçüm araçlarıdır. Standart testlerde güvenilirlik katsayıları için Yu (2002) aşağıdaki değerlerin kullanılmasını önermiştir:[99]

1. İç tutarlılık ,95.
2. İstikrarlılık ,90.
3. Eş değerlilik ,85.

Güvenilirlik konusunda en fazla alıntı yapılan yazarlardan biri olan Nunnally 1967'de güvenilirlik katsayısının ,50 ilâ ,60 arasında değişebileceğini söylerken, 1978'de herhangi bir açıklama yapmadan katsayıyı ,70'e çıkarmıştır (aktaran Peterson).[100] Nunnally, güvenilirlik için araştırmanın türüne göre farklı katsayılar önermiştir: Başlangıç seviyesinde ölçek geliş-

[a] Yetenek, potansiyel kabiliyeti ortaya çıkarmaya yönelik iken; başarı bir kişinin sahip olduğu bilgi ve beceriyle ilgilidir.

tirmeye yönelik olarak yapılan pilot araştırmalar için ,60; temel araştırmalar için ,80 ve uygulamalı araştırmalar için ,90–,95 oranının bulunması gerektiğini belirtmiştir. Peterson (1994) tarafından yapılan bir araştırmada 4286 makaledeki alfa değerleri incelenmiş ve bu değerlerin %75'inin ,70 ve daha yüksek olduğu, %49'unun ,80 ve daha yüksek olduğu ve %14'ünün ise ,90 ve daha yüksek olduğu bulunmuştur.[101] Bu araştırmada ortalama değer ,77 ve medyan değeri ,79 çıkmıştır. Öte yandan Boyle (1991) ölçeklerde ,90'ın üzerindeki yüksek güvenilirlik katsayılarından kaçınılması gerektiğini, çünkü bu tür ölçeklerde söz konusu yüksek rakamların ölçeğin güvenilirliğine değil fazladan, lüzumsuz maddelerin bulunduğuna işaret edebileceğini ifade etmiştir (aktaran Peterson).[102] Bir test alt testlerden oluşmuşsa alfa güvenilirlik değerini her bir alt test için ayrı ayrı hesaplamak gerekir. Testin tamamı veya test bataryasının tamamı için güvenilirlik değeri daha düşük çıkabilir. Bileşik test ve ölçeklerde toplam veya genel güvenilirlik rakamının en az ,60 olması gerektiği iddia edilmiştir.[103]

Güvenilirlik katsayıları araştırmanın türüne göre de değişiklik gösterebilmektedir. İlk kez yapılan keşfedici araştırmalardan elde edilen veriler için aranılan güvenilirlik katsayısı, teyit edici ve nedenselliği test eden araştırma bulguları için aranılan katsayıdan daha düşüktür.[104] Keşfedici araştırmalar için ,60 katsayısı yeterli görülebilirken teyit edici araştırmalarda bu oranın en az ,80 olması gerekir.

Güvenilirlik rakamlarının değerlendirilmesinde bir başka sorun alfa güvenilirlik rakamı yüksek çıkmışken paralel form güvenilirlik rakamının düşük çıkmış olması halidir. Böyle bir durumda paralel formun aynı özelliği ölçüp ölçmediği dikkatli bir biçimde incelenmelidir.

Okullarda sınıf ortamında yapılan bilgi ve öğrenme testleri için yüksek derecede güvenilirlik rakamları gerekli değildir. Sadece kritik okul yerleştirmelerinde yüksek güvenilirlik rakamları aranır. Öğrenciler kendilerine anlatılan konulara hakim oldukça, puanlar arasındaki değişkenlik ve sonuçta testin güvenilirliği azalır. Sınıf ortamında uygulanan testlerde ,50 veya ,60 güvenilirlik oranı yeterlidir.[105] Güvenilirlik testin/ölçeğin ve uygulandığı örneklem grubunun ortak puanıdır. Sadece araca veya örnekleme ait değildir. Bu nedenle belirli bir örneklemde ,90 olan güvenilirlik katsayısı başka bir örneklemde daha düşük veya daha yüksek çıkabilir.

Değişik Analiz Yöntemlerine Göre Güvenilirlik Katsayılarındaki Farklılıklar

Güvenilirlik analizi yöntemlerinde farklı ölçüm hataları dikkate alındığından güvenilirlik rakamları da farklı çıkar. Genel olarak ifade etmek gerekirse "genellenebilirlik" (*G*) kuramına göre hesaplanan güvenilirlik katsayıları "gerçek puan" kuramına göre hesaplanan güvenilirlik katsayılarından daha düşüktür. Genellenebilirlik kuramına göre yapılan hesaplamada sistematik hatalar ve tüm hata kaynakları dikkate alındığından güvenilirlik rakamları daha düşük çıkar. Genellenebilirlik kuramında küme içi korelasyon (KİK) ve varyans analizi değerleri dikkate alınır.

Yine çeşitli gerçek puan güvenilirlikleri içinde de iç tutarlılık güvenilirliği değeri, alternatif formlar ve test-yeniden test güvenilirliği katsayılarından daha yüksektir.[106] Öte yandan bir araştırma raporunda farklı yöntemlere göre yapılan hesaplamalar sonucunda elde edilen değişik güvenilirlik katsayılarının aritmetik ortalamasını vermenin bir anlamı yoktur.

Ölçek ve testler bireyler arasındaki farklılıkları ölçmek için geliştirilmiştir. Araştırma yapılan örneklemde bireyler araştırılan yetenekler ve özellikler açısından önemli ölçüde farklılık göstermiyorsa güvenilirlik analizinin sonuçları da düşük çıkar. Ölçüm sonuçlarının varyansı düşükse test sonuçları daha düşük bir güvenilirliğe sahiptir. Varyans değeri (değişkenlik) arttıkça testin/ölçeğin güvenilirliği artar. Farklı güvenilirlik katsayılarının nispî etkililiğini değerlendirmek için ya testin veya ölçeğin madde sayısı sabitlenir veya güvenilirlik katsayıları sabitlenerek buna uygun bir biçimde madde sayısında artış yapma yoluna başvurulur.[107]

Farklı yöntemlere göre yapılan hesaplamalar sonucunda elde edilen güvenilirlik rakamları araştırmanın amacına ve ölçüm aracına göre uygun veya uygun olmayabilir. Geniş bir ölçüm aralığı temel alınarak test edilen (standart hatasının yüksek tutulduğu) çalışmalarda güvenilirlik katsayılarının düşük çıkması fazla bir sorun teşkil etmez.

Değişik Ölçüm Sonuçlarına Göre Güvenilirlik Katsayılarındaki Farklılıklar

Ölçek ve testlerin aynı yöntem kullanılarak elde edilen güvenilirlik analizi sonuçları, ölçümler arasında da farklılık gösterebilir. Standart testlerde, bu testleri yayımlayan şirketlerin el kitaplarında belirtilen güvenilirlik katsayıları ile kullanıcıların elde ettikleri güvenilirlik katsayıları tam olarak uyuşmayabilir. El kitaplarında belirlenen güvenilirlik katsayıları bazen daha dikkatli koşullarda elde edilmiştir ve bazen de ölçümün dayandığı örnek kütle hakkında hiç bir bilgi verilmediğinden hangi koşullarda elde edildiği konu-

sunda bilgi edinilemez. Bu nedenle geliştirilen test ve ölçekler başka kişiler tarafından aynı ana kütleye ait başka örnek kütlelerde uygulandığında değişik güvenilirlik katsayılarıyla karşılaşılabilir. Bu gibi durumlarda Fisher z' puanları yöntemi kullanılarak farklı güvenilirlik katsayılarının birleştirilmesi yöntemine başvurulabilir. Bilim adamı eğer test veya ölçeği farklı koşullarda uygulamışsa ve bu koşulların güvenilirlik katsayıları üzerindeki etkisini görmek istiyorsa meta analizi yaklaşımından yararlanarak etki büyüklüğü hesaplaması yöntemine başvurur.

Güvenilirlik Katsayılarının Analiz Sonuçları Üzerindeki Etkileri

Düşük güvenilirlik katsayılarına sahip değişkenlerle yapılan istatistikî analizlerin değişik etkileri söz konusudur. Örneğin düşük güvenilirlik katsayıları, regresyon analizinde yanlı ve etkili olmayan regresyon katsayısının elde edilmesine neden olur (Bohrnstedt ve Carter, 1971, aktaran Ping).[108] Bu tür ölçeklerle yapılan hipotez testlerinde yanlış negatif (Tip I) ve yanlış pozitif (Tip II) hatası yapma olasılığı artar. Düşük güvenilirlik katsayıları, verilerin standart hatasını arttırarak elde edilen sonuçların önemini ve değerini azaltır.

Güvenilirliğin düşük olması aynı ana kütleden çekilecek diğer örneklemlerde varyansın yüksek olacağı ve sonuçta verilerin gerçeği yansıtmayacağı anlamına gelir. Oysa değişik gruplarda yapılacak ölçüm ortalamalarının ana kütle ortalamasını yansıtması gerekir. Düşük güvenilirlik katsayısı yukarıdaki nedenlerle yapılan tahminlerin kuşkulu olmasına yol açar.

ALINTI YAPILAN KAYNAKLAR

[1] R.K. Henson, "Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha [İç Tutarlılık Tahmin Değerlerinin Anlaşılması: Alfa Katsayısı Üzerine Kavramsal Bir İnceleme]," *Measurement and Evaluation in Counseling and Development,* Oct 2001, 177-189.

[2] D.J. Ozer ve S.P. Reise, "Personality Assesment [Kişilik Değerlendirmesi]," <http://www.psyc.uow.edu.au/subjects/GHMC976/perspaper1.html> (18.10.2002).

[3] D. Bartram, P. Lindley, I. Coyne, "P Manuel [P Elkitabı]," <http://www.viewai.com/docs/PManual.pdf.> (18.10.2002).

[4] J.J. Ray, "Can There be "Too Much" Internal Consistency in a Scale? Rejoinder to Boyle [Bir Ölçekte Çok Fazla İç Tutarlılık Olabilir mi? Boyle'ye Katılma]," <http://members.optusnet.com.au/~jonjayray/boyle.html> (18.10.2002).

[5] Gary D. Gaddy, "Assessing the Reliability Of Scales [Ölçeklerin Güvenilirliğini Değerleme]," <http://www.irss.unc.edu/irss/shortcourses/gaddyhandouts/ReliabilityHandouts/reliabilityhandout.pdf.> (06.09.2002).

[6] A. Field, "Reliability Analysis [Güvenilirlik Analizi]," <http://www.cogs.susx.ac.uk/users/andyf/teaching/rm2/reliability.pdf.> (19.09.2002).

[7] UMASS, "Indices and Scales [İndekler ve Ölçekler]," <http://www-sociology.sbs.umass.edu/courses/soc210/Indices_and_Scales.htm> (26.10.2002).

[8] J. M. Lebreton, "Test Theory [Test Kuramı]," <http://www.google.com.tr/search?hl=tr&ie=UTF-8&q=McDonald+L3+cronbach+G-C+alpha+&btnG=Arama&meta=> (23.05.2004).

[9] A. Yu, "Using SAS For Item Analysis and Test Construction [Madde Analizi İçin SAS Programının Kullanılması ve Test Oluşturma]," <http://seamonkey.ed.asu.edu/~alex/teaching/assessment/alpha.html> (26.10.2002).

[10] C. Ho Yu, "An introduction to computing and interpreting Cronbach Coefficient Alpha in SAS [SAS'ta Cronbach Alfa Değerinin Hesaplanması]," <http://www2.sas.com/proceedings/sugi26/p246-26.pdf.> (19.09.2002).

[11] Yu, "An introduction."

[12] J. Esteves, "Using the Partial Least Squares Methods ... [En Küçük Kareler Yönteminin Kullanılması]," <http://www.lsi.upc.es/dept/techreps/ps/R02-23.pdf> (21.01.2004).

[13] Aynı.

[14] Aynı.

[15] "Statistical Information [İstatistiksel Bilgi]," <http://www.facs.gov.au/internet/facsinternet.nsf/vIA/nshs/$file/Appendix%20C.rtf> (21.01.2004).

[16] Aynı.

[17] nschen@cc.nsysu.edu.tw, "Assessment Of E-Learning Satisfaction [E-öğrenme Tatmininin Değerlendirilmesi]," <http://www.im.fju.edu.tw/conference/%E5%85%A8%E6%96%87%E6%AA%94/ASSESSMENT%20OF%20elearning%20satisfaction%20from%20criticalincidents%20.pdf> (21.01.2004).

[18] N. Schmitt, "Uses and Abuses of Coefficient Alpha [Alfa Katsayısının Kullanılması ve Suiistimal Edilmesi]," *Psychological Assesment*, 8 (4), 350-353.

[19] W.M. Rogers, N. Schmidt ve M.E. Mullins, "Correction For Unreliability Of Multifactor Measures: Comparison Of Alpha And Parallel Forms Approaches [Çok Faktörlü Ölçümlerde Düşük Güvenilirlik Rakamlarını Düzeltme] <http://io.psy.msu.edu/Schmitt/ormalpha.htm> (19.10.2002).

[20] Rogers, Schmidt ve Mullins, "Correction for Unreliability."

[21] J.R. Hills, "Alpha and Test-Retest [Alfa ve Test-Yeniden Test Yöntemi]," <http://www.math.yorku.ca/Who/Faculty/Monette/Ed-stat/0233. html> (26.10.2002).

[22] A. Ferligoj ve A. Mrvar, "Assesment of Reliability [Güvenilirliğin Değerlendirilmesi]," <http://mrvar.fdv.uni-lj.si/pub/mz/mz15/socan.pdf.> (26.10.2002).

[23] Rogers, Schmidt ve Mullins, "Correction For."

[24] Garson, "Structural Equation Modelling [Yapısal Eyitlik Modeli], <http://www2.chass.ncsu.edu/garson/pa765/structur.htm> (28.10.2002).

[25] K. Vehkalahti, "Reliability of Measurment Scales, [Ölçeklerin Güvenilrliği]," <http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/reliabil.pdf.> (26.10.2002).

[26] Ferligoj ve Mrvar, "Assesment of Reliability."

[27] Metric, "Congeneric Measurement Model [Konjenerik Ölçüm Modeli]," t.y., <http://www.measurementexperts.org/learn/theories/theories_ct.asp> (23.05.2004).

[28] Gregor Sočan, "Assessment of Reliability when Test Items are not Essentially tau-Equivalent [Test Maddeleri Yaklaşık Tau Eşitliğine Sahip Olmadığı Durumda Güvenilirlik Değerlendirmesi]," <http://mrvar.fdv.uni-lj.si/pub/mz/mz15/abst/socan.htm> (23.05.2004).

[29] Ferligoj ve Mrvar, "Assesment of Reliability."

[30] R. C ve Diğerleri, "Education, Weschler's Full Scale IQ and g [Eğitim, Wescheler'in IQ ve g.Değeri]," <http://216.239.39.100/search?q=cache:37Ak7UH03lQC:www.udl.es/usuaris/e7806312/grup/rcolom-archi/articulos/Education,%2520IQ%2520and%2520g.pdf+%22general+factor%22+reliability+composite&hl=tr&ie=UTF-8> (19.10.2002).

[31] R. Bobocel, "Reliability [Güvenilirlik]," <http://www.arts.uwaterloo.ca/psychology/courses/Psych492/lect4_492w03.pdf>

[32] RAND, "Mos Psychometric Philosophy For Hrqol Measurement [Mos, Hrqol Ölçümlerinde Psikometrik Felsefe]," <http://www.rand.org/publications/MR/MR162/MR162.ch4.pdf> (25.05.2004).

[33] D.O. Segall, "General Abiliy Measurement [Genel Yetenek Ölçümleri]," <http://segalld.home.netcom.com/GMAT_Segall.PDF.> (19.10.2002).

[34] L.M. Rudner, "Informed Test Compenent Weighting [Test Ögellerini Ağırlıklandırma]," <http://marces.org/mdarch/pdf/m030833.pdf.> (19.09.2002).

[35] Rudner, "Informed Test."

[36] L. M. Rudner, "Informed Test Component Weighting [Test Bileşenlerini Ağırlıklandırma]," <http://marces.org/mdarch/pdf/m030833.pdf> (23.05.2004).

[37] C. Dröge, "How Valid Are Measurements [Ölçümler Ne Ölçüde Geçerlidir?]," <http://www.decisionsciences.org/Newsletter/vol27/27_5/res27_5.htm> (21.09.2002).

[38] A. Yu., "An Introduction to Computing and Interhreting Cronbach Coefficient Alpha in SAS [SAS Yazılımında Cronbach Alfa'nın Hesaplanması ve Yorumlanmasına Giriş];" <http://seamonkey.ed.asu.edu/~alex/pub/cronbach.doc> (12.10.2002).

[39] Schmitt, "Uses and."

[40] J.A. Gliem ve R.R. Gliem "Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales [Likert Tipi Ölçeklerde Cronbach Alfa Güvenilirlik Katsayısının Hesaplanması, Yorumlanması ve Raporlanması]," < http://www.alumni-osu.org/midwest/midwest%20papers/Gliem%20&%20Gliem--Done.pdf> (20.01.2004).

[41] Field, "Reliability Analysis."

[42] A. Christmann ve V. Aelst, "Robust Estimation of Cronbach's Alfa [Güçül Alfa Değeri]," <http://www.pims.math.ca/science/2002/icors/abstracts/AndreasChristmann.pdf> (22.09.2002).

[43] Karl L. Wuensch, "Cronbach's Alpha and Maximized Lambda4 [Cronbach Alfa ve Maksimum Lambda4]," <http://core.ecu.edu/psyc/wuenschk/MV/Alpha.doc> (23.10.2002).

[44] R.A. Yafee, "Common Correlation and Reliability Analysis with SPSS for Windows [SPSS'te Ortak Korelasyon ve Güvenilirlik Analizi]," <http:// www.nyu.edu/its/socsci/Docs/correlate.html> (16.10.2002).

[45] K. Vehkalahti, "Reliability of Measurements Scales [Ölçeklerin Güvenilirliği]," <http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/reliabil.pdf> (24.05.2004).

[46] P. Barrett, "Rejoinder to The Eysenckian Personality Structure [Eysenck'in Kişilik Yapısına Yeniden Katılım]," <http://www.pbarrett.net/Barrett-Ng-Rejoinder-PAID-1999.pdf> (24.05.2004).

[47] Vehkalahti. "Reliability of Measurements."

[48] R. Ho, "Homogeneity, Reliability, Generalizability [Türdeşlik, Güvenilirlik, Genellenebilirlik]," <http://www.psych.uiuc.edu/~rykhlevs/sum/sumchp6.pdf> (12.12.2002).

[49] J. Kehoe, "Basic Item Analysis [Temel Madde Analizi]," <http://ericae.net/pare/getvn.asp?v=4&n=10.> (16.10.2002).

[50] "Reliability [Güvenilirlik]," <http://www.ahs.cqu.edu.au/psysoc/units/53306/pdf/lecturenote2.pdf> (19.05.2003).

[51] "Ways in Which the Realiability of a Measure can be Estimated [Güvenilirlik Ölçülerinin Tahmini]," <http://www-iea.fmi.uni-sofia.bg/Module6/WAYS.HTM> (06.09.2002).

[52] R. Gebotys, "Handout on Reliability [Güvenilirlik Üzerine Broşür]," <http://www.wlu.ca/~wwwpsych/gebotys/book/reliability.pdf> (06.09.2002).

[53] P. Kline, *Handbook of Psychological Testing* [Psikolojik Test Elkitabı], (New York: Routledge, 1993), 11.

[54] K. R., Murphy ve C.O. Davidshofer, *Psychological Testing: Principals and Applications* (Üçüncü Baskı). (Englewood Cliffs, New Jersey: Prentice Hall International. 1994), aktarılan kaynak

<http://www.chandlermacleod.com.au/AssessmentTools_HUMM_Reliability.html> (18.10.2002).

[55] Kline, 11.

[56] R. Gebotys, "Handout On Reliability [Güvenilirlik Üzerine Notlar]," <http://www.wlu.ca/~wwpsych/gebotys/book/reliability.pdf.> (19.04.2003).

[57] D. Garson, "Scales and Standart Measures [Ölçekler ve Standart Ölçümler]," <http://www2.chass.ncsu.edu/garson/pa765/standard.htm> (29.10.2002).

[58] W.E. van Schoor ve Jan C.P.M. Vis, "Political Knowledge of Dutch Parlementary Journalists the Gender Gap Revisited [Hollanda Parlemento Muhabirlerinin Politik Bilgileri: Cinsiyet Farklılığı], <http://www.essex.ac.uk/ecpr/jointsessions/ Copenhagen/papers/ ws17/ schuur.pdf.> (18.10.2002).

[59] J. J. Mondak ve Mary R. Anderson, "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge [Bilgi Açığı: Politik Bilgide Cinsiyet Temelli Farklılıklar]," <http://journalofpolitics.org/Contents/Vol66/arts662/mondak.pdf> (24.05.2004).

[60] B.W. Junker, "Some Applications and Perspectives in Item Response Modelling [Madde Yanıt Modelinde Bazı Uygulamalar ve Görüşler]," <http://www.stat.cmu.edu/~brian/twente/talk1.ps.> (18.12.2002).

[61] Clarion University, "Measures Of Internal Consıstency [İç Tutarlılık Ölçümleri]," <http://spsp.clarion.edu/mm/RDE3/C3/C3Handout32.html> (18.10.2002).

[62] A. J. Fairchild, "Instrument Reliability and Validity [Araç Güvenilirliği ve Geçerliliği]," <http://www.jmu.edu/assessment/wm_library/Reliability_validity.pdf> (25.05.2004).

[63] HumRRO, "Reliability [Güvenilirlik]," "<file:///C:/WINDOWS/Temporary%20Internet%20Files/Content.IE5/C0CJ0T6G/ptcslides%5B1%5D.ppt#364,6,Reliability and Agreement> (10.10.2002).

[64] B. Sekula, "Reliability and Objectivity [Güvenilirlik ve Nesnellik]," <http://pdts.uh.edu/~bsekula/kin4310/Ch%203%20Jackson.htm> (23.10.2002).

[65] Kline, 6.

[66] E. Innes ve L. Straker, "Reliability of Work Related Assesments [İşle İlgili Değerlendidrmelerin Güvenilirliği]," <http://www.curtin.edu.au/curtin/dept/physio/pt/staff/straker/publications/1999Work4Reliability.html> (08.06.2003).

[67] W.G. Hopkins, "Precision of Measurement [Ölçümün Hassaslığı]," <http://www.sportsci.org/resource/stats/precision.html> (27.11.2002).

[68] Kline, 7.

[69] Aynı.

[70] R.A. Charter, "Sample Size Requirements for Precise Estimates of Reliability, Generalizability, and Validity Coefficients [Güvenilirilik Tahminlerinde Örneklem Büyüklüğü Gerekliliği]," <http://www.szp.swets.nl/szp/journals/jc214559.htm> (13.12.2002).

[71] Questionmark, "Testing and Assesment of Glossary of Terms [Test ve Değerleme Terimleri]," <http://www.questionmark.com/uk/glossary.htm> (24.09.2002).

[72] Aynı.

[73] Clarion, "What Different "Reliability" Coefficients Assess [Farklı Güvenilirlik Katsayıları Neyi Değerliyor?]," <http://spsp.clarion.edu/mm/RDE3/c3/Handout31Reliability.html> (14.09.2002).

[74] D.N.M. de Gruijter ve L.J. Th. Van der Kamp, "Statistical Test Theory For Education and Psychology [Eğitim Bilimleri ve Psikoloji İçin Test Kuramı]," < http://icloniis.iclon.leidenuniv.nl/gruijter/statistical%20test%20theory%20for%20education %20and%20psychology.pdf> (19.05.2003).

[75] P.F. Sanders ve A.J. Verschoor, "Parallel Test Construction Using Classical Item Parameters [Kasik Madde Parametrelerini Kullanarak Paralel Test Oluşturma]," <http://download.citogroep.nl/pub/pok/reports/Report96-4.pdf> (25.08.2002).

[76] Aynı.

[77] Kline, 12-13.

[78] StatSoft, "Reliability and Item Analysis [Güvenilirlik ve Madde Analizi],"<http://www.statsoftinc.com/textbook/streliab.html>(23.09.2002).

[79] Modelin uygulanma biçimi için *bk.*, "Creating Parallel Forms Using IRT [Madde Yanıt Kuramına Göre Paralel Formlar Yaratma]," <http://work.psych.uiuc.edu/irt/par_forms.asp> (10.10.2002).

[80] HumRRO, "Reliability [Güvenilirlik]," "<file:///C:/WINDOWS/Temporary%20Internet%20Files/Content.IE5/C0CJ0T6G/ptcslides%5B1%5D.ppt#364,6,Reliability and Agreement> (10.10.2002).

[81] Vocational and Rehabilitional Institute, "CET Reliability Assesment [CET Güvenilirlik Değerlendirmesi], <http://www3.gov.ab.ca/pdd/docs/prov/CET%20Report.pdf.> (08.10.2002).

[82] J.Wilde , "Assessment Strategies for Professional Development Activities [Mesleki Gelişme Etkinlikleri İçin Değerleme Stratejileri]," <http://www.ncela.gwu.edu/miscpubs/eacwest/profdev/part2.htm#To%20Ensure> (26.09.2002).

[83] HumRRO, "Reliability."

[84] Aynı.

[85] Trochim, "Types of Reliability [Güvelirlik Türleri]," <http://trochim.human.cornell.edu/kb/reltypes.htm> (08.09.2002).

[86] Nunnally, 210.

[87] "Household Foot Security in USA [ABD'de Hanehalkının Gıda Güvenliği]," <http://www.fns.usda.gov/oane/MENU/Published/FoodSecurity/TECH_RPT.PDF> (21.09.2002).

[88] Nunnally, 214.

[89] "Household Foot Security ."

[90] Nunnally, 215.

[91] R. Ho, "Reliability Theory For Total Test Scores [Toplam Test Puanları İçin Güvenilirlik Kuramı]," <http://www.psych.uiuc.edu/~rykhlevs/sum/sumchp5.pdf> (13.12.2002). Ayrıca *bk.*, <http://eval1.crc.uiuc.edu/edpsy392/lecture4/lecture4.ppt>,

[92] DSS Research, "Validity and Reliability [Geçerlilik ve Güvenilirlik]," <http://www.dssresearch.com/library/general/validity.asp> (19.05.2003).

[93] Yaşar Baykul, *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*, (Ankara: ÖSYM, 2000), 143.

[94] T.E. Dawson, "Basic Concepts in Classical Test Theory: Relating Variance Partitioning in Substantive Analyses to the Same Process in Measurement Analyses [Klasik Test Kuramında Temel Kavramlar], <http://128.8.182.4/ft/tamu/Dawson.pdf> (19.05.2003).

[95] Hal S. Kopeikin, "Correlation and Regression [Korelasyon ve Regresyon Analizi]," 2000, <http://www.psych.ucsb.edu/~kopeikin/121lec04.htm> (19.05.2003).

[96] ERIC, "How High Should Reliability Be [Güvenilirlik Katsayısı Ne Kadar Yüksek Olmalıdır]," <http://www.ed.gov/databases/ERIC_Digests/ed458213.html> (14.09.2002).

[97] L.A. Becker, "Reliability and Validity [Güvenilirlik ve Geçerlilik]," <http://web.uccs.edu/lbecker/Psy590/relval_I.htm> (14.09.2002), Reliability Standards.

[98] J.C. Nunnally ve I.H. Bernstein, *Psychometric Theory*, (New York: MacGraw-Hill, 1994).

[99] A. Yu, "Reliability and Validity of Standardized Tests [Standart Testlerin Güvenilirlik ve Geçerliliği]," <http://seamonkey.ed.asu.edu/~alex/teaching/assessment/reliability_standard.html> (28.12.2002).

[100] Peterson, 381.

[101] Aynı.

[102] Aynı.

[103] "Appendix C: Statistical Information [Ek C: İstatistiksel Bilgi]," <http://www.facs.gov.au/internet/facsinternet.nsf/VIA/nshsph2001/$File/AppC.doc>(10.10.2002).

[104] C. Dröge, "How Valid Are Measurements [Ölçümler Ne Ölçüde Geçerlidir?]," <http://www.decisionsciences.org/Newsletter/vol27/27_5/res27_5.htm> (21.09.2002).

[105] ERIC, "How High Should."

[106] L.S. Wang, "Reliability [Güvenilirlik]," <http://psy.ccu.edu.tw/testroom/Reliability_Part_Two.doc> (07.10.2002).

[107] Wang, "Reliability."

[108] R. Ping, "Testing Latent Variable Models With Survey Data [Alan Araştırmalarında Gizli Değişkenin Test Edilmesi]," <http://www.wright.edu/~robert.ping/lv/v.doc> (12.12.2002).

GİRDİ KALİTESİNİN DEĞERLENDİRİLMESİ İÇİN VERİ TARAMASI

Güvenilirlik yöntemi belirlendikten sonra bilim adamı verilerin analizi için matematiksel veya istatistiksel çözümleme yöntemlerine başvurur. Çoğunlukla her iki yöntem birlikte kullanılır. Ancak güvenilirliği doğrudan test edecek veya hesaplayacak teknikler kullanılmadan önce verilerin genel olarak kalitesinin değerlendirilmesi gerekir. Veri kalitesini değerlendirme çalışmalarına *veri taraması* veya *veri incelemesi* adı verilir. Veri taramasında veri tashihi, istatistiksel özet analizleri ve normallik testleri yapılır, eksik verilerin ve ayrık değerlerin bulunma durumu araştırılır, verilerin türdeşselliği, ayrısallığı ve çoklu doğrusallık özelliği incelenir. Bu tür araştırma, inceleme ve iyileştirmelerin sonucunda bilim adamı nitelikli verilerle çalışma olanağına kavuşmuş olur.

VERİ TASHİHİ

Değişik ölçüm araçlarıyla toplanan verilerin güvenilirlik analizleri için ya istatistik kitaplarındaki formüllerden veya bu konuda özel olarak hazırlanmış bulunan istatistikî analiz programlarından yararlanılır. Bu bölümde formüllerin uygulanması yerine, istatistiksel analiz yazılımlarının kullanılması konusuna ağırlık verilmiştir. İstatistikî analiz programlarından bazıları genel amaçlıdır ve bu programlarda *güvenilirlik analizleri* sınırlı ölçüde yapılır. SPSS, SAS, Systat, NCSS ve Statistica gibi yaygın kullanılan yazılımlar bu gruba girer. Diğerleri ise *klasik test kuramına* göre yapılan madde analizi, alfa, KR-20 gibi yöntemlerin yanı sıra *modern test kuramına* göre de maddelerin geçerlilik ve güvenilirlik analizlerini yapmak amacıyla geliştirilmiştir. Özel istatistik yazılımları adını verdiğimiz bu programlar kitabın daha sonraki bölümlerinde ayrıntılı olarak tanıtılmıştır.

Çalışmalarında istatistiksel analiz programlarını kullanarak güvenilirlik analizi yapmak isteyen araştırmacılara, veri taraması işleminden önce aşağıdaki adımları atmalarını öneririz.

1. Geliştirdiğiniz ölçüm modeliyle ilgili olarak bağımsız veya tesadüfî değişkenlerinizi (demografik değişkenler veya diğer ölçüm değişkenleri olabilir) programın veri yükleme çizelgesine giriniz.
2. Geliştirdiğiniz ölçüm modeliyle ilgili olarak varsa sabit (kontrol) değişkenlerini programın veri yükleme çizelgesine giriniz.
3. Ölçeğe ait tüm değişkenlerin (maddelerin) her birisiyle ilgili sayısallaştırmış olduğunuz kodları programın veri yükleme çizelgesine tanıtınız.
4. Ölçekte alt boyutlar (faktörler) varsa bu faktörlere ait toplam puanları ayrı bir değişken olarak tanımlayınız veya bu puanları bilgisayar ortamında ayrıca toplatınız.
5. Ölçeğin *genel toplam puanını* ayrı bir değişken olarak (bağımlı veya duruma göre bağımsız değişken olabilir) tanımlayınız.
6. Verilerin doğru girilme durumunu kontrol ederek hataları düzeltiniz ve verileri işlem yapacak duruma getiriniz.

Araştırmacı, verilerini bilgisayara tanıttıktan sonra, istatistiksel analizlere geçmeden önce verilerin doğru girilip girilmediğini belirlemek üzere veri kontrolü yapmalıdır. Veri kontrolü, *çift giriş* veya *tek giriş* sistemine göre yapılır. Tek giriş sisteminde değişkenler için min – maks analizi[a] ve toplam puanlar için ise bilgisayar ortamında toplama veya kağıt üzerindeki verilerle bilgisayardaki verilerin karşılaştırması yapılır. Çift giriş işleminde ise veriler bilgisayara girildikten sonra özel bir yazılım kullanılarak ikinci bir kez daha girilir. İkinci giriş sırasında veya birinci girişte bir aykırılık varsa yazılım otomatik olarak bu hataları göstermekte ve hatanın giderilmesine imkan sağlamaktadır.[1] Veri tashihi aşamasında negatif maddeler eğer daha önceden tersine çevrilmemişse bilgisayar ortamında tersine çevrilir.[b] Veri giriş hatalarının düzeltilmesinden sonra bilim adamı yapacağı güvenilirlik analizleri için belirli bir plana göre hareket eder. Bu plan üç aşamadan oluşur: (a) veri taraması, (s) güvenilirlik ve geçerlilik analizleri,

---

[a] Bazı kaynaklarda bu uygulama *ranj testi* (range test) olarak isimlendirilmiştir.

[b] Verileri bilgisayar ortamında tersine çevirmek için SPSS'te şu işlemlere başvurulur: Transform mönüsünden, Recode alt mönüsüne girilir. Buradan "into same variables" şıkkı seçilir. Daha sonra "old and new values" penceresine girilerek eski ve yeni değerler belirlenir ve Add tuşuna basılır. Ters dönüştürme işleminde bir diğer yaklaşım, dönüştürme formülünden yararlanmaktır. Yeni değer = (en yüksek değer + 1) – original değer.

(c) hipotez testleri. Bu bölümde özellikle ve sadece veri taraması konusu üzerinde durulmuştur.

Veri taraması, birbirinden farklı birkaç aşamada gerçekleştirilir. Bu aşamalar şunlardır: (a) özet analizleri, (b) normallik analizleri, (c) eksik veri analizleri, (ç) türdeşsellik ve doğrusallık analizi, (d) koşutluk analizi, (e) ayrık değer analizleri, (f) veri standardizasyonu.

ÖZET ANALİZLERİ

Veri kümelerinin sıklık değerlerini, merkezî dağılım değerlerini (aritmetik ortalama, mod, medyan) ve değişkenliklerini (frekans dağılımları, minimum ve maksimum değerler, değişim aralığı, kartiller arası aralık, standart sapma, varyans ve kovaryans) açıklayan istatistikî teknikler *özet analizleri* olarak isimlendirilir. İstatistikî özet analizleri, tanımlanan ölçüm değişkenleri tek tek ele alınarak yapılır. Tek bir değişkene dayalı olarak yapılan basit istatistikî teknikleri uygulamadan bir araştırma problemini çözmeye çalışmak veya maddelerin toplam puanlarını, ortalama puanlarını, güvenilirlik ve geçerlilik analizleri sonuçlarını raporlamak doğru değildir.

Frekans Dağılımları

Frekans dağılım tabloları, verilerin güvenilirliği hakkında bilgi vermez. Ancak en önemli yararı verilerin dağılımı hakkında araştırmacıya belirli ipuçlarını sağlamasıdır. Dikkatli bir gözlemci frekans tablosuna bakarak verilerin dağılım biçimini anlayabilir.

Değişkenlerin frekans tabloları, ayrıca *ayrık* değerlerin tespit edilmesine imkan sağlar. Ayrık değerler, kontrol edemediğimiz tesadüfî hata öğeleridir. Ayrık değerleri otomatik olarak eleyecek, ortadan kaldıracak bir yöntem bulunmamakla birlikte *kutu grafiği* ile veya *nokta dağılım grafiği* üzerinde bu değerlerin hangi değişkenlerde bulunduğunu kolaylıkla saptayabiliriz.

Frekans dağılımlarının bir diğer yararı, değişkenlerdeki eksik verileri göstermesidir. Böylece hangi değişkende yüzde kaç oranında eksik veri bulunduğu saptanmış olur.

Merkezî Eğilim ve Değişkenlik Ölçüleri

Araştırmacının özet analizleri kapsamında kullanabileceği bir diğer hesaplama yöntemi, *merkezî eğilim* ve *değişkenlik* ölçüleridir. Eşit aralıklı ölçek verilerinde bu ölçülerin en önemli ikisi aritmetik ortalama ve standart sapmadır. Bu değerler (a) maddeler için ve (b) ölçek/test toplam puanları için ayrı ayrı hesaplatılır. İlk aşamada her bir maddenin önce aritmetik ortalama-

sı ve daha sonra standart sapması incelenir. Örneğin, *i*'ninci maddenin aritmetik ortalaması Eşitlik 4-1'deki gibi hesaplanır:

$$M_i = \frac{\sum X}{n}. \qquad (4\text{-}1)$$

Merkezî dağılım ölçüsü, nominal verilerde mod ve sıralı ölçek verilerinde ise medyandır. Aritmetik ortalama, sadece eşit aralıklı ve oranlı ölçek verilerinde kullanılır. Ancak eşit aralıklı ve oranlı ölçek verileri önemli ölçüde çarpık ise aritmetik ortalama yerine medyan değeri temel alınır. Beş veya yedi dereceli Likert tipi maddelere ait veriler *sıralı ölçek* verisi niteliğindedir. Bu maddeler normal dağılım özelliği ile ilgili varsayımları karşılıyorsa eşit aralıklı ölçek verileri için uygun olan istatistikî analizlerin bu veriler için de uygulanabileceği belirtilmiştir.[a] Bununla birlikte böyle bir analiz uygulandığında, elde edilen sonuçlar araştırmacı için yararlı bir bilgi ve tahmin imkanı sağlıyor olmalıdır.[2] Likert tipi maddelerde aritmetik ortalama ve standart sapma hesaplamaları incelenen sosyal ve davranışsal olay hakkında yargıya varmak için değil, veri kalitesini değerlendirmek için kullanılırsa anlamlıdır.

Sürekli veri niteliğindeki maddelerin değişkenlik ölçüsü standart sapmadır. Maddelerin ve test puanlarının standart sapma değerlerinin hesaplanma amacı, değerlerin ortalama etrafındaki dağılımını, değerlerin minimum ve maksimum puanlar arasındaki değişkenliğini görmektir. Standart sapma, veriler normal dağılım özelliği gösteriyorsa anlamlıdır. Standart sapmanın yüksek olması maddelerin veya testin kişileri daha iyi ayırt ettiğini gösterir ve ölçüm yapılan kişilere ait puanların birbirinden önemli ölçüde farklı olduğu anlamına gelir. Standart sapma, varyansın karekökü alınarak bulunur. Varyans değerinin tersine kişilerin puanlarını aritmetik ortalama değeriyle karşılaştırma yaparak değerlendirme imkanı sağlar. Standart sapma, dizideki ayrık değerlerden büyük ölçüde etkilendiğinden hesaplama yapmadan önce frekans dağılımı ile ayrık değer bulunma durumu kontrol edilmelidir. Araştırmacı standart sapma ile puanların, %68,26'sının 1 standart sapma, %95,44'ünün 2 standart sapma ve %99,73'ünün 3 standart sapma diliminde

---

[a] *Journal of Agricultural Education* isimli derginin 27. ilâ 32. ciltlerinde yayımlanan 188 araştırmanın 95'inde Likert tipi maddelerin kullanıldığı ve bunların da %54'ünde, ölçek maddelerinin analizinde aritmetik ortalama ve standart sapma gibi tanımlayıcı istatistiksel analizlerden yararlanıldığı belirtilmiştir. (*bk.*, D. L. Clason ve T. J. Dormody, "Analyzing Data Measured by Individual Likert–Type Items," <http://pubs.aged.tamu.edu/jae/pdf/Vol3-5/35-04-31.pdf.> (07.02.2003).

hangi değerleri alabileceğini görme şansına sahip olur. Standart sapma, Eşitlik 4-2'deki formül ile hesaplanır.

$$SS_i = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}. \qquad (4\text{-}2)$$

Maddelerin ortalamaları ve standart sapmaları yapılan ölçümün niteliğine göre değişik nedenlerle hesaplatılabilir. Bunlardan birincisi, başarı testlerinde maddelerin zorluk durumunu saptamaktır. İkincisi ölçeklerde maddelerin tau eşitliğine veya paralel yapıya sahip olup olmadıklarını anlamaktır. Ölçeklerde eğer maddeler yaklaşık olarak paralel ise aritmetik ortalamaları ve standart sapmaları birbirine yakın çıkar. Başarı testlerinde değişik zorluk derecelerindeki maddelerin aritmetik ortalamaları ve standart sapmaları Tablo 4-1'deki gibi belirlenir:[3]

Tablo 4-1. Ölçüm Verilerinin Niteliğine Göre Güçlük Değerleri

	İkili veriler	Sürekli veriler
Ortalama güçlük derecesi Ortalama başarı	,30 ilâ ,70 arasında	5 dereceli Likert □=3 $SS \cong 1{,}0$
Güç maddeler Düşük başarı	<,30	5 dereceli Likert □=1,5 $SS \cong 0{,}50$
Kolay maddeler Yüksek başarı	>,70	5 dereceli Likert □=4,5 $SS \cong 1{,}50$

Kaynak. "Item Analysis [Madde Analizi]," <http://www.sph.uth.tmc.edu: 8053/behsci/lmasse/ph1130/Item%20analysis.rtf> (28.12.2002).

Tablo 4-1'e göre 5 dereceli bir ölçekte madde ortalamaları eğer 3,0 ilâ 3,5 gibi değerler arasında kalıyorsa bu maddelerin ayırt etme gücü yüksektir. Madde ortalamaları 4,5 veya 1,2 gibi uç değerlere yakın çıkmışsa serideki dağılımın sağa veya sola dayalı olduğu anlaşılır. Bu tür maddeler sorunludur, yanıtlayıcıları iyi bir şekilde ayırt etmiyor demektir. Söz konusu maddeler eğer dikkatli bir şekilde işaretlenmişse anket ya iyi bir şekilde tasarlan-

mamıştır veya seçilen örneklem yanlıdır. Aritmetik ortalama değeriyle standart sapma değerleri arasında da bir ilişki vardır. Bazı yazarlar bu ilişkiyi belirli oranlarla ifade ederler. Ölçekteki derece sayısının 1/6'sı maksimum standart sapma değeri olarak kabul edilir. Örneklem verilerinin normallikten sapma gösterdiğini belirlemede kullanılan bir diğer yaklaşım standart sapmanın, aritmetik ortalama değerinin 1/3'ünden daha yüksek çıkmasıdır.

Tek örneklem verilerine dayalı olarak yapılan incelemelerde aritmetik ortalama ve standart sapma değerleri yeterince bilgi vermez. Maddelerin standart sapma değerleri birden fazla örneklem verileriyle veya norm değerleriyle karşılaştırma yapıldığı zaman daha anlamlıdır.

Bilgi testlerinde aritmetik ortalama sınıfın genel başarı düzeyini gösterir ve bu değer sınıfın üst çeyreğinde (veya %27'lik diliminde) yer alan başarılı öğrencilerin ve daha sonra alt çeyreğinde (veya %27'lik diliminde) yer alan başarısız öğrencilerin ortalamalarıyla karşılaştırılır. Başarılı öğrenciler bir maddeyi doğru yanıtlamışlarsa bu maddenin ortalaması genel sınıf ortalamasından yüksek; başarısız öğrenciler bir maddeyi doğru yanıtlamışlarsa o maddenin ortalaması genel başarı ortalamasından düşük çıkmalıdır. Bu koşul sağlanabiliyorsa maddenin ayırt etme yeteneğine sahip olduğu söylenir. Maddenin standart sapmasının düşük çıkması, cevaplayıcıların hepsinin benzer yanıtlar verdiği anlamına gelir ve bu durum istenmez.

Ölçeğin/testin toplam puanlarının ortalaması ve standart sapması da birkaç nedenle hesaplatılır. Araştırmacı test-yeniden test uygulamalarında istikrarlılık güvenilirliğini aritmetik ortalamaların değişmezliği ile takip eder. Ana kütleden seçilen çok sayıda örneklemden elde edilebilecek ortalamalar x'in *beklenen değeri* olarak adlandırılır. Beklenen değer ana kütle ortalamasına (μ) yakındır. Standart sapmanın düşük çıkması toplam puanların birbirine benzer olduğunu ve puan dağılımlarının ortalama etrafında yoğunlaştığını gösterir. Bir test ve ölçeğe ait standart sapmanın düşük değil, yüksek çıkması arzulanır. Standart sapmanın yüksek olması puanların heterojen bir dağılıma sahip olduğunu, bireylerin farklılaştığını gösterir. Standart sapma, ayrıca *ölçümün standart hatasını* belirlemek için kullanılacak olan bir değerdir. Yüksek güvenilirlik için, ana kütleden seçilecek değişik örneklemlere ait ortalama ve standart sapma değerlerinin birbirine benzer çıkması istenir. Örneklemlerin çoğu, farklı istatistikî değerlere sahipse güvenilirlik düşüktür.

Bilim adamı çıkış noktası olarak madde-yanıt kuramını temel almışsa bu kuramda maddelerin aritmetik ortalama ve standart sapma değerlerinin hesaplanması istenmez. Eğer maddelerin aritmetik ortalama değerleri raporlanmak isteniyorsa ya medyan değerleri temel alınır veya ham puanlar standart z puanlarına dönüştürüldükten sonra bu puanların ortalaması alınır. Pu-

anlar arasındaki değişkenliği görmek için ise, *çeyrek dilimler arası aralık* (interkartil) değeri hesaplanır.

Varyans Değerleri

İlk kez Fisher (1918) tarafından kullanılan varyans terimi, puanların yayılım veya dağılım ölçüsünü verir. Varyans değerinin büyüklüğü puanların ortalama etrafında geniş bir dağılıma sahip olduğu anlamına gelir. Varyans değerinin küçüklüğü ise ortalama etrafındaki daha dar bir alanı temsil eder. Küçük varyans değeri, tahmin açısından ortalamanın daha *kesin* bir değer ifade etmesi anlamına gelir. Ancak, varyans değeri standart sapmada olduğu gibi ham puanların %68'inin hangi puanlar arasında yer alabileceği konusunda bilgi vermez. Bu açıdan standart sapmaya göre teorik bir değerdir ve bilimsel araştırmalar ile güvenilirlik analizlerinde daha çok varyans değerleri kullanılır. Varyansı hesaplamak için Eşitlik 4-3'teki formülden yararlanılır.

$$V = \sigma^2 = \frac{\sum x^2}{n}. \qquad (4\text{-}3)$$

V = Varyans.
x = $X - M$ (Aritmetik ortalamadan sapmalar).
x^2 = Sapmaların karesi.
Σx^2 = Sapmalara ait kareler toplamı.
X = Gözlem değerleri.
M = Aritmetik ortalama.

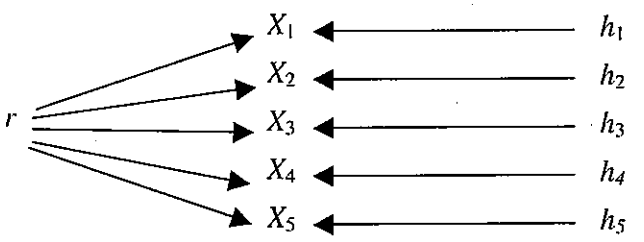
Varyans türleri. Literatürde varyans değerleri değişik şekillerde karşımıza çıkar. Hata varyansı, varyans analizi, varyans bileşenleri, ana kütle varyansı terimleri bunlardan birkaçıdır. Bilim adamının, kullanım amacına ve yerine göre varyans değerlerine doğru anlam vermesi gerekir. Varyans türleri literatürde beş grupta ele alınmıştır.

Ana kütle ve örneklem varyansı. Ana kütle varyansı belirli bir özellikle ilgili olarak ana kütledeki tüm bireylerde ölçüm yapılması halinde ortaya çıkabilecek değişkenliktir. Böyle bir ölçümü yapma işleminde zorluk bulunması nedeniyle ana kütle varyansı, örnek kütlede yapılan ölçümlere bağlı olarak saptanır. Örnek kütle verilerinden elde edilen değişkenlik, örnek kütle

varyansıdır. Ana kütleden seçilecek her bir örnek kütlede varyans değerleri farklı çıkar. Değişik örneklem grupları tesadüfî olarak seçilmişse ve yeterli büyüklüklere sahipse *örneklem ortalamalarının varyansı* nispeten küçük çıkmalıdır. Ancak ana kütleden çok sayıda örneklem seçilememesi nedeniyle ana kütle varyansı tek bir örneklem verilerine bağlı olarak tahmin edilir. Bunun için örneklem verilerinde *ortalamanın standart hatası* değeri hesaplatılır. Ortalamanın standart hatası, örneklem verilerinden hareket ederek ana kütle varyansını verir.[4] Ana kütle varyansı σ^2 ve örneklem varyansı ise s^2 simgesiyle gösterilir.

Sistematik varyans. Bilinen veya tahmin edilebilen nedenlerle puanların sağa veya sola çarpık çıkmasıdır. Doğal nedenler veya insandan kaynaklanan tahmin edilebilir nedenler bazen puanlarda belirli bir şekilde değişkenliğe yol açar. Örneğin, üniversitelerde birinci öğretimde okuyan öğrencilerle ikinci öğretimde okuyan öğrencilerin başarı puanları aynı değildir. Birinci öğretime devam eden öğrencilerin başarı puanları sistematik bir şekilde daha yüksektir. Sistematik varyansın pek çok nedeni olabilir. Bilim adamı araştırma sonuçlarını etkileyecek sistematik varyans etkenini göz önünde bulundurmalı ve bu varyans faktörüne dikkat etmelidir. Güvenilirlik analizlerinde sistematik varyans, düşük güvenilirlik nedenidir.

Hata varyansı. Ölçümlerde şans faktörü veya dış etkenler nedeniyle ortaya çıkan dalgalanmaları, puan değişikliklerini ifade eder. Puan farklılıkları bütünüyle tesadüfîdir. Değişiklikler küçük, önemsiz, tekrar etmeyen nitelikte ve birbirini nötr haline getiren türdendir. Hata varyansının pratik bir tanımı, bilinen bütün sistematik varyanslar ölçümden çıkarıldıktan sonra geriye kalan varyanstır. Literatürde bu varyans aynı zamanda "tesadüfî hata" olarak isimlendirilir. Herhangi bir ölçümdeki toplam varyans, sistematik varyans ve hata varyansından oluşur. Güvenilirlik katsayılarının 1'den çıkarılmasıyla elde edilen değer hata varyansını verir. Örneğin, güvenilirlik katsayısı ,85 çıkmışsa bu ölçümde ,15 hata varyansı bulunduğu anlamına gelir. Tau eşitliğine sahip testlerde gerçek puanlar eşit olmakla birlikte hata varyansları farklıdır. Hata varyanslarının farklı olması nedeniyle gözlem puanlarının varyansları da eşit değildir. Kişiler arasında hata varyansının beklenen değeri sıfırdır. Ayrıca hata varyansları gerçek puanlarla ve diğer maddelerin hata terimleriyle de ilişkili değildir (*bk.*, Şekil 4-1).



Şekil 4-1. Hata varyansları-gerçek puan ilişkisi.

Kaynak. "Measurement Theory and Measure Evaluation [Ölçüm Kuramı ve Ölçüm Değerlendirmesi]," <https://depts.washington.edu/hservdoc/ HSERV_590_H/ 3RA E_ survey_sessions.pdf> (28.12.2002).

Gerçek puan varyansı. Gerçek puanlarda gözlenebilecek değişkenliktir. Gözlem puanları, gerçek puan varyansı ile hata varyansını birlikte içerir. Gerçek varyans, test edilen örnekleme aittir, hata varyansı ise daha çok ölçüm aracıyla ilgilidir. Güvenilirlik; teorik olarak "gerçek puan varyansının gözlem puanları varyansına olan oranı" şeklinde tanımlanmıştır.

Eğer iki test birbirine tam olarak paralel ise, ana kütle varyansı ile gözlem puanlarının varyansı birbirine eşittir. Çünkü paralel testlerde maddelerin gerçek puan varyansları da birbirlerine eşittir.

Gruplar arası varyans. Değişik gruplarda yapılan ölçümlerde grup puanları arasındaki değişkenliği tanımlar. Bir gruba ait puanların diğer gruba ait puanlardan ne ölçüde farklı olduğu hakkında bilgi verir. Teknik anlamda ifade etmek gerekirse grupların toplam puanlarının, ortalamadan sapmalarının karelerinin toplamının grup sayısına bölünmesidir.

Varyans ve testler. Bilgi ve yetenek testlerinin, tutum ve kişilik envanterlerinin esas amacı bireyler arasındaki farklılıkları ortaya çıkarmaktır. Bu nedenle, bu testlerde ölçümün toplam veya ortalama puanları önemlidir. Klasik test kuramında ölçüm verilerinin güvenilirlik ve geçerliliğine içerdiği tek tek maddelerin varyansından çok toplam (veya ortalama) puanların

varyansına bakılarak karar verilir. Bu nedenle araştırmacı klasik test kuramında toplam/ortalama puanların varyansını maksimize etmeye çalışır.

Bir ölçek veya testte madde sayısı arttıkça (negatif korelasyona sahip olmadığı sürece) toplam test puanının varyansı artar. Testin geçerlilik ve güvenilirliği bu varyansa bağlıdır. Bireysel maddelerin varyanslarıyla toplam puanların varyansları arasında bir şekilde bir ilişki olmalıdır. Toplam test puanlarının varyansı, maddelerin zorluk dereceleri eşit olduğu zaman artar. Çünkü böyle bir durumda maddeler arasındaki ortak varyans (kovaryans), ortak değişkenlik artar. Maddelerin "orta" güçlük derecesinde olması varyansı artırır (*bk.*, Tablo 4-2). Varyansın büyüklük derecesinin tam olarak ne olması gerektiğini belirleyen bir kriter / standart yoktur. Varyansın büyüklüğü ancak ölçekten maddeler çıkarılarak veya paralel ölçek varyansları arasında karşılaştırmalar yapılarak değerlendirilebilir.

■ Toplam test puanının varyans formülü Eşitlik 4-4'teki gibidir.

$$\sum_{j=1}^{J} \sigma_j^2 + 2 \sum_{i<j}^{J(J-1)/2} \sigma_{ij}^2 \quad . \tag{4-4}$$

Maddelerin varyanslarının toplamı + (2 x Madde kovaryanslarının toplamı).

Tablo 4-2. Maddelerin Güçlük Oranı ile Varyans Arasındaki İlişkiler

Güçlük oranı (p)	Varyans (S^2)
,10	,09
,20	,16
,30	**,21**
,40	**,24**
,50	**,25**
,60	**,24**
,70	**,21**
,80	,16
,90	,09
1,0	,00

Eşitlik 4-4'teki formülü açarak daha net bir şekilde Eşitlik 4-5'teki gibi ifade edebiliriz.

$$S_{x+y+z} = S_x^2 + S_y^2 + S_z^2 + 2(C_{xy} + C_{yz} + C_{xz}). \quad (4\text{-}5)$$

S_{x+y+z} = Üç maddeden oluşan bir ölçeğin varyansı.
S_x^2 = x maddesinin varyansı.
C_{xy} = xy maddeleri arasındaki kovaryans.

Bilim adamı hesaplama sonucunda maddelerin ve teste ait toplam puanın aritmetik ortalama, standart sapma ve varyans değerlerini hazırlayacağı bir tablo üzerinde okuyucuların bilgisine sunmalıdır.

Kovaryans Değerleri

Kovaryans, iki ölçüm verisi, madde veya değişken puanları arasındaki değişkenliğin paylaşılma derecesini tanımlar. Literatürde, iki değişkenin eşleşmesindeki sapma olarak da ifade edilmiştir. İki değişken arasındaki korelasyon yüksekse kovaryans değeri de yüksek çıkar (tersi de doğrudur). Fakat, bilimsel araştırmalarda rakamları daha anlamlı hale getirmek için kovaryans değerlerinden çok korelasyon rakamları kullanılır. Bir ölçümde X ve Y değişkenlerinin standardize edilmiş puanları arasındaki kovaryans değeri korelasyon katsayısını verir. Diğer bir deyişle kovaryans, standardize edilmemiş korelasyon katsayısıdır. Çok maddeli bileşik bir ölçümün varyansı, madde varyansları toplamıyla maddeler arası kovaryansların toplamının birleşimine eşittir. Bir testin veya ölçeğin toplam puanlarının varyansı tek tek maddelerin varyansından değil, maddelerin ikişerli karşılaştırılmasına dayanan kovaryans değerlerinden etkilenir. Kovaryans test maddelerindeki ortak özü gösterir. Testin varyansını etkileyen bir diğer faktör testteki madde sayısıdır. Alfa hesaplamasında, ikişerli olarak maddeler arasındaki kovaryans zayıf ise, alfa değeri negatif çıkar.

Varyans grafiksel olarak normal dağılım eğrisi ile gösterilirken iki madde arasındaki kovaryans üç boyutlu bir grafik üzerinde bir tepe şeklinde gösterilir. Maddelerin puanları standardize edildikten sonra tepenin uç noktası sıfır veya sıfıra yakın ise ve etekleri dört bir taraftan dengeli bir şekilde dağılıyorsa "bu iki madde birbiriyle tutarlıdır" denir. Maddeler arasındaki kovaryans Eşitlik 4-6 ile hesaplanır.

$$\text{cov}_{xy} = \frac{\sum xy}{n}. \qquad (4\text{-}6)$$

cov_{xy} = x ve y maddelerinin kovaryansı (ortak değişkenliği).
x = x maddesindeki ortalamadan sapmalar ($X - \overline{X}$).
y = y maddesindeki ortalamadan sapmalar ($Y - \overline{Y}$).
n = Toplam ölçüm yapılan kişi sayısı.

Kovaryansın belirli özellikleri vardır. Eğer madde i ve madde j birlikte artış eğilimi gösteriyorlarsa, biri artarken diğeri de artıyorsa $\text{cov}(i,j) > 0$'dır. Madde i azalış gösterirken madde j artış eğilimi gösteriyorsa $\text{cov}(i,j) < 0$'dır. Eğer madde i ve madde j birbirinden bağımsız ise $\text{cov}(i, j) = 0$'dır. Buna göre maddeler arasındaki kovaryans pozitif ise veya diğer bir deyişle sıfırdan büyükse testteki madde sayısı arttıkça testin varyansı da artış gösterir. Ölçek ve testlerde hesaplanan varyansın büyük bir bölümü maddelerin varyansından değil kovaryansından gelir. Beceri testlerine göre, tutum ölçeklerinde maddeler arasındaki kovaryans değerleri daha yüksektir. Kovaryans değerlerinin yüksek olması ölçeğin arka planda gizli bir yapıyı ölçtüğünü gösterir. Kovaryansın, korelasyon katsayılarında olduğu gibi alt ve üst sınır değerleri yoktur ve bu nedenle bu belirlisizlik en önemli eksikliğini oluşturur. İstatistiksel analiz programı SPSS'te kovaryans farklı iki alt mönüden hesaplatılabilir: birincisi Correlations ve ikincisi ise Reliability mönüleri. Klasik güvenilirlik analizlerinde kovaryans matrisini hesaplatmanın pratik bir faydası yoktur, ancak gizli kavramsal yapıları ortaya çıkarmak isteyen veya yapısal eşitlik modelini kullanmak isteyen araştırmacıların maddelerle gizli yapı arasındaki doğrusal olmayan ilişkileri ve etkileşimleri görebilmeleri için kovaryans matrisini kullanmaları gerekir. Ortak faktör analizi yöntemini kullanarak bir ölçüm aracının arka planındaki gizli yapıları ortaya çıkardığını iddia eden bir araştırmacının raporunda ölçeğin söz konusu maddelerine ilişkin kovaryans matrisini vermesi gerekir. Kovaryans matrisi kaba ölçülerle de olsa ölçüm aracının faktöriyel yapısını ortaya koyar.

PUANLARIN NORMAL DAĞILIM ÖZELLİĞİ

Ölçüm verilerinin normal dağılım özelliğine sahip olup olmadıklarının belirlenmesi normallik testleri ve normallik grafikleriyle araştırılır. Aşağıdaki bölümde önce testler ve daha sonra grafikler ele alınmıştır.

Normallik Testleri

Puanların normal dağılım özelliği normallik testleri ile araştırılır. Veriler; (a) ayrık değerlere sahip olabilir, (b) çarpık olabilir, (c) basık olabilir veya (ç) çarpıklık ve basıklık karakterinin her ikisini de gösteriyor olabilir. Bu tür özellikler nedeniyle normal dağılım özelliğini kaybeden verilerin güvenilirliği düşüktür. Normallik testleri dört faktör açısından önemlidir. Bunlardan üçü güvenilirlik analizleri, sonuncusu ise hipotez testleri için gereklidir.

Normallik testleri, öncelikle toplam puanların ana kütleyi temsil etme derecesi açısından önemlidir. Analizler, seçilen örneklem verilerinin ana kütleye paralel olup olmadığı hakkında bilgi verir ve ölçek verilerinin ana kütledekine benzer bir şekilde ne ölçüde normal dağılım özelliği gösterdiğini ortaya koyar. Örneklem verileri ana kütle özelliklerini yansıttığı ve normal dağılım özelliğine sahip olduğu ölçüde ölçüm sonuçları ana kütleye genellenebilir. Seçilen örneklem büyüklüğü $n > 30$ olduğu sürece verilerin normal dağılım özelliğine sahip olduğu varsayılır. Verilerin ana kütledeki dağılımı aslında sağa veya sola dayalı, çarpık olabilir. Fakat belirli bir büyüklüğe[a] sahip örneklem verilerinde ana kütledeki dağılımına bakılmaksızın *merkezî limit teoremi* işler ve verilerin normal dağılım eğrisine uygun olduğu düşünülür veya verilerin normal dağılım eğrisine uygun olması amaçlanır. Örnek kütle-ana kütle karşılaştırmalarında verilerin normal dağılım özelliği için ölçek veya testlerin toplam/ortalama puanları temel alınır.

Normallik testlerine başvurmanın ikinci amacı, maddelere ait puanların dağılımı hakkında bilgi edinmektir. Maddeler, *kavramsal yapı alanını* temsil ediyor olmalıdır. Klasik test kuramında (KTK) bütün test / ölçek maddelerindeki *gerçek değerlerin* normal dağılım özelliğine sahip olduğu varsayıldığından, uygulanan testlerde maddelerin bu koşulu sağlayıp sağlamadığını belirlemek gerekir. Normal dağılım özelliği gösteren maddeler ikili korelasyonlarda doğrusal bir eğime sahiptirler. Bunun için ikişerli olarak verilerin *nokta dağılım grafiği* çizilir.[5] Maddeler arasındaki çarpıklık çok fazla değilse veya normal dağılım özelliği çok fazla bozulmamışsa ölçümün güçlü olduğu varsayılır. Değişkenliği yeterince içermeyen, çarpıklık ve basıklık katsayıları yüksek olan maddeler ölçekten çıkarılır.

Literatürde, bir ölçeğe / teste ait birden fazla maddenin ikişerli olarak hep birlikte normal dağılım özelliği göstermesine *çok değişkenli normal*

---

[a] Örneklem büyüklükleri ile ilgili kaba sınıflandırma ölçüleri şu şekilde ele alınabilir: $n < 30$ küçük örneklemler, $n = 30$–100 orta büyüklükteki örneklemler, $n > 100$ büyük hacimli örneklemler.

dağılım adı verilir. Çok değişkenli normal dağılım özelliğinde tüm değişken çiftleri arasındaki ilişkiler doğrusal bir niteliğe sahiptir. Çok değişkenli normal dağılımda, maddelerin sadece ortalama ve varyansları değil, birbirleriyle olan korelasyonları da önem kazanır. Bir ölçekte tek tek maddelerin her biri normal dağılım özelliğine sahip olsa bile çoklu değişkenler normal dağılım özelliği göstermiyor olabilir.[6] Bu nedenle verilerin normal dağılım özelliğini saptamak için, *çok değişkenli basıklık ve çarpıklık analizi*[a] sonuçlarını incelemek gerekir. Çok değişkenli çarpıklık ve basıklık analizi için Lisrel isimli yazılımdaki Mardia Testi'nin kullanılması önerilmiştir. Bu testte *Mardia PK* değerinin 3'ten küçük çıkması çok değişkenli normalliğin sağlandığı anlamına gelir.[7] Ayrıca faktör analizi için de maddelerin her birinin normal dağılım özelliğine sahip olması gerekir. Bazı bilim adamlarına göre, çok değişkenli normal dağılım özelliği, aslında *gizli değişkenle* ilgilidir. Gözlem değerlerinin kendileri normal dağılım özelliğine sahip olmak zorunda değildir. Psikolojik ölçeklere ait puanların çoğu normal dağılım özelliği göstermez. Puanların 1 ilâ 5 gibi belirli değerler arasında kalması verilerin normal dağlım özelliğini sınırlandırır. Maddelere ait gözlem puanlarının normal dağılım özelliği göstermesi değil, normale yakın çıkması önemlidir.[8]

Bilim adamı ölçüm çalışmalarında *madde-yanıt kuramını* temel almışsa, bu kuramda maddelerin ana kütleyi temsil etmesi gibi bir zorunluluk yoktur. "Yansız madde" özelliği, ana kütleyi temsil etmeyen örnek kütlelerden de elde edilebilir.[9]

Verilerin normal dağılım özelliğinin araştırılmasında etkili olan üçüncü faktör, norm temelli test/ölçek verilerinin standartlaştırılma amacına uygunluğunu belirlemektir. Veriler normal dağılım özelliği göstermiyorsa ölçüm yapılan bir grup için norm değerini oluşturmak güçleşir. Herhangi bir yaş, cinsiyet veya meslek grubu için ham veriler genelde normal dağılım özelliği göstermez. Veriler çarpık, sivri veya basıktır. Ancak bu çarpıklık insan yapısından kaynaklanması nedeniyle çarpıklığı gideren veya nötralize eden verilerin normalleştirilmiş standart puanları kullanılır. Veriler standart z puanlarına dönüştürülerek normalleştirilir. Öte yandan test puanlarının normal dağılım özelliğini tek bir örneklemde yapılan araştırmalara dayandırmak yetersizdir. Ana kütleden tesadüfî yöntemle seçilecek en az birkaç örneklem üzerinde daha verilerin (toplam / ortalama puan) dağılım özelliği araştırılmalı ve nihaî karara ondan sonra varılmalıdır. Se-

[a] Çok değişkenli çarpıklık ve basıklık analizleri SAS ve SPSS'in önceden tanımlanmış mönülerinde bulunmaz. Kimi araştırmacılar SPSS'te çalıştırmak üzere bu amaçla özel makrolar hazırlamışlardır. Çok değişkenli çarpıklık ve basıklık analizleri Yapısal Eşitlik Modeli'ni test eden AMOS ve PRELIS2 gibi istatistiksel analiz programlarında vardır.

çilen örneklemlerden birinde veriler normal dağılım özelliği gösterirken diğer örneklemlerde normal dağılım özelliği göstermemesi halinde *etki büyüklüklerinin* farklı olması nedeniyle güvenilirlik katsayıları benzer çıksa bile bu rakamları ihtiyatlı olarak yorumlamak gerekir. Norm temelli test ve ölçeklerde verilerin normal dağılım özelliğine sahip olması amaçlanırken kriter referanslı testlerde verilerin özellikle çarpık olması hedeflenir.

Normallik testlerini yapmanın dördüncü amacı, hipotez testlerinin sınanmasında parametrik veya nonparametrik testlerden hangisinin seçileceğiyle ilgilidir. Bilim adamı güvenilirlik ve geçerlilik analizi yapılmış ölçüm aracıyla hipotezlerini sağlıklı bir şekilde test edebilmesi için parametrik nitelikteki verilerin normal dağılım özelliğine sahip olmasını arzu edecektir. Parametrik verilerle belirli istatistikî analizlerin yapılabilmesi verilerin normal dağılmasına bağlıdır. Normallik koşulunun sağlanması gereken istatistikî analizler aşağıdaki gibidir:

1. *t*-Testi.
2. *z*-Testi.
3. Varyans analizi.
4. Pearson korelasyon analizi.
5. Regresyon analizinde bağımlı değişken.
6. Faktör analizi (hipotez testlerinde).
7. Kümeleme analizi.
8. Diskriminant analizi.
9. Yapısal eşitlik modeli.

Verilerin normal dağılım özelliği, istatistiksel hesaplamalarla veya grafik yöntemlerle saptanır. Aşağıdaki bölümde sık kullanılan normallik testleri hakkında bilgiler verilmiştir.

Sık kullanılan normallik testleri. Bu testler örneklem verilerinin normal dağılıma sahip bir ana kütleden gelip gelmediğini belirlemek için kullanılır. Normallik testlerinin bir bölümü her örneklem büyüklüğü için uygun iken diğerleri sadece küçük örneklemlerle çalışıldığı zaman uygulanır. Ayrıca kullanılan istatistikî analiz programına göre de gerekli olan örneklem büyüklükleri değişebilir. Normallik testlerinin bazıları belirli koşullarda daha güçlüdür. Araştırmacı diğer dağılımların yanında normallikten sapmayı, sağa veya sola çarpıklığın derecesini belirleme amacına göre bu testlerden birini seçer. Bu tür testler için sıfır ve alternatif hipotezleri aşağıdaki gibi belirlenir:

H_0: Örneklem, normal dağılma sahip bir ana kütleden çekilmiştir. Örneklem verilerinin ortalamasıyla ana kütle verilerinin ortalaması arasında fark yoktur.

H_a: Örneklem, normal dağılma sahip olmayan bir ana kütleden çekilmiştir. Örneklem verilerinin ortalamasıyla ana kütle verilerinin ortalaması arasında fark vardır.

Bilim adamı analiz sonucunda *p* değerlerine bakarak kararını verir. Verilerin normal dağılım özelliği gösterdiğini söyleyebilmek için bu değerin ,05'ten büyük olması gerekir. Aşağıdaki bölümde sıfır hipotezini test eden normallik testlerinden sık kullanılanlara yer verilmiştir.

Kolmogorov-Smirnov testi. Gözlem verilerinin herhangi bir dağılıma uygunluğunu ölçmek için kullanılan Kolmogorov-Smirnov testinin *özel olarak sadece normalliği ölçmek için belirlenmiş* testlere göre daha zayıf kaldığı belirtilmiştir. Bu testin uygulanabilmesi için ana kütle ortalaması ve varyansı önceden belirlenmiş olmalıdır. Aritmetik ortalama ve varyans eğer örneklem verilerinden hareket edilerek saptanmışsa Kolmogorov-Smirnov test sonucu *tutucu* çıkar ve sıfır hipotezinin ret edilme ihtimali daha düşüktür.[10] Diğer normallik testlerinde ise, aritmetik ortalama ve varyansın önceden belirlenmiş olması gibi bir şart yoktur. Testin araştırmacıyı kısıtlaması nedeniyle bazı istatistikî analiz programlarında Lilliefor'un yaklaşımına benzer formüller geliştirilmiştir. Kolmogorov-Smirnov testi herhangi bir örneklem büyüklüğünde uygulanabilir. Analiz sonucunda $p \leq ,05$ çıkmışsa örneklem verilerinin ve bu örneklemin dayandığı ana kütlenin normal dağılım özelliğine sahip olmadığı söylenir. Ancak KS testinin az sayıda ayrık/uç puandan büyük ölçüde etkilenmesi nedeniyle *p* değeri her zaman sağlıklı bir sonuç vermeyebilir. Bu nedenle araştırmacı KS testinin yanında, grafik analizlerine bakarak son kararını vermelidir.

Büyük örneklem verilerinde (pragmatik bilim adamlarına göre $n > 100$ ve daha duyarlı davranan bilim adamlarına göre ise $n > 400$) test sonuçları normallik şartının sağlanmadığını gösterse de bunun pratikte çok fazla bir önemi yoktur. Böyle bir durumda *histogram, kutu grafiği* ve *normal olasılık grafikleri* incelenir. Dağılım, normale yakın bir özellik gösteriyorsa, *p* değerinde normallik koşulunun sağlanmaması çok fazla önemli değildir.[11]

Shapiro-Wilk testi. S.S. Shapiro ve M.B. Wilk[a] tarafından önerilen bu test özellikle normallikten sapmayı belirlemek için geliştirilmiştir. Ana kütle ortalaması ve varyansın önceden bilinmesini gerektirmez. Fakat normallikten sapmanın ne yönde olduğu, dağılımın sağa veya sola çarpıklığı hakkında kesin bir bilgi vermez. Bunun için dağılım grafiklerinin ayrıca incelenmesinde yarar vardır. İstatistiksel analiz programı SPSS, Shapiro-Wilk testini $n = 50$'ye kadar olan örneklemler için hesaplar. Daha büyük örneklemler için SPSS'te Kolmogorov-Smirnov testi uygulanır. İstatistiksel analiz programı SAS'ta ise Shapiro-Wilk testi örneklem hacmi 2000'e kadar olan araştırmalarda kullanılır. Örneklem büyüklüğüne bağlı olarak uygulanabilmesi nedeniyle test W_n simgesiyle gösterilir. Wilk istatistiği her zaman sıfırdan büyük ve 1'den küçüktür ($0 < W_n \leq 1$).[12] İstatistiksel analiz programı SPSS'te Shapiro-Wilk testi Explore mönüsü altında tanımlanmıştır.

Lilliefors testi. Hubert Lilliefors (1967) tarafından geliştirilen bu test Kolmogorov-Smirnov testi gibi herhangi bir sayıdaki örnekleme uygulanabilir. Nonparametrik normallik testidir. Ana kütlenin varyansı ve ortalaması bilinmediği zamanlarda dahi uygulanabilir. Öncelikle iki yönlü Kolmogorov-Smirnov D istatistiği hesaplanır ve D katsayısı elde edilir. Bu katsayı Lilliefors tablosundaki kritik değerlerle karşılaştırılır. Hesaplanan değer kritik değerden küçük çıkmışsa H_0 hipotezi kabul edilir. İstatistikî analiz programı SPSS, vak'a sayısı 50 veya daha fazla ise Lilliefors düzeltme faktörüyle birlikte otomatik olarak KS testini uygular.

D'Agostino ve Stephens testi. R.B. D'Agostino ve M.A. Stephens (1986) tarafından geliştirilen bu yaklaşım, Saphiro-Wilk testinin değişik bir şeklidir. Test başlangıçta orta büyüklükteki örneklemler için önerilmiştir. İstatistiksel analiz programı SPSS'te bulunmayan bu teknik SAS'ta örneklem hacmi 2000'den büyük olan araştırmalarda kullanılır. Bu yaklaşım, Anderson-Darling testi, ki-kare ve Kolmogorov-Smirnov testlerinin alternatifidir. Test, Kolmogorov-Smirnov istatistiğine dayanır. Stephens istatistiğinin yayımlanmış olan kritik değerleri ,01 ilâ ,15 arasında değişir.[13]

Anderson-Darling testi. Gözlenen kümülatif dağılım fonksiyonunun beklenen kümülatif dağlım fonksiyonuna uygunluğunu karşılaştırmak için kullanılan bir testtir.[14] Anderson-Darling testi, verilerin normal dağılıma sahip olduğunu göstermez, tersine normal dağılım özelliği göstermeme olasılığı

---

[a] Bilim adamının adı literatürde farklı şekillerde yazılmıştır. Bazı yazarların *Wilks* şeklinde yazdıkları görülmektedir. Ancak *Wilk* şeklindeki yazım biçimi daha yaygındır.

hakkında bilgi verir. Büyük ölçüde finansman verilerinin analizinde kullanılan bu testte *ampirik dağılım fonksiyonu*[a] (empirical distribution function) göz önünde bulundurulur. Dağılımdaki çarpıklığı test etmek için kuyrukların uzunluğuna KS testinden daha fazla önem verilir. Anderson-Darling testi ki-kare ve KS testinin alternatifi olarak kullanılır.[15] Bu test $n < 10$ ve $n > 40$ olan örneklemler için uygulanmaz.[16] Test tek yönlüdür ve kritik değerleri dağılım biçimlerine göre ayrı ayrı tablolar halinde verilmiştir. Analiz sonucunda istatistiksel analiz programı tarafından öngörülen dağılım biçimine uygun kritik değerleri hesaplanır.[17]

■ Anderson-Darling testi kritik değerleri.

α	,10	,05	,025	,01
A^2_{kritik}	,631	,752	,873	1,035

Ki-kare uygunluk testi. Ki-kare uygunluk testi, örnek kütle dağılımının varsayılan dağılama uygun düşüp düşmediğini belirlemek için kullanılır. Örnek kütle dağılımı, gözlem değerlerinden (frekans dağılımlarından) ve varsayılan dağlım ise, beklenen değerlerden oluşur. Ki-kare uygunluk testinde ya maddelerin frekans değerleri veya toplam/ortalama puanları gruplandırılarak bu gruplara ait frekans değerleri temel alınır.

■ Maddelerin frekans değerleri.

Değer	1	2	3	4	5
Frekans	8	16	40	22	9

[a] Ampirik dağılım fonksiyonu (ADF). Tahminler örneklem verilerine dayalı olarak yapıldığından belirli sayıda veriye bağlı olarak çizilen dağılım eğrisi düz bir kavis şeklinde değil, küçük küçük basamaklar yaparak oluşur. Örneklem büyüklüğü arttıkça bu basamakların sayısı azalarak kırıklar gerçek eğri haline gelmeye başlar. Fonksiyonu basamaklı yapıdan kurtarmak için kümülatif yüzde değerlerinde hareketli ortalamalar yöntemi kullanılır. Bu uygulama herhangi bir parametre tahminine dayanmadığı için nonparametrik yaklaşım olarak isimlendirilir ve bu nedenle ampirik dağılım fonksiyonu kısaca, *kümülatif dağılım fonksiyonunun nonparametrik tahmin değeri* şeklinde tanımlanır. Bu konuda *bk.*, D. Stephenson, "Theoretical Continuous Distributions [Kuramsal Sürekli Dağılımlar]," <http://www.met.rdg.ac.uk/cag/courses/Stats/course/node48.html> (07.2.2003).

■ Toplam puan gruplarının frekansları.

Değer	10 - 20	21 - 30	31 - 40	41 - 50
Frekans	8	16	60	22

Beklenen dağılım, frekansların normal dağılım özelliği göstermesi için ne şekilde olması gerektiği ile ilgili düşüncelere dayanır. Araştırmacı bunun için "nasıl dağılsaydı normal dağılım özelliği gösterirdi?" sorusunu sorar. Örneğin, beş dereceli bir ölçekte puanlara gelen cevapların sıklıkları 10, 20, 40, 20 ve 10 şeklinde gerçekleşseydi veriler normal dağılmış olurdu.

Ki-kere uygunluk testinin, eşit aralıklı ölçeklerde normalliği belirlemek için kullanılması tavsiye edilmemiştir. Bu test de Kolmogorov-Smirnov testi gibi varsayılan dağılımın aritmetik ortalama ve varyans değerlerinin önceden bilinmesini gerektirir. Test, ölçüm verilerinin kategorilere bölünmesini gerektirdiğinden kesikli veriler (nominal, ordinal) için uygundur. Ölçek ve testlerin toplam puanlarının normalliğini belirlemek için kullanılamaz. Toplam ve ortalama puanların normalliği için KS testi kullanılır. Ki-kare uygunluk testinin uygulanabilmesi için beklenen değerin en az 5 olması ve çalışılan örneklem büyüklüğünün $n > 50$ olması gerekir.

Basıklık ve çarpıklık testleri. Birim sayısı 50'den büyük olan örneklemlerde uygulanır (*bk.*, Tablo 4-4). Basıklık ve çarpıklık değerleri, eşit aralıklı ve oranlı veriler için uygundur. Tutumların araştırıldığı 5'li veya 7'li derecelerne ölçekleri aslında sıralı ölçek niteliğindedir. Ancak bu ölçek maddeleri arasındaki mesafenin *yaklaşık olarak eşit olduğu* varsayıldığından eşit aralıklı ölçek gibi değerlendirilir. Bu açıdan basıklık ve çarpıklık katsayıları hem ayrı ayrı madde puanları için hem de ölçeğin toplam puanı için hesaplanabilir. Basıklık ve çarpıklık değerlerinin yüksek veya düşük olmasına göre eğri normal dağılıma yakın veya uzaktır. Basıklık ve çarpıklık değerlerinin her ikisi de ±1 standart sapma değerinden küçük olmalıdır.[18] Ölçek ve testlerde tek tek maddelerin çarpıklık ve basıklık katsayılarını hesaplamak gerekir fakat bu yeterli değildir. Çok değişkenli ölçek ve testlerin analizlerinde, *çok değişkenli basıklık ve çarpıklık* değerlerinin de hesaplatılması uygun olur.[19] Bu tür hesaplamalar AMOS gibi özel amaçlı istatistiksel analiz programlarında bulunur. Basıklık ve çarpıklık testleri sonucunda belirlenen sınır değerinin hafif bir şekilde üzerinde kalan basıklık değerleri de kabul edilebilir bulunmuştur.[20] Basıklık ve çarpıklık değerlerinin bir anlam ifade edebilmesi için örneklem hacminin büyük olması gerekir. Küçük örneklem-

lerde bu değerler yeterince bilgi vermez. Sadece verilerin aşırı ölçüde çarpık olması halinde çarpıklık ve basıklık değerlerinden anlamlı sonuçlar çıkarılabilir.[21]

Dağılımın simetriden uzaklaşması anlamına gelen çarpıklık, değerlerin sağ veya sol yarıda yoğunlaşmasıdır. Test veya ölçek oluşturma aşamasında her bir maddenin çarpıklık değeri hesaplatılarak 1,0'in üzerinde değer içeren değişkenler ölçekten çıkarılır. Çarpıklık katsayısı 1,00 – ,50 arasındaki değerler orta derecede kabul edilebilir bir sapma olduğunu, ,50 – ,00 arasındaki rakamlar ise sapmanın önemsiz olduğunu ortaya koyar. Bir maddenin çarpıklık değeri yüksekse, aynı gizli yapıyı ölçüyor olsa bile bu maddenin diğer maddelerle olan korelasyonu düşüktür. İstatistiksel analiz programı SPSS'te çarpıklık değeri hesaplandıktan sonra bu değer ya olduğu gibi veya çarpıklığın standart hatasına bölünerek kullanılır. Bölme işlemi çarpıklık değerini standart *z* puanına dönüştürür. Hesaplanan *z* değeri ±2 sınırları arasında kalıyorsa yaklaşık olarak normal dağılım özelliğine sahip olduğu söylenir.[22] Hipotez testi için eğer sonuç ±2,53'ten daha büyük ise %99 güven aralığında normallikten ayrılmanın anlamlı olduğuna karar verilir.[23] Örneğin, *X* değişkeninin çarpıklık değeri –,416 ve çarpıklığın standart hatası ise ,113 çıkmış olsun. Söz konusu çarpıklığın şiddetli bir çarpıklık olup olmadığını belirlemek için *çarpıklık katsayısı* hesaplanır [ÇK = (–,413-0)/,113]. Hesaplama sonucunda –3,681 değeri elde edilir. Bu değer 2,0'den büyük olduğundan verilerin önemli ölçüde sola çarpık olduğuna karar verilir. Sonuçlar raporlanırken, çarpıklık değeri ile çarpıklığın standart hatası değerinin her ikisi de parantez içinde gösterilir.

■ SPSS'te çarpıklık değerinin yorumlanması:

Skewness değeri : ±1'den küçükse normal dağılım.
z değeri : ±2'den küçükse normal dağılım.

Çarpıklık özelliği gösteren verilerle ilgilenmek için birkaç yönteme başvurulabilir. Birincisi verilerin doğru girilme durumunun kontrol edilmesidir. Bu kontrol yapıldıktan sonra ayrık ve uç değerler araştırılır. Duruma göre bu ayrık değerleri içeren vak'alar ya bütünüyle çıkarılır veya yerine normal değerler konarak hesaplama yeniden yapılır. Örneklem sayısı 50'nin altında ise daha sağlıklı bir değerlendirme için örneklem hacminin artırılması yoluna başvurulur. Son olarak araştırmacı verileri standart değerlere dönüştürmeyi düşünebilir.

Basıklık[a], ölçüm değerlerinin sivri veya yatık olmasıdır. Normal dağılımda eğri, ne yatık ne de sivridir ve bu nedenle sıfır basıklık değerine sahiptir. İstatistikî analiz programı SPSS'te basıklık değeri hesaplatıldıktan sonra ya olduğu gibi kullanılır veya standart *z* puanı temel alınır. Standart *z* puanı basıklık değerinin, *basıklığın standart hatasına* bölünmesi suretiyle hesaplanır. Basıklıkta da *z* değerlerinin +2,00 ilâ –2,00 aralığında bulunması gerekir. Sonuçlar raporlanırken basıklık değeriyle birlikte basıklığın standart hatası da birlikte gösterilir (*bk.*, Tablo 4-3).

■ SPSS'te basıklık değerinin yorumlanması:

Basıklık değeri : ±1'den küçükse normal dağılım.
z değeri : ±2'den küçükse normal dağılım.

Bazı yazılımlarda, hesaplanan basıklık katsayısı 3 olarak saptanmıştır ve bu rakam temel alınır. Fakat SPSS ve LISREL gibi istatistikî analiz programlarında orijin, 3'ü temsil etmek üzere 0 olarak belirlenmiştir. Normal dağılımı 0 merkezli olarak hesaplayan yaklaşım *Fisher basıklığı* ve 3 merkezli olarak hesaplayan yaklaşım ise *Pearson basıklığı* olarak adlandırılır. SPSS, Fisher basıklığını temel almıştır.[24]

Tablo 4-3. Çarpıklık ve Basıklık Değerlerin Tabloda Gösterilmesi

	N	*Min.*	*Maks.*	*Ort.*	*SS*	*Çrp.*		*Bsk.*	
						ÇD	*ÇSH*	*BD*	*BSH*
M1	93	1	5	3,64	1,11	–,91	,072	0,03	,121
M2	92	1	5	3,28	1,17	–,03	,072	0,90	,121
M3	89	1	5	3,19	1,13	–,82	,072	0,92	,121
M4	93	1	5	3,08	1,12	–,33	,072	0,84	,121

Not. Simgeler ve anlamları: *ÇD* = Çarpıklık değeri, *ÇSH* = Çarpıklığın standart hatası, *BD* = Basıklık değeri, *BSH* = Basıklığın standart hatası.

[a] Aslında *basık* ve *yatık* sözcükleri benzer anlamlara gelmektedir. Türkçede *çarpıklık* kelimesinde olduğu gibi normalden aşağıya ve yukarıya doğru her iki yöndeki sapmayı gösterecek daha iyi bir sözcük bulunamadığından bu kelime tercih edilmiştir.

İstatistiksel analiz programı SPSS'te basıklık ve çarpıklık katsayılarını hesaplamak için öncelikle eksik veri içeren vak'aların analizden çıkarılması gerekir. Bunun için Data mönüsünden Select Cases tuşu ile If kartına girilir ve bu bölümde not missing ($v1$) and ($v2$) şeklinde programın eksik veri içeren vak'aları elemesi sağlanır. Daha sonra Analyze mönüsünden Descriptives düğmesi tıklatılarak Options seçeneği ile Skewness ve Kurtosis kutucukları seçili hale getirilir.

Tablo 4-4. Örneklem Büyüklükleri ve Uygulanabilecek Normallik Testleri

Örneklem büyüklüğü	*Normallik testi*	*Açıklama*
$n < 50$	Shapiro – Wilk testi	SPSS
$n > 50$	Lilliefors testi	SPSS
Her büyüklük için	Kolmogorov-Smirnov testi	SPSS
$n < 2000$	Shapiro – Wilk testi	SAS
$n > 2000$	Kolmogorov-Smirnov testi	SAS
$n > 50$	Çarpıklık ve basıklık testi	SPSS

Çarpık dağılımlarda, merkezî dağılım ölçüleri verilirken aritmetik ortalamanın yanında aynı zamanda medyan değerleri de gösterilerek okuyucuların daha sağlıklı bilgi edinmelerine olanak sağlanmalıdır. Çünkü bu dağılımlarda aritmetik ortalama kişileri yanıltabilir.

Veriler normal dağılım özelliği göstermiyorsa. Ölçüm yapılan verilerin normallik testleri sonucunda normal dağılım özelliğine sahip olmadığı anlaşılmışsa bu verilerle yapılan diğer istatistiksel analizler ve yorumlar yanlış ve yönlendirici olur. Bu gibi durumlarda araştırmacı değişik yöntemlere başvurabilir. Bunlardan en önemli ikisi *tabakalaştırma* ve *dönüştürme* yaklaşımlarıdır.[25]

Tabakalaştırma. Tabakalaştırma yaklaşımında örneklem, bir veya daha fazla özelliğe bağlı olarak alt örneklemlere bölünür. Alt örneklemlerdeki dağılım fonksiyonu farklı ise, örneğin alt örneklemdeki veriler normal dağılım özelliğine sahipse bölme özelliğinin önemli olduğu anlaşılır. Ancak bu yöntemin sakıncası alt örneklemlerin küçük olması halinde testin güvenilirliğinin düşmesidir.

Dönüştürme. Veriler önemli ölçüde çarpık veya basıksa dönüştürme yöntemine başvurulabilir. Bu konuda çarpıklığın derecesi ve yönü göz önünde bulundurulur.[26] Dönüştürmeye karar verirken z değerleri dikkate alınır (*bk.*, Tablo 4-5). Verilerin dönüştürülmesiyle Tip II hatası yapma olasılığı azalır. Dönüştürme işlemi daha çok basıklıkta değil, verilerin çarpıklığı durumunda uygulanır.

Tablo 4-5. Dönüştürme Yöntemleri

	Ilımlı ölçüde çarpık	Orta derecede çarpık	Büyük ölçüde çarpık
z değerleri	$z < 1{,}96$	$1{,}96 \leq z \leq 2{,}33$	$z > 2{,}58$
Sağa çarpık (pozitif)	Karekök	Logaritmik	Ters dönüşüm
Sola çarpık (negatif)	Karekök	Logaritmik	Ters dönüşüm

Tabachnick ve Fidell'e göre ılımlı ölçüde sağa çarpık maddelerde *karekök dönüşümü*, orta ölçüde sağa çarpık verilerde *logaritmik dönüşüm* ve şiddetli ölçüde çarpık verilerde ise *ters dönüşüm* formülü uygulanmalıdır. Negatif çarpık dağılımlarda tüm değerler, dizinin en yüksek değerine 1 rakamı ilave edilerek elde edilen değerden çıkarılır ve ondan sonra karekök, logaritmik veya ters dönüştürme işlemleri yapılır (aktaran Garson).[27] İstatistiksel analiz programı SPSS'te çarpık verilerin dönüştürülmesi ve normalleştirilmesi için *Blom dönüştürme yöntemi* uygulanır.[28] Diğer yöntemler için programın Transform, Compute komutları ile yeni hedef değişkenler belirlenir ve Numeric Expression bölümünden uygun dönüştürme formülü seçilerek gerekli hesaplama yaptırılır. Örneğin, daha çok ayrık değerlerde kullanılan karekök dönüşümü için SQRT fonksiyonu, logaritmik dönüşüm için LG10 veya LOGE, iki taraflı ters dönüşüm için ise $-1/X$ fonksiyonu kullanılır. Burada X değişkeni veri matrisindeki ölçüm değişkenini tanımlar. Oransal verilerin dönüştürülmesinde ise ARCSIN fonksiyonu kullanılır.

Normallik ihlalini yenmek için dönüştürme yöntemine başvurmak her zaman istenen sonucu vermeyebilir. Örneğin bir ankette, ılımlı ve uç yanıtlar belirli bir zihin yapısını ortaya koyuyorsa ölçek maddeleri birbirleriyle tutarlı olabilir. Bu gibi durumlarda verileri dönüştürme yoluna başvurmak ölçümün gücünü ve maddelerin arka planındaki doğrusallığı azaltır. Araştırmacı

eğer yorum gücünü azaltmayacaksa dönüştürme yöntemini düşünmelidir. Ayrıca tek tek değişkenlerin normalliğini sağlamak, çok değişkenli normalliği garanti altına almaz, fakat gerçekleşme olasılığını artırır.

Test ve ölçeklerin güvenilirlik analizleriyle ilgili olarak yüksek lisans tezlerinde nokta dağılım grafiklerini ve diğer normal dağılım test sonuçlarını verme zorunluluğu getirilmemelidir.[29] Çünkü verilerin normalliğine ilişkin test sonuçları ile her zaman yeterli ve geçerli bilgiler elde edilemez. Örneğin KS testi arka plandaki ana kütlenin normal dağılımdan farklı olduğunu söyleyebilir, ancak ne ölçüde ve ne derecede farklı olduğu hakkında bir bilgi vermez. Veriler normallik koşulunu tam olarak karşılamıyorsa hipotez testlerinde güvenilirlik düzeyi, yüzde beşten ,025'e veya ,01'e çıkarılır.[30]

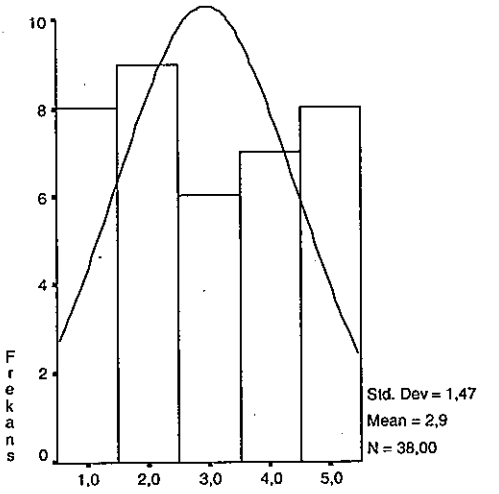
Normallik Grafikleri

Verilerin normal dağılım özelliğine sahip olma durumu istatistiksel testlerin dışında grafik yöntemlerle de izlenebilir. Bu bölümde normalliği inceleme imkanı sağlayan grafik teknikler tanıtılmıştır.

Puanların histogram grafiği ile incelenmesi. Sürekli (eşit aralıklı ve oranlı) veriler için kullanılan bu grafikte puanların dağılımının yaklaşık olarak normal dağılım özelliği gösterip göstermediğine bakılır. Testlerde ve tutum ölçeklerinde duruma göre hem maddelerin hem de toplam puanların histogram grafiği çizilir. Tam olarak normal dağılıma sahip olmasa bile, önemli ölçüde sağa veya sola çarpık değilse verilerin normal dağılım özelliğine sahip olduğu varsayılır. İstatistikî analiz programı SPSS'te histogram grafiği aşağıdaki komutlar çalıştırılarak çizilir:

1. Analyze mönüsünden Descriptive Statistics bölümüne giriniz ve buradan Frequencies başlığını seçiniz.
2. Açılan kartta, Display frequency tables kutucuğunu seçili olmaktan kurtarınız.
3. Analiz etmeyi düşündüğünüz değişkenleri sağ taraftaki alana taşıyınız.
4. Taşıma işleminden sonra Charts düğmesine basınız. Burada Histograms seçeneğini With normal curve seçeneğiyle birlikte işaretli hale getiriniz ve geriye dönünüz.
5. Daha sonra Statistics düğmesine basınız. Burada Dispersion bölümünde bulunan Range ve Distribution bölümünde bulunan Skewness ve Kurtosis kutucuklarını seçili hale getiriniz.

6. Continue düğmesine basınız ve daha sonra OK düğmesiyle hesaplamayı yapınız (*bk.*, Şekil 4-2).



Şekil 4-2. Histogram grafiği.

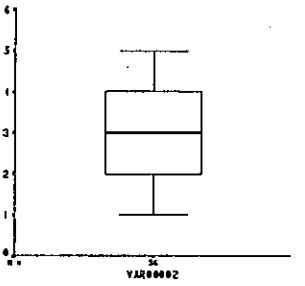
Histogram grafiği özellikle küçük örneklemlerde çubukların sayısına ve genişliğine bağlı olması nedeniyle normalliği göstermede iyi bir yöntem olarak değerlendirilmemiştir. Küçük örneklemlerde diğer grafik yöntemlerin kullanılması daha doğru olur.

Puanların kutu grafiği ile incelenmesi. Araştırmacı nominal ve eşit aralıklı ölçek niteliğindeki verilerin normal dağılım özelliğini belirlemek için SPSS yazılımının Graphs mönüsündeki Boxplot veya Analyze mönüsündeki Plots seçeneğinden yararlanabilir. Test veya ölçeklerde, maddelerin normal dağılım özelliği kutu grafiği ile belirlenir. Kutu grafiği, bazı kitaplarda uzantılara sahip olması nedeniyle "kutu-bıyık grafiği" olarak isimlendirilmiştir. Aşağıdaki bölümde önce grafik ve daha sonra analiz mönüsünün kullanılması tanıtılmıştır.

1. Graphs mönüsünden Boxplot düğmesini seçiniz.
2. Açılan kartta, Simple düğmesini seçili hale getiriniz. Değişkenlerin ayrı ayrı grafikleri çizilmek isteniyorsa Summaries of separete variables kutucuğunu seçiniz.
3. Analiz etmeyi düşündüğünüz değişkeni sağ taraftaki alana taşıyınız ve OK tuşuna basınız.

Analiz mönüsünü kullanmak isteyen araştırmacılar aşağıdaki sıra içinde gerekli işlemleri uygulamalıdırlar.

1. Analyze düğmesine tıklayarak Descriptive Statistics bölümüne gidiniz. Burada Explore düğmesine basınız.
2. Kartın sol alt köşesinde bulunan Plots şıkkını seçili hale getiriniz.
3. Daha sonra sağ taraftaki Plot düğmesine tıklayınız ve Boxplots bölümündeki Factor Levels Together seçeneğini işaretleyiniz. Descriptives bölümü altındaki Stem-and-Leaf düğmesini seçili hale getiriniz.
4. Gruplama değişkeniniz yoksa Factor list bölümünü boş bırakınız.
5. OK tuşunu tıklayınız (*bk.*, Şekil 4-3).



Şekil 4-3. Kutu-bıyık grafiği.

Kutu grafiğinin içindeki orta çizgi dağılıma ait medyan değerini gösterir. Çizgi, kutunun tam ortasında değilse veriler çarpıktır. Grafikte ayrık değerler, *o* harfi ve uç değerler ise, * işaretiyle gösterilir. Ayrık değerler %75'lik

sınır değerinden (percentil) itibaren kutunun 1,5 katı kadar bir genişlikte yer alan değerleri gösterir. Uç değerler ise, %75'lik sınır değerinden itibaren kutunun üç katı kadar bir genişlikte yer alan değerlerle ilgilidir. Kutunun uzunluğu, *birinci ve üçüncü çeyrek dilimler arasındaki mesafeyi* tanımlar ve bu mesafeye "kartiller arası değişim aralığı" (KADA) adı verilir. Verilerin normal dağılım özelliğini belirlemek için başvurulabilecek bir diğer yöntem KADA değerini verilerin standart sapmasına bölmektir. Çıkan değerin 1,3 civarında olması verilerin yaklaşık olarak normal dağıldığı anlamına gelir.[31]

Puanların sap-yaprak grafiği ile incelenmesi. Sap-yaprak grafiğinin en önemli özelliği serideki ayrık değerleri kolayca ortaya koymasıdır. Ayrık değerler veri girişinde yapılan hatalar veya bilinçli olarak tercih edilen ekstrem değerlerdir. Sap, iki veya üç haneli herhangi bir sayıda en önde gelen rakamı temsil eder. Örneğin 92 rakamında sap 9 ve yaprak ise 2'dir. Üç haneli rakamlarda en sağdaki rakam genellikle ihmal edilir. Saptaki sıfır belirli bir hanenin altındaki rakamların sayısını gösterirken frekansı sıfır olan saptaki rakamlar değer bulunmadığı anlamına gelir. Örneğin Şekil 4-4'te sapta görülen 5 rakamının frekansının sıfır gözükmesi ilk hanesi 5 olan bir değer bulunmadığı şeklinde yorumlanır.

ÜCRET Stem-and-Leaf Plot

Frequency	Stem - Leaf
4,00	0 . 1113
5,00	1 . 22245
11,00	2 . 11122245555
4,00	3 . 2226
2,00	4 . 55
,00	5 .
3,00	6 . 233
2,00	Extremes (>=654)

Stem width: 100,00
Each leaf: 1 case(s)

Şekil 4-4. Sap-yaprak grafiği.

Yan dönmüş histograma benzeyen bu grafikte sap ile yapraklar birbirlerinden düşey bir çizgi (|) veya nokta (.) ile ayrılır. Sap-yaprak grafiğinde

medyanı bulmak için yapraklar sayılır ve yarısı alınır. Çıkan değer hangi satıra denk gelmişse medyan bu satır tarafından temsil ediliyor demektir. Sap-yaprak grafiği histogram grafiğine göre değerlerin sıralanış biçimini de göstererek daha fazla bilgi verir.

Puanların nokta dağılım grafiği ile incelenmesi. Nokta dağılım grafiği, test veya ölçek maddelerinin normallik derecesini belirlemek için uygulanır. Grafik bir çift maddeye ait çok sayıda kişiden elde edilmiş değerleri iki boyutlu bir düzlemde noktalar halinde gösterir. Noktalar, ölçüm yapılan her bir kişinin *X* ve *Y* maddelerinden (ölçümlerinden) aldıkları puanın yerini gösterir. Bilim adamı bu grafiği inceleyerek; (a) düzlemde diğerlerinden önemli ölçüde sapma gösteren ayrık bir nokta olup olmadığını, (b) noktaların yuvarlak bir biçimde mi yoksa doğrusal bir şekilde mi dağıldığını, (c) değişkenler arasındaki ilişkinin ne ölçüde güçlü olduğunu araştırır. Nokta dağılım grafiği SPSS'te aşağıdaki adımlar çerçevesinde çizilir:

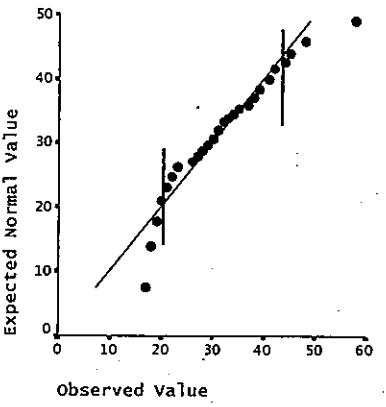
1. Graphs mönüsünden Scatter düğmesine tıklayınız.
2. Karışınıza dört farklı nokta dağılım grafiği seçeneği çıkacaktır: simple, matrix, overlay, and 3-D. Bunlardan Simple seçeneğini seçili hale getiriniz. Daha sonra Define komutu ile değişkenlerinizi tanımlayınız.
3. Dikey boyutta *Y* bağımlı değişkenini ve yatay boyutta *X* bağımsız değişkenini tanımlayınız.
4. Daha sonra OK düğmesine basınız.
5. Çıktı sayfasında (output) grafiğin üzerine çift tıklayınız ve yeni açılan pencerenin üzerine geliniz.
6. Chart başlığı altında Options düğmesine geliniz.
7. Buradan Fit line bölümünde Total şıkkını seçiniz ve daha sonra fit options düğmesine basınız.
8. Fit method bölümünde linear regression şıkkını seçiniz ve regresyon seçeneklerinden include constant in equation ve display R-square maddelerini seçili hale getirerek continue tuşuyla çıkıp OK düğmesine basınız.

Bilim adamı grafiği noktaların dağılımı, noktaların regresyon doğrusu üzerine düşmesi ve ayrık noktaların bulunup bulunmadığı açısından inceler. Noktalar regresyon doğrusuna yakın bir şekilde sıralanmışsa veriler doğrusal bir dağılım gösteriyor demektir.

Puanların normal olasılık grafiği ile incelenmesi. Normal olasılık grafiğinde gözlem verileri normal dağıldığı varsayılan beklenen değerlerle karşılaştırılır. Eğer gözlem verileri normal dağılım özelliği gösteriyorsa doğruya yakın bir çizgi şeklinde gözükür. Aksi halde çizgi dalgalı bir şekilde yatay *S* harfi görünümüne sahiptir. Normal olasılık grafiği $n > 50$ olan örneklemlerde daha iyi çalışır. Küçük örneklemlerde tek bir maddede ortaya çıkan değişiklik, dağılımın yığışımlı yüzde değerlerinde önemli ölçüde değişikliğe neden olması nedeniyle istenen sonucu vermez. İstatistiksel analiz programı SPSS'te bu grafiği çizmek için Graphs mönüsünden P-P düğmesi seçilir. Eğer grafikteki noktalar test dağılımına uygun düşmüyorsa verilerin standardizasyonu yöntemine başvurulabilir.

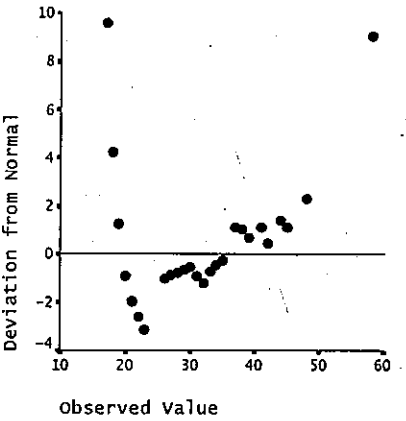
Puanların normal kantil grafiği ile incelenmesi. Bu grafik, medyanla alt kartil ve medyanla üst kartil arasındaki dağlımı gösterir. Normal kantil grafiği, histogram ve kutu grafiğine göre normalliği daha hassas bir şekilde görmemize imkan sağlar. Eşit aralıklı ve oranlı ölçekler için kullanılan bu grafik, *gözlem değerlerini* normal dağılıma sahip *beklenen değerlere* göre karşılaştırır. Gözlem değerleri beklenen değerlerle uyumlu ise grafikteki noktalar 45 derece açılı tek bir doğru şeklinde gözüken çizginin üzerinde veya yakınında dizilir. Grafiğin yorumlanması için araştırmacı grafikteki çizginin orta bölümündeki üçte ikilik bölümü dikkate alır ve buradaki noktaların dağılımını değerlendirir.[32] Grafiğin SPSS istatistiksel analiz programında çizilmesi için aşağıdaki adımlar atılır.

1. Bir veya daha fazla değişkeni seçerek Variables listesine alınız.
2. Dağılım biçimi olarak Normal (ön tanımlı), seçeneğini işaretleyiniz.
3. Eğer gerek duyuyorsanız veya sonuçlar normal dağılım özelliği göstermiyorsa Transform şıkkı ile puanları dönüştürünüz.
4. Eğer SPSS'in grafikteki değerleri standartlaştırmasını istiyorsanız Standardize values şıkkını seçili hale getiriniz.
5. Proportion Estimation Formula beklenen normal dağılımın kantil veya yüzdelik dilime göre hesaplanma şeklini belirler. Bu bölümde önceden tanımlı Blom's seçeneğini bırakabilir veya değer seçeneklerden birini seçebilirsiniz
6. Daha sonra grafiği çizmek için OK düğmesine basınız (*bk.*, Şekil 4-5).



Şekil 4-5. Q-Q grafiği ve ölçüm verilerinin normal dağılım özelliği.

Puanların detrended normal olasılık grafiği ile incelenmesi. Bu grafik, istatistiksel analiz programı SPSS'te bir önceki başlıkta açıklanan Q ve Q grafiği ile birlikte çizilir. Grafik, dikey eksen üzerinde orijinden bir veya iki basamak yukarıda belirlenen 0 noktasından çizilen bir yatay çizgi etrafında ölçüm verilerine ait noktaların ne şekilde dağıldığını gösterir. Noktalar çizgiye yakın ve paralel bir şekilde dağılmışsa verilerin normal dağılım özelliği gösterdiğine karar verilir. Veriler, yatay bir S harfi, J harfi, V harfi veya yuvarlak topa benzer belirli biçimlere sahipse normal dağılım özelliği göstermiyor demektir (*bk.*, Şekil 4-6).



Şekil 4-6. Detrended grafiği ve ölçüm verilerinin normal dağılım özelliği.

DEĞİŞKENLERİN BAĞIMSIZLIĞI

Değişkenlerin birbirinden farklı ve ilişkisiz olması anlamına gelir. Bu olguya aynı zamanda "dikeysellik" adı verilmiştir. Değişkenler ilişkisiz ise, $r = 0$ çıkar. Tek bir boyutu veya faktörü ölçen test ve ölçeklerde değişkenlerin belirli bir oranda birbirleriyle ilişkili olmaları gerekir. İkili serilerde ilişkiyi göstermek üzere küçük r simgesi, çoklu korelasyonlarda ise büyük R harfi kullanılır. İkili serilerdeki korelasyon katsayısı –1,0 ilâ +1,0 arasında değişirken çoklu korelasyonlarda R katsayısı 0 ilâ +1,0 arasında değişir ve R değeri birden fazla bağımsız değişkenin bağımlı değişken üzerindeki kombine etkisini gösterir. Test maddeleri arasında bağımlılık özelliği araştırılırken, bağımsız değişkenlerle bağımlı değişkenler arasındaki ilişkilerin araştırıldığı ölçüm modellerinde dikeysellik özelliği aranır.

ŞÜPHELİ VE EKSİK VERİLERİN DEĞERLENDİRİLMESİ

Veri kalitesinin değerlendirilmesinde göz önünde bulundurulması gereken bir diğer husus, şüpheli verilerin varlığı ile eksik değerlerin miktarı ve dağılım biçimidir. Verilerin dağılım özelliğinden önce kuşkulu ve eksik veriler incelemeye alınır.

Şüpheli Veriler

Verilerin bir bölümü; çift işaretleme, işaretin hangi şıkka ait olduğunun bilinememesi, işaretin onay mı ret mi olduğunun anlaşılamaması gibi nedenlerle kuşkulu bir niteliğe sahiptir. Kuşkulu veriler, mantıksal olarak net bir şekilde yorumlanamıyorsa farklı bir rakamla kodlanır. Bilim adamı kuşkulu verileri kodlamak için örneğin, aşağıdaki gibi bir plan geliştirebilir:

1. –990 yanıtın hangi şıkka ait olduğunun saptanamaması.
2. –991 birden fazla şık işaretlemesi.
3. –992 operatör hatası, ilgisiz bir değer girilmiş olması.
4. –993 anlamsız bir işaret yapılmış olması.

Şüpheli veriler belirli göstergelere göre mantık yürütme sonucunda bazı değerleri alabiliyorsa bu değerler verilir. Ancak mantık yürütmeyi destekleyecek güçlü kanıtlar olmalıdır.

Eksik Veriler

Veri yokluğu veya veri bulunmaması anlamına gelen eksik veriler değişik şekillerde sınıflandırılmıştır. Eksik verilerle ne şekilde ilgileneceği bu verilerin türüne, yoğunluğuna ve nedenlerine göre değişir.

Eksik verilerin türleri. Eksik veriler; maddelere, anket formunun belirli bir bölümüne veya anketin tamamına ilişkindir. Bir veya birkaç maddeye yanıt verilmemesi *madde bazında eksik veri* olgusunu gündeme getirir. Anketin anlamlı bir bölümüne veya tamamına hiç yanıt verilmemişse; sosyal araştırmalarda hane halkı üyelerinden bir veya bir kaçına anket uygulanamamışsa bu olgu *birim bazında eksik veri* olarak değerlendirilir. Eğer anket iptal edilmemişse "birim bazındaki" eksik verilerin yerine de değer ataması yapılabilir ve bu yaklaşıma *kitlesel atama* adı verilir. Büyük hacimli araştırmalarda özel yazılımların kullanılması halinde birim bazındaki eksik verilerin yerine yeni değerlerin atanması yanlış bir uygulama değildir.[33] Ancak, yanıtlama oranı %85'in altına düşmüşse "cevaplama oranı" yanlılığını da göz önünde bulundurmak gerekir.[34]

Öte yandan eksik veriler, kullanıcı tanımlı veya sistem tanımlıdır (*bk.*, Tablo 4-6). Eksik veriler yanıtlayıcıların cevap vermemelerinden kaynaklanmışsa kullanıcı tanımlıdır. Eğer veri dönüştürme uygulamaları sonucunda bazı hücrelerde değer ortaya çıkmamışsa buna sistem tanımlı eksik veri denir.

Eksik verilerin kodlanması. Kullanıcı tanımlı eksik verilerin bilgisayara tanıtılmasında, istatistiksel analiz yazılımının niteliğine göre karakterler veya rakamlar kullanılır. Bazen yıldız veya nokta imi boş veri anlamında kullanılır. Karakter (string) verilerdeki eksiklikler için ise tırnak işareti içinde tek bir aralık boşluk (' ') bırakılır. Eksik verilerin tanımlanmasında boş bırakma veya sıfır değerini girme karışıklıklara neden olduğundan rakamsal kodlama sisteminden yararlanmak daha doğrudur. Eksik verilerin kodlanması araştırmacının amacına ve kullanılan istatistiksel analiz programına göre değişir. Araştırmacı, bir maddeye kasıtlı olarak veya unutularak yanıt verilmemesiyle, yanıtın kişiyle ilgili olmaması olguları arasında ayırım görmek istiyorsa ayrı kod biçimlerinden yararlanır. Sık kullanılan eksik veri kodları aşağıdaki gibidir:

1. 9 Yanıt vermemiş, unutmuş veya kasıtlı olarak boş bırakmış.
2. –9 Yanıt kendi pozisyonuyla veya durumuyla ilgili değil.
3. –99 Birden fazla işaretleme yaptığı için eksik veri olarak kodlama.
4. –999 Kontrol sorusunu yanıtsız bırakmış.

5. –9999 Seçme sorusu gereği maddenin boş bırakılması gerekiyor.

Eksik veriler gelir, yaş ve eğitim durum gibi demografik değişkenlere veya bir tutum ölçeğinin maddelerine ilişkin olabilir. Literatürde daha çok demografik değişkenlerdeki eksik veriler üzerinde durulmuştur. Tutum ölçeklerindeki maddeler büyük ölçüde birbirine benzer olduğundan bu maddelerden birine işaretleme yapılmamış olması *gerçek değerin* ortaya çıkarılmasını zorlaştırır. Bu nedenle demografik değişkenlerde olduğu gibi, ölçeklerdeki eksik verilerin yerine de ikame değerlerinin atanması gerekir. Bir araştırmada anketlerin yanıtlama oranı ,70'in altında kalmış ve eksik yanıtların yerine atama yöntemi kullanılmışsa bu durum araştırma metninde rapor edilmelidir. Öte yandan bir maddenin yanıtlama oranı yine ,70'in altına düşmüş ve atama yapılmışsa bu durum o maddeye ilişkin tablonun altında dipnot olarak belirtilir.[35]

Tablo 4-6. Eksik Verilerin Sınıflandırılma Biçimleri

Yoğunluk açısından	*Değişkenlerin niteliği açısından*	*Kaynakları açısından*	*Kodlama biçimi açısından*
Anketin tamamında eksik veri	Nominal değişkenlerde eksik veri	Tam tesadüfî eksik veri	Kullanıcı tanımlı eksik veri
Anketin bir bölümünde eksik veri	Sıralı değişkenlerde eksik veri	Tesadüfî eksik veri	Sistem tanımlı eksik veri
Anketin bazı maddelerde eksik veri	Sürekli değişkenlerde eksik veri	Tesadüfî olmayan eksik veri	

Eksik verilerin nedenleri. Eksik veriler değişik nedenlerden kaynaklanabilir. Farkına varmadan işaretlememe olgusuyla cevaplayıcıların belirli türdeki sorulara kasıtlı olarak yanıt vermemeleri farklı iki olgudur. Birincisine *tesadüfî eksik veriler* (TEV) ikincisine ise, *sistematik eksik veriler* (SEV) adı verilir. Tesadüfî eksik veriler de kendi içinde üç grupta incelenir. Birinci grupta *tam tesadüfi eksik veriler* (TTEV) yer alır. Bu tür eksik verileri tanımlamak için gelir ve yaş gibi iki değişken düşünelim. Bunlardan gelir değişkeninde eksik veriyle karşılaşma olasılığı, gelir düzeyi ve yaş grubuyla ilintili olarak artıp azalmıyorsa *tam tesadüfi eksik veri* olgusundan söz ederiz. İkincisi *tesadüfî eksik* verilerdir. Gelir değişkeninde eksik veriyle karşılaşma olasılığı, gelir düzeylerinden etkilenmezken yaş düzeyinden etkileniyorsa tek ayaklı bu realiteye *tesadüfi eksik veri* (TEV)

adı verilir. Üçüncüsü ise tesadüfi olmayan eksik verilerdir. Bu tür eksik verilerde cahillik, yanılma, unutma gibi faktörler söz konusu değildir.

Değişkenlerin özellikleri ve eksik veriler. Eksik veri içeren değişkenler nominal, sıralı ve sürekli ölçek verisi niteliğinde olabilir. Değişkenlerin niteliğine göre yapılacak eksik veri işlemi de farklılık gösterir. Bunun dışında değişkenler araştırmacı açısından taşıdığı öneme göre iki grupta değerlendirilir: *anahtar değişkenler* ve *diğer değişkenler*. Anahtar değişkenler analize baz olan, bulunmadıkları takdirde ölçümün veya araştırmanın anlamını yitirdiği değişkenlerdir. Bunlar bazen demografik değişkenlerdir, bazen de diğer ölçüm değişkenleri olabilir. Örneğin bir araştırmada cinsiyet, yaş, eğitim değişkenleri anahtar değişken olarak saptanmış olabilir. Demografik değişkenler anahtar değişken olarak belirlenmişse bu verilerdeki eksiklikler aritmetik ortalama gibi istatistiksel atama yöntemleriyle doldurulamaz. Bu gibi durumlarda sıcak-deste yöntemine başvurulur. Atama yapılmayan anahtar değişkenlerdeki eksik veri oranı %15'ten fazlaysa bu bilgi tablonun altında dipnot olarak gösterilir.

Eksik verilerin iyileştirilme nedeni. Eksik veriler mantıksal veya istatistiksel atama yöntemleriyle ikame edilmedikleri durumda önemli ölçüde bilgi kaybına neden olur. Yapılan bir araştırmada 1000 vak'a ve 20 değişken içeren bir veri matrisinde değişkenlerde %5'lik eksik veri bulunması halinde liste bazında yapılan silme işlemi sonucunda vak'a sayısının 350'ye düştüğü görülmüştür. Böyle bir araştırmada 1000 vak'anın 350 vak'ayla temsil edilmesi, sonuçlarda yanlılığa neden olur. Veri toplama işleminin maliyetli olması ve sonuçta eldeki çok az veriyle yetersiz sonuçlara ulaşma tehlikesinin bulunması nedeniyle verilerin iyileştirilmesi istatistiksel bir zorunluluk olarak görülür.

Eksik verilerin iyileştirilmesi. İşleme tâbi tutulmamış eksik veriler ciddî tahmin hatalarına yol açtığından bu verilerde gerekli iyileştirme veya temizlik çalışmalarına başvurulur. Bu çalışmaların en önemlileri aşağıdaki gibidir.

Değişkenleri iptal etme. Araştırmacı, ölçüm aracındaki bir değişkenin sürekli olarak veya büyük ölçüde boş bırakıldığını saptamışsa ve bu değişken ölçüm açısından hayatî bir öneme sahip değilse, o değişkeni bütünüyle iptal eder.

Vak'aları iptal etme. Bu uygulamada, bir anketteki maddelerin %15'inden fazlasına cevap verilmemişse ve cevap verilmeyenler anahtar

değişkenler ise bu anket formu iptal edilebilir. Ancak anketlerin %20'sinde veya %30'unda iki veya üç sorunun/maddenin boş bırakılmış olması bu anketlerin iptal edilmesini gerektirmez. Değişken açısından bakarsak, veri matrisinde bir değişkendeki eksik verilerin oranı %5'ten az olduğu durumda vak'a iptali düşünülebilir. Bir değişkendeki eksik verilerin oranı %5'ten fazlaysa iptal yöntemi yerine atama yöntemine başvurulmalıdır. Veriler bir şekilde bilgisayara girilmişse iptal işlemi bilgisayar ortamında da yapılabilir. İstatistiksel analiz yazılımlarında iptal işlemi için iki seçenek belirlenmiştir. *Liste bazında* silme (listwise deletion) ve *çiftli silme* (pairwise deletion). Liste bazında silme, vak'a temelli iptal seçeneğidir ve SPSS'te ön tanımlı olarak belirlenmiştir. Liste bazlı ve çiftli silme yöntemleri, araştırmalarda başvurulan en kötü iki uygulama olarak tanımlanmıştır.[36] Eksik verilerin liste temelli olarak iptal edilmesi sonuçta analizi yapılacak vak'a sayısını azaltacağından sonuçların güvenilirliğini büyük ölçüde düşürür. Öte yandan her bir değişkende eğer az sayıda eksik veri varsa *çiftli silme* yöntemine başvurulur. Bu yöntemde eksik veri içeren vak'a bütünüyle iptal edilmediğinden eksik veri içermeyen değişkenlerde aritmetik ortalama, varyans değeri ve kovaryans analizleri hesaplatılabilir. Sadece korelasyon analizleri eksik veri içermeyen değişkenler arasında yapılır. Bu yöntemin sakıncası çoklu korelasyon analizlerinin sonuçlarında tutarsızlık göstermesidir.

Eksik verilerin veri olarak yorumlanması. Bazı kişilerin ölçek veya testteki bazı maddelere yanıt vermemeleri, başka bir davranışın göstergesi olarak yorumlanır ve eksik veriler başlı başına nev-i şahsına özgü değerler olarak kabul edilir.

İkame ölçüm yöntemi. Örneklemde iptal edilen anketlerin yerine aynı ana kütleden seçilen başka kişilere yeni anketlerin uygulanması ve analizde bu anketlerden elde edilen değerlerin kullanılmasıdır.

Atama yöntemi. Atama yönteminde eksik veya tutarsız verilerin yerine istatistiksel hesaplamalara dayalı olarak elde edilen yeni değerler atanır. Atama yöntemi iki grupta değerlendirilir: mantıksal atama ve istatistiksel atama yöntemi. Mantıksal atama yönteminde eksik verilerin yerine önceki veriler, benzer sorulara verilen yanıtlar, diğer sorulardan kolaylıkla anlaşılan değerler atanır. Mantıksal atama yöntemi kullanılamıyorsa bu kez istatistiksel atama yöntemine başvurulur. İstatistiksel atama yönteminde regresyon analizi, aritmetik ortalama gibi belirli istatistikî teknikler kullanılır. Literatürde atama amacıyla kullanılan istatistiksel tekniklerin sayısı oldukça geniştir. Bu tekniklerin bir bölümü basit ve kolay anlaşılır türden

iken, diğerleri çok daha karmaşıktır. Karmaşık nitelikteki teknikler daha çok bilgisayar yazılımları aracılığıyla uygulanır. Bu kitapta söz konusu tekniklerden sadece birkaçı üzerinde durulmuştur.

İstatistiksel atama yöntemi de kendi içinde iki grupta değerlendirilir: deterministik ve stokastik atama yöntemi.[37] Deterministik yöntemde insan iradesinden bağımsız bir biçimde belirli hesaplamalara dayalı olarak bir değişkendeki eksik verilerin yerine sürekli olarak hep aynı değer atanır. Deterministik yöntemde *artık değer'in* (veya hata payının, tesadüfî parazitin) sıfır olduğu varsayılır. Bu yaklaşım, eksik değer için çok iyi bir tahmin değeri verir, fakat verilerin dağılımını bozar ve değişkenin varyansını zayıflatır. Stokastik yöntemde ise belirlilik değil, tesadüfîlik söz konusudur. Stokastik yaklaşımda bir değişkendeki eksik verilere atanan değerler rasgele belirlendiğinden ortaya çıkması muhtemel hatalar da (parazitler de) tesadüfîdir. Böylece değişken daha gerçekçi bir dağılıma sahip olur. Bu yaklaşımda, veri bulunmayan bir hücreye hangi değerin geleceği önceden bilinmez. Brick ve Kalton (1996, aktaran NCES) deterministik yaklaşımlara göre stokastik yaklaşımların daha gerçekçi sonuçlar vermesi nedeniyle genelde tercih edildiğini bildirmişlerdir.[38] Literatürde sık kullanılan atama yöntemlerinden bazıları aşağıdaki gibidir

Global sabit bir değerin ikame değeri olarak atanması. Bu yaklaşımda herhangi bir hesaplama yapılmadan belirli bir mantığa göre belirlenen sabit bir değer tüm eksik verilerin yerine atanır. Örneğin, beş dereceli Likert ölçeklerinde bu amaçla tam orta noktaya gelen 3 değeri kullanılabilir.

Sıcak-deste (hot-deck). Yöntem adını ilk çıkan bilgisayarlarda veri tanımlama amacıyla kullanılan delikli kartlar destesinden almıştır. Terimdeki *sıcak* sözcüğü "aynı veri dosyası" anlamına gelir. Bu yaklaşımda eksik veri içeren vak'aya en çok benzer başka bir vak'a bulunur (benzer yanıt modeline sahip) ve bu vak'adaki *Y* değerinin aynısı eksik *Y* verisi için atanır. Sıcak-deste yöntemi PRELIS isimli yazılımda mönü destekli olarak yapılır. Bu yöntemin ayrıca Imputer2, Solas isimli yazılımlarla da hesaplanabildiği belirtilmiştir. Sıcak-deste yaklaşımının değişik bir şekli, *ardışık sıcak-deste* ve *tesadüfî sıcak-deste* uygulamalarıdır. Ardışık yöntemde eksik veriye en yakın komşu vak'adaki veriler temel alınırken, tesadüfî yöntemde rasgele seçilen bir vak'adaki veri göz önünde bulundurulur.

Soğuk-deste yöntemi. Eksik verilerin yerine atanacak değerler, önceki araştırma sonuçlarına bakılarak belirlenir. Kuramsal temeli zayıf olan bu

yaklaşımın uygulamada nadiren kullanıldığı bildirilmiştir. Yaklaşım aynı zamanda *tarihsel ortalama* adı ile bilinir.

Ortalama yöntemi. Deterministik nitelikteki bu yöntemde eksik veri bulunan değişkenin genel olarak aritmetik ortalaması (veya duruma göre sınıf içi ortalaması) alınarak bu ortalama değeri eksik verilerin yerine yazılır. Aritmetik ortalama yerine nonparametrik verilerde medyan değeri de kullanılabilir. Bilim adamları ortalama yönteminin parametre tahminlerinde yanlılığa neden olduğunu bildirmişlerdir. Eksik veriler için ortalama yönteminin kullanıldığı değişkenlerde hesaplanan varyans, gerçek varyansı olduğundan daha düşük gösterir. Korelasyonlarda negatif değerler elde edilir ve yeni değerlerin dağılımı ana kütleyi iyi bir şekilde temsil etmez.

Downey ve King (1995) tutum ölçeklerindeki eksik verilerin giderilmesi için *madde ortalaması ikame değeri* (MOİD) ve *kişi ortalaması ikame değeri* (KOİD) olmak üzere iki yaklaşım önermişlerdir. Bu bilim adamları yaptıkları hesaplamanın sonucunda her iki yöntemin de ekstrem veri eksikliği durumlarında dahi orijinal veri yerine kullanılabileceğini göstermişlerdir. Bu araştırmada orijinal verilerle atanmış veriler arasındaki korelasyon katsayısı ,90'ın üzerinde çıkmıştır. Ancak KOİD kullanıldığında varyans ve güvenilirlik katsayıları bir ölçüde şişkin çıkmıştır. Downey ve King eğer güvenilirlikle daha fazla ilgileniliyorsa MOİD yöntemine başvurulmasının daha doğru olacağını belirtmişlerdir (aktaran King, 2003).[39] King kendi yaptığı araştırmasında da benzer sonuçlara ulaşmış ve cevaplayıcıların %20'sini ve maddelerin %30'unu aşmadığı ölçüde eksik verilerin yerine atama yapılabileceğini belirtmiştir.[40]

İstatistiksel analiz yazılımı SPSS'te eksik verilerin yerine aritmetik ortalama değerini atamak için Transform mönüsünden Replace Missing Values düğmesi seçilir ve açılan diyalog kutusunda atama modeli belirlenir.

Regresyon yöntemi. Eksik verilerin regresyon analizi yöntemi uygulanarak bu analiz sonucunda elde edilen değerlerle doldurulması anlamına gelir. Bu yöntemde eksik veri bulunan değişkenle yüksek derecede korelasyon ilişkisi gösteren diğer yardımcı değişkenler regresyon analizine tâbi tutulur. Atanan verilerin doğru veya iyi olması, tahmin etmek için kullanılan regresyon modelinin güçlü olmasına bağlıdır. Garson bu yöntemi, verilerde gerçekçi olmayan bir biçimde düşük düzeyde hata içererek ikame değerini olması gerekenden daha yüksek gösterdiği için eleştirmiştir. Bu nedenle bazı çoklu regresyon modellerinde atanan tahmin değerlerine kullanıcı tanımlı tesadüfî hata birimlerinin eklenmesine izin verilir. Örneğin,

SPSS'te regresyon tahmini yapılırken bu amaçla tesadüfen alınan veya seçilen bir vak'anın *artık değeri* hesaplanan tahmin değerine ilave edilir.[41]

Veri ataması amacıyla değişik regresyon yöntemlerinden veya modellerinden yararlanılır: doğrusal regresyon modeli, genelleştirilmiş doğrusal model, birikimli model, genelleştirilmiş birikimli model. Regresyon yönteminin ancak küçük veri yapılarında istenen sonucu verdiği bildirilmiştir. Büyük veri yapılarıyla çalışan araştırmacıların diğer atama yöntemlerini de incelemelerinde yarar vardır.

Stokastik regresyon (stochastic regression). Regresyon analizi ile tahmin edilen değer, *tesadüfî artık bir değerle* toplanır.

Enterplasyon ve ekstrapolasyon yöntemi. Aynı kişiye ait gözlem değerlerinden hareket edilerek ileriye yönelik bir tahmin yapılması esasına dayanır.

Maksimum olasılık tahmini. Bu yöntemde herhangi bir veri ataması yapılmaz, ancak bunun yerine her bir vak'a için gözlem verileri kullanılarak maksimum olasılık tahmininde bulunulur. Bu yöntem çoklu atama yaklaşımında olduğu gibi standart hatayı daha yansız olarak gösterir. Analiz AMOS ve SPSS'in eksik veri analizi modüllerinde bulunur.

Beklenti maksimizasyonu yöntemi. Smith'e göre, (2003) "Tekrarlamalı regresyon tekniği olan bu yaklaşımda, eksik veri bulunan değişkenler mevcut olan diğer verilerle regresyona tâbi tutulurlar ve analiz sonucunda ek değişkenler ortaya çıkarılır. Birinci aşamada bütün verilere dayalı olarak vektör ortalamaları ve kovaryans matrisi hesaplanır. Daha sonra ortalama değerleri her bir değişkene ikame değeri olarak atanır. Atanmış ortalama değerleri nihaî atama işlemi için bir başlangıç veya çıkış değeri olarak hizmet eder. İkinci aşamada ise, eksik veri içeren değişkenler diğer değişkenlerle birlikte regresyona tâbi tutulur. Bu kez, atanmış ortalama değerleri regresyon eşitliği ile hesaplanmış tahmin değerleriyle yer değiştirir. Yeni atanmış değerlere bağlı olarak aritmetik ortalama ve kovaryans değerleri yeniden hesaplatılır. Bundan sonra regresyon eşitliği ve değer atama işlemlerine aritmetik ortalama ile kovaryans matrisi değerleri bir noktada birleşinceye kadar devam edilir."[42]

Çoklu atama yöntemi. Donald B. Rubin (1970) tarafından geliştirilen çoklu atama yöntemi ölçümü yapılan parametrelerin tahmin değerlerini tekli atama yöntemine göre daha yansız olarak belirleyen bir tekniktir. Adından anlaşıldığı gibi çoklu atama yönteminde bir değer birkaç defa

(tipik olarak 3 veya 5 defa) atanır (*bk.*, Şekil 4-7). Her bir atama tesadüfen seçilen farklı bir hata terimiyle yüklenir. Çoklu atama yöntemi parametrik ve nonparametrik veriler için ayrı ayrı yapılır.[43] Bu yöntemde TYVA ve lojistik regresyon analizi gibi teknikler kullanılır ve daha sonra elde edilen sonuçlar birleştirilir. Çoklu atama yönteminin SAS, S-Plus ve Solas isimli yazılımlarda bulunduğu bildirilmiştir. İstatistik yazılımı SAS'ın bir alt modülü olan PROC MIANALYZE çok değişkenli verilerde görülen eksik verilerin yerine çoklu atama prosedürünü uygulayarak yeni değerler atar (*bk.*, Tablo 4-7).

Tablo 4-7. Çoklu Atama Yöntemi

Vak'alar	*V*1	*V*2	*V*3	*V*4	Atanan değerler
1	?				1, 2, 3 ... *m*
2			?		1, 2, 3 ... *m*
3		?			1, 2, 3 ... *m*
4			?		1, 2, 3 ... *m*
5				?	1, 2, 3 ... *m*
:	:	:	:	:	:
k		?			1, 2, 3 ... *m*

Veri atamasının uygun olmadığı durumlar. Her tür veri eksikliği için atama yöntemine başvurmak uygun olmayabilir. Özellikle küçük örneklemlerde veya tam sayıma dayanan ana kütle ölçümlerinde atama yöntemini kullanmanın bazı sakıncaları vardır. Örneğin, öğrencilerin öğretmenlerini, meslektaşların birbirlerini ve yöneticilerin personelini değerlendirdikleri çalışmalarda ortaya çıkan eksik veriler için atama yöntemi kullanılmaz.

Eksik veri içeren dosyaların korunması. Bilgisayarda ham verilerin bulunduğu dosya ile veri ataması yapılmış dosyalar ayrı isimlerle kayıt edilmelidir. Çünkü değişik amaçlarla ham verilerin bulunduğu dosyadan yararlanmak gerekebileceğinden bu dosyanın ayrı bir isimle saklanmasında yarar vardır. Bilim adamı ölçek ve testin güvenilirliğini eksik verili dosya ve atama yöntemiyle düzeltilmiş dosyaların her ikisinde de test ederek sonuçlarda anlamlı bir değişiklik olup olmadığını gözlemelidir.

Eksik veri yazılımları. İstatistiksel analiz programı SPSS'te eksik veriler, bu amaçla hazırlanmış olan SPSS Missing Value Analysis adlı bir modülle incelenerek eksik verilerin yerine yeni değerlerin ataması yapılır. Standart modülde ise eksik veriler veri giriş tablosuna üç şekilde tanımlanır: ayrık değerler olarak, eksik veriler dizisi olarak veya her iki yöntemi de içerecek şekilde. Eksik veri analizi yaparak eksik verilerin yerine yeni değerleri ikame eden diğer yazılımlar şunlardır: NORM, SOLAS, AMELIA, LISREL, PRELIS, R, IVEware, HLM, CAT, MIX, PAN. Avrupa Birliği istatistik komisyonun EUREDIT projesi kapsamında bilgisayar programlarıyla otomatik hale getirilmiş atama yöntemleri altı kategoride toplanmıştır: deterministik atama, model temelli atama, deste temelli atama, karma atama yöntemi, uzman sistemler atama yöntemi ve yapay sinir ağları atama yöntemi. Son iki yaklaşımda yapay zeka sistemleri kullanılarak değer ataması yöntemine başvurulur. Bu konuda daha fazla bilgilenmek isteyen okurlar bu yazılımlar hakkında ayrıntılı bilgiyi literatürden ve İnternet'teki Ağ kümelerinden edinebilirler.

TÜRDEŞSELLİK VE DOĞRUSALLIĞIN TEST EDİLMESİ

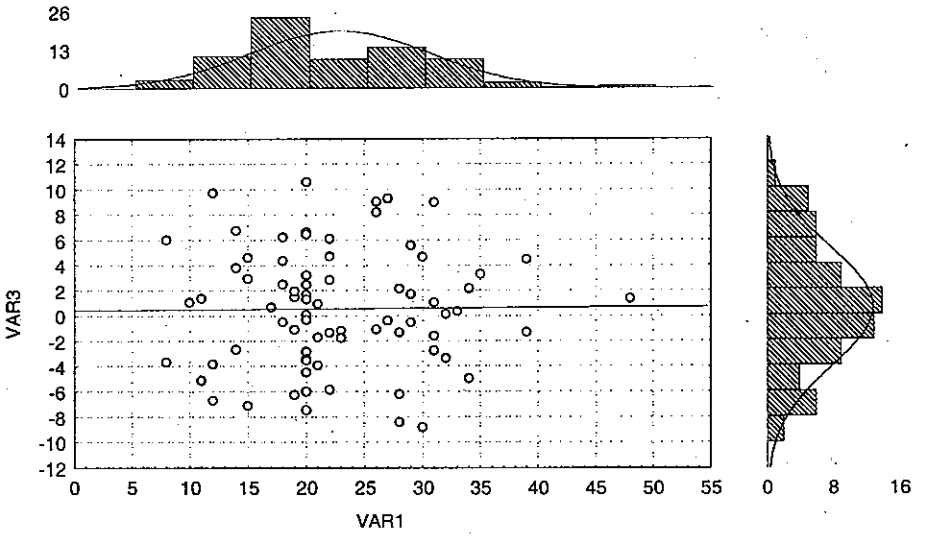
Türdeşsellik[a] (homoscedasticity) aralarında ilişki aranılan iki değişkenden birindeki değişkenliğin ikinci değişkendeki değişkenlikle aynı veya büyük ölçüde benzer olması anlamına gelir. Her iki değişkendeki veriler normal dağılım özelliği gösteriyorsa türdeşsellik sağlanmıştır denilir. Değişkenlerin her ikisi de veya biri normal dağılım özelliğinden uzaklaşmışsa türdeşsellik bozulur ve bu duruma ayrısallık (heteroscedasticity) adı verilir. Ayrısallık özelliğinde veriler bütünüyle geçersiz olmasa bile büyük ölçüde zayıflar. Verilerin türdeşsellik ve ayrısallık özelliği nokta dağılım grafiği ile test edilir ve dönüştürme yöntemleri kullanılarak düzeltilir.[44] Türdeşsellik ve ayrısallık, özellikle korelasyon analizi ve regresyon analizi bulgularının geçerliliği açısından önemlidir. Veriler ayrısallık özelliği gösteriyorsa korelasyon katsayısı *r* iyi bir tahmin göstergesi değildir (*bk.*, Şekil 4-7).

Korelasyon analizinde veriler arasındaki ilişkilerin doğrusal nitelikte olduğu varsayılır. İki değişken arasındaki ilişkiler nokta dağılım grafiğinde karmaşık bir şekilde dağılmış bir görüntü ortaya çıkarmışsa doğrusallık özelliği sağlanamamış demektir. Böyle bir durumda doğrusal olmayan ilişkilerden söz edilir. İki değişken arasındaki ilişkiler çok güçlü olsa bile

---

[a] Kelimenin *türdeşlik* yerine *türdeşsellik* olarak çevrilmesinin nedeni türdeşliği tek bir veri dizisinde değil aynı zamanda iki veri dizisinde birden araştırıyor olması nedeniyledir. Bazı bilim adamları türdeşselliği *sabit varyans* ve ayrısallığı ise *değişken varyans* terimleriyle karşılamışlardır.

bu ilişkiler örneğin, karmaşık veya eğrisel bir niteliğe sahipse korelasyon katsayısı *r* çok küçük bir değerdir veya sıfır çıkar. Verilerin karmaşık dağılım özelliğine doğrusallık teriminin karşıtı olarak *karmaşıklık* veya *doğrusuzluk* (nonlinearity) adı verilir. Doğrusuzluk verilerin eğrisel biçim de dahil olmak üzere herhangi bir biçimde dağılacağı anlamına gelir. Doğrusuzluk aynı zamanda *kaos* veya *belli bir düzenin bulunmaması* anlamında da kullanılmıştır. Veriler doğrusallıktan uzaklaşmışsa korelasyon analizi yöntemini uygulamak doğru olmaz. Doğrusallık ve türdeşsellik özelliği göstermeyen verilerin *nokta-dağılım grafiğinde* noktaların dağılım biçimi, futbol topu görünümüne sahip olur. Bu tür verilerin yorumlanmasında sadece *r* değeri değil aynı zamanda *X* değişkenin ortalaması ve standart sapmasıyla *Y* değişkenin ortalaması ve standart sapması da göz önünde bulundurulur.[45] Doğrusal olmayan ilişkilerin yorumlanmasında yararlanılabilecek bir diğer yöntem eta katsayısına başvurmaktır. Eta katsayısı *r* katsayısından daha büyük çıkar.



Şekil 4-7. Verilerde türdeşsellik ve ayrısallık özelliği.

Şekil 4-7'de üç numaralı değişkene ait verilerin normal dağılım özelliğine sahip olduğu anlaşılmaktadır. Bir numaralı değişkene ait veriler ise sağa çarpıktır. Bu nedenle ilişki katsayısı $r = ,01$ olan bu verilerin türdeşsellik koşulunu karşılamadığına karar verilir. Değişkenlerden birinde doğ-

rusallık özelliğinin sağlanmış olması fakat diğerinde sağlanamaması verilerin ayrısallık (yeknesak olmayan varyans) özelliğini gösterir.

ÇOKLU DOĞRUSALLIK VE TEKİLLİĞİN TEST EDİLMESİ

Çoklu doğrusallık veya koşutluk[a], özellikle korelasyon analizlerinde, regresyon analizinde ve çok değişkenli istatistiksel analizlerde önemlidir. Koşutluk, değişkenler arasında veya regresyon analizinde bağımsız değişkenler arasında kabul edilemeyecek yüksek ilişkililik oranını tanımlar. İlişki katsayısı ,90 veya bu rakamın üzerindeyse koşutluk özelliğinden şüphelenilir. Tekillik ise iki değişken arasında tam bir ilişki bulunması anlamına gelir. Koşutluk ve tekillik değişken sayısındaki fazlalığa işaret eder. Şişkin özgünlük yaratan maddeler, koşutluğa neden olur. Regresyon analizinde bağımsız değişkenler arasındaki korelasyon katsayılarının düşük olması gerekir. İlişkililik derecesi yüksekse bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini saptamak güçleşeceğinden sağlıklı bir sonuç alınamaz. Bağımsız değişkenler arasında ,80'in üzerindeki bir ilişkililik derecesi, *koşutluk* özelliği ile karşı karşıya olduğumuz anlamına gelir.[46]

Koşutluk özelliğini test etmenin bir diğer yöntemi regresyon analizi sonucunda elde edilen *tolerans değerine* bakmaktır. Tolerans değeri, $1–R^2$ olarak bilinir. Formüldeki R^2 bağımsız değişkenlere ait çoklu korelasyon katsayısının karesidir. Tolerans değeri eğer ,20 gibi bir değerin altında ise bağımsız değişkenler analize alınmaz. Bu yöntem ,80 kuralından daha iyidir çünkü bağımsız değişkenlerin etkisini birlikte ele alıp değerlendirmektedir. Verilerin koşutluk özelliği taşıyıp taşımadığını test etmek için yararlanılan bir başka değer regresyon analizi sonucunda elde dilen *varyans şişme faktörü*'dür (variance inflation factor – VIF). Ölçüm yapılan X değişkeni eğer fazla/gereksiz bilgi içermiyorsa varyans şişme faktörü (VŞF) 1'e eşit çıkar. Fakat ortak doğrusallık özelliğine sahipse VŞF değeri 1'den yüksek çıkar.[47] Varyans şişme faktörü 4,0'ten büyük olduğunda verilerde koşutluk sorununa işaret eder.

---

[a] Literatürde istatistikî terimler genellikle tek bir kelime ile ifade edilmeye çalışılır. Bu nedenle İngilizce *multicollinearity* kelimesinin karşılığı olarak *çoklu doğrusallık* veya *koşutluk* kelimelerini öneriyoruz. Yabancı dillerdeki terimlere kolay anlaşılır bir karşılık bulmak her zaman mümkün olmamaktadır. Okuyucunun, kelimenin sözlük anlamından çok yüklenen anlamı öğrenmesi doğru olur.

AYRIK DEĞER ANALİZİ

Ayrık değerler bir dizide diğer rakamlardan farklı olan ve bu nedenle araştırmacının dikkatini çeken verilerdir. Ayrık değerler gerçeği yansıtabileceği gibi, hatalı veya gerçek dışı rakamlar da olabilir. Tek değişkenli bir dağılımda ayrık değerler aritmetik ortalamayı ve dizinin varyansını etkiler. İki değişkenli dağılımlarda ise özellikle korelasyon ve regresyon hesaplamalarında etkili olur. Korelasyon hesaplamalarında bir taraftan korelasyon katsayısını sunî bir şekilde arttırırken öte yandan verilerin niteliğine göre bazen gerçek korelasyon rakamını olduğundan daha düşük gösterir.[48] Verilerde eğer ayrık değerler varsa korelasyon katsayısı *r* iki değişken arasındaki ilişkilerin gücünü göstermekte kullanılacak iyi bir değer değildir.

Ayrık Değerlerin Saptanması

Ayrık değerlerin saptanması birkaç şekilde yapılabilir. Birincisi st*a*ndart *z* puanlarından yararlanmaktır. Örneklemdeki vak'a sayısı 50'den az ise –2,5 ilâ +2,5 değerinin dışında kalan değerler ayrık değer olarak değerlendirilir. Vak'a sayısı daha büyük ise bu kez –3,3 ilâ +3,3 değerinin dışında kalan değerler ayrık değer olarak kabul edilir. Vak'a sayısı 1000'den fazla ise ±3,3'ün üzerindeki değerler dahi normal değer olarak yorumlanır.[49] Ayrık değerleri saptamakta yararlanılabilecek ikinci yöntem grafik çizimlerine başvurmayı gerektirir. Tek bir değişkendeki ayrık değerler için histogram ve kutu-bıyık grafikleri; iki değişkendeki ayrık değerlerin saptanması için ise nokta dağılım grafiği kullanılır.

Ayrık Değerlerin İyileştirilmesi

Ayrık değerlerin varlığı saptandıktan sonra bu değerlere ne yapılacağı konusu bilim adamın zihnini meşgul eder. Ayrık değerin niteliğine göre bazı önlemler alınabilir. Bunlardan birincisi ayrık değerlerin belirli değişkenlerde ortaya çıkıp çıkmadığının araştırılmasıdır. Eğer hep aynı değişkende ortaya çıkıyorsa ve bu değişken de analiz açısından hayatî bir öneme sahip değilse değişken analizden çıkarılabilir. İkinci yöntem ayrık değer içeren vak'anın veri matrisinden çıkarılmasıdır. Üçüncü yöntem ayrık değerlerin araştırma veya ölçüm yapılan belirli bir alt gruba ait olup olmadığının araştırılmasıdır. Eğer belirli bir gruptaki kişiler ayrık değerler vermişlerse bu kişilerin analize alınmamasının ne gibi etki doğuracağı araştırılır. Dördüncü yöntem ayrık değerin belirli tahminlere dayalı olarak değiştirilmesidir. Örneğin, bu çerçevede ayrık değer yerine grubun ortalama değeri atanır. Beşinci yöntem ise verilerin dönüştürülmesidir.[50] Örneğin, karekök dönüş-

türme yöntemi ayrık değerleri diğer değerlere yakınlaştırır fakat bu kez dönüştürülmüş değerlerin yorumlanmasında güçlükle karşılaşılır.

Bilim adamı, ayrık değerleri temizleme konusunda dikkatli olmalıdır. Kimi araştırma konularında veri "temizliği" yapmak kesinlikle gereklilik haline gelir. Psikoteknik testlerde reaksiyon zamanı ölçülürken katılımcıların puanları 300 ilâ 700 milisaniye arasında değişir. Bu tür ölçümlerde 10–15 saniyelik uç reaksiyon zamanları ölçüm profilini bütünüyle değiştireceğinden bu değerler *ayrık rakam* olarak kabul edilir ve değerlendirme dışı bırakılır. Uç veya ayrık değer tanımlaması sübjektif bir değerlendirmedir. Araştırmacı bu değerlendirmeyi yaparken verilerin dağılımını, deneysel paradigmaları, önceden kabul edilmiş standartları göz önünde bulundurur.[51]

PUANLARIN STANDARTLAŞTIRILMASI

Güvenilirlik analizlerini yapmak için ham puanların standart puanlara dönüştürülmesi gibi standart bir uygulama söz konusu değildir. Ancak, test ve ölçeklerden elde edilen ham puanlarla (toplam veya ortalama olabilir) çalışarak güvenilirlik analizleri yapmak her zaman doğru olmayabilir. Örneğin, gözlemciler yaptıkları değerlendirmelerde farklı puan baremleri kullanmışlarsa, yarıya bölme güvenilirliğinde testin iki yarısında farklı sayıda madde varsa; paralel formlar farklı sayıda madde içeriyorsa, ölçek dereceleri farklıysa veya maddelere ağırlık verilmesi nedeniyle toplam puanlar arasında önemli ölçüde farklılık varsa bu ölçüm araçları arasındaki güvenilirliği hesaplayabilmek için toplam veya ortalama ham puanların standart puanlara çevrilmesi gerekir. Madde sayıları, puanlama sistemleri ve puanlama dereceleri farklı iki test doğrudan karşılaştırılamaz.[52]

Ham puanların standart z puanlarına dönüştürülmesiyle puanların dağılım şeklinde bir değişiklik ortaya çıkmaz. Ham puanlar eğer sağa dayalı ise z puanlarına dönüştürmeyle dağılım eğrisi yine sağa dayalı olarak çıkar. Dönüştürme, puanları normal dağılım eğrisiyle ifade etme imkanı sağlar, fakat puanların dağılım biçimini etkilemez. Dönüştürme normal dağılım eğrisini elde etmek için değil, birikimselliği sağlamak için yapılır. Birikimsellik aynı puan cinsinden toplama yapmaya imkan verme anlamındadır. Bu işlem sonucunda birikimsellikle birlikte yan ürün olarak normalleşme de sağlanmış olur.

Sınav temelli bilgi ve yetenek testlerinden elde edilen ham puanların da güvenilirlik analizini yapmadan önce belirli ölçütler doğrultusunda düzeltilerek standardize edilmesi gerekir. Puanların dönüştürülmesi, sadece istatistikî

test koşullarının sağlanması açısından değil, güvenilirlik ve geçerlilik analizleri için de düşünülmelidir.

Dönüştürme işlemine karar vermek için göz önünde bulundurulması gereken temel ölçüt, istikrarlılık ve eş değerlilik ölçümlerinde puanların farklı bir niteliğe sahip olup olmadığıdır. Bir diğer kriter, puanların normal dağılım eğrisine sahip olmamasıdır. Öte yandan iç tutarlılık işlemi için puanları standardize etmeye gerek yoktur. Ancak bir test/ölçek birden fazla faktör içermesi nedeniyle alt testlerden veya ölçeklerden oluşuyorsa, alt testlerin her birinde farklı sayıda madde varsa araştırmacı puanları karşılaştırılabilir duruma getirmek için standartlaştırma yöntemine başvurmalıdır. Standartlaştırma işleminde sık kullanılan yöntemler, formül puanları, z puanları, T puanları, standart dokuz, standart on ve norm standardı yaklaşımlarıdır. Aşağıdaki bölümde bu yaklaşımlar ele alınmıştır.

Formül Puanları

Çoktan seçmeli bilgi testlerinde, kültürden bağımsız psikometrik yetenek testlerinde en önemli sorun testi alan kişilerin bilemedikleri soruların yanıtları için tahmin yürütmeleridir. Bu arada kişiler bazı soruları veya test maddelerini yanıtsız bırakmış olabilirler. Teste ait toplam puanlar belirlenirken bu iki olguya ait hata faktörünün düzeltilmesi ve güvenilirlik analizlerinin daha sonra yapılması yoluna başvurulur (*bk.*, Eşitlik 4-7).

$$X_t = \left(\frac{k-1}{k}\right)X + \frac{N-t}{k}. \qquad (4\text{-}7)$$

X_t = Tahmin faktörü için düzeltilmiş gözlem puanı.
k = Her bir maddedeki cevap şıkkı, derece sayısı.
N = Toplam madde sayısı.
t = Yanlış cevap sayısı.

Standart z Puanları

Standart z puanlarında veriler, ortalaması 0 ve standart sapması 1 olacak şekilde yeniden düzenlenir. Ancak bazı araştırmacılar bu yöntemi yorumlama ve değerlendirme açısından yeterince pratik bulmadıklarından standart T puanlarıyla çalışmayı yeğlerler. Örneğin, bir gruba test uygulanmış ve grubun aritmetik ortalaması 70 ve standart sapması 2,8 çıkmışsa bu grupta 68 alan bir kişinin ortalamadan kaç standart sapma uzağa düştüğünü bulmak

için şu formül uygulanır: $(X - M) / SS$. Hesaplama sonucunda –,71 gibi bir değer elde edilir ve kişinin yaklaşık 1 standart sapmaya yakın negatif alanda yer aldığı tespit edilir.

Standart *T* Puanları

Standart *T* puanları çalışılması daha kolay ve anlamlı olan bir görünüme sahiptir. Uygulamada standart *z* puanlarını *T* puanlarına dönüştürmek için Eşitlik 4-8'deki formül kullanılır.

$$T = 10z + 50. \tag{4-8}$$

Norm Standardı

Bu uygulamada norm grubunun ortalama ve standart sapma değerleri kriter olarak kabul edilir. Örneğin TEOFL testinde kişilerin aldıkları puanlar norm grubunun ortalama ve standart sapma değerleriyle karşılaştırılır. Ulusal ve uluslar arası düzeyde düzenlenen bütün büyük test uygulamalarının kendi norm standartları vardır. Norm standardı, *z* puanları elde edildikten sonra Eşitlik 4-9'daki formülle hesaplanır.

$$Y_i = \sigma^* \left(\frac{X_i - \mu_x}{\sigma_x} \right) + \mu^*. \tag{4-9}$$

σ^* = Norm grubunun standart sapma değeri.
μ^* = Norm grubunun aritmetik ortalama değeri.
X_i = Bireysel gözlem değeri.
μ_x = Gözlem değerlerinin ortalaması.
σ_x = Gözlem değerlerinin standart sapması.

Standart Dokuz Puanları

Bu uygulamada ham puanlar 1 ilâ 9 arasındaki sayı birimlerine göre standartlaştırılır. Normal dağılım özelliği gösteren bir dizide standart dokuz puanlarının aritmetik ortalaması 5,0 ve standart sapması ise 1,96'dır.[53]

Standart On Puanları

Bu uygulamada ham puanlar 1 ilâ 10 arasındaki sayı birimlerine göre standartlaştırılmıştır. Normal dağılım özelliği gösteren bir dizide standart on puanlarının aritmetik ortalaması 5,5 ve standart sapması ise 2,00'dir.[54]

Diğer Standart Puanlar

Literatürde, yukarıda açıklananların dışında daha farklı standartlaştırma yöntemleri de söz konusudur. Yaş puanları, 11'li ölçek biriminin temel alındığı C ölçekli puanlar (c-scaled scores), sınıf eşlik değerleri (grade equivalent), harfli puanlandırmalar, normal eğri eşlik değeri (normal curve equivalent), yüzdelik sırası, yetkinlik düzeyi (acemi, çırak, kalfa, usta, uzman), IQ puanları gibi yöntemler bunlar arasındadır.

ALINTI YAPILAN KAYNAKLAR

[1] B. Trochim, "Data Preparation [Veri Hazırlama]," <http://trochim.human.cornell.edu/kb/statprep.htm> (14.06.2003).

[2] Shizuka, "Statistical Concepts For Test Theory [Test Kuramı için İstatistiksel Kavramlar]," <http://www2.ipcku.kansai-u.ac.jp/~shizuka/class/POSTGRAD/ L02%20Stat%20 Concepts.rtf> (02.02.2003).

[33] "Item Analysis [Madde Analizi]," <http://www.sph.uth.tmc.edu:8053/behsci/lmasse/ph1130/Item%20analysis.rtf> (28.12.2002).

[4] Fred.N. Kerlinger, *Foundations of Behavioral Research,* İkinci Baskı, (New York:, Holt, Rinehart and Winston, 1973), 74.

[5] R.A. Hannemen, "Cronbach's Alpha Reliability Analysis With SAS [SAS İle Cronbach Güvenilirlik Analizi]," <http://faculty.ucr.edu/~hanneman/soc203b/examples/alpha.html> (02.12.2002).

[6] E. Rigdon, "Important Continuous Statistical Distributions [Önemli Sürekli İstatistiksel Dağılımlar]," <http://www.gsu.edu/~mkteer/continuo.html> (25.01.2003).

[7] D. Garson, "Testing of Assumptions [Varsayımların Test Edilmesi]," t.y., <http://www2.chass.ncsu.edu/garson/pa765/assumpt.htm> (25.01.2003).

[8] Monash University, "Guidelines For Grading Honours And Masters Theses In Relation To Statistics [Master Tezlerinde İstatistik Test Sonuçlarının Raporlanması]," <http://www.med.monash.edu.au/psych/research/rda/Honours%20grades.htm> (07.01.2003).

[9] Rasch Measurement Transactions "New Rules of Measurement [Ölçümün Yeni Kuralları]," <http://www.rasch.org/rmt/rmt132e.htm> (18.01.2003).

[10] Prophet Stat Guide, "Normal Distribution Tests [Normal Dağılım Testi]," <http://www.basic.nwu.edu/statguidefiles/n-dist.html> (12.01.2003).

[11] BBN Corporation, "Do Your Data Violate Normality Test Assumptions? [Verileriniz Normallik Test Varsayımlarınızı İhlal Ediyor mu?]," 1996, <http://www.basic.nwu.edu/statguidefiles/n-dist_ass_viol.html> (12.01.2003).

[12] SAS Institute, "Test for Normality [Normallik Testi]," <http://mathstat.carleton.ca/~help/sashtml/qc/chap1/sect19.htm> (25.01.2003).

[13] BBN Corporation, "Examining Normality Test Results [Normallik Test Sonuçlarının Gözden Geçirilmesi]," 1996, <http://www.basic.nwu.edu/statguidefiles/n-dist_exam_res.html#Stephens> (18.01.2003).

[14] Glossary, "Anderson-Darling Test," <http://sunsite.univie.ac.at/textbooks/statistics/glosa.html> (15.01.2003).

[15] "Anderson-Darling Test," <http://www.math.auc.dk/~svante/VS7/l4/esh/ad0.html> (15.01.2003).

[16] Glossary, "Anderson

[17] C. Annis, "Goodness-of-Fit tests for Statistical Distributions [İstatiksel Testlerde Uygunluk Testleri]," <http://www.statisticalengineering.com/goodness.htm> (15.01.2003).

[18] Aktarılan kaynak, Den Norske Dataforening, "Implementering av IT-strategi", <http://dataforeningen.no/publikasjoner/fagartikler_rapporter/rapport_19980101b.php> (11.01.2003).

[19] Aynı.

[20] Aynı.

[21] B.Shi, "Vavelet and Fractal" <http://www.isye.gatech.edu/~bshi/research.htm> (11.01.2003).

[22] "Module Four [Bölüm Dört]," <http://www.unb.ca/courses/mhodgins/n6051/module3b.htm> (11.01.2003).

[23] A. Heathcote ve S. Paolini, "Laboratory Notes [Laboratory Notes [Laboratuvar Notları]," <http://216.239.37.100/search?q=cache:n3DPVKBvuHUC:sbs.newcastle.edu.au/~jbowman/2070%25202002%2520Experimental%2520Methodology/Lab%2520Notes/PSYC207-2002lab2notes.doc+kurtosis+spss+interpretation&hl=tr&ie=UTF-8> (25.01.2003).

[24] D. Garson, "Testing of Assumption [Varsayımların Test Edilmesi]," <http://www2.chass.ncsu.edu/garson/pa765/assumpt.htm> (20.01.2003).

[25] BBN Corporation, "Prophet Statguide: Possible Alternatives If Your Data Violate Normality Test Assumptions [Veriler Normallik Test Varsayımlarını İhlal Ediyorsa Muhtemel Çözümler]," <http://www.basic.nwu.edu/statguidefiles/n-dist_alts.html> (07.01.2003).

[26] C. Rorden, "Statistics [İstatistik]," <http://www.psychology.nottingham.ac.uk/staff/cr1/anova2g.pdf> (20.01.2003).

[27] Garson, "Testing of Assumption."

[28] M.R. Hyman, "A Critique and Revision of the Multidimensional Ethics Scale [Çok Boyutlu Etik Ölçeğinin Gözden Geçirilmesi ve Kritiği]," <http://www.empgens.com/Pubs/jems/MES/MES.html> (01.02.2003).

[29] Monash University, "Guidelines For."

[30] Aynı.

[31] Netfirms, "Normal Distribution [Normal Dağılım]," <http://lunney.netfirms.com/vita/stat231/labs/STATLAB5_an.html> (18.01.2003).

[32] "Module Four [Bölüm Dört]." <http://www.unb.ca/courses/mhodgins/n6051/module3b.htm> (11.01.2003).

[33] J. Laiho ve d., "Statistical Editing and Imputation [İstatistiksel Tashih ve Atama]," <http://www.google.com.tr/search?q=cache:GCal3W8pyZsJ:www.stat.fi/tk/tt/laatuatilastoissa/lm020900/sp_en.html+missing+value+Imputation&hl=tr&ie=UTF-8> (17.06.2003).

[34] National Center for Education Statistics, "Nonresponse Bias Analysis [Cevaplamama Oranı Yanlılığı]," <http://nces.ed.gov/statprog/2002/std4_4.asp> (22.06.2003).

[35] National Center for Education Statistics, "Processing and Editing of Data [Verileri Düzeltme ve İşleme]," <http://nces.ed.gov/statprog/2002/std4_1.asp> (22.06.2003).

[36] A.L. Cool, "Dealing With Missig Data [Eksik Verilerle İlgilenme]," <http://ericae.net/lt/tamu/cool1.pdf.> (21.06.2003).

[37] National Center For Education Statistics, "Evaluating The Impact of Imputations For Item Nonresponse [Eksik Verilerin Yerine Atama Yapmanın Etkisi]," <http://nces.ed.gov/statprog/2002/appendixb3.asp.> (17.06.2003).

[38] National Center For Education Statistics, "Evaluating The Impact."

[39] C.V. King, "Mean Subtitution for Missing İtems [Eksik Maddeler İçin Ortalama İkamesi]," <http://www.populus.com/tech_papers/mean_substitution.pdf.> (21.06.2003).

[40] Aynı.

[41] D. Garson, "Data Imputation for Missing Values [Eksik Veriler İçin Değer Ataması]," <http://www2.chass.ncsu.edu/garson/pa765/missing.htm> (21.06.2003).

[42] B.L. Smith, "Exploring Imputation Techniques for Missing Data [Eksik Veriler İçin Atama Tekniklerinin Araştırılması]," <http://www.wsdot.wa.gov/ppsc/research/TRB_Special/TRB2003-000894.pdf> (22.06.2003).

[43] J.J. Hox, "A Review of Current Software for Handling Missing Data [Eksik Verileri Ele Alan Yazılımların Genel Bir Değerlendirmesi]," <http://www.fss.uu.nl/ms/jh/publist/misrevkm.pdf> (17.06.2003).

[44] Canadian Forest Service, "Homoscedasticity [Türdeşsellik]," <http://www.pfc.forestry.ca/profiles/wulder/mvstats/homosced_e.html> (15.06.2003).

[45] P.BB. Stark, "Corralation and Association [Korelasyon ve İlişki]," <http://stat-www.berkeley.edu/users/stark/SticiGui/Text/ch4.2.htm> (15.06.2003).

[46] D. Garson "Testing of Assumptions [Ön Kabullerin Testi]," <http://www2.chass.ncsu.edu/garson/pa765/assumpt.htm#transforms> (15.06.2003).

[47] InStat, "Is multicollinearity a problem? [Koşutluk Bir Sorun mudur?]," <http://www.graphpad.com/instatman/Ismulticollinearityaproblem_.htm> (15.06.2003).

[48] StatSoft Inc., "Basic Statistic [Temel İstatistik]," <http://www.statsoftinc.com/textbook/stbasic.html#Correlationse> (09.10.2002).

[49] "Outliers [Ayrık Değerler]," <http://www.gseis.ucla.edu/courses/ed231a/lecture/Course5N.doc> (15.06.2003).

[50] Aynı.

[51] Aynı.

[52] Penn State University, "Academic Testing Test Design and Construction [Akademik Test Tasarımı ve Test Oluşturma]," <http://www.uts.psu.edu/Test_construction_frame.htm> (23.12.2002).

[53] Colorado Center for Teaching, Learning and Technology, "Student Data Handbook [Öğrenci Veri El Kitabı]," <http://www.ctlt.org/documents/us_oec/student_Data_Handbook.doc> (11.08.2003).

[54] Aynı.

CRONBACH ALFA GÜVENİLİRLİK ANALİZLERİ

Alfa değerleri, klasik test kuramına göre hazırlanan çoklu veri yapısına sahip test ve ölçekler için uygundur. Madde-yanıt kuramına göre hazırlanan testlerde Cronbach alfa hesaplaması yapılmaz. Cronbach alfa güvenilirlik analizleri, istatistiksel bir test olmadığından matematiksel hesaplama yöntemlerine dayanır. Ancak günümüzde pek çok istatistiksel analiz programının mönülerine bu hesaplama yöntemi de dahil edilmiştir. Cronbach alfa, güvenilirliğin sadece tek bir yönünü, iç tutarlılığı hesaplar. Bu açıdan çok boyutlu bir ölçek veya testin *genel güvenilirliğini* belirlemede duruma göre, yetersiz kalabilen bir değerdir. Bu bölümde alfa indeks değerinin matematiksel formüller ve istatistiksel analiz programları kullanılarak hesaplanması konuları üzerinde durulmuştur.

FORMÜL ARACILIĞIYLA HESAPLAMA

Cronbach alfa, maddelerin varyans değerlerini dikkate alır. Maddelerin varyans değerlerinin toplamı, toplam puanların varyansından küçük çıktığı ölçüde alfa değeri büyük çıkar. Diğer bir deyişle bir katılımcıya ait yanıtlar ölçeğin kendi içinde büyük ölçüde tutarlı olmalı, örneklemdeki katılımcılar arasındaki değişkenlik ise büyük olmalıdır. Anketlerde belirli sayıda maddeye işaretleme yapılmamışsa maddelerin toplamları alınıp hesaplama yapılamayacağından varyans değeri de hesaplanamaz. Bu nedenle eksik veri içeren değişkenlere atama yöntemiyle belirli bir değer atfedilir. Ölçek tek boyutlu ise ölçeğin aritmetik ortalama/medyan puanı, çok boyutlu ise bu kez faktörlerin aritmetik ortalama değeri veya medyan puanı *atama değeri* olarak belirlenir. Alfa değerinin hesaplanması iki formül modelinden yaralanılarak yapılır. Araştırmacı bu formüllerden kendisine uygun geleni kullanır.

Korelasyon Matrisi Verilerinden Hareket Ederek Hesaplama

Birinci model, alfa değerinin korelasyon matrisi verilerine dayalı olarak hesaplanmasıdır. Bunun için öncelikle maddeler arasındaki korelasyon katsayıları elde edilir. Korelasyon katsayıları için MS-Excell veya istatistiksel analiz yazılımlarından yararlanılır. Daha sonra alfa değeri Eşitlik 5-1'deki formülle hesaplanır.

$$\alpha = \frac{k \cdot \bar{r}}{1 + (k-1)\bar{r}}. \qquad (5\text{-}1)$$

Formüldeki k simgesi, ölçekteki madde sayısını; üzeri çizgili olan $\bar{r}$ sembolü ise maddeler arasındaki korelasyon katsayılarının ortalamasını gösterir. Formülden de anlaşılacağı gibi ölçekteki madde sayısı belirli bir düzeye kadar arttıkça ve korelasyon katsayılarının ortalaması büyük çıktığı ölçüde alfa güvenilirlik seviyesi de artar. Maddeler arasındaki korelasyonların ortalaması düşükse güvenilirlik de düşük çıkar.

■ Örnek: Sekiz ifadeden oluşan bir ölçekte maddeler arasındaki korelasyon katsayılarının ortalaması ,65 çıkmış olsun. Böyle bir durumda alfa güvenilirlik katsayısı Eşitlik 5-2'deki gibi hesaplanır.

$$\alpha = \frac{8 \cdot 0{,}65}{1 + (8-1)\ 0{,}65}, \qquad \alpha = \frac{5{,}20}{5{,}55}, \qquad \alpha = 0{,}936. \qquad (5\text{-}2)$$

Maddelerin Varyans Değerlerinden Hareket Ederek Hesaplama

İkinci model, alfa indeks değerinin maddelerin varyans değerlerinden hareket edilerek hesaplanmasıdır. Bunun için her bir maddenin ve daha sonra en son sütundaki toplam puanların varyans değerleri hesaplanır. Bu hesaplama için Excell programı veya istatistiksel analiz yazılımları kullanılabilir. Varyans, gözlem verilerindeki değişkenlik ölçüsüdür ve Eşitlik 5-3'deki formülle hesaplanır.

$$s_x^2 = \frac{\sum (x - \bar{x})^2}{n-1}, \quad \textit{formüldeki} \quad \bar{x} = \frac{\sum x_i}{n} \text{ dir.} \tag{5-3}$$

s^2 = Bir maddenin varyans değeri.
n = Örneklemdeki kişi sayısı.
x_i = Bir bireyin puanı.
$\bar{x}$ = Değişkenin aritmetik ortalama puanı.

Maddelerin ve toplam puanın varyans değerleri saptandıktan sonra ölçeğin alfa güvenilirlik değeri Eşitlik 5-4'teki formülle belirlenir:

$$\alpha = \frac{k}{k-1}\left(\frac{\sigma_t^2 - \sum \sigma_i^2}{\sigma_t^2}\right). \tag{5-4}$$

k = Ölçüm / madde sayısı.
σ_t^2 = Toplam sütununun varyansı.
σ_i^2 = Değişkenlerin her birinin varyansı.
$\sum \sigma_i^2$ = Değişken varyanslarının toplamı.

Bu formüle göre, üç maddeden oluşan bir ölçeğin alfa güvenilirlik analizi için önce verilerin aritmetik ortalaması, standart sapması ve varyans değerleri hesaplanır. Daha sonra hesaplanan değerler formülde yerine konarak alfa katsayısı belirlenir (*bk.*, Tablo 5-1).

Tablo 5-1. Alfa Katsayısının Hesaplanması

Kişiler	Madde 1	Madde 2	Madde 3	Toplam puan
1	2	3	3	8
2	3	4	3	10
3	3	2	2	7
4	2	4	4	10
5	4	4	1	9
O (Aritmetik ortalama)	2,8	3,4	2,6	8,8
SS (Standart sapma)	,84	,89	1,14	1,30
S^2 (Varyans, σ^2)	,7	,8	1,3	1,7

$$\alpha = \frac{5}{4}\left(\frac{1{,}7 - 2{,}8}{1{,}7}\right), \quad \alpha = \frac{5}{4}\left(\frac{-1{,}1}{1{,}7}\right), \quad \alpha = 1{,}25 \times -0{,}64, \quad \alpha = -0{,}80. \qquad (5\text{-}5)$$

Maddeler kavramsal yapıya ilişkin olarak gerçek puanı yeterli ölçüde içermiyorsa veya maddeler sadece hata puanını içeriyorsa, hata puanı ağırlıkta ise, *toplam puanın varyansı* maddelerin varyanslarının toplamına eşit olur veya ondan küçük çıkar. Böyle bir durumda alfa değeri sıfır veya negatif bir değer olarak gözükür. Toplam puanın varyansı, maddelerin varyans değerlerinin toplamından büyük olduğu ölçüde ise, alfa değeri pozitif işaretli ve yüksek bir değer olarak çıkar. Araştırmacının amacı tek tek maddelerin varyans değerlerini minimize etmek ve toplam puanların varyansını ise maksimum seviyeye çıkarmaktır. Yukarıdaki örnekte, maddelerin varyans değerleri toplamının, toplam puan değerlerine ait varyanstan (ölçek varyansından) büyük olması nedeniyle alfa değeri negatif çıkmıştır.

Maddelerin Kovaryans Değerlerinden Hareket Ederek Hesaplama

Alfa değerini hesaplamanın bir diğer yöntemi, maddelerin kovaryans katsayıları ortalamasının dikkate alınmasıdır. Bu uygulamada Eşitlik 5-6'daki formül kullanılır.

$$\alpha = \frac{\dfrac{k \times ort\,(\text{kov})}{ort\,(\text{var})}}{1 + \dfrac{(k-1) \times ort(\text{kov})}{ort\,(\text{var})}}. \qquad (5\text{-}6)$$

k = Madde sayısı.
ort (kov) = Ortalama kovaryans değeri.
ort (var) = Ortalama varyans değeri.

Formülden de anlaşılacağı gibi alfa değeri, madde sayısına ve maddeler arasındaki korelasyonlara (kovaryans değerlerine) bağımlıdır. Korelasyon ortalamaları küçük olsa bile madde sayısı yeterince büyük ise alfa değeri yine büyük çıkar.[1]

Madde Sayısının Artırılmasıyla veya Azaltılmasıyla Alfa Güvenilirlik Katsayısını Tahmin Etme

Alfa güvenilirlik katsayısı ölçekteki/testteki madde sayısına bağımlıdır. Madde sayısı arttıkça güvenilirlik de artar. Madde sayısını belirli oranda artırmanın veya azaltmanın güvenilirlik katsayısını ne kadar yükselteceği veya belirli bir güvenilirlik oranına ulaşmak için madde sayısını ne oranda artırmak/azaltmak gerektiği, yarıya bölme güvenilirliğinde sık kullanılan Spearman-Brown kehanet formülünden yararlanılarak hesaplanır. Spearman-Brown formülüyle elde edilen değer gerçek bir güvenilirlik katsayısı değildir. Bilim adamına ölçek geliştirme çalışmaları sırasında fikir veren ve onu yönlendiren bir değer olarak görülmelidir.

Madde sayısının tahmini. Bilim adamı yaptığı pilot araştırma sonucunda istediği güvenilirlik rakamını elde edememişse madde sayısını artırarak bu

rakama ulaşabilir. Madde sayısının hangi oranda arttırılacağı Spearman-Brown kehanet formülünden yararlanılarak belirlenir (*bk.*, Eşitlik 5-7).

■ Örnek: Diyelim ki, bir araştırmacı 20 maddeden oluşan bir ölçeğin alfa güvenilirlik katsayısını ,70 olarak saptamıştır. Ancak bu rakamı yeterli görmeyerek katsayıyı ,80 düzeyine çıkarmak istemekte ve bunu sağlamak için ölçeğe kaç madde daha ilave edilmesi gerektiğini merak etmektedir.

$$m = [r_i(1 - r_m)] / [r_m(1 - r_i)] . \qquad (5\text{-}7)$$

m = Mevcut maddelere ilave edilmesi gereken oran.
r_i = İstenen güvenilirlik düzeyi.
r_m = Mevcut değer, hesaplanan güvenilirlik düzeyi.

$$m = [,80\,(1 - ,70)] / [,70\,(1 - ,80)] ,$$
$$m = 1,714 . \qquad (5\text{-}8)$$

Gerekli toplam madde sayısı = k .
k = mevcut madde sayısı x gerekli oran .
$k = 20$ x 1,714 ,
$k = 34$.

Buna göre bilim adamı ölçekteki madde sayısını 20'den 34'e çıkardığında ,80 güvenilirlik katsayısını elde etme ihtimaline sahip olacaktır. Klasik test kuramının öncülerinden olan Lord ve Novick (1968) *Statistical Theories of Mental Test Scores* adlı kitaplarında testin uzunluğu ile güvenilirlik arasındaki ilişkileri gösteren bir tablo hazırlamışlardır.Bu tablo incelendiğinde belirli bir madde sayısına kadar güvenilirlik rakamlarının arttığı ve bu rakamdan sonra ise önemli bir yükselme ortaya çıkmadığı görülür (aktaran Shuford, 2004).[2]

Güvenilirlik katsayısının tahmini. Ölçeğe, mevcut sayının belirli bir oranı kadar madde ilave edildiği zaman alfa güvenilirlik katsayısının ne olacağını tahmin etmek için ise Eşitlik 5-9'daki formül kullanılır.[3]

$$r_{sb} = \frac{m \times r}{1 + (m-1) \times r}. \qquad (5\text{-}9)$$

r_{sb} = Spearman-Brown güvenilirlik katsayısı.
m = İlave edilmesi gereken oran.
r = Mevcut güvenilirlik katsayısı.

■ Örnek: Diyelim ki, bir araştırmacı 12 maddeden oluşan bir ölçeğin alfa güvenilirlik katsayısını ,70 olarak saptamıştır. Madde sayısını 20'ye çıkardığında güvenilirlik katsayısının ne olacağını merak etmektedir. Bu işlemde Spearman-Brown formülünü çalıştırabilmek için öncelikle ilave edilmesi gereken oran veya ölçeğin kaç katı artırılması gerektiği bulunur.

İlave edilmesi gereken oran veya kat = m .
$m = 20 / 12$,
$m = 1{,}66$.

Çıkan sonuç daha sonra Eşitlik 5-10'daki formülde yerine konarak gerekli hesaplama yapılır.

$$r_{sb} = \frac{1{,}66 \times ,70}{1 + (1{,}66 - 1) \times ,70}, \quad r_{sb} = \frac{1{,}162}{1{,}462}, \quad r_{sb} = 0{,}79. \qquad (5\text{-}10)$$

Spearman-Brown formülü tek boyutlu ölçeklerde farklılığı ölçebilir. Eğer ölçekte birden fazla boyut varsa bu formül her bir faktör için ayrı ayrı çalıştırılmalıdır. Brown'ın kendisi, ölçekte birden fazla boyut varsa genellenebilirlik kuramından yararlanılmasını önermiştir.

İSTATİSTİKSEL ANALİZ PROGRAMINDA HESAPLAMA

Ön Tetkikler

Cronbach alfa değerini istatistiksel analiz programı kullanarak hesaplamak isteyen bilim adamları alfa analizinden önce belirli ön tetkikleri yapmalıdır-

lar. Ön tetkikler kapsamında faktör analizi yöntemi kullanılarak ölçeğin veya testin tek boyutlu olup olmadığı belirlenmeye çalışılır.

Alfa güvenilirlik analizinin faktör analizinden önce mi yoksa sonra mı yapılması gerektiği literatürde tartışmalı bir konudur. Bazı bilim adamlarına göre 8-10 madden oluşan ölçek ve testler için önce faktör analizi yöntemini uygulamaya gerek yoktur. Eğer güvenilirlik analizi sonuçları yüksek çıkmışsa bu maddeler bünyesinde kaç faktör içeriyor olursa olsun tutarlıdır ve belirli bir kavramsal yapıyı ölçüyor demektir.

Yapılan ilk analizde eğer alfa değeri düşük çıkmışsa bilim adamı böyle bir durumda ölçek veya testin farklı faktörlere sahip olabileceğinden şüphelenmeli ve faktör analizi yöntemini uygulamalıdır. Alfa güvenilirlik katsayısı düşük çıkan ölçek ve testlerde faktör analizli yöntemi uygulanarak kavramsal yapıya ilişkin alt boyutlar veya faktörler belirlenmeye çalışılır. Eğer ölçek/test iki veya üç faktörü ortaya çıkaran bir niteliğe sahipse alfa güvenilirlik analizi her bir faktör için ayrı ayrı yapılır ve bu yöntem daha doğrudur.

SPSS Yazılımı Aracılığıyla Alfa Değerinin Hesaplanması

Alfa güvenilirlik testleri, istatistiksel analiz programı SPSS'te Scale mönüsü altında bulunur. Araştırmacı bu mönüdeki Reliabilitiy Analysis düğmesi ile açılan pencerede önce değişkenlerini tanımlar. Daha sonra Options düğmesine basarak açılan kartta Descriptives for alanındaki item, scale, scale if item deleted şıklarını; Summaries alanından ise means, variances, covariances kutucuklarını ve Inter-item bölgesindeki Correlations ve Covariances seçeneklerini işaretler. Bunlar güvenilirlik analizini yapmak ve sonuçlarını yorumlamak için gerekli olan temel testlerdir. Araştırmacı *F* testi sonucunu ve madde ortalamalarının eşit olma durumunu da görmek istiyorsa ayrıca Hotelling's T-squared ve Friedman ki-kare analizi gibi diğer istatistiksel tekniklere ait kutuları da seçili hale getirmelidir. Bu bölümdeki güvenilirlik analizlerinden hangilerinin uygulayacağına Model mönüsüne bakılarak karar verilir. Bu bölümde (a) alfa, (b) Guttman, (c) yarıya bölme, (ç) paralel formlar ve (d) tam paralel formlar yöntemine ilişkin seçenekler vardır.

Alfa yöntemi. Alfa yönteminde sadece *Cronbach alfa* formülü temel alınır. Bu bölümde SPSS hem standartlaştırılmış, hem de ham[a] alfa değerini hesaplar. Çıktılarda İngilizce "alpha" ve "standardized item alpha" sözcükleriyle ifade edilen bu katsayılar güvenilirliğe ilişkin tahmin değerlerini verir.

Standartlaştırılmış alfa değeri, maddelerin varyanslarının (değişkenliklerinin) eşit olduğu varsayımı altında yapılan güvenilirlik ölçüsüdür. Bu varsayıma göre bütün maddeler birbirlerine tam olarak paralel bir niteliğe sahiptir. Bilim adamı ölçekteki maddelerin tam paralel olduğunu, varyans değerlerinin benzerliğiyle kanıtlamışsa standartlaştırılmış alfa değerini kullanır. Ham alfa değeri, gözlem puanlarının temel alındığı maddeler arasındaki korelasyonlara dayanırken standartlaştırılmış alfa değeri ham puanların *z* puanlarına dönüştürülmesi sonucunda maddeler arasındaki kovaryans değerlerine dayanır. Kovaryans, iki maddenin birlikte değişkenliği veya dağılımı anlamındadır. Bir değişkene ait varyans değeri grafiksel olarak normal dağılım eğrisi ile gösterilirken; iki madde arasındaki kovaryans üç boyutlu bir grafik olarak, tepe şeklinde gösterilir. Standartlaştırılmış alfa değerinin bir başka açıklaması, *k* sayıda yarıya bölme güvenilirliğinin Spearman-Brown formülüyle düzeltilmiş katsayılarının ortalamasıdır. Bazı araştırmacılar verileri standart *z* puanlarına dönüştürmesi nedeniyle standartlaştırılmış alfa değerinin ham alfa değerinden daha güçlü olduğunu düşünürler, ancak bu görüş doğru değildir. Standartlaştırılmış alfa değeri; farklı sayıda madde içeren ve birbirine benzer yapıları ölçen ölçüm araçlarının güvenilirlik katsayılarını karşılaştırmak için kullanılır. Bu nedenle, "tam paralel test maddeleri varsayımını" temel alır. Kimi bilim adamları ise, maddelerin varyans değerleri arasında önemli ölçüde farklılıklar varsa, ham alfa değeri yerine standartlaştırılmış alfa değerini vermenin daha doğru olacağını ifade etmişler ve madde çıkarılması halinde meydana gelebilecek değişiklikleri standartlaştırılmış alfa değeriyle takip etmenin daha doğru olacağını belirtmişlerdir. Standartlaştırılmış alfa değerinde maddelerin varyansları eşittir. Maddelerin varyansları birbirine eşitse standartlaştırılmış ve ham alfa değerleri de birbirine yakın çıkar. Gliem ve Gliem'e (2003) göre, standartlaştırılmış alfa değeri sadece bireysel maddelerin ölçek dereceleri aynı olmadığı zaman kullanılır.[4]

---

[a] Ham alfa değeri hesaplanırken cevaplayıcıların işaretledikleri gözlem puanları temel alınır. Standartlaştırılmış alfa değerinde ise ham puanlar önce standardize edilir ve daha sonra standardize edilmiş puanlar üzerinden alfa değeri hesaplanır.

Ham alfa değeri ise, her bir maddenin varyansının farklı olduğu varsayımı altında hesaplanmıştır. Bu nedenle, "maddelerin varyanslarının eşit olduğu" gibi bir varsayımı dikkate almadığından veya "maddelerin varyanslarının yaklaşık olarak eşit olduğu" varsayımından hareket edilmesi nedeniyle daha pratiktir. Ham alfa değeri aynı zamanda genellenebilirlik kuramında "tek yüzeyli ölçüm (madde) tasarımında *G* katsayısına ve "Hoyt küme içi korelasyon katsayısına" eşittir.[5] Standartlaştırılmış alfa değeri ile ham alfa değeri arasında önemli ölçüde farklılık varsa, maddelerin değişkenliklerinin (varyanslarının) farklı olduğu sonucuna varılır.

Uygulamalarda nadiren birbirinden önemli ölçüde farklı çıksa da raporda öncelikle ham alfa değeri verilir, standartlaştırılmış alfa değeri ise parantez içinde gösterilir. Araştırmacı raporunda bu değerlerin her ikisini de vermelidir. Böylece okuyucular ölçek maddelerinin tau eşitliğine sahip olup olmadığı konusunda bir fikir edinirler. Rakamlar birbirinin aynı veya benzeri ise maddelerin tau eşitliğine sahip olduğu söylenir.

Yarıya bölme yöntemi. İstatistiksel analiz programı SPSS'te model olarak Split-half yöntemi seçilmişse program değişkenleri otomatik olarak yarıya böler ve birinci yarıyla ikinci yarı arasında gerekli hesaplamaları yapar. Araştırmacı, yarıya bölme uygulamasını anketin ilk yarısıyla ikinci yarısı arasındaki korelasyon yerine tek rakamlı ve çift rakamlı maddeler arasında yapmak istiyorsa SPSS'te bunun için küçük bir kod değişikliği yapması gerekir.[6] Yazılımda istenen istatistiksel testler seçildikten sonra güvenilirlik analizinin ana penceresine dönülür ve buradaki Paste tuşuna basılır. Açılan Syntax (komut kodları) penceresinde aşağıdaki gibi bir kod dizisi görülecektir:

```
RELIABILITY
 /VARIABLES=s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12
 /FORMAT=LABELS
 /SCALE(SPLIT)=ALL/MODEL=SPLIT
 /STATISTICS=DESCRIPTIVE SCALE CORR COV
 /SUMMARY=TOTAL MEANS VARIANCE COV CORR.
```

Bu kod dizisinin dördüncü satırı, tek ve çift rakamlı değişkenler arasında hesaplama yapılabilmesi için aşağıdaki şekilde yeniden düzenlenir.

/SCALE(SPLIT)=s1 s3 s5 s7 s9 s11 s2 s4 s6 s8 s10 s12
/MODEL=SPLIT

Diğer satırlarda hiçbir değişiklik yapılmaz. Dizimde tek rakamlı ve çift rakamlı değişkenler belirli bir sıra içinde ele alınmıştır. Daha sonra, açılan Syntax penceresinden Run mönüsüne girilerek All şıkkı seçildiğinde bu kez hesaplamanın tek – çift sıralamasına göre yapıldığı görülür.

Guttman yarıya bölme modeli. Yarıya bölme güvenilirliğinin bir diğer türü, Guttman güvenilirlik katsayısıdır (G). Bu hesaplamada temel alınan varsayımlar şunlardır: (a) Testin her bir yarısının alfa güvenilirlik katsayıları diğerine eşit değildir. (b) Testin birinci yarısının varyansı ile ikinci yarısının varyansı birbirine eşit değildir. Araştırmacılar Guttman ölçekleriyle Guttman güvenilirliğini birbirine karıştırmamalıdırlar. Bu güvenilirlik katsayısı, Guttman tarafından geliştirilen ve ölçeklerin güvenilirliğini hesaplamak için kullanılan başka bir formüle dayanır.

Paralel formlar modeli. Paralel formlar yönteminin iki önemli varsayımı vardır: (a) Her iki ölçekteki veya testteki maddelerin gerçek puan varyansları eşittir. (b) Her iki ölçek veya testin aynı zamanda hata varyansları da eşittir. Araştırmacı maddelerin ortalamalarının farklı, fakat varyanslarının ve hata varyanslarının eşit olduğunu düşünüyorsa bu durumda *maksimum benzerlik güvenilirliği* hesaplama yönteminden yararlanır. İstatistiksel analiz programı SPSS'te bu güvenilirlik Parallel seçeneği ile belirlenmiştir.

Tam paralel formlar modeli. Bu test *paralel form* varsayımlarına ilave olarak, her iki ölçekteki maddelerin ortalama değerlerinin de eşit olduğunu varsayar. Cronbach, bu yöntemi madde havuzundaki ifadeler ortak faktöre işaret etmediği fakat, bir kaç tane grup faktörü ortaya çıkarma özelliğine sahip olduğu zaman uygulamıştır. Bu testte maddelerin ortalamalarının eşit olup olmadığını belirlemek için Hotelling T-square test for equality of means kutusu seçili hale getirilir.

Çıktıların yorumlanması. İstatistiksel analiz sonucunda SPSS'ten değişkenlerin ortalama ve standart sapma değerleri, korelasyon ve kovaryans matrisleri, ölçeğin toplam puanlarının ortalama, standart sapma ve varyans değe-

ri, madde-toplam puan istatistikleri ve son olarak da alfa değeri elde edilir. Madde-toplam puan istatistikleri arasında özellikle üç grup değer önemlidir.

Düzeltilmiş madde-toplam puan korelasyonu (Corrected item-total correlation). Bu bulgular, tek tek maddelerin toplam puanla olan korelasyonlarını gösterir. Maddeler, toplam puanla yüksek bir korelasyona sahip olmalıdır. Toplam puan-madde korelasyon katsayısı ,30'un altında olan maddeler ölçekten çıkarılır. Cronbach alfa hesaplaması, tek tek maddelerin güvenilirliğini belirlemeye yönelik olarak yapılmaz. Buradaki hesaplama daha çok maddelerin kavramsal yapıya yaptıkları katkıyı belirlemeye yöneliktir.

Madde iptal edilirse alfa değeri (Alpha if item is deleted). SPSS bir maddenin çıkarıldığında alfa katsayısının ne şekilde etkileneceğini göstermek üzere böyle bir hesaplama yapmıştır. Bir madde çıkarıldığında alfa değeri önemli ölçüde yükseliyorsa bu yönteme başvurulur. Önemsiz sayılabilecek değişiklikler için bu yöntem uygulanmaz.

Cronbach alfa (Cronbach's Alpha). Bu bölümde son olarak alfa ve standartlaştırılmış alfa değerleri verilir.

Alfa değerinin negatif ve birden büyük çıkması. Alfa değeri bazen negatif ve bazen de hem negatif hem de 1'den büyük bir değer olarak çıkabilir.[7] Bunun değişik nedenleri vardır: (a) Bireysel maddelere ait varyansların toplamı, toplam puanlara ait varyanstan daha büyük ise sonuç eksi çıkar. (b) Maddeler gerçek değer açısından tau eşitliğine sahip olsalar bile araştırmada ciddî ölçüm hataları varsa alfa yine negatif çıkar. (c) Ters yönlü ifadelerin kodlanmasına dikkat edilmemesi halinde ortak varyans pozitif olduğu halde negatif çıkar. (ç) Örneklem hacmi küçükse veya ölçekteki madde sayısı azsa alfa negatif çıkar. (d) Bazen maddeler gerçekten ortak pozitif varyansa sahip olmayabilir ve bu nedenle alfa katsayısı negatif çıkar. (e) maddeler arası kovaryans değerlerinin ortalaması negatif ise alfa değeri de negatif çıkar. (f) Negatif çıkmasının bir diğer nedeni, testin birden fazla boyut/faktör içermesi ve bu boyutların birbiriyle negatif yönde ilişkili olmasıdır.

Negatif alfa değeriyle karşılaşılmışsa yapılacak ilk iş ifadelerde kodlama hatasının yapılıp yapılmadığını kontrol etmektir. Ters yönlü ifadelerin puanları yeniden gözden geçirilir ve ölçekteki madde sayısının yeterli olma durumu araştırılır. Negatif alfa değeri ölçeğin güvenilmez olduğu anlamına

gelir. Wiersma ve Jurs (1990) hesaplama sonucunda alfa değeri negatif çıkmışsa güvenilirlik katsayısının sıfır olarak rapor edilmesini önermişlerdir (aktaran Krus ve Helmstadter, 1993).[8]

İkili veri yapılarında alfa güvenilirlik hesaplaması. İkili veri yapılarında *toplam varyans* katılımcıların yarısı bir maddeye doğru ve diğer yarısı ise yanlış yanıt vermişlerse maksimum seviyeye çıkar. İkili veri yapılarında alfa güvenilirlik katsayısı ile KR-20 aynı sonucu verir. Fakat, istatistiksel analiz programı SPSS'te KR-20 hesaplaması yapılmadığından ikili veriler için de alfa güvenilirlik katsayısı hesaplatılır ve aynı sonuçlar elde edilir.

SYSTAT Yazılımı Aracılığıyla Alfa Değerinin Hesaplanması

Güvenilirlik hesaplamalarında kullanılabilecek bir diğer yazılım SYSTAT'tır. Grafik özellikleri ağırlık basan bu programda Statistics mönüsündeki Scale düğmesi altında bulunan Cronbach's alfa seçeneği ile gerekli hesaplama yapılır. Alfa değerinin hesaplanabilmesi için en az iki değişkenin bulunması gerekir. SYSTAT'ta madde analizi komutu, yarıya bölme ve tek/çift güvenilirliklerini hesaplamak üzere ayrı bir düğme üzerinde konuşlandırılmıştır.

STATISTICA Yazılımı Aracılığıyla Alfa Değerinin Hesaplanması

STATISTICA ülkemizde sık kullanılan istatistik yazılımlarından bir diğeridir. Bu yazılımda güvenilirlik analizleri Reliability/Item analysis mönüsü altında toplanmıştır. Maddeler, güvenilirlik hesaplaması yapılacak *değişkenler* olarak tanıtıldıktan sonra OK tuşuna basılarak gerekli hesaplama yapılır. STATISTICA yazılımının çıktıları grafik özellikleriyle zenginleştirilmiştir. Ayrıca çıktı panosunda Spearman-Brown kehanet formülünü ve *zayıflığı giderme* formüllerini çalıştıran düğmeler eklenmiştir. Böylece belirli bir güvenilirlik düzeyi için kaç maddeye ihtiyaç duyulabileceği veya belirli sayıda maddeyle hangi düzeyde güvenilirlik katsayısının elde edilebileceği kolaylıkla hesaplanabilmektedir. Bu bölümde göze çarpan bir diğer özellik, katılımcılar arasında ve maddeler arasında karşılaştırma yapmaya imkan sağlayan varyans analizinin bulunmasıdır.

RAPORLAMA

Metin İçinde

Bilim adamı alfa değerlerini raporlarken birkaç değişik yöntemden yararlanabilir. Ölçek tek boyutlu ise sadece standartlaştırılmış ve ham alfa değerlerinin verilmesi yeterlidir, ayrıca tablo yapmaya gerek yoktur. Bilimsel dergi editörlerinin bir bölümü alfa değerlerinin yanında her bir faktörle ilgili olarak korelasyon ve kovaryans tablolarının, alfa değerinin standart hatasına ilişkin bilgilerin verilmesi de isterler. Editörlerin veya hakemlerin böyle bir istekte bulunmalarının nedeni, alfanın standart hatası ve kesinlik derecesi hakkında bir fikir edinmek istemeleridir. Hazırlanan bilimsel metnin bir tez veya makale olmasına göre sunum biçimi değişir. Makalelerde yer sorunu olması nedeniyle güvenilirlik hesaplamalarına ilişkin bilgiler özet olarak verilirken tezlerde daha ayrıntılı bir sunum yapılır. Tezlerde her bir maddenin toplam puanla olan korelasyon değerlerini, maddelerin birbirileriyle olan korelasyon ve kovaryans değerlerini de vermek gerekir. Seçilen alfa modeline göre, maddelerin ve toplam puanın varyans değerlerinin gösterilmesi inandırıcılığı ve seçilen modelin doğruluğu hakkında bir fikir verir. Maddelerin ve toplam puanın aritmetik ortalama ve standart sapma değerlerinin gösterilmesi ise isteğe bağlıdır. Cronbach alfa değerleri verilirken virgülden sonra iki hanenin gösterilmesi yeterlidir.

■ Makalelerdeki sunum biçimine ilişkin örnek.

Yapılan güvenilirlik analizleri sonucunda *Stres Ölçeği*'nin ham Cronbach alfa değeri ,75 ve standartlaştırılmış alfa değeri ise ,78 çıkmıştır.

Tablolaştırarak Sunum

Geliştirilen ölçek birden fazla boyut içeriyorsa, aynı veya farklı gruplarda birden fazla ölçüm yapılmışsa her bir boyutun alfa değeri ayrı ayrı gösterilir. Bunun dışında maddeler arası korelasyon ve kovaryans değerleri, TYVA tablosu değerleri, madde-toplam puan analizi sonuçları yine tablo oluşturularak raporlanır (*bk.*, Tablo 5-2).

Tablo 5-2. Stres Ölçeğinin Faktörlerine Ait Alfa Değerleri

Stres faktörleri	Pilot araştırma sonuçları		Son araştırma sonuçları	
	Ham alfa değeri	Standardize edilmiş alfa	Ham alfa değeri	Standardize edilmiş alfa
İşin kendisinden	,71	,72	,75	,74
Yöneticilerden	,74	,73	,75	,75
Örgütsel yapıdan	,75	,74	,81	,75
Genel olarak stres ölçeği	,72	,73	,74	,72

Pek çok durumda alfa değerlerinin sadece tablolaştırılması yeterli olmayabilir. Araştırmacı daha sonra tablo içeriğine ilişkin yorumlar yapmalı ve tablo çerçevesinde gözden geçirdiği maddeler ve yaptığı değişiklikler hakkında bilgi vermelidir.

İYİLEŞTİRME

Dikkatli bir şekilde yapılan ölçüm uygulamasının sonunda elde edilen bulgularda maddeler arasındaki korelasyon katsayıları ve alfa değerleri beklenenden düşük çıkmışsa veya negatif değerler elde edilmişse iyileştirme amacıyla aşağıdaki işlemlere başvurulur.

1. Maddeler arası korelasyon ve kovaryans matrisi çıkarılır. Negatif korelasyona sahip maddeler "bayrak" simgesiyle işaretlenir.
2. Negatif işaretli ölçek maddelerinin tersine çevrilme durumu kontrol edilir.
3. Az sayıda negatif korelasyon katsayısına sahip madde varsa bu maddeler çıkarılarak güvenilirlik yeniden hesaplanır.
4. Negatif işaretli ölçek maddelerinin örneklemin küçük olmasından kaynaklanıp kaynaklanmadığı araştırılır.
5. Seçilen maddelerin kavramsal yapıyla ilgili olarak mümkün olduğu kadar geniş bir perspektife sahip olup olmadığı araştırılır.
6. Az sayıda düşük ($r < 30$) korelasyon katsayısına sahip madde varsa

bu maddeler çıkarılarak güvenilirlik yeniden hesaplanır

7. Test maddelerinin homojen işaretlenme durumu kontrol edilir.
8. Örneklemdeki kişilerin toplam puanlarının heterojen olması koşulunun sağlanma özelliği araştırılır.
9. Maddelerin büyük çoğunluğunun orta güçlük derecesinde (,40-,60) olup olmadığına dikkat edilir. Ölçek 5 veya 7 dereceli bir Likert ölçeği ise maddelerin büyük çoğunluğunun işaretleme sıklığının ortadaki derecelerde yoğunlaşma durumu araştırılır.
10. İşaretlemelerin 5 dereceli bir ölçekte 1-2 ve 4-5 şıkları üzerinde yoğunlaşma durumu incelenir. Eğer uç noktalarda yoğunlaşma varsa tavan-taban etkisinin ortaya çıktığına karar verilir.
11. Faktör analizi yöntemiyle, ölçeğin veya testin faktöriyel yapısı sorgulanır. Faktöriyel bir yapı varsa güvenilirlik analizi her bir faktör için ayrı olmak üzere yeniden hesaplanır.
12. Ölçekteki madde sayısının yeterliliği incelenir. Madde sayısı yetersizse arttırılması sağlanır.
13. Ölçeğin uygulandığı örnek kütle büyüklüğü incelenir. Örnek kütle büyüklüğü yeterli değilse artırılır.
14. Ölçeğin uygulandığı örnek kütlenin demografik özellikleri ve bireylerin kişisel özellikleri incelenir. Büyük ölçüde birbirine benzer kişilerden oluşuyorsa ve benzer düşüncelere sahiplerse örnek kütle çeşitlendirilir veya genişletilir.
15. Ölçeğin uygulanması sırasında katılımcılara yeterli zamanın verilip verilmediği ve anket uygulamasının belirli bir ciddiyet içinde yapılıp yapılmadığı araştırılır.
16. Ölçeğin tüm katılımcılar tarafından *standart uygulama prosedürü* içinde doldurulma durumu araştırılır.
17. Ölçek, diğer iç tutarlılık analizi yöntemleriyle sınanarak düşük güvenilirlik katsayısının yöntemden mi yoksa maddelerin kendisinden mi kaynaklandığı araştırılır.
18. Çok boyutlu ölçeğin bir bütün olarak alfa güvenilirlik katsayısına ulaşılmak isteniyorsa bu amaçla üretilmiş özel yazılımlar kullanılarak ayrıntılı hesaplama yöntemlerine başvurulur.

ALINTI YAPILAN KAYNAKLAR

[1] SYSTAT, "Cronbach Main Dialog Box [Cronbach Ana Etkileşim Kutusu]," Help Mönüsü.

[2] Emir H. Shuford, "Atlantic Fleet Training Command - Quartermaster Course [Atlantik Filosu Eğitim Komutanlığı- Eğitim Kursları]," <http://www.p-mmm.com/pioneers/quartermaster.htm> (24.01.2004).

[3] Bethany, "Reliability [Güvenilirlik]," <http://www.geolog.com/gmsmnt/gmrel.htm> (20.04.2003).

[4] J.A. Gliem ve R.R. Gliem "Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales [Likert Tipi Ölçeklerde Cronbach Alfa Güvenilirlik Katsayısının Hesaplanması, Yorumlanması ve Raporlanması]," < http://www.alumni-osu.org/midwest/midwest%20papers/Gliem%20&%20Gliem--Done.pdf> (20.01.2004).

[5] David E. Conroy ve Jonathan N. Metzler, "Factorial Invariance and Latent Mean Stability of Performance Failure Appraisals [Başarı ve Başarısızlık Değerlendirmelerinde Faktöriyel Değişmezlik ve Örtük Puanların İstikrarlılığı]," <http://www.personal.psu.edu/faculty/d/e/dec9/lab/reprints/sem03_FFstability.pdf> (24.01.2004).

[6] R. Gebotys, "Using SPSS for Windows to Implement Reliability Analyses [Güvenilirlik Analizi İçin SPSS'in Kullanımı]," <http://www.wlu.ca/~wwwpsych/gebotys/book/relspss.pdf.> (20.04.2001).

[7] David P. Nichols, "My Coefficient a is Negative [Katsayım Negatif]," <http://www.ats.ucla.edu/stat/spss/library/negalpha.htm> (10.07.2004).

[8] David J. Krus ve Gerald C. Helmstadter, "Problem of Negative Reliabilities [Negatif Güvenilirlik Sorunu]," 1993, <http://www.visualstatistics.net/Publications/Negative%20Reliability/negative_reliability.htm> (20.04.2003).

GÜVENİLİRLİK VE KORELASYON ANALİZLERİ

Klasik test kuramı kapsamında ele alınan güvenilirlik, değişik nitelikteki korelasyon analizleriyle yakından ilgilidir. Güvenilirliği belirlemeye yönelik olarak yapılan korelasyon analizleri; (a) maddeler arasındaki, (b) toplam puanlar arasındaki, (c) toplam puanla maddeler arasındaki ve (ç) gözlemci değerlendirme puanları arasındaki ilişkileri belirlemeye yöneliktir. İlişkinin güçlü olması güvenilirliğe işarettir. Bu bölümde ölçümlerin güvenilirliğini saptamak için sık uygulanan korelasyon analizi yöntemleri üzerinde durulmuştur.

MADDELER ARASINDAKİ KORELASYON

Araştırmacı, ölçümün güvenilirliğini saptamak için klasik test kuramını temel almışsa; maddeler arasındaki ilişkilerin güçlü veya zayıf olup olmadığını görmek, maddeler arasındaki tutarlılığı belirlemek veya maddelerin arka palanındaki gizli değişkeni ortaya çıkarmak için maddeler arası korelasyon analizi yöntemine başvurabilir. Maddeler arasındaki tutarlılık, *pozitif yönlü yüksek ilişki derecesi* ile belli olur. Maddeler arasındaki değişkenlik (varyans) fazla olduğu ölçüde korelasyon katsayısı r büyük çıkar Bir maddenin varyansı düşük ise bu madde muhtemelen yanlı bir şekilde ifade edilmiş veya ölçüm yanlı bir örneklem üzerinde yapılmıştır. Bu maddelerin diğer maddelerle olan korelasyonu düşüktür.

İlişki analizinde Pearson korelasyon katsayısı örneklem verilerinde r, ana kütle verilerinde ise ρ (rho) simgesi ile gösterilir. Örneklem verilerine dayalı olarak yapılan *küme içi*[a] *korelasyon katsayısı* için ise büyük R harfi kullanılır.

Maddeler arasındaki korelasyon katsayısının yüksekliği veya düşüklüğü bir ölçek veya testin güvenilirliğini tek başına belirlemez. Maddeler arası

---

[a] Bazı kaynaklarda "sınıf içi korelasyon" terimiyle karşılanmıştır.

korelasyon değerleri, ölçekten madde çıkarılarak alfa katsayısına yaptığı katkı açısından ayrıca incelenmelidir. Nihaî ölçeğe sadece korelasyon katsayısı yüksek olan maddeler alınır, ancak bu uygulamanın yanı sıra maddelerin toplam puanla olan korelasyonlarına da bakılır. Bilim adamı, değişik açılardan inceleme yaparak alfa değeri yanında maddeler arası korelasyon değerlerinin ortalamasına da bakabilir. Maddeler arası korelasyon değerlerinin ortalaması güvenilirlik katsayısını verir.

Bir ölçekte maddelerin birbirleriyle olan korelasyon katsayılarının hepsinin yüksek olmasının ölçeğin bütününün güvenilirliğini düşüreceği ifade edilmiştir. Bunun için araştırmacı nihaî ölçekte olması gerektiği kadar yüksek korelasyonlu madde alıkoyar, diğerlerini ölçekten çıkarır. Fakat bunu yaparken değişik büyüklükteki korelasyon değerlerine sahip maddelerin ölçekte yer alması sağlanır. Maddelerin hepsi birbiriyle yüksek derecede ilişkili ise, madde sayısında azaltmaya gidileceğinden sonuçta bir taraftan ölçek küçülecek, diğer taraftan nispeten düşük korelasyon katsayısına sahip maddeler de çıkarıldığı için bazı ayırt edici özelliği olan maddelerin kapsam dışında kalması durumu söz konusu olabilecektir. Bu tür ölçeklerde maddeler arasındaki korelasyon yüksek gözükse bile ölçeğin genel güvenilirliği düşüktür.

Maddeler arası korelasyon, ölçülmek istenen kavramsal yapının basit veya karmaşık olmasına göre değişiklik gösterir. Basit kavramsal yapılarda bütün maddeler birbirleriyle yüksek derecede ilişkili iken karmaşık kavramsal yapılarda birden fazla alt boyut / faktör olması nedeniyle sadece tek bir boyutun içindeki maddeler arasındaki korelasyon katsayıları yüksek çıkar. Farklı faktörlere / alt boyutlara ait maddeler arasındaki korelasyon görece daha düşüktür. Bilim adamı ölçüm aracından tek bir puan mı yoksa alt ölçekleri temel alarak birden fazla puan mı elde etmek istediğine bakarak korelasyon analizlerini yapmalıdır.

Araştırmacı; ölçümün düzeyine (kategorik, sıralı, aralıklı, oranlı), maddelerin ikili veya çok dereceli olmasına, verilerin normal dağılım özelliği gösterip göstermemesine göre değişik korelasyon analizi yöntemlerinden yararlanır. Örneğin, gerçek sıralı veya aralıklı ölçek verilerinde Pearson ve Spearman korelasyon analizlerini; yapay sıralı ölçek verilerinde ise polikorik ve tetrakorik korelasyon analizlerini kullanmayı tercih edebilir. Aşağıdaki bölümde bu teknikler üzerinde durulmuştur.

Pearson korelasyon analizi. Pearson, aynı zamanda *momentler çarpımı korelasyonu* olarak isimlendirilir. Beş veya yedi dereceli sıralı ölçek maddelerinde, dereceler arasındaki mesafenin *yaklaşık olarak eşit olduğunun* varsayılması nedeniyle, Pearson korelasyon analizi uygulanabilir. Jaccard ve

Van'a göre (1996, aktaran Broderick, 1999) veriler normal dağılım özelliği gösteriyorsa ve derecelerin eşit dağılma özelliği önemli ölçüde bozulmamışsa Likert maddeleri "eşit aralıklı ölçek olarak" kabul edilip parametrik istatistiksel analizler uygulanır.[1] Likert ölçeği dışında ölçek olmayıp, bağımsız maddelerden oluşan fakat *Likert tipi etiketlerle* derecelendirilen ifadelerin de eşit aralıklı olabileceği ileri sürülmüştür. Crask ve Fox'a göre (1987, Grisaffe'den alınmıştır, 2002) pazarlama araştırmalarında kullanılan "*Zayıf, Orta, İyi, Oldukça İyi, Çok İyi*" şeklindeki bütün derecelendirme ölçeklerinin yaklaşık olarak eşit aralıklı olduğu bulunmuştur.[2] Sıralı ölçek verilerinin eşit aralıklı ölçek olarak kabul edilmesi bilim adamları arasında tartışmalı olan bir konudur. Yukarıda sıralanan lehteki görüşlerin yanında aleyhte görüşlere sahip olan bilim adamları da vardır. Bazı yazarlara göre Likert maddeleri ölçek dereceleri arasındaki mesafenin eşit olmaması yüzünden eşit aralıklı değil, sıralı ölçek niteliğindedir. Sıralı ölçek verileri sürekli ölçek verileri gibi değerlendirilmemelidir. Sıralı değişkenler herhangi bir orijine ve ölçüm birimine sahip değildir. Bu nedenle bu verilerle aritmetik ortalama, varyans ve kovaryans hesaplamalarını yapmanın bir anlamı yoktur.[3] Jöreskog'a (1990) göre ise, ölçek maddeleri normal dağılım özelliği göstermediği ve sürekli veri olmadığı durumda korelasyon katsayıları tahminlerde negatif yanlılığa neden olur. Böyle bir durumda normallik koşulu aranmayan Yapısal Eşitlik Modeli tekniği kullanılmalıdır. İstatistiksel analiz programı LISREL'de normal dağılım özelliği göstermeyen maddelerin analizi için *Ağırlıklandırılmış En Küçük Kareler Yöntemi* adı verilen özel bir hesaplama yöntemi geliştirilmiştir (aktaran, Young, 1998).[4]

Sosyal bilimler alanında çalışan bilim adamlarının önemli bir bölümü sadece belirli sayıda maddeye ait derecelerin toplamını (veya ortalamasını) *eşit aralıklı ölçek* niteliğinde kabul etmişlerdir. Toplam puanların eşit aralıklı ölçek niteliğinde olabilmesi için asgarî derece sayısı konusunda değişik görüşler söz konusudur. Bazı bilim adamları 11 veya 15 derece kavramını öne sürerken bazıları da derece sayısının 20 olması gerektiğini ifade etmişlerdir.[5] Bu bilim adamları, yanıtlayıcıların belirlenen ölçek dereceleri üzerinde *Gaussiyen dağılıma* uygun bir dağılım göstermeleri konusuna önem vermişlerdir. Diğer bazı bilim adamları ise derece sayısından çok madde sayısı üzerinde durmuşlar ve bu kez ölçekteki madde sayısının 20 veya daha fazla olması gerektiğini belirtmişlerdir.[6] Normalde Pearson korelasyon analizinin uygulanabilmesi belirli ön kabullere veya koşullara bağlıdır:[7]

1. Verilerin eşit aralıklı veya oranlı ölçek olması.
2. Doğrusallık.

3. İki maddenin de normallik koşulunu sağlamış olması.
4. Türdeşsellik (homoscedasticity)[a], madde varyanslarının eşit olması.
5. Gözlemlerin bağımsız olması.
6. Temsil edici örneklem.

Korelasyon analizinde yukarıda sıralanan koşulların sağlanması ve doğrusallığın test edilmesi için değişkenlerin *nokta-dağılım grafiği* çizilir ve normallik testleri yapılır. Serilerde ayrık değerler varsa bu değerler çıkarılır veya revize edilir. Beş veya yedi dereceli Likert ölçeği maddeleri eğer normallik koşulunu ve diğer ön kabulleri sağlamamışsa, veriler sağa veya sola çarpık ise verilerin dönüştürülmesi yöntemine başvurulur veya Pearson yerine nonparametrik Spearman rho korelasyon analizi uygulanır. Likert ölçeği maddeleri genelde normal dağılım özelliği göstermediğinden bu gibi durumlarda Spearman rho analizinin uygulanması daha doğrudur. Aslında korelasyon analizi sonuçları karşılaştırıldığında Pearson *r* ile Spearman rho değerlerinde önemli bir farklılık çıkmaz.[8] Dikkatli bir araştırmacının bu gibi durumlarda her iki test sonuçlarını karşılaştırmalı olarak incelemesinde yarar vardır. Diğer bir deyişle hem dönüştürülmüş değerler arasındaki Pearson korelasyon analizi sonuçları hem de ham verilere dayalı olarak yapılan Spearman korelasyon analizi sonuçları birlikte değerlendirmeye alınmalıdır.

Korelasyon katsayısı ile ilgili bir diğer kavram *belirlilik katsayısıdır.* Belirlilik katsayısı (r^2) iki değişken arasındaki ortak varyansın oranını gösterir. Bu katsayı eğer iki paralel maddeye ait ise, söz konusu iki maddenin örneğin ,49 oranında ortak bir öze sahip olduğunu; gizli yapıyı içerdiğini ve maddelerin ,49 oranında birlikte hareket ettiğini gösterir.

İstatistiksel analiz programı SPSS'te korelasyon analizi yapmak için Analyze mönüsünden Correlate düğmesi ve bu düğmenin altındaki Bivariate şıkkı seçilir. Bivariate korelasyonlar iki değişken arasındaki ilişkileri tanımlar. Açılan pencereden değişkenler / maddeler belirlenir. Daha sonra verilerin yapısına uygun olarak Pearson kutucuğu seçili hale getirilir.[b]

Araştırmacı isterse ayrıca anlamlılık testi için iki yönlü test two-tailed significance kutusunu işaretleyebilir. Böyle bir durumda anlamlı korelasyonları göstermek üzere ilgili rakamların yıldız imiyle işaretlenmesini sağlayacak Flag significan corelations kutusu da seçili hale getirilir. Programın

---

[a] Sabit varyans veya eşit varyans.

[b] SPSS'teki Correlates düğmesi altında bulunan Bivariate modülünün, analizi yapılacak verilerin ikili veya çok dereceli olmasına bakmaksızın korelasyon analizini doğru bir şekilde hesapladığı bildirilmiştir.

Options düğmesinden Means and Standart Deviations hesaplaması tik işareti ile seçilir. Kovaryans hesaplatılmasına başvurulmaz. Kovaryans, iki değişken arasındaki standardize edilmemiş ilişki katsayısını verir. Aynı şekilde Cross-Product Deviations[a] hesaplaması da yapılmaz. (Bu hesaplama, eksik değerlerin liste bazında silinmesini sağlar.) Daha sonra geri dönülerek hesaplama yapılır.

Spearman korelasyon analizi. Spearman, "büyüklük sırası korelasyon katsayısı", sıralı ve eşit aralıklı ölçeklerde maddeler arasındaki ilişkilerin gücü hakkında bilgi verir. Maddelere ait derecelerin eşit aralıklı olduğu varsayılmıyorsa (3, 4, 5, 6 veya 7 dereceli olabilir), eşit aralıklı olduğu varsayılmakla birlikte örneklem hacmi belirli bir büyüklüğe sahip değilse (n<30) veya örneklem hacmi büyük olmakla birlikte veriler sağa veya sola çarpıksa bu teknik uygulanır. Spearman rho tek başına hatanın oranı hakkında bilgi vermez, fakat *ro kare* (rho^2) ortak varyans, ortak değişkenlik hakkında daha sağlıklı bir bilgi verir.

Pearson veya Spearman korelasyon katsayısı iki değişken arasındaki ilişkinin gücünü gösterir. İki değişken arasındaki ilişki her zaman doğrusal olmayabilir, bazen iki değişken arasında, "eğrisel *U* ilişkisi" bulunur. Maddeler arasında doğrusallık ön koşulu sağlanamamışsa bu maddeler ölçeğe alınmaz. Korelasyon katsayıları Tablo 6-1'deki gibi yorumlanır:

Tablo 6-1. Korelasyon Katsayılarının Yorumu

Değer	Güvenilirlik
r > ,80	Yüksek
r = ,60 – ,80	Güçlü ilişki
r = ,40 – ,59	Orta derecede ilişki
r = ,20 – ,39	Düşük ilişki
r < ,20	Zayıf ilişki (tesadüfe bağlı olabilir)

[a] Sapmaların çarpımı, *X* ve *Y* değişkenlerindeki değerlerin ortalamadan farklarının çarpımı anlamındadır. Varyans formülünün payında yer alan işlemi tanımlar ve kısaca "sapmaların çarpımı" veya "farkların çarpımı" olarak isimlendirilir $(X - \overline{X})(Y - \overline{Y})$. Sapmaların çarpımı, tek başına yorum yapma amacıyla kullanılmaz.

Kendall tau-b. Büyüklük sırasına sokulmuş veriler arasındaki birebir uyuşmayı dikkate alarak ilişkinin gücünü belirler. Kendall tau Spearman sıra korelasyonunda olduğu gibi, bir değerin her iki seride de aynı sırada yer alma durumuna göre pozitif veya negatif değerler alır. Pozitif değerler, her iki değerin büyüklük sırasının birlikte artış gösterdiğini ortaya koyar. Kendall tau, –1 ilâ +1 arasındaki değerleri alır. Bu analiz, SPSS'te Crosstabs ve Correlations mönüleri olmak üzere iki farklı çıkış noktasında hesaplatılabilir. Birinci yöntemde Crosstabs mönüsünün altında Ordinal bölümündeki ilgili kutu seçili hale getirilerek hesaplanır. İkinci yöntemde ise Correlations mönüsündeki Biserial düğmesi ile açılan karttaki Kendall tau-b seçeneği kullanılır. Kendall *tau-b*, $n < 30$ olan örneklemlerde uygulanır. Kendall tau'nun avantajı, gözlemlenen uyuşma ve uyuşmama oranları hakkında doğrudan yorum yapmaya imkan veriyor olmasıdır.[9] Bu hesaplama yöntemiyle ilgili olarak ayrıca "Kendall tau a, b, c" başlığına bakınız.

Ki-kare. Ölçek / test verileri nominal veya sıralı ölçek verisi olarak değerlendirildiğinde iki madde arasındaki ilişkiler ki-kare analizi ile test edilebilir. Bu uygulamaya aynı zamanda *ki-kare bağımsızlık testi* adı verilir. Ki-kare sonucu ,05'ten küçük çıktığında, H_0: $\rho_{xy} = 0$ şeklinde belirlenen sıfır hipotezi reddedilerek maddelerin birbirinden bağımsız olmadığına karar verilir. Ancak ki-kare analizi, ilişkinin yönü (pozitif veya negatif olması) ve gücü (düşük, orta derecede veya güçlü olduğu) hakkında bilgi vermez. Ki-kare testinin uygulanabilmesi için iki ön koşulun sağlanmış olması gerekir; (a) her bir hücrede beklenen frekans en az 5 olmalı ve (b) veriler tesadüfî bir örneklemden elde edilmiş bulunmalıdır.

Phi korelasyonu. Nominal ölçek niteliğinde *Doğru, Yanlış* gibi *gerçek iki şıkka* sahip iki madde arasındaki ilişkileri veya tutarlılığı belirlemek için kullanılır. Ancak phi (ϕ) korelasyon katsayılarına bakılarak tek başına bir maddenin yetersiz olduğuna karar verilmez. Bu testin tamamlayıcısı olarak ayrıca *nokta-iki serili korelasyon analizi* yöntemine başvurulmalıdır. Test, SPSS'te Correlate veya Crosstabs mönüsü ile hesaplatılır. Diğer korelasyon analizlerinde olduğu gibi phi-kare (phi^2) iki madde arasındaki paylaşılan ortak varyansı gösterir.

Phi korelasyonu ve katsayısının kullanıldığı bir diğer alan *performans testlerinde* bireysel maddelerin başarısı ile genel test puanlarına ait başarı arasındaki ilişkileri saptamaktır. Bu uygulamada bireysel maddeler 1 = *Doğ-*

ru, 0 = *Yanlış;* test puanlarının başarısı ise *kesim puanı*[a] dikkate alınarak 1 = *Standart puana sahip* veya *Standardın üzerinde* ve 0 = *Standardın altında* şeklinde kodlanır. Böylece her bir madde için phi katsayısı hesaplanmış olur. Yetkin ve yetkin olmayan kişilerin dağılımı büyük ölçüde sola çarpık olması nedeniyle phi değerlerinde daha düşük katsayılarla çalışılır. Kriter referanslı testlerde phi katsayıları ,30 ilâ ,70 arasında değişir. Yüzde 70 katsayısı, madde ile toplam başarı arasında güçlü bir ilişki olduğunu gösterirken ,30 katsayısı bu ilişkinin düşük olduğu anlamına gelir. Teste alınacak maddelerin toplam test başarısı ile olan korelasyonu en azından orta derecede (,40 ve üzeri) bir ilişkiyi ortaya koymalıdır.[10]

Cramer V. Ki-kare istatistiğinin türevi olan ve iki kategorik değişken arasındaki ilişkileri analiz eden bu testte 2x2'den daha fazla kontenjan (2x3 gibi) vardır. Diğer bir deyişle *Doğru-Yanlış* gibi iki şıklı değişkenlerin dışında üç, dört veya beş şıklı değişkenler arasındaki ilişkileri incelemeye müsait olan bir testtir. Nominal değişkenlere dayalı olarak yapılan Cramer *V* ki-kare analizinin tersine örneklem büyüklüğünden etkilenmez ve bu nedenle şüpheli ki-kare analizlerinin yerine kullanılır. Bu analiz Correlates mönüsü altında değil, Crosstabs mönüsü altında bulunur. Cramer *V* katsayısı sıfıra yaklaştığı ölçüde iki değişken arasında ilişki olmadığı kararına varılır. Katsayı 1'e yaklaştığı ölçüde ise korelasyon katsayısında olduğu gibi ilişkinin güçlü olduğu düşünülür. Cramer *V* katsayısı negatif değer vermez.

Goodman ve Kruskal gamma. Gamma, ilişki aranan değişkenlere ait şıkların sıra büyüklüğü içinde düzenlenmiş olduğunu varsayar ve bu nedenle sıralı veriler için uygundur. Gamma, eşleşmiş çiftli değerleri ihmal ederek eşleşmemiş değerlere dayalı olarak hesaplama yapar. Gamma eşleşmiş çiftlerden eşleşmemiş çiftelerin çıkarılması ve bu değerin eşleşmiş çiftler artı eşleşmemiş çiftler toplamına bölünmesiyle bulunur.[11] Gamma değeri ilişkinin gücü hakkında bilgi verir. Artı 1'e yakın bir değer güçlü bir ilişki olduğunu sıfıra yakın bir değer ise ilişkinin zayıf olduğunu gösterir.

Tetrakorik korelasyonu. Puanların geçici veya yapay olarak 1 ve 0 şeklinde kodlandığı iki gözlem değişkenine ait değerlerin, kendilerini temsil eden *arka plandaki gizli değişkende normal dağıldığı* varsayımı altında yapılan korelasyon analizi tekniğidir. Tetrakorik korelasyonun özellikleri aşağıdaki gibidir.

---

[a] Kesim puanı. Uzmanlık veya yetkinliği belirleyen sınır değer. Örneğin 10 üzerinden 7 puan almak yetkinlik sınır değeri olarak belirlenmiş olabilir.

1. Her iki değişken de veya analize alınan tüm değişkenler ikili veri yapısına sahiptir.
2. Sürekli değişken verileri, belirli bir sınır değerinin üzerinde olanlar 1 ve altında olanlar ise 2 (veya 1, 0) şeklinde kodlanmıştır.
3. Arka plandaki veriler sürekli veri niteliğindedir veya bu verilerin normal dağılım özelliğine sahip olduğu varsayılır.

Sürekli değişken verilerinden belirli bir sınır değerinin üzerinde olanların 1 ve altında olanların 0 olarak kodlanmasının faktör analizi sürecinde yapay sonuçlara yol açmayacağı belirtilmiştir. Maddelerinin güçlük derecelerinin farklı olması faktör analizinin varsayımlarını veya ön koşullarını etkilemez.[12] Karmaşık formül yapısı nedeniyle el ile hesaplanması zor olan bu korelasyon için istatistiksel analiz yazılımları kullanılır. Bilim adamları bu konuda popüler istatistik yazılımlarından biri olan SYSTAT adlı programı kullanabilirler. Yazılımın Correlations diyalog kutusunda, önce Binary şıkkı seçili hale getirilir ve daha sonra açılır mönü listesinde Tetra şıkkı ile gerekli hesaplama yapılır. Yazılım tetrakorik korelasyon matrisini ayrı bir dosya halinde kayıt edecektir. Faktör analizi yapabilmek için kayıt edilen *tetrakorik korelasyon matrisi* dosyası açılarak faktör analizi yapılacak değişkenler bu dosyadan belirlenir. Verilerin arka planda normal dağılıma sahip olduğu varsayıldığından ortak faktör analiz (iterated principal axis) yöntemi tercih edilir.[13]

Pearson korelasyon katsayısının standart hatasına göre, tetrakorik korelasyon katsayısının standart hatasının çok daha büyük olması nedeniyle sadece büyük örnek kütlelerde çalışıldığı zaman güvenilir bir değer verir.[14] Tetrakorik korelasyon analizi, ikili değişkenlerde *ortak faktör analizinin* ön aşaması olarak kullanılır. Diğer bir deyişle ortak faktör analizi tetrakorik korelasyon matrisi verilerine dayalı olarak yapılır. Bir ölçekteki veya testteki düşük korelasyona sahip maddeleri ayıklamak için kullanılmaz.

Polikorik korelasyon. Çok dereceli ve yapay sıralı ölçek verisi niteliğine sahip iki değişkenin arka plandaki gizli değişkende normal dağıldığı varsayımı altında yapılan korelasyon analizi tekniğidir. Yapay sıralı ölçek verisi, IQ puanlarının 1 = yüksek, 2 = vasat, 3 = düşük olarak gruplandırılmasıyla elde edilir. Likert ölçeği toplam puanları da benzer şekilde gruplandırılabilir. Araştırmacı ölçüm modelini sınamak için, *yapısal eşitlik modeli*'ni kullanmışsa, çok dereceli yapay sıralı ölçek verilerinde maddelerin gizli yapıyla

ilgili olup olmadığını belirlemek için bu yönteme başvurur. Bu hesaplama yöntemi SPSS'te bulunmaz, araştırmacılar bu konuda yapısal eşitlik modelini test eden PRELIS, SAS ve EQS gibi yazılımları kullanabilirler. Polikorik korelasyon analizi, çok dereceli/şıklı değişkenlerde ortak faktör analizinin veya yapısal eşitlik modelinin ön aşaması olarak kullanılır.

TOPLAM PUAN İLE MADDE PUANLARI ARASINDAKİ KORELASYON

Toplam puan ile madde puanları arasındaki korelasyon analizi maddelerin güvenilirliklerini belirlemeye yöneliktir. İstatistiksel analiz programları bu testleri değişik düzeylerde içermiştir. Bu bölümde, PRELIS isimli yazılımda yer alan *nokta-iki serili* korelasyon analizi ile *iki serili korelasyon* analizi ve SPSS yazılımında yer alan *Pearson ve Spearman korelasyon analizi* yöntemleri üzerinde durulmuştur.

Nokta-iki serili korelasyon analizi. Bu yöntem, Pearson moment çarpımları korelasyonunun değişik bir formülü ile hesaplanır. Aynı zamanda *ayırma indeksi* (*A*) olarak bilinen bu analiz, bir yetenek veya başarı testindeki maddelerin başarılı kişilerle başarısızları ayırt etme gücünü ortaya koyar. Korelasyon analizi yapılacak iki değişkenden biri *gerçek ikili* (1 = doğru, 0 = yanlış), diğeri sürekli veri niteliğinde ise uygulanır. Bir testte maddeler ikili veri[a] özelliğinde iken toplam puanları sürekli veri olarak kabul edilir. Ancak bu yöntemde, maddelerin güçlük dereceleri farklı ise iki seride hiçbir zaman 1,00 gibi yüksek bir korelasyon katsayısı elde edilmez. Test maddeleri aşamalı olarak zorlaşan (örneğin progresif matris testinde olduğu gibi) ölçümlerde güçlük dereceleri farklı olduğundan *nokta-iki serili korelasyon* analizi yerine sadece *iki serili korelasyon* analizi uygulanır. Nokta-iki serili korelasyon analizi formülü Eşitlik 6-1'deki gibidir.

$$\rho_{nis} = \frac{(\mu_+ - \mu_x)}{\sigma_x}\left(\sqrt{p/q}\right). \qquad (6\text{-}1)$$

μ_+ : $Y = 0$ olduğunda X değerlerinin ortalaması.
μ_x : $Y = 1$ olduğunda X değerlerinin ortalaması.

[a] İkili verilerin 0, 1 veya 1, 2 şeklinde kodlanmasının hesaplama üzerinde bir etkisi yoktur.

σ_x : X değerlerinin standart sapması.
p : $Y = 1$ olduğunda değerlerinin oranı.

Nokta-iki serili korelasyon analizinde maddeler 1 ve 0 şeklinde kodlanmıştır. Sürekli veriler ise ölçümün toplam puanlarından oluşur. *Nokta-iki serili korelasyon* analizinde sonuç eğer negatif çıkmışsa, puanları düşük olan gruptaki cevaplayıcıların teste doğru yanıt verdiklerini veya başarılı gruptaki kişilerin yanlış cevap verdiklerini gösterir ki bu durum genelde istenmez. Eğer korelasyon katsayısı 0 ilâ ,20 arasında çıkmışsa söz konusu maddelerin iyi öğrencilerle kötü öğrencileri iyi ayırt etmediğini gösterir. Bu maddeler testin güvenilirliğini düşürdüğünden nihaî test uygulamasına alınmamalıdır. Korelasyon katsayısı ,21-,40 arasında olanlar ayırt edicilik özelliği *iyi* olan maddelerdir. Korelasyon katsayısı ,40'ın üzerinde olanlar ise ayırt edicilik açısından *çok iyi* olarak değerlendirilir. Bir testten iyi puan alan ilk %27'lik dilimle düşük puan alan son %27'lik dilim arasında *nokta-iki serili korelasyon analizinin* yapılabilmesi için örneklem verilerinin en az 50 vak'dan oluşması gerekir. Nokta-iki serili korelasyon analizi r_{nis} simgesiyle gösterilir. İstatistiksel analiz programı SPSS'te özel olarak düzenlenmiş nokta-iki serili korelasyon analizi yoktur. Bu konuda Internet'teki Ağ kümelerinde bulunan SPSS için yazılmış nokta-iki serili korelasyon analizi program kodlarından yararlanılabileceği gibi SPSS'in Pearson veya Spearman korelasyon analizi seçenekleri de kullanılabilir. Korelasyon analizi yapılırken ilişki kurulan maddeye ait puanın, toplam puandan çıkarılması daha sağlıklı bir hesaplama yapılmasını temin eder. Bu konuda yararlanılabilecek bir diğer yaklaşım SPSS'in Reliability mönüsü altında bulunan Interclass correlation coefficient ve scale if item deleted kutucuklarının seçili hale getirilerek hesaplama yapılmasıdır. Bu hesaplama sonucunda "düzeltilmiş madde-toplam puan korelasyon değerleri" elde edilir. Nokta-iki serili korelasyon analizinin PRELIS isimli istatistiksel analiz yazılımında bulunduğu bildirilmiştir.

Nokta-iki serili korelasyon analizi Rasch ölçüm yönteminde çok dereceli ölçeklerin ve kısmî kredili modellerin tek boyutlu olup olmadığını belirlemeye yönelik olarak kullanılır. Maddelerin korelasyon katsayısı ,70'in üzerinde ise bu maddelerin tek bir boyutu ölçtüğü sonucuna varılır.

İki serili korelasyon analizi. Bu yöntemde *yapay iki serili*[a] (yüksek puanlar = 1, düşük puanlar = 2) maddelerin yanında çok dereceli / şıklı maddeler de toplam puanlar ile korelasyona tâbi tutulur. Nokta-iki serili korelasyonunun düzeltilmiş biçimidir. İki serili korelasyonda, sıfırdan 1'e kadar tüm değerler elde edilebilir. Bu yöntemin, *nokta-iki serili* korelasyon analizi uygulamasından farkı, maddelere ait derecelerin ikili, üçlü, beşli, yedili olabilmesidir. Maddelerin ve toplam puanların temelinde yatan gizli özelliğin normal dağıldığı varsayılan iki serili korelasyon analizinde korelasyon katsayısının bazı durumlarda 1'den yüksek çıkabileceği bildirilmiştir. Ayrıca iki serili korelasyon analizinde H_0 hipotezini test eden anlamlılık testi yapılmaz.[15] İstatistiksel analiz programı SPSS'te iki serili korelasyon analizi yoktur. Bunun için Correlations veya Reliability mönüsündeki korelasyon analizlerinin kullanılması önerilmiştir. İki serili korelasyon analizi formülü Eşitlik 6-2'deki gibidir.[16]

$$\rho_{is} = \frac{(\mu_+ - \mu_x)}{\sigma_x}(p/Y). \qquad (6\text{-}2)$$

μ_+ = Maddeye doğru yanıt veren kişilerin puan ortalaması (kriter değer).
μ_x = Bütün kişilerin puan ortalaması.
σ_x = Bütün kişilerin puanlarının standart sapması.
p = Doğru yanıt verenlerin oranı.
Y = Doğru yanıt veren kişilerin oranını gösteren ordinat değeri.

Nokta-iki serili ve iki serili korelasyon analizleri test ve ölçeklerin içeriğine en yüksek korelasyon değerine sahip maddeleri alıyor olması nedeniyle eleştirilmiştir. Bu uygulamanın yüksek ve düşük puan alan kişiler arasında yüksek puan alanlar lehine bir yanlılık yarattığı, azınlık gruplarının bu tür testlerden başarılı olma şanslarının düşük olduğu ifade edilmiştir. Güvenilirliğin artmasıyla birlikte daha az şanslı olan gruplara yönelik yanlılığın artması bu uygulamanın temel sakıncası olarak görülür.[17]

---

[a] Yapay iki serili veri. Sürekli veya nicel verilerin bilim adamının isteği doğrultusunda iki serili veri haline dönüştürülmesidir. Örneğin, 130'dan düşük puan alanların 1 ve 130'dan yüksek puan alanların ise 2 olarak kodlanması. Yapay *iki serili verilerin* arka planındaki yapı gerçek sürekli değişken niteliğindedir.

SPSS yazılımındaki korelasyon analizi. Madde puanlarıyla toplam puanlar arasındaki korelasyonlar istatistiksel analiz programı SPSS'te iki şekilde hesaplanabilir; Reliability ve Correlates mönüsü kullanılarak.

Birinci yöntemde Reliability Analysis mönüsü ile açılan pencerede Scale if Item Deleted şıkkı seçili hale getirilerek gerekli hesaplatma yapılır. Daha sonra Output penceresinde Item - Total Statistics başlığı altında madde – toplam korelasyon katsayıları incelenir. Bu yöntemde madde puanlarının ikili, çoklu sıralı ölçek niteliğinde veya eşit aralıklı olması dikkate alınmaz.

İkinci yöntemde, Analysis mönüsü altında Correlete düğmesinden yararlanılır. Burada Bivariate mönüsüne girilerek duruma göre Spearman veya Pearson korelasyon katsayısı hesaplatılır. Pearson korelasyon analizinin uygulanabilmesi için madde puanlarının eşit aralıklı veri kabul edilmesi veya varsayılması gerekir. Bunun yanında veriler normal dağılım özelliğine sahip ve büyük bir örneklem grubundan sağlanmış olmalıdır. Likert tipi ölçeklerde madde puanlarının sıralı mı yoksa eşit aralıklı mı olduğu tartışmalı bir konudur. Bazı kaynaklarda elde edilen veriler eğer parametrik verilerin taşıması gereken koşullara sahipse kökenine bakılmaksızın bu verilere parametrik istatistiksel analizlerin uygulanabileceği belirtilmiştir.[18]

Likert ölçeklerinin madde puanları eşit aralıklı olarak değerlendirilmese bile toplam puanları eşit aralıklı ölçek olarak kabul edilir. Sıralı ölçek verileri 11 veya 15 dereceden fazlaysa bu verilerin dağılımı normale yakın çıkar. Nunnally ve Bernstein (1994) sıralı ölçek verileri 11 veya daha fazla ise Pearson korelasyon analizinin uygulanabileceğini belirtmişlerdir (aktarılan kaynak, U. Of New Brunswick).[19] Joreskog ve Sorbom ise 15 veya daha fazla sıralı verinin, *sürekli veri* olarak değerlendirilebileceğini iddia etmişlerdir (aktaran, Yaffee, 1999).[20] Bu kitapta araştırmacılara önerimiz, eğer ölçüm yapılan konuyla ilgili olarak kişiler veya objeler hakkında karar veriliyorsa Likert maddelerinin sıralı ölçek verisi olarak düşünülmesi, toplam puanların ise eşit aralıklı ölçek olarak kabul edilmesi yönündedir. Likert maddeleri karar vermek için değil de, veri taraması yapmak için veya bir ölçeğin geçerlilik ve güvenilirliğini saptamak için ele alınmışsa belirli bir büyüklüğe ve yaklaşık normal dağılım özelliğine sahip olması koşuluyla eşit aralıklı olarak değerlendirilebilir. Bilim adamı eğer bu yaklaşımdan hareket ederse, madde-toplam puan korelasyon analizi için Pearson veya maddelerin normal dağılım özelliği göstermemesi halinde

Spearman korelasyon analizi yöntemini uygulayabilir. Araştırmacı veri yapılarına uygun istatistiksel analiz türünü belirlemek için Tablo 6-2'den yararlanabilir.[21] Spearman korelasyon analizinde, madde puanları ve toplam puanlar SPSS'in arka planında otomatik olarak büyüklük sırasına sokulur ve analiz bundan sonra gerçekleştirilir. Bu analiz bilim adamına önemli ölçüde uygulama esnekliği sağlamasına karşın Pearson'a göre daha zayıf bir test olarak değerlendirilmiştir. Spearman korelasyon analizinde veri yapısında normallik, türdeşsellik ve doğrusallık aranmaz. Veriler büyük ölçüde çarpık olsa bile uygulanabilir. Analiz ayrık değerlere daha az duyarlıdır. Spearman korelasyon analizi 7 ilâ 30 arasında değişen örneklem büyüklükleri için uygundur.

Poliseriyal korelasyon. Araştırmacı, nedensel ilişkileri belirlemeye yönelik olarak yapısal eşitlik modelini (YEM) temel almışsa, sıralı ölçek niteliğindeki madde puanları ile eşit aralıklı ölçek niteliğindeki toplam puanlar arasındaki ilişkiyi *poliseriyal korelasyon* analizi tekniği ile inceleyebilir. Bu teknik PRELIS isimli istatistiksel analiz programında bulunur. Poliseriyal korelasyon analizi, maddeleri ayıklamak için değil, geliştirilen modelin güvenilirliğini test etmek amacıyla kullanılır.

Tablo 6-2. Verilerin Niteliği ve Korelasyon Analizi

	1. değişken	2. değişken	Test	Açıklama
Nominal	Nominal (ikili)	Nominal (ikili)	Phi katsayısı	Gerçek ikili
	Nominal (ikili)	Nominal (ikili)	Tetrakorik	Yapay ikili
	Nominal (çoklu)	Nominal (ikili)	Kontenjan kats.	
	Nominal	Nominal	Lambda	Bağımlı/bağımsız
	Nominal (ikili/çoklu)	Eşit aralıklı/oranlı	Biserial	Yapay ikili
	Nominal (ikili)	Eşit aralıklı/oranlı	Point biserial	Gerçek ikili
	Nominal (ikili)	Nominal (ikili/çoklu)	Cramer V	
Sıralı	Sıralı	Sıralı	Spearman rho	Yapay ikili
	Sıralı	Sıralı	Kendall tau	$n<10$
	Sıralı	Sıralı	Goodman	
	Sıralı	Sıralı	Somer *D*	
Eşit aralıklı	Sıralı	Eşit aralıklı/oranlı	Pearson *r*	Madde analizi
	Eşit aralıklı/oranlı	Eşit aralıklı/oranlı	Pearson *r*	Yapay ikili
	Sıralı	Eşit aralıklı/oranlı	Spearman rho	Madde analizi
	Nominal, sıralı	Eşit aralıklı/oranlı	Eta	Eğrisel ilişkiler

TOPLAM PUANLAR ARASINDA KORELASYON

Bilim adamı; (a) test-yeniden test, (b) yarıya bölme, (c) paralel formlar ve (ç) gözlemciler arası güvenilirlik analizleri için toplam/ortalama puanlar arasındaki ilişkileri belirlemeye yönelik olarak yine korelasyon analizi yönteminden yararlanır. Likert ölçeğine ve diğer indekslere ait toplam/ortalama puanlar eşit aralıklı ölçek olarak kabul edilir. Cliff'e göre (1992, Barrett'ten) bir ölçüm aracında birbiriyle tutarlı sıralı ölçek niteliğinde en az üç maddenin bulunması eşit aralıklı ölçek olarak tanımlanması için gerekli ve yeterlidir.[22]

Ölçüm birimleri nedeniyle toplam puanlar arasında farklılıklar varsa, puanlar normal dağılım özelliği göstermiyorsa korelasyon analizi yapmadan önce rakamlar standart *z* puanlarına dönüştürülür. Örneğin, ölçeklerden biri 5 dereceli ve diğeri 7 dereceli ise, ölçeklerdeki madde sayıları farklı ise puanları standartlaştırmak gerekir. Standart *z* puanlarına dönüştürme gereği duyulan bir diğer alan, sürekli verilerle çalışırken ekstrem değerlerle karşılaşılması halidir. Puanlarda eğer ayrık değerler varsa, dizideki değerler normal dağılım özelliği göstermiyorsa daha doğru bir hesaplama için yine standart *z* puanları kullanılır. Çünkü testlerde korelasyon katsayısı kullanılan birimlerden, madde ve ölçeklerin derece sayılarından bağımsız olmalıdır. Kendall ve Stuart'a göre (1958) toplam puanlar eğer $z = \pm 2{,}0$'den daha fazla bir çarpıklık değerine sahipse korelasyon katsayıları Pearson korelasyon katsayısının gerçek değerini göstermez (aktaran Barett, 2003).[23] Bilim adamı verilerin niteliğine, örneklem büyüklüğüne ve verilerin dağılım özelliğine göre korelasyon analizlerinde Pearson veya Spearman yöntemini uygular. Pearson yönteminin uygulanabilmesi için bu analizle ilgili varsayımların karşılanmış olması gerekir.

Korelasyon analizi test-yeniden test, yarıya bölme, paralel formlar ve gözlemciler arası güvenilirliği belirlemeye yönelik olarak uygulanırken elde edilen güvenilirlik katsayısının ,75 veya ,80 gibi bir değer olması tek başına bir anlam ifade etmez. Bu katsayının aynı zamanda %95 güven aralığında anlamlı olup olmadığına bakmak ve *p* değerini de vermek gerekir. Anlamlılık düzeyi, etki büyüklüğü ve örneklem hacmi çerçevesinde değerlendirilir. Elli kişilik bir örneklem hacminden elde edilecek ,80 güvenilirlik katsayısı ile 500 kişilik bir örneklem hacminden elde edilecek ,80 güvenilirlik katsayısı aynı değildir.

Toplam puanlar arasındaki korelasyon, istatistiksel analiz programı SPSS'te Analysis mönüsü altında Correlete düğmesinden yararlanılarak yapılır. Buradan Bivariate mönüsüne girilerek duruma göre Spearman veya Pearson korelasyon katsayısı hesaplatılır. Korelasyon katsayısı aynı zaman-

da güvenilirlik katsayısıdır. Araştırmacı isterse bu rakamın karesini alarak iki ölçüm arasındaki ortak varyansı da görebilir.

Bazı bilim adamları *istikrarlılığı* ve *eş değerliliği* belirlemek için korelasyon analizinin kullanılmasının doğru olmadığını ifade etmişlerdir. Bu bilim adamlarına göre, korelasyon analizi iki ölçüm değeri arasındaki ilişkiyi gösterir, tutarlılığı değil. Analizin iki değişken arasında yapılması puanların bütün olarak görülmesini engeller. Güvenilirlik, tek değişken üzerinde yapılırsa anlamlıdır. Ayrıca bir çok araştırmada ikiden fazla zamanda ölçüm yapılır veya ikiden fazla paralel form kullanılır. Bu tür çoklu ölçümlerde korelasyon analizi kullanılırsa bu analiz ile değişkenliğin kaynağı saptanamaz. Pearson momentler çarpımı korelasyon katsayısı (PMÇKK) test puanlarındaki dalgalanmalara karşı duyarlı değildir ve bu nedenle güvenilirlik rakamlarını olduğundan daha büyük gösterir. Bu bilim adamlarına göre, *kümeler arası korelasyon analizi* anlamına gelen PMÇKK yerine, *küme içi korelasyon analizi* (KİK) yöntemini uygulamak daha doğrudur. Çünkü, Pearson uygun bir şekilde kullanıldığında güvenilirliği değil, geçerliliği ölçer.[24] Bu iddia çerçevesinde KİK analizini uygulamayı düşünen bilim adamı bilgisayara v_1 sınıflama değişkenini birinci ölçüm = 1, ikinci ölçüm = 2, üçüncü ölçüm = 3 olarak ve ölçüm sonuçlarını ise v_2 = toplam/ortalama puanlar olarak girmelidir. İstatistiksel analiz programı SPSS'te *tek yönlü varyans analizi* (TYVA) mönüsü altında Factor hücresine sınıflama değişkeni olan v1 tanıtılır.

GÖZLEMCİ PUANLARI ARASINDAKİ KORELASYON

Gözlemciler arasındaki değerlendirmenin güvenilirliği konusuna "Toplam Puanlar Arasındaki Korelasyon Analizi" başlığında değinilmişti. Ancak kendi içinde farklı hesaplama yöntemleri içermesi nedeniyle bu bölümde ayrı bir başlık halinde ele alınmasının doğru olacağı düşünülmüştür. Gözlemcilerin verdikleri puanlar arasındaki ilişkiyi tespit etmek için değişik istatistiksel analizlerden ve matematiksel hesaplama yöntemlerinden yararlanılır.

Gözlemcilerin verdikleri puanlar arasındaki uyuşmayı belirlemek için öncelikle verilerin niteliği göz önünde bulundurulur. Veriler; (a) sınıflandırılmış, (b) ikili, (c) çok dereceli veya (ç) sürekli olabilir (*bk.*, Tablo 6-8). Bu dörtlü sınıflama kısaltılarak iki grupta toplanır: kesikli veriler ve sürekli veriler. Gözlemcilerin yaptıkları değerlendirme sonucunda verdikleri puanların güvenilirliğini saptamak için aşağıdaki hesaplama yöntemleri ve istatistiksel analizler uygulanır:

1. Kesikli.
   a. Nominal verilerde *Uyuşma yüzdesi.*
   b. Nominal ikili verilerde *Phi katsayısı.*
   c. Nominal ikili verilerde (2x2) *kappa istatistik tekniği.*
   d. Kendall W uyuşma katsayısı.
   e. Sıralı verilerde *Spearman rho*
   f. Sıralı verilerde *Kendall tau*

2. Sürekli (normal dağılım özelliğine sahip).
   a. Pearson korelasyon analizi.
   b. Küme içi korelasyon analizi.

3. Sürekli (normal dağılım özelliği göstermeyen).
   a. *Spearman rho.*
   b. *Kendall tau.*

Bilim adamı hesaplama yapmaya karar vermeden önce verilerin niteliğini, örneklem büyüklüğünü, verilerin büyüklük sırasına sokulma ve normal dağılım özelliği gösterme durumunu araştırmalıdır (*bk.*, Tablo 6-3).

Uyuşma indeksi. Uyuşma indeksi, gözlemcilerin veya değerlendiricilerin uyuştukları madde sayısının toplam değerlendirme veya gözlem sayısına olan oranıdır. Bu yöntemin olumsuz yönü, tesadüfî uyuşmaları dikkate almamasıdır. İstatistiksel test değil, bir tür hesaplama yöntemidir.

■ Uyuşma indeksi formülü.

Uİ = ((Toplam uyuşma sayısı) / (Toplam değerlendirme sayısı))*100.
Toplam değerlendirme sayısı = Uyuşan + Uyuşmayan değerlendirme sayıları.

Gözlemciler arası değerlendirme sonuçlarının güvenilir sayılabilmesi için *Uİ* değerinin %75'in üzerinde olması gerekir. Daha düşük bir oran, değerlendirmede gözlemcilerin önemli ölçüde farklı düşündükleri anlamına gelir.

■ Örnek: Üç gözlemci, 10 kişilik bir sınıftaki öğrencilerin el yazılarının düzgün ve okunaklı olma durumunu 3 dereceli bir ölçek üzerinde (1 = iyi, 2 = vasat, 3 = zayıf biçiminde) değerlendirmiş olsun. Bu uygulamada gözlemcilerin aynı puanı verdikleri (her üçünün de 1, 2 veya 3) öğrenci sayısı sekiz

ise bu rakam toplam değerlendirme sayısı 10'a bölündüğünde %80 rakamı elde edilir.

Phi katsayısı. Phi, ikili veri yapısına sahip (0, 1 veya 1, 2 gibi) iki değişken arasındaki ilişkileri belirlemek için kullanılır ve bir tür Pearson çarpım momentleri korelasyon katsayısıdır. Nominal nitelikteki ikili veri yapıları, kukla değişken olarak belirlenmiş olabilir. Örneğin iki hakemin kişileri aldıkları puanlara bakarak *Başarısız* veya *Başarılı* şeklinde değerlendirmeleri arasındaki güvenilirlik phi katsayısıyla saptanır. Phi katsayısı 0 ilâ 1,0 arasında değişir ve 1,0 rakamı hakemlerin yaptıkları değerlendirmenin güvenilir olduğu anlamına gelir. Nominal ölçek değerleri kullanıldığından *negatif phi katsayısı* elde edilmez. Araştırmacı isterse iki değişken arasındaki ilişkiyi ki-kare testi ile de analiz edebilir. Ancak ki-kare istatistiği küçük örneklemlerle çalışıldığında istenen sonucu vermez. Beklenen değer matrisinde herhangi bir hücredeki değer 5'ten küçükse Yates'in düzeltme faktörü (Fisher kesin testi) kullanılır. İstatistiksel analiz programı SPSS'te phi katsayısı tanımlayıcı istatistiksel analiz mönüsünde bulunan Crosstabs seçeneği kullanılarak yapılır.

Cohen kappa. Cohen (1960) tarafından geliştirilen Kappa istatistiği, iki veya daha fazla gözlemcinin yaptığı değerlendirmeler arasındaki uyuşmayı belirlemek için kullanılır. İstatistik Cohen'den sonra Scott (1955), Maxwell ve Pilinler (1968) ve Bangdiwala (1987) tarafından geliştirilmiş ve son yıllarda Gaccione (1993) tekniği SAS istatistiksel analiz yazılımına uyarlamıştır (aktaran Gren).[25] Kappa katsayısının uyuşma indeksinden farkı, şans faktörünün etkisini ortadan kaldırmasıdır. Gözlemcilerin verdikleri kategorik nitelikteki (nominal) puanlar istatistiksel analiz programı SPSS'te Crosstabs mönüsü altında Kappa seçeneği ile hesaplatılır. Ancak bu bölümde sadece iki değerlendiriciye ait veriler için hesaplama yapılabilmektedir. İkiden fazla değerlendirici veya gözlemci için özel olarak yazılmış SPSS makrolarından[a] yararlanmak veya el ile hesaplama yapmak gerekir. İstatistiksel analiz yazılımı SAS'ı kullanan araştırmacılar bu amaçla yazılmış program kodlarından yararlanabilirler. Ayrıca İnternet ortamında hesaplama yapabilen çevrimiçi yazılımları kullanmak da mümkündür. Siegel ve Castellan (1988) ikiden fazla değerlendirici arasındaki

[a] Bilgisayar programlarında yapılacak ardışık işlemleri, tanımlandıktan sonra bir tıklama veya bir komutla kendiliğinden yapan otomat. "Otomatik çalışan yazılım kodu" anlamında "otokod".

güvenilirliği hesaplamak için Eşitlik 6-3'teki formülü önermiştir (aktaran Yaffee, 1999).[26]

$$K = \frac{\text{Gözlenen Uyuşma Oranı} - \text{Beklenen Uyuşma Oranı}}{1 - \text{Beklenen Uyuşma Oranı}} . \qquad (6\text{-}3)$$

Formülü çalıştırmak için *karşılaştırma matrisi* (confusion matrix) düzenlenir ve bu matris üzerinden gerekli hesaplamalar yapılır (*bk.*, Tablo 6-3). Karşılaştırma matrisine; örneğin, değerlendirmeler dört dereceli ölçek verilerine göre (A, B, C, D) yapılmışsa gözlemcilerin yapmış olduğu değerlendirmelerin niteliği yazılır.

■ Değerlendirme verileri.

Kişiler	1	2	3	4	5	6	7	8	9	10	.....	*n*
Gözlemci 1:	A	C	D	A	B	B	C	C	A	B	.....	*n*
Gözlemci 2:	B	C	D	D	B	C	C	C	B	B	.....	*n*

Değerlendirme verilerinde gözlemcilerin uyuştukları puanlar karşılaştırma matrisinin köşegenindeki hücrelere çentik atılarak işaretlenir. Uyuşmadıkları değerler ise köşegenin dışındaki harflerin kesiştiği hücrelere yine çentik atılarak doldurulur. Daha sonra her bir hücredeki çentik sayısı sayılarak uyuşma ve uyuşmama sayıları saptanır. Araştırmacının gözlenen uyuşma oranı ve beklenen uyuşma oranlarını bulabilmesi için satır ve sütun toplamlarını alması gerekir. Satır ve sütun toplamlarından sonra ayrıca genel toplam rakamı elde edilir. Gözlenen uyuşma oranı, matrisin köşegeninde yer alan uyuşum sayılarının toplam değerlendirme sayısına bölünmesiyle bulunur. Bu rakam şans faktörünü içermez. Beklenen uyuşma oranı ise, her bir hücrede şans faktörünü de içerdiğinden tablonun çaprazında kalan toplam değerlerin birbirleriyle çarpılarak genel değerlendirme sayısına bölünmesi ve çıkan rakamların toplanıp tekrar genel rakama bölünmesiyle bulunur.

■ *Gözlenen uyuşma oranı (GUO)*: Her iki gözlemcinin benzer puanlar vermesi veya değerlendirmeler yapması.

■ *Beklenen uyuşma oranı (BUO)*: Çaprazdaki satır toplamları ile sütun toplamları çarpımlarının toplam değerlendirme sayısına bölünmesiyle elde edilen hata oranları toplamının tekrar toplam değerlendirme sayısına bölünmesi.

Tablo 6-3. Dört Dereceli Bir Değerlendirme İçin Uyuşma Oranlarına Dayalı Kappa İstatistiği

		Birinci gözlemci				
	Dereceler	A	B	C	D	Toplam
İkinci gözlemci	A	9	2	8	15	34
	B	2	18	8	6	34
	C	3	4	8	1	16
	D	2	2	1	11	16
	Toplam	16	26	25	33	100

GUO: (9 + 18 + 8 + 11) / 100 = ,46 ,

BUO: [(16*34) /100 + (26*34) /100 + (25*16) / 100 + (33*16) /100] /100 ,

BUO: [5,44 + 8,84 + 4,00 + 5,28] / 100 ,

BUO = ,23 ,

Kappa: (,46 – ,23) / (1– ,23) = ,29 .

Kappa katsayısı –1 ilâ +1 arasında değişir. Sıfır değeri tesadüfî uyuşmayı, negatif değerler tesadüfî olmaktan daha kötü bir uyuşmayı, artı 1 ise mükemmel uyuşmayı temsil eder. Kappa katsayısı ,40 ilâ ,75 arasında ise makul bir uyuşma, ,75'ten büyük ise mükemmel bir uyuşma olduğu anla-

mına gelir.[27] Uyuşma oranı çok yüksekse[a] değerlendirmelerin doğru olmama ihtimali de göz önünde bulundurulmalıdır.[28]

Bazı araştırmacılar, çok gözlemcili ölçümlerde ikili olarak hesaplanan normal korelasyon katsayılarının ortalamasından hareket ederek güvenilirlik katsayısı hesaplama gibi bir yönteme başvurabilmektedirler, ancak bu uygulama gözlemciler arası güvenilirliği olduğundan daha büyük göstermesi nedeniyle tercih edilmemeli onun yerine *Çok Gözlemcili Kappa istatistiği* kullanılmalıdır. Gözlemciler arasındaki değerlendirmelerde kullanılan puanlar sürekli veri niteliğinde ise verilerin güvenilirliğini belirlemek için Kappa istatistiği yerine küme içi korelasyon analizi yöntemini kullanmak daha doğru olur.

Ölçüm değerlerinin güvenilirliğini belirlemeye yönelik olarak Kappa istatistiğini kullanmayı düşünen araştırmacıların bu yöntemin yetersiz kaldığı hususları da göz önünde bulundurmalarında yarar vardır. Posner ve Sampson (1990) bu yöntemin gözlemcilerin güvenilirliğini değerlendirmek için kullanılmasının genellikle tatmin edici olmayan sonuçlar verdiğini bildirmiştir (aktaran, Rudner, 2003).[29] Yöntem uyuşmazlığın şiddetini dikkate almamaktadır. Ayrıca sıralı ölçek verileri için yöntemin orijinal biçimi değil, Fleiss tarafından geliştirilen *ağırlıklı Kappa* formülü uygulanmalıdır.

Kendall uyuşma katsayısı W. Çok sayıda hakemin vermiş olduğu puanlar arasında ne ölçüde uyuşma[b] bulunduğunu belirler. Hakemlerin veya gözlemcilerin verdikleri puanlar sıralı ölçek verisi niteliğindedir. Bu test aynı zamanda üç ve üçten fazla ölçüm değişkeni arasındaki ilişkileri incelemek için de kullanılmıştır. Örneğin üç farklı gruptan elde edilen test sonuçları arasında anlamlı bir farklılık bulunup bulunmadığı Friedman ki-kare testi ile incelenebilir.

Çok sayıda hakem arasındaki uyuşma, aslında alternatif yöntem olarak Pearson, Spearman ve Kendall tau ikili korelasyon analizleriyle belirlenip

---

[a] Uyuşma oranlarının yüksekliği veya düşüklüğüne ilişkin standartlar farklı bilim adamları tarafından değişik şekillerde açıklanmıştır. Burada Fleiss'in standardı temel alınmıştır. Landis ve Koch ile Altman, Kappa katsayılarını beşerli gruplar halinde açıklamışlardır. Landis ve Koch'a göre ,00 - ,20 düşük, ,21-,40 makul ,41 - ,60 orta, ,61 - ,80 önemli, ,81 – 1,00 mükemmel uyuşma olarak nitelendirilmiştir.

[b] Avrupa Komisyonu İstatistik Bürosunun hazırladığı *ISI İstatistik Terimler Sözlüğü*'nde kavram *uyum katsayısı* biçiminde çevrilmiştir. Türkçede *görüşlerin uyumundan* değil *uyuşmasından* söz edilmesi nedeniyle *uyuşma katsayısı* biçimindeki çevirinin daha doğru olduğunu düşünüyoruz.

daha sonra bunların ortalaması alınarak da saptanabilir. Ancak hakemler sıralamada mutabık olmakla birlikte puanların büyüklüğü açısından farklı görüşlere sahip olduklarında bu tür korelasyon değerleri çok fazla anlamlı değildir. Puanların büyüklükleri de dikkate alınmak isteniyorsa küme içi korelasyon analizi uygulanır.

Bilim adamı, sadece puanları büyüklük sırasına sokarak bunlar arasındaki uyuşmayı araştırıyorsa böyle bir durumda puanların büyüklüğünü dikkate almaz ve *Kendall uyuşma katsayısı W* formülünden hareket eder. Bunun için önce hakemlerin vermiş oldukları puanlar büyüklük sırasına sokulur ve daha sonra formül uygulanır. İstatistiksel analiz programı SPSS'te Analyze, Nonparametric Tests, K Related Samples, Friedman ve Kendall's W düğme ve kutuları seçili hale getirilerek gerekli hesaplama yaptırılır. Ancak bu testi uygularken dikkat edilmesi gereken husus hakemlerin veya gözlemcilerin sütunlarda değil satırlarda yer almasıdır.[30] Sütunlarda ise değerlendirilen kişiler, nesneler veya olgular yer alır. Friedman testi ile Kendall W birbiriyle yakından ilgili olduğu için aynı sonuçlar elde edilir. SPSS'te rakamları ayrıca büyüklük sırasına sokmaya gerek yoktur, program bu işlemi arka planda otomatik olarak kendisi yapar. Kendall uyuşma katsayısı W değerleri 0 ilâ 1,0 arasında değişir. Sonuçlar raporlanırken Kendall W değeri yanında Friedman ki-kare değerleri de gösterilir [χ_F^2 (2, n :17) = 10.94, p = ,895].

Pearson. İki gözlemcinin yaptığı değerlendirmeler eşit aralıklı ölçek verilerine[a] dayanıyorsa, verilerin büyüklük sırası önemli değilse ve veri sayısı 30'un üzerinde ise ilişkiyi belirlemek amacıyla Pearson korelasyon analizi yönteminden yararlanılabilir.[b] Ancak, Pearson korelasyon analizi sadece iki puan dizisi arasındaki ilişkiyi belirler. Bu rakam bir uyuşma katsayısı değildir. Gözlemcilerin *yüzde kaç oranında* uyuştuklarını göstermez. Pearson *r* gözlemciler arasındaki varyansı dikkate almadığından *gözlemciler arasındaki değişkenliğe* karşı duyarsızdır. Bu nedenle giriş cümlesinde "yararlanılabilir" sözcüğü kullanılmıştır. Pearson, kümeler arası ilişki analizidir. Gözlemciler arasındaki değişkenliği daha iyi yansıttığı için bilim adamları kümeler arası ilişki analizi yerine, *küme içi ilişki analizini* kullanmayı tercih ederler. Gözlemci sayısı ikiden fazla ise ve gözlemciler arasındaki uyuşma-

[a] Değerlendirmeler 5, 7 veya 10 dereceli ölçek verileri şeklinde ise bu verilerin yaklaşık olarak eşit olduğu varsayılır.

[b] Pearson korelasyon analizi ile Spearman korelasyon analizi sonuçları arasında önemli bir farklılık yoktur. Yapılan araştırmalar, sıralı ve eşit gibi görünen ölçek verilerine Pearson ve Spearman korelasyon teknikleri uygulandığında sonuçların ,91 oranında benzer çıktığını göstermiştir.

ya önem veriliyorsa küme içi korelasyon analizi yöntemine başvurmak daha doğrudur.

Spearman. Gözlemcilerin verdikleri puanların kendileri değil de büyüklük sırasına sokulmuş (rank) değerleri temel alınmışsa, sürekli veriler normal dağılım özelliği göstermiyorsa veya Likert tipi ölçek verileri kullanılıp bu veriler *sıralı ölçek* niteliğinde değerlendirilmişse bilim adamı Spearman sıra korelasyonunu temel almalıdır (*bk.*, Tablo 6-4). Spearman sıra korelasyonunda veriler önce büyüklük sırasına sokulur. Ancak hesaplama istatistiksel analiz programı SPSS'te yapılıyorsa büyüklük sırasına sokmaya gerek yoktur, program bu işlemi arka planda otomatik olarak kendisi yapar. Bilim adamı gözlemciler arası değerlendirme sonuçlarını *p* değeriyle sıfır hipotezini test edecek şekilde verebileceği gibi, korelasyon katsayısını %95 güvenilirlik düzeyinde aralık tahmini yaparak da gösterebilir ve bu ikinci yaklaşım daha doğrudur. Spearman korelasyon katsayısının güven aralığını tahmin etmek için Fisher z´ dönüştürme yöntemi uygulanır.

Tablo 6-4. Spearman Sıra Korelasyonu Testinde Verilerin Büyüklük Sırasına Sokulması

Ham veriler		Büyüklük sırasındaki veriler	
X	*Y*	*X*	*Y*
6	2	2	1
3	6	1	2
7	10	3	4
8	8	4	3

Kendall tau a, b, c. M.G. Kendall (1938) tarafından geliştirilen Kendall tau katsayısı, ikili ve sıralı ölçek verileri arasındaki ilişkileri belirler. İki farklı gözlemcinin yaptığı değerlendirmeler puan büyüklüğü sırası içinde verilmişse veya bilim adamı ham puanları daha sonradan puan büyüklüğü sırasına sokmuşsa örneklem büyüklüğüne bağlı olarak Spearman korelasyon analizinin yanında Kendall *tau-b* analizi de yapılabilir. Kendall tau, Spearman testinin benzeridir. Ancak Kendall tau ve Spearman rho, büyüklük açısından birbirine eşit sonuçlar vermez. Çünkü formüllerin temelinde yatan mantık ile hesaplama biçimleri ve bu nedenle yorumları da farklıdır. Spearman rho büyüklük sırasına sokulmuş değerler arasındaki Pearson mo-

mentler çarpımını temsil ederken, Kendall tau *olasılığı* temsil eder. Diğer bir deyişle iki değişkende, aynı sırada yer alan (uyuşan) verilerin gözlenme olasılığı ile farklı sırada yer alan (uyuşmayan) verilerin gözlenme olasılığı arasındaki farktır.[31] Kendall *tau-b*, bağlı sıralılığa[a] sahip ölçek verileri için uygundur. Bu analizin özellikle verilerin normal dağılım özelliği göstermediği küçük örneklemlerde ($n < 20$, bazı yazarlara göre ise $n < 10$) daha iyi sonuçlar verdiği bildirilmiştir. Örneklem hacmi 20'den büyük ise ve veriler normal dağılım özelliğine sahipse korelasyon ölçüsü olarak Pearson r (Lehner, 1996, aktaran Packard),[32] veriler normal dağılım özelliği göstermiyorsa Spearman rho tercih edilir. İki serideki veriler birebir tam olarak uyuşuyorsa $\tau_b = +1$ değeri elde edilir. Diğer taraftan, veriler eğer tam bir uyuşmazlık içinde ise bu kez $\tau_b = -1$ sonucuyla karşılaşılır. (*Tau-b* simgesi Lâtin harfleriyle *tb* şeklinde veya Grek simgesiyle τ_b şeklinde gösterilir.)

Kendall istatistiklerinde, veri çiftlerinin aynı kategoriye düşüp düşmediğine ve hesaplama yöntemine[b] bakılarak *tau-a*, *tau-b* ve *tau-c* gibi adlar verilmiştir. *Tau-a*, uyuşan ve uyuşmayan çiftler arasındaki farkın toplam çift sayısına bölünmesiyle bulunur. Uyuşma istatistiklerinde Gamma testi en yüksek değeri verirken *tau-a* en küçük değere sahiptir. Öte yandan *tau-b* ve *tau-c* hesaplamalarından ise orta derecedeki değerler elde edilir.[33] Kendall *tau-b* "kare tablolar" için, *tau-c* ise "dikdörtgen" veya geniş tablolar için kullanılır. *Kendall tau-c* aynı zamanda *Stuart tau-c* veya *Kendall-Stuart tau-c* olarak da isimlendirilir ve *tau-b*'nin eşitidir. Pek çok vak'ada *tau-b* ve *tau-c* sonuçları birbirine benzer çıkar. Eğer farklı çıkmışsa daha emniyetli yaklaşım, düşük olan değere bakarak yorum yapmaktır.[34] Kendall tau değerlerinin yorumlanmasında aşağıdaki standartlar temel alınır:[35]

- Kendall tau katsayıları.

> ,50	Yüksek ilişki.
,36 – ,49	Önemli ilişki.
,20 – ,35	Orta derecede ilişki.
,10 – ,19	Düşük ilişki.
< ,10	İlişki yok.

---

[a] Bağlı sıralılık (tied ranks). Dizideki büyüklük sırasında bir ham değerden iki tane varsa ve bu değerler örneğin 4. ve 5. sırada yer almışsa her ikisinin de büyüklük sırası toplanıp ikiye bölünerek 4,5 olarak belirlenir ve bu şekilde sıralanan değerler "bağlı sıralılığa sahip" olarak adlandırılır.

[b] Hesaplama formülleri için *bk.*, D. Garson, "Ordinal Association [Sıralı İlişki]," t.y., <http://www2.chass.ncsu.edu/garson/pa765/assocordinal.htm> (24.02.2001).

Kendall *tau-a* istatistiği, SPSS'te tanımlanmamıştır. Onun yerine Crosstabs ve Correlations mönüleri altında Gamma, *tau-b* ve *tau-c* istatistikleri verilmiştir.

Küme içi korelasyon analizi. Patrick E. Shrout ve Joseph Fleiss (1979) tarafından geliştirilen küme içi[a] korelasyon analizi (KİK)[b], daha çok gözlemcilerin veya değerlendiricilerin yaptıkları puanlamanın tutarlılık güvenilirliğini belirlemek için kullanılır. Bunun yanında yapılan ölçümlerde X ve Y değişkenlerinden hangisinin bağımlı ve hangisinin bağımsız olduğunun saptanamadığı durumlarda da bu teknikten yararlanılabilir.[36] Bazı bilim adamları KİK'i aynı zamanda test-yeniden test, yarıya bölme güvenilirliği ve paralel formlar güvenilirliğinde de uygulama eğilimindedirler. Test-yeniden test ölçümlerinde $n < 15$ olduğunda veya ikiden fazla ölçüm yapıldığında KİK yönteminin uygulanması önerilmiştir.[37] Onlara göre, normal korelasyon analizi *geçerliliği* hesaplar, *güvenilirliği* değil. Güvenilirliği hesaplamak için KİK yöntemine başvurmak gerekir.

Küme içi korelasyon analizi en çok gözlemci değerlendirmelerinin güvenilirliğini belirlemek için kullanılmıştır. Gözlemci değerlendirmeleri değişik şekillerde yapılabilir:

1. Öğretmenlerin öğrencilerini değerlendirmeleri.
2. Öğrencilerin eğitimi ve öğretmenlerini değerlendirmeleri.
3. Yöneticilerin sistemi değerlendirmeleri.
4. Yöneticilerin astlarını veya astların yöneticilerini değerlendirmeleri.
5. Doktorların hastalarını değerlendirmeleri.
6. Gözlemcilerin sporcuları değerlendirmeleri.

Bu tür uygulamaların hepinde değerlendirme yapan kişilerin verdikleri puanların tutarlı olması aynı zamanda verilerin güvenilir olduğunu ortaya koyar. Küme içi korelasyon analizi, bir kişiye ait gözlem değerlerinin diğer kişilerin gözlem değerlerine ne ölçüde benzer olduğunu gösterir. Literatürde

---

[a] Kavram, Avrupa Birliği'nin yayımladığı istatistik sözlüğünde "sınıf içi" sözcüğüyle tanımlanmıştır.

[b] Küme içi korelasyon analizi ifadesinde *korelasyon* sözcüğünün kullanılmasının yanlış bir isimlendirme olduğu, çünkü bu hesaplamanın *varyans analizi* yöntemine dayandığı bildirilmiştir. Varyans analizi yöntemi kullanılmış olsa da, ilişki saptamaya yönelik içeriği nedeniyle bu bölümde ele alınmıştır.

kişiye ait gözlem değerleri, "grup içi" değerlendirmeler ve diğer kişilere ait değerlendirmeler ise "gruplar arası" terimiyle tanımlanmıştır (*bk.*, Tablo 6-5). Eğer, grup içi ve gruplar arasındaki puanların benzerlikleri yüksekse varyans düşük çıkar ve verilerin güvenilir olduğuna karar verilir.

Tablo 6-5. Grup İçi ve Gruplar Arası Gözlem Değerleri

Kişiler / Nesneler	1. gözlemci	2. gözlemci	3. gözlemci
1	12,00	18,00	16,00
2	18,00	18,00	17,00
3	19,00	14,00	16,00
4	20,00	15,00	17,00
5	12,00	12,00	11,00

Küme içi korelasyon analizinde, *n* sayıdaki kişiler/vak'alar ana kütleden tesadüfî olarak seçilmişlerdir. Öte yandan sütunlardaki *k* sayıdaki değişken ise gözlemcileri, ölçüm sayısını veya ölçüm yapılan sınıfları gösterir. Küme içi korelasyon analizi sonuçları, örnek kütle veya ana kütleyi tanımlamak için kullanılır. Ana kütle, küme içi korelasyon katsayısı rho (ρ_K) simgesiyle, örneklem küme içi korelasyon katsayısı ise büyük *R* harfiyle gösterilir. "Gruplar arasındaki varyansın toplam varyansa olan oranı" biçiminde tanımlanan küme içi korelasyon analizi Eşitlik 6-4'deki formülle hesaplanır.

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_h^2} . \qquad (6\text{-}4)$$

σ_a^2 = Gruplar arası değerlerin varyansı.

σ_h^2 = Grup içindeki değerlerin varyansı.

$\sigma_a^2 + \sigma_h^2$ = Toplam varyans.

Model türleri. Küme içi korelasyon katsayıları, bilim adamının yaptığı ölçümün niteliğine veya araştırmanın modeline göre değişir. Shrout ve Fleiss küme içi korelasyon analizi için üç farklı model öngörmüşlerdir. Araştırmacı, küme içi korelasyon analizinde bu üç modelden birini temel

alabilir. Her bir model daha sonra kendi içinde tekrar ikiye ayrılmıştır. Buna göre araştırmacı güvenilirliği tek bir gözlemci için veya gözlemcilerin ortalama puanlarına dayalı olarak hesaplar. Seçilen modeller parantez içinde KİK (1,1), KİK (1,3) şeklindeki simgelerle gösterilir. Parantezdeki birinci rakam ölçümün kaçıncı modele ait olduğunu ikinci rakam ise ölçümün formunu belirler. Ölçümün formu, güvenilirlik katsayısının tek bir gözlemcinin değerlerine mi ait olduğunu[a] yoksa gözlemcilerin puan ortalamasına mı dayandığını gösterir. Birden fazla gözlemci kullanılmış veya birden fazla ölçüm yapılmışsa bunların ortalamasının temel alınması güvenilirliği arttırır. Hangi yöntemin seçileceği, araştırmanın amacına, gözlemcilerin seçilme durumuna ve ölçümün yapılma biçimine bağlıdır. Bilim adamı bireysel puanların çok fazla belirsizlik taşıdığını düşünüyorsa, belirli sayıda hakemin veya gözlemcinin grup ortalamasını temel alır. Kaç tane gözlemciden yararlanacağını belirlemek için ise pilot araştırma KİK sonuçlarını kullanır. Ön araştırmadan elde edilen verilere dayalı olarak; $GS = KİK^*(1 - RL)/RL\ (1 - KİK^*)$ formülü ile gerekli gözlemci sayısını saptar. Formülde GS, gözlemci sayısını; RL, alt güvenilirlik düzeyini ve KİK*, kabul edilebilir minimum güvenilirlik katsayısını gösterir.[38] Küme içi korelasyon analizi türleri aşağıdaki gibidir:[39]

KİK (1, 1). *Tek yönlü varyans analizi* modeline[b] dayanır. Rasgele seçilen kişiler / nesneler, yine rasgele seçilen çok sayıda gözlemci tarafından değerlendirilirler. Bu modelde çok sayıda gözlemci kişilerin değerlendirilmesi için tesadüfî olarak atanmışlardır. Gözlemcilerin her biri bütün kişileri veya nesneleri değerlendirmez. Kişiler veya nesneler farklı gözlemci grupları tarafından değerlendirilirler.[40] Tesadüfî değişkenler, kişilere veya nesnelere farklı hakemler tarafından verilen puanlarıdır. Tek yönlü varyans analizi puan ortalamaları arasında fark bulunup bulunmadığını belirler. Gözlemcile-

---

[a] İstatistiksel analiz programında "tek ölçüm KİK" değerinden söz edilir. Bazı yazarlar bu kavramı "tek gözlemciye ait güvenilirlik değeri" olarak takdim etmişlerdir. Böyle olunca hesaplanan güvenilirlik değerinin birinci, ikinci ve üçüncü gözlemciden hangisine ait olduğu gibi bir soruyla karşılaşılır. Oysa burada, bilgisayar verileri yeniden düzenlemekte ve tek bir gözlemci verisine dönüştürmektedir. Hesaplamada gözlemcilerin ortalama puanları değil, ham birim puanları temel alındığından doğru ifadelendirme biçimi "tek ölçüm KİK değeridir". Eğer, *tek gözlemciye ait güvenilirlik değeri* ifadesi tercih edilmişse bunun herhangi bir gözlemciye ait olmadığı unutulmamalıdır.

[b] Tek yönlü varyans analizi / tek yönlü tesadüfî etki modelinde, "varyans" ölçülen kişiler veya nesneler arasındaki değişkenlikle ilgilidir. TYVA modelinde ölçüm yapılan kişilerin/nesnelerin puanına aynı zamanda "tesadüfî faktör" adı verilir. Kişileri değerlendiren gözlemciler, rasgele atanmış/belirlenmiş olduklarından ve sayılarının da birden fazla olmasından dolayı bu yöntemde puanların gözlemcilere göre değişip değişmediğine bakılmaz.

rin/hakemlerin rasgele atanmasının zorluğu nedeniyle yaygın kullanımı olmayan bu yöntemden zaman zaman tıp bilimlerinde klinik ölçümlerin güvenilirliği için yararlanılır.[41] Güvenilirlik katsayısı SPSS'te "Single Measure Intraclass Correlation" başlığı altında verilmiştir.

KİK (1, *k*). Rasgele seçilen kişiler, çok sayıda farklı gözlemciler tarafından değerlendirilir. Çok sayıda gözlemcinin yaptığı değerlendirmelerin ortalamaları alınarak hesaplama yapılır. Örneğin, personel seçim mülakatlarında ve testlerinde bu yönteme başvurulur. İstatistiksel analiz programı SPSS'te "Average Measure Intraclass Correlation" başlığı altında gözlemcilerin ortalama değerlendirme güvenilirlik katsayısı elde edilir.

KİK (2,1): *İki yönlü tesadüfî etki* modeline[a] dayanır. Bu modelde öncekinden farklı olarak kişilerin/nesnelerin tamamı hep aynı gözlemciler tarafından değerlendirilir. Gözlemciler daha büyük bir ana kütlenin temsilcileridir ve ana kütleden tesadüfî olarak seçilmişlerdir. Güvenilirlik katsayısı SPSS'te "Single Measure Intraclass Correlation" başlığı altında verilir.

KİK (2, *k*). Her bir kişi aynı gözlemciler tarafından değerlendirilir. Gözlemciler daha büyük bir ana kütlenin temsilcileridir. Güvenilirlik katsayısı, *k* sayıdaki gözlemcinin vermiş olduğu puanların ortalaması alınarak hesaplanır. Güvenilirlik katsayısı istatistiksel analiz programı SPSS'te "Average Measure Intraclass Correlation" başlığı altında verilir.

KİK (3,1): Bu yöntem *iki yönlü karma etki modeline* dayanır. Tüm kişiler aynı gözlemciler tarafından değerlendirilir. Bunun yanında, gözlemciler ana kütleyi oluşturan kişilerin tamamıdır. Diğer bir deyişle değerlendirme yapacak gözlemciler sadece o kişilerden ibarettir. Bu uygulamada diğer değerlendiricilere genelleme yapılmaz. Güvenilirlik katsayısı SPSS'te "Single Measure Intraclass Correlation" başlığı altında verilir.

---

[a] İki yönlü tesadüfî etki modelinde, (iki yönlü varyans analizi), belirli sayıda gözlemci vardır ve bu kişiler tüm nesneleri/kişileri değerlendirirler. Böyle olunca, puanların varyansı sadece kişiler arasındaki farklılıklardan değil aynı zamanda gözlemcilerin eğilimlerinden de etkilenir. *Gözlemciler* ve *kişiler* tesadüfî olarak seçildiklerinden bu yönteme iki yönlü tesadüfî etki modeli adı verilir. Eğer gözlemciler belirli bir ana kütleden tesadüfî olarak seçilmemiş, iradî olarak araştırmacı tarafından belirlenmişse "iki yönlü karma etki modeli (two-way mixed model) yöntemi uygulanır. Modelin "karma" olarak adlandırılmasının nedeni, kişilerin tesadüfî olarak gözlemcilerin ise sabit bir biçimde belirlenmesi sebebiyledir. İradî olarak belirlenen gözlemciler ise, bilgisayara "sabit faktör" olarak tanıtılır. İki-yönlü karma model sosyal araştırmalarda daha çok kullanılır.

KİK (3, *k*). Tüm kişiler aynı gözlemciler tarafından değerlendirilir. Gözlemciler ana kütleyi oluşturan kişilerin tamamıdır. Güvenilirlik katsayısı *k* sayıdaki gözlemcinin vermiş olduğu puanların ortalaması alınarak hesaplanır. Araştırmada dört gözlemci varsa, bu durum simgesel olarak KİK (3, 4) şeklinde gösterilir. Güvenilirlik katsayısı SPSS'te "Average Measure Intraclass Correlation" başlığı altında verilir.

Uygulamada, belirli bir uzmanlar grubunu temsil etmek üzere tesadüfî olarak seçilmeleri halinde gözlemciler arasındaki güvenilirliği belirlemek için KİK (2, 1) ve (2, *k*) modelleri kullanılır. Eğer gözlemciler iradî olarak belirlenmişse KİK (3, 1) ve (3, *k*) modellerine başvurulur.

Garson'a göre (2003), "grup içinde değişkenlik yoksa fakat grup ortalamaları farklı ise *küme içi korelasyon analizi* büyük ve pozitif bir değer verir. Grup ortalamaları aynı, fakat grup içinde büyük değişkenlik varsa sonuç negatif çıkar. Küme içi korelasyon analizinin maksimum değeri 1,0'dir."[42]

SPSS mönüleriyle hesaplama. Küme içi korelasyon analizi, istatistiksel analiz programı SPSS'te üç farklı mönüden hareket edilerek hesaplanabilir:

1. Birinci yöntemde, General Lineear Model mönüsü kullanılır.
2. İkinci yöntemde Compare Means mönüsü altındaki One-Way ANOVA düğmesi aracılığıyla gerekli hesaplama yapılır.
3. Üçüncü ve daha basit olan yöntem ise, Reliability mönüsünü kullanmaktır.

İlk iki hesaplama yöntemine ilişkin bilgiler bundan sonraki "Güvenilirlik ve Varyans Analizi" bölümünde ele alınmış, bu bölümde ise sadece Reliability mönüsündeki hesaplama yöntemi üzerinde durulmuştur. Reliability mönüsüyle hesaplama yapabilmek için veri tablosundaki sütunlara ya, birinci , ikinci ve üçüncü gözlemcinin değerleri (verdikleri puanlar) veya duruma göre t_1, t_2 ve t_3 zamanındaki ölçüm sonuçlar girilir. Değişkenler *k* sayıda olabilir. Hesaplama yapılabilmesi için en az iki dizi veriye sahip olunmalıdır.

Araştırmacı "tek yönlü tesadüfî etki modelinden" hareket ediyorsa hakemlerin / gözlemcilerin daha büyük bir ana kütleden tesadüfî olarak seçildiğini varsayıyor demektir. *Ancak hakemlerin / gözlemcilerin değişik sayıda olabilmesi nedeniyle verilerin varyansı üzerindeki etkisi araştırılmaz.* Veri matrisinin sütunlarında gözlemciler satırlarda ise vakalar / kişiler veya nesneler yer alır. Vakalar da *n* sayıda olabilir ve ana kütleden tesadüfî olarak

seçilmişlerdir. Veri hücrelerinde ise ölçüm değerleri veya eğer Likert ölçeği kullanılmışsa ölçeğin toplam/ortalama puanları gösterilir.

Küme içi korelasyon analizini yapmak için Reliability mönüsü altındaki gözlemcilerin yaptıkları değerlendirme puanları analiz penceresine alınır. Daha sonra Statistics mönüsüne girilir ve buradaki iki göz seçili hale getirilir. Bunlardan birincisi ANOVA Table başlığı altında None şıkkı ve ikincisi ise, Intraclass correlation coefficient seçeneğidir.

Yazılımın Statistics mönüsünde yer alan Model bölümünde ise, araştırma tasarımına göre üç seçenekten biri seçilir. Birinci model "One-Way ANOVA"dır. Bu yöntemde, örneğin bir araştırmaya 10 kişi katılmışsa ve bu 10 kişiye üç farklı gözlemci tarafından farklı puanlar verilmişse puanlar arasındaki tutarlılık araştırılır. Uygulama simgesel olarak (1,1) şeklinde gösterilir. Tek yönlü tesadüfî etki modelinde eğer (1, *k*) tasarımı tercih edilmişse SPSS'teki veri giriş tablosunda bir değişiklik yapılmaz. Yazılım aynı düzenleme biçiminden hareket ederek (1, *k*) tasarımının sonuçlarını "Average Measure Intraclass Correlation" başlığı altında verir. Bilim adamı böyle bir durumda güvenilirlik değerlendirmesi için (1, *k*) hesaplaması sonuçlarını esas alır.

İki yönlü tesadüfî etki modelinde (Two-Way Random), önceden belirlenmiş belirli sayıda gözlemci/hakem/teknisyen/değerlendirici vardır. Bu gözlemciler/hakemler daha büyük bir gözlemciler grubundan tesadüfî olarak seçilmişlerdir. Tek yönlü varyans analizinden farkı, puanlar üzerinde gözlemcilerin ve kişilerin birlikte etkili olabileceğidir. Her iki faktörün de tesadüfî etkileri söz konusudur (*bk.*, Tablo 6-6).

Tablo 6-6. İki Yönlü Varyans Analizi Modelinde Çapraz Kodlama

Kişiler / Hedef	Tesadüfî olarak belirlenen hakemler		
	1. hakem	2. hakem	3. hakem
Murat	12,00	18,00	16,00
Serkan	18,00	18,00	17,00
Ahmet	19,00	14,00	16,00
Nermin	20,00	15,00	17,00

İki yönlü karma etki modelinde ise (two-way mixed), yine belirli sayıda gözlemci/hakem/teknisyen/değerlendirici vardır. Ancak bu gözlemciler /

hakemler / değerlendiriciler daha büyük bir gözlemciler grubundan tesadüfî olarak seçilmemişlerdir (*bk.*, Tablo 6-7). Araştırmacı, kendisinin uygun görmesi nedeniyle veya erişebilirliğinin yüksek olması nedeniyle *X*, *Y* ve *Z* şahıslarını hakem / değerlendirici olarak seçmiştir. Bu kişiler araştırmaya katılan tüm deneklere puan verirler. Bilimsel araştırmalarda ve ölçümlerde hakemlerin/gözlemcilerin tesadüfî olarak atanmaları nadirdir. Bu kişiler daha çok iradî olarak belirlenirler ve bu nedenle böyle bir durumda *iki yönlü karma etki modeli* tercih edilir. İstatistiksel analiz programı SPSS'te ön tanımlı olarak birinci sırada *iki yönlü karma etki modeli*, ikinci sırada *iki yönlü tesadüfî etki modeli* ve üçüncü sırada ise *tek yönlü tesadüfî etki modeli* yer almıştır.

Tablo 6-7. İki Yönlü Karma Varyans Analizi Modeli (Two-Way Mixed ANOVA).

Kişiler / hedef	İradî olarak belirlenen hakemler		
	1. hakem	2. hakem	3. hakem
Murat	12,00	18,00	16,00
Serkan	18,00	18,00	17,00
Ahmet	19,00	14,00	16,00
Nermin	20,00	15,00	17,00

Yazılımın Statistics mönüsünde yer alan Type seçeneği, *iki yönlü modellerde* çalışır ve araştırmacının ölçüm yaparken daha çok hangi tür hata faktörünün etkisinde kaldığını belirlemek, bu hata faktörünün etkisini ortadan kaldırmak için kullanılır. Örneğin araştırmacı, "gözlemciler arasındaki değerlendirme hataları ile bir gözlemcinin farklı kişilere puan verirken yapabileceği sistematik hataları" denkleştirmek veya nötrleştirmek istediği zaman Absolute agreement şıkkını seçili hale getirir. Fakat ölçümün yanlılığa neden olabilecek sistematik hata içermediğini düşünüyorsa yazılımdaki ön tanımlı Consistency şıkkını olduğu gibi bırakır. Araştırmacı, one-way random modelini kullanmayı tercih etmişse sadece absolute agreement seçeneği söz konusudur.[43]

SPSS analiz çıktısı. Yapılan hesaplama sonunda küme içi korelasyon katsayıları ile birlikte alfa güvenilirlik katsayıları elde edilir. Eğer (1,1), (2,1) veya (3,1) modelleri kullanılmışsa çıktılardaki "tek ölçüm küme içi

korelasyon katsayısı" (Single Measure Intraclass Correlation) değeri dikkate alınır. Eğer (1, *k*), (2, *k*) veya (3, *k*) modelleri kullanılmışsa bu kez, ortalama küme içi korelasyon katsayısı (Average Measure Intraclass Correlation) dikkate alınarak yorum yapılır. Bu değer iki, üç veya duruma göre dört gözlemciye ait puan ortalamalarının güvenilirliğini gösterir. Güvenirlik esas olarak gruplar arasındaki varyansa bağlıdır.

■ Küme içi korelasyon analizi SPSS çıktısı.

Intraclass Correlation Coefficient
One-way random effect model: People Effect Random

Single Measure Intraclass Correlation = ,6171

95,00% C.I.: Lower = ,2606 Upper = ,8718
F= 5,8353 DF= (9, 20,0) Sig.= ,0005 (Test Value = ,000)

Average Measure Intraclass Correlation = ,8286

95,00% C.I.: Lower= ,5139 Upper = ,9533
F= 5,8353 DF= (9,20,0) Sig.= ,0005 (Test Value = ,000)

Örnekte de görüldüğü gibi Reliability mönüsündeki KİK'in kullanılması, diğer mönülerdeki hesaplamadan farklı olarak araştırmacıya tahminin güven aralığını hesaplama imkanı da sağlar.

Fleiss (1981) Cicchetti ve Sparrow'a (1981) göre, hesaplanan küme içi korelasyon katsayılarından $R < ,40$ oranı zayıf, $R = ,40$ ilâ ,59 arasındaki oranlar orta, $R = ,60$ ilâ ,74 arasındaki oranlar iyi ve $R > ,75$ den yüksek olan oranlar mükemmel olarak kabul edilir (aktaran Barlett, 1999).[44] Alfa katsayısına göre küme içi korelasyon katsayıları daha düşük çıkar. Bilimsel araştırmalarda, görece düşük katsayılar kabul edilebilir bulunurken uygulamalı testlerde bu oranın en az ,70 olması gerektiği belirtilmiştir.

Tablo 6-8. Gözlemci Puanlarında Güvenilirlik Analizleri

Ölçüm düzeyi	*İki gözlemci*	*İkiden fazla gözlemci*
Nominal	Uyuşma yüzdesi. Cohen Kappa.	Uyuşma yüzdesi Cohen Kappa
Sıralı	Spearman sıra korelasyonu Kendall uyuşma katsayısı.	Kendall *W* uyuşma katsayısı
Eşit aralıklı - Oranlı	Pearson korelasyon analizi	Küme içi korelasyon analizi Pearson korelasyon analizi

KORELASYON KATSAYILARINDAKİ ZAYIFLIĞIN DÜZELTİLMESİ

Hesaplanan korelasyon katsayıları değişik nedenlerle gerçeği tam olarak yansıtmıyor olabilir. Bilim adamı tek başına hesaplama sonucuna bakarak elde edilen korelasyon katsayısının "zayıf" olup olmadığını bilemez. Zayıf korelasyon katsayısı, ana kütledeki ilişkilerin olduğundan düşük veya olması gerekenden daha büyük tahmin edilmesine neden olur. Korelasyon katsayısının zayıflığından kuşkulanılacak haller şunlardır: (a) örneklem hacminin küçük olması, (b) ölçüm sırasında yapılan hatalar, (c) testteki maddelerin istatistiksel ölçüm güçlerinin düşük olması, (ç) ranj kısıtlaması.

Bilim adamları, yukarıda sıralanan haller olmasaydı ana kütlede gerçek güvenilirliğe sahip bir korelasyon katsayısının ne olacağını belirlemeye yönelik olarak "zayıflığı yenme" formülünü geliştirmişler ve bu formülle gerçek güvenilirliği en azından teorik olarak tahmin etmeye çalışmışlardır (*bk.*, Eşitlik 6-5).

■ Zayıflığı yenme formülü.

$$\rho_{\theta_x\theta_y} = \left(\frac{1}{\sqrt{\rho_{XX'}\,\rho_{YY'}}}\right)\rho_{XY} \quad . \qquad (6\text{-}5)$$

$\rho_{XX'}$ = Birinci ölçümün/ölçeğin güvenilirlik katsayısı.
$\rho_{YY'}$ = İkinci ölçümün/ölçeğin güvenilirlik katsayısı.

ρ_{XY} = Birinci ölçümle ikinci ölçüm arasındaki düzeltilmemiş korelasyon katsayısı.

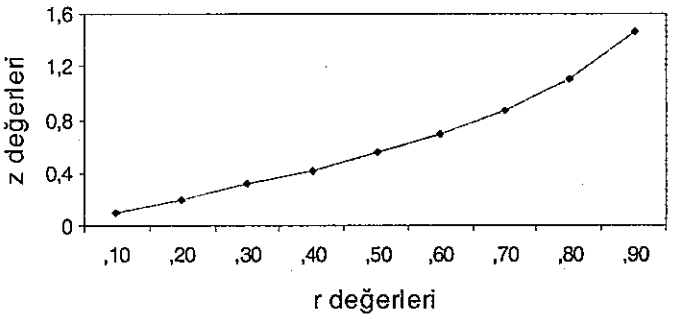
Paralel formlarda, testi yarıya bölme uygulamalarında ve test yeniden test uygulamalarında elde edilen korelasyon katsayısının ana kütleye genellenmek istenmesi halinde zayıflığı yenme formülünden yararlanılabilir. Bu formül SPSS'te bulunmaz. İstatistiksel analiz yazılımı STATISTICA'da "ana kütle korelasyon katsayısı" olarak tanımlanmıştır.

KORELASYON KATSAYILARININ FİSHER Z PUANLARINA DÖNÜŞTÜRÜLMESİ

Korelasyon analizine dayalı güvenilirlik katsayıları normal dağılım özelliği göstermez. Bu değerler, sıralı ölçek verisi niteliğindedir. R.A. Fisher (1915) bu katsayıların "eşit aralıklı" ölçek verisi niteliğinde, normal dağılım özelliğine sahip olmasını sağlayacak bir hesaplama yöntemi geliştirmiş ve elde edilen değerler "Fisher z' puanı" olarak adlandırılmıştır(*bk.*, Şekil 6-1). Normal dağılım özelliği gösteren Fisher z' değerlerinin "standart z puanlarıyla" karıştırılmaması gerekir. Her iki yöntemin hesaplama formülleri farklıdır. Bilim adamı korelasyon analizine dayalı güvenilirlik katsayılarını değişik amaçlarla Fisher z' puanlarına dönüştürmeyi düşünebilir. Bunlar aşağıdaki gibidir:

1. Güvenilirlik (korelasyon) katsayısının, hipotez testi ile ana kütle için ne ölçüde anlamlı olduğunu belirlemek ve sonuçta hesaplanan *örneklem güvenilirlik katsayısını* ana kütleye genellemek.
2. Aynı ölçek veya teste ait farklı güvenilirlik katsayılarını (cinsiyet, yaş, meslek gruplarına ait olabilir) karşılaştırmak.
3. Elde edilen tek bir güvenilirlik (korelasyon) katsayısının güven aralığını belirlemek.
4. Farklı güvenilirlik katsayılarını tek bir güvenilirlik katsayısı haline getirmek ve daha sonra bu güvenilirlik katsayısının güven aralığını belirlemek.
5. Ara değişkenlerin güvenilirlik katsayısını etkileyip etkilemediğini etki büyüklüğü ile saptamak.

6. Korelasyon katsayısı *r* değerinin *etki büyüklüğü* olarak kullanıldığı yerlerde, bu değeri Fisher *z´* puanlarına dönüştürerek ana kütle etki büyüklüğünü daha doğru bir şekilde ölçmek.



Şekil 6-1. Güvenilirlik katsayılarının Fisher *z´* dönüşümü.

Korelasyon (güvenilirlik) katsayılarının Fisher *z´* puanlarına dönüştürülmesi işlemi, test veya ölçek sonuçlarının daha sağlıklı karşılaştırılmasına imkan sağlar. Bilim adamları böylece güvenilirlik katsayılarını yansız ve nesnel bir biçimde değerlendirme olanağı elde ederler.

Korelasyon Katsayısının Anlamlılığını Belirleme

Araştırmacı, test-yeniden test veya paralel test uygulamalarından[a] elde ettiği korelasyon katsayısının; (a) ne ölçüde "yüksek güvenilirliğe" sahip olduğunu veya (b) elde edilen güvenilirlik katsayısının ne ölçüde ana kütleye genellenebileceğini belirlemek için *r* değerini *Fisher z´* puanlarına dönüştürebilir. Bunun için sıfır hipotezi, H_0: $\rho = 0$ ve alternatif hipotez ise H_1: $\rho \neq 0$ şeklinde saptanır. Sıfır hipotezinin anlamı, dönüştürülmüş verilerde ana kütle korelasyon katsayısının teorik olarak, "sıfıra eşit olacağı" veya güven aralığı değerlerinde 0 rakamını içereceğidir. Güven aralığı değerleri sıfır rakamını içeriyorsa anlamlı bir ilişki yoktur, test ve yeniden

[a] Bilim adamları yarıya bölme güvenilirliğinden elde edilen korelasyon katsayılarının SB formülü ile düzeltmesi yapıldıktan sonra elde edilen korelasyon katsayısının ayrıca Fisher *z´* puanlarına dönüştürülmesinin uygun olmayacağını belirtmişlerdir. Benzer şekilde, örneklemlere ait alfa katsayısı dağılımlarının moment çarpımları korelasyon katsayısı değerleriyle bir tutulamayacağı ifade edilmiştir.

test değerleri önemli ölçüde birbirinden farklıdır. Eğer ana kütleye genelleme yapılmak isteniyorsa, sıfır hipotezinin anlamı örneklem korelasyon katsayısının ana kütle korelasyon katsayısına eşit olduğudur. Araştırmacı esas olarak sıfır hipotezini reddederek alternatif hipotezin doğru olduğunu kanıtlamaya çalışır. Ana kütlede korelasyon katsayısı her halde sıfırdan farklı ve güven aralığında sıfırı içermeyen bir değer olmalıdır. Bunun için önce, r değeri Eşitlik 6-6 formülü ile z' puanlarına çevrilir.

$z' = 0{,}5\ [\ln (1 + r) - \ln (1 - r)]$. veya,

$z' = 0{,}5\ [\ln (1 + r) / (1 - r)]$. veya,

$$z' = 0{,}5 \log_e \left(\frac{1+r}{1-r}\right). \quad \text{Ana kütlede } \zeta_0 = 0{,}5 \log_e \left(\frac{1+\rho}{1-\rho}\right). \qquad (6\text{-}6)$$

Fisher z' değerleri Eşitlik 6-7'deki standart hata formülü ile normal dağılım özelline sahip olur. Bu eşitlik aynı zamanda, z' değerlerinin standart sapması olarak bilinir. Birden fazla ölçüme ait farklı örneklem büyüklükleri ve bunlara ait korelasyon katsayıları varsa iki veya daha fazla ölçümün standart hatası serbestlik derecesi, $(n - 3)$ değerlerinin toplamıyla gösterilir.

$$\sigma_{z'} = \frac{1}{\sqrt{n-3}}. \qquad \sigma_{z'} = \frac{1}{\sqrt{(n_1 - 3) + (n_2 - 3) \ldots + (n_k - 3)}}. \qquad (6\text{-}7)$$

Fisher z'puanlarının standart sapması belirlendikten sonra istenilen güvenilirlik düzeyinde *güven aralığı* hesaplanabilir.

■ Güven aralıklıları ve karşılık gelen standart z değerleri.

GA	%50	%60	%70	%75	%80	%90	%95	%99	%99,9
Değer	0,67	0,84	1,03	1,15	1,28	1,65	1,96	2,58	3,30

Yüzde 95 güvenilirlik düzeyinde güven aralığını hesaplamak için, Eşitlik 6-8'deki formül kullanılır.

$GA = z' \pm 1{,}96 \times$ Karekök (z' değerinin standart sapması).

$GA = z' \pm 1{,}96 \times$ Karekök $(1 / (n - 3))$. (6-8)

$$GA = z' \pm 1{,}96 \times \sqrt{\frac{1}{n-3}}.$$

Hesaplama sonucunda alt ve üst limitleri göstermek üzere iki değer elde edilir. Bu değerler korelasyon katsayısının ana kütlede hangi değerler arasında oynayacağını gösterir. Eğer elde edilen yüksek *GA* ve düşük *GA* değerleri 0 değerini içeriyorsa o zaman "%95 güven aralığında test-yeniden test uygulamaları arasında istatistiksel olarak anlamlı bir ilişki olmadığı" söylenir ve H_0 hipotezi kabul edilir. Sıfır hipotezinin kabul edilmesi birinci ölçümdeki verilerle ikinci ölçümdeki verilerin birbirinden önemli ölçüde farklı olduğu anlamındadır. Eğer 0 değerini içermiyorsa bu kez "%95 güven aralığında iki test uygulaması (test-yeniden test veya paralel formlar) arasında istatistiksel olarak anlamlı bir ilişki olduğu" ifade edilir ve H_0 hipotezi reddedilir. İkincisinde güven aralığı değerleri sıfırı içermeyecek şekilde daha dar bir erime sahiptir; alt ve üst ρ değerleri birbirine yakın çıkar. Böyle bir durumda korelasyon katsayısı yüksek güvenilirliğe sahip demektir. Bilim adamının amacı, H_0 hipotezinin reddedilmesini sağlayarak, test-yeniden test güvenilirlik katsayısının istatistiksel olarak anlamlı olduğunu vurgulamaktır. Bir başka şekilde yorumlamak gerekirse, ana kütlede test-yeniden test korelasyon değeri her hâlde sıfırdan farklıdır. Elde edilen korelasyon rakamı güven aralığının alt ve üst limitleri dahilinde ana kütleye genellenebilir. Ancak ana kütledeki gerçek korelasyon değerinin ne olduğunu tam olarak söylemek imkansızdır.

Güvenilirlik Katsayılarının Karşılaştırılması

Bilim adamı, eğer birden fazla örneklemde araştırma yapmışsa elde ettiği güvenilirlik katsayıları arasında önemli bir farklılık bulunup bulunmadığını yine Fisher z' dönüşümü ile test edebilir. Güvenilirlik katsayıları Cronbach alfa hesaplamasına[a], küme içi korelasyon katsayısına veya Pearson kore-

[a] Bazı bilim adamları Cronbach alfa için bu hesaplamanın uygun olmadığını düşünmektedirler.

lasyon katsayısına dayanıyor olabilir. Alfa değeri ve korelasyon katsayıları eşit aralıklı ölçek verisi değildir. Bu nedenle hesaplama sonucunda elde edilen r = ,10 ile r = ,20 arasındaki mesafe; bir başka araştırmada elde edilen r = ,80 ile r = ,90 arasındaki mesafeye eşit değildir. Fisher z' değeri, güvenilirlik katsayıları arasındaki büyüklüğün konumunu daha doğru bir şekilde değerlendirme imkanı sağlar.[45] Karşılaştırma işlemi Cronbach alfa değerleri arasında yapılıyorsa Fisher z' dönüşümünden sonra Behrens tarafından geliştirilen (1997, aktaran Yu) Q istatistiği veya Feldt, Woodruff, ve Salih (1987, aktaran Yu) tarafından geliştirilen UX testi uygulanır.[46] Tek bir güvenilirlik katsayısı test veya ölçeğin güvenilirliğini açıklamakta yetersiz kalacağından araştırmacı birden fazla güvenirlik katsayısı arasında karşılaştırmalar yaparak sonuçların birbirini teyit edip etmediğine bakmalıdır.

Fisher z' için, korelasyon katsayısı r (veya α) değerinin e tabanına göre logaritmik dönüşümü, önce belirtildiği gibi, $z' = 0{,}5[\ln (1 + r) / (1 - r)]$ formülü ile hesaplanır. İstatistiksel analiz programı SPSS'te bulunmayan bu hesaplama yöntemi sayesinde iki güvenilirlik katsayısı arasında anlamlı bir farklılık bulunup bulunmadığını test etmek mümkündür (*bk.*, Eşitlik 6-9).[47]

■ Örnek: Korelasyon katsayısı r değerinin, z' puanlarına dönüştürülmesi.

$$
\begin{aligned}
&r = ,45\ ,\\
&z' = ,5\ [\ln (1 + ,45) / (1 - ,45)]\ ,\\
&z' = ,5\ [\ln (1{,}45) / (\ ,55)]\ ,\\
&z' = ,5\ [\ln (2{,}636)]\ ,\\
&z' = ,5\ [,969]\ ,\\
&z' = ,484\ .
\end{aligned}
\qquad (6\text{-}9)
$$

Eşitlik 6-9'daki logaritmik değer, bilimsel hesap makineleriyle veya MS-Excell'de LN ( ) formülüyle hesaplanabilir. Daha kolay bir yöntem bu konuda özel olarak hazırlanmış bulunulan z' dönüşüm tablolarından yararlanmaktır (*bk.*, Ek-A, Fisher z' Dönüşüm Tablosu).

Korelasyona bağlı olarak elde edilen *güvenilirlik katsayıları* iki ölçüme veya daha fazla ölçüme ait olabilir. İki bağımsız örneklemden elde edilen

korelasyon (güvenilirlik) katsayılarının eşitliğini belirlemek için sıfır hipotezi aşağıdaki gibi belirlenir:

$H_0 : \rho_1 = \rho_2$ (Birinci çalışmadaki korelasyon katsayısı [rho], ikinci çalışmadaki korelasyon katsayısına eşittir.)

$H_1 : \rho_1 \neq \rho_2$ (Korelasyon katsayıları eşit değildir, farklıdır.)

Hesaplanan z' değeri ,05 anlamlılık düzeyinde 1,96'dan küçük ise H_0 hipotezi kabul edilir ve birinci çalışmadaki korelasyon katsayısının ikinci çalışmadaki korelasyon katsayısına eşit olduğuna karar verilir. Bilim adamı iki çalışmaya ait dönüştürülmüş puanlar arasındaki farkı, Eşitlik 6-10'daki formülle hesaplar.[48]

z'_f = (z'1 − z'2) / karekök (1/ (n_1 − 3) + 1/ (n_2 − 3)).

$$z'_f = \frac{z'_1 - z'_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} . \qquad (6\text{-}10)$$

z'_f = Dönüştürülmüş z'değerleri arasındaki fark.
z'_1 = Birinci ölçüme ait güvenilirlik katsayısının dönüştürülmüş değeri.
z'_2 = İkinci ölçüme ait güvenilirlik katsayısının dönüştürülmüş değeri.
n_1 = Birinci ölçümdeki örneklem büyüklüğü.
n_2 = İkinci ölçümdeki örneklem büyüklüğü.

■ Örnek: Bir araştırmada 100 kişide uygulanan test-yeniden test uygulamasında güvenilirlik katsayısı ,80 çıkmıştır. Altmış kişiden oluşan başka bir grupta yapılan ikinci bir test-yeniden test uygulamasında ise bu kez güvenilirlik rakamı ,70 çıkmıştır. İki güvenirlik katsayısı birbirlerinden önemli ölçüde farklı mıdır (*bk.*, Eşitlik 6-11).

$$z' = \frac{,5 \log_e \frac{1+,80}{1-,80} - ,5 \log_e \frac{1+,70}{1-,70}}{\sqrt{1/(100-3) + 1/(60-3)}} , \qquad (6\text{-}11)$$

$$z' = \frac{1{,}09 - 0{,}86}{0{,}16}, \qquad z' = 1{,}43 .$$

Sonuç olarak, Eşitlik 6-11'deki hesaplamalara dayalı olarak iki güvenilirlik katsayısının %95 güvenilirlik düzeyinde birbirinden önemli ölçüde farklı olmadığı söylenir ($z' = 1{,}43 < 1{,}96$).

Güvenilirlik Katsayısının Güven Aralığını Belirleme

Test ve ölçek uygulamaları sonucunda elde edilen güvenilirlik katsayısı örnek kütleye ilişkin "nokta tahminini" verir. Nokta tahmini, tek bir istatistiksel değerdir. Örneğin, test-yeniden test uygulaması sonucunda elde edilen r = ,85 değeri bir nokta tahminîdir. Fakat, nokta tahmini bir rakamın hangi değerler arasında oynayabileceğini gösteren aralık tahmini kadar güçlü olmadığından elde edilen güvenilirlik katsayısının aralık tahminini hesaplamak gerekir. Bunun için önce dönüştürülmüş z' puanlarının standart hatası, $\sigma z_r = 1 / \sqrt{(n-3)}$ formülü ile hesaplanır. Daha sonra %95 güven aralığında Fisher z' puanının alt ve üst güven aralığı $z' = z' \pm (z_{\alpha/2} / (\sigma z_r)$ formülü ile belirlenir.

■ Örnek: Dönüştürülmüş z' değeri 0,484 ve n = 80 olan bir ölçümün güven aralığı aşağıdaki gibidir.

Alt limit:

$z' = z - (z_{\alpha/2} / \sqrt{n-3})$,
$z' = 0{,}484 - (1{,}96 / \sqrt{80-3})$,
$z' = 0{,}484 - (1{,}96 / 8{,}71)$,
$z' = 0{,}484 - 0{,}225$,
$z' = 0{,}259$.

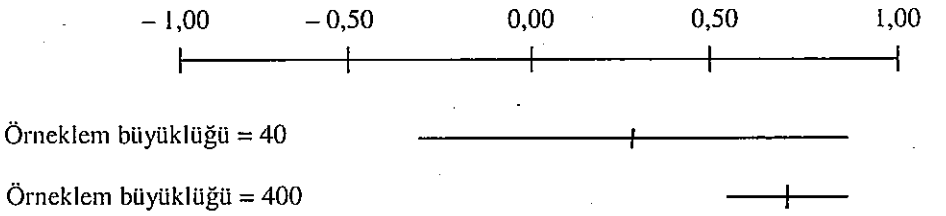
Üst limit:

$z' = z + (z_{\alpha/2} / \sqrt{n-3})$,
$z' = 0{,}484 + (1{,}96 / \sqrt{80-3})$,
$z' = 0{,}484 + (1{,}96 / 8{,}71)$,
$z' = 0{,}484 + 0{,}225$,
$z' = 0{,}709$.

$GA = 0{,}259 - 0{,}709$.

Güven aralığı, ölçüm yapılan kişilerin sayısına bağlıdır. Örneklem büyüklüğü (n) artıkça güven aralığının sınırları daralır ve daha kesin değerler elde edilir. Kırk kişilik bir örneklemden elde edilen güvenilirlik katsayısının güven aralığına göre 400 kişilik bir örneklemden elde edilen güvenilir-

lik katsayısının güven aralığı daha dardır ve kesine yakın bir sonuç verir (*bk.*, Şekil 6-2).



Şekil 6-2. Örneklem büyüklüğü ve Fisher *z'* puanlarının güven aralığı.

Katsayıların Birleştirilmesi

Araştırmacı birden fazla örneklemde test-yeniden test uygulaması yapmışsa veya birden fazla paralel form arasında ilişki aramışsa elinde aynı teste ait birden fazla korelasyon katsayısı olacaktır. Bu korelasyon katsayılarının ortalamasını alarak tek bir katsayıya dönüştürmek mümkündür. Ancak böyle bir uygulamada potansiyel üç hata faktörünün etkisi göz önünde bulundurulmalıdır:[49] (a) örneklem hatası, (b) ölçüm hatası ve (c) ara değişkenlerin etkisi. Test farklı büyüklükteki örneklemlerde uygulanmışsa istatistiksel olarak örneklem hatasına dikkat etmek gerekir. Küçük hacimli örneklemlerde örneklem hatası daha yüksektir. Öte yandan her bir ölçümün kendi koşulları ve tesadüfî ölçüm hataları söz konusudur. Bir uygulamada ölçüm hataları büyük ölçüde kontrol altına alınmışken diğerinde değişik faktörler nedeniyle hatalar yeterince kontrol altında tutulamamış olabilir. Ara değişkenler ise, testin uygulandığı gruplardaki değişikliklerle ilgilidir. Testin farklı yaş gruplarına, farklı bölgedeki kişilere, farklı okullardaki kişilere uygulanması ara değişkenlerin etkisini ortaya çıkarır. Korelasyon katsayıları birleştirilirken sözü edilen potansiyel hata kaynaklarının etkisi minimum düzeye düşürülmelidir.

Eğer elde edilen korelasyon katsayıları aynı ana kütleden çekilen eşit büyüklükteki örneklemlere ait ise, basit bir şekilde korelasyon katsayılarının ortalaması alınabilir. Korelasyon katsayıları farklı örneklem büyüklüklerine aitse bu rakamlar ya örnek kütle büyüklükleriyle çarpılarak toplam örneklem büyüklüğüne bölünmek suretiyle veya güvenilirlik katsayıla-

rı, Fisher z' puanlarına dönüştürülmek suretiyle birleştirilir (*bk.*, Tablo 6-9). Literatürde Fisher z´ puanlarına dönüştürme konusu tartışmalıdır. Kimi yazarlar dönüştürme işleminin gerekli olduğunu (Rosenthal 1991, Wolf, 1986, aktaran, Field, 2003)[50] kimi yazarlar da sonuçların birbirine oldukça yakın çıkıyor olması nedeniyle buna gerek olmadığını belirtmişlerdir (Lax, 1995, aktaran Lind).[51]

■ Örnek: Basit ağırlıklandırma yöntemi.

$$M_{\rho} = \frac{\Sigma nr}{\Sigma n} . \qquad (6\text{-}12)$$

M_{ρ} = Korelasyon katsayılarının ortalaması.
n = Örneklem büyüklüğü.

Tablo 6-9. Korelasyon Katsayısı r Değerlerinin Fisher z' Puanlarına Dönüştürülerek Birleştirilmesi

	Birinci çalışma	İkinci çalışma	Üçüncü çalışma	Dördüncü çalışma
n	90	80	120	45
r	,651	,745	,857	,684
$1 + r$	1,651	1,745	1,857	1,684
$1 - r$	,349	,255	,143	,316
$(1 + r) / (1-r)$	4,730	6,843	12,986	5,329
$Log_n (1 + r) / (1-r)$	1,553	1,923	2,563	1,673
$0,5\ Log_n (1 + r) / (1-r)$	,776	,961	1,280	,836

$$\bar{Z}_r = \frac{\Sigma Z_{r_i}}{n} = \frac{,776+,961+1,280+,836}{4} = ,963. \qquad (6\text{-}13)$$

Araştırmayı yapan bilim adamı, isterse z' değerlerini araştırmada kullanılan örneklem büyüklükleriyle çarparak ağırlıklandırabilir (*bk.*, Eşitlik 6-14).

$$\bar{Z}_r = \frac{\Sigma(N_i Z_{r_i})}{\Sigma n_i} = \frac{(90\times,77)+(80\times,96)+(120\times1,28)+(45\times,83)}{90+80+120+45} = \frac{337,94}{335} = 1,00. \qquad (6\text{-}14)$$

Meta Analizi ve Etki Büyüklüğü Tahmini

Gene V. Glass (1976) tarafından geliştirilen ve "analizlerin analizi" anlamına gelen *meta analizlerinde* aynı veya farklı bilim adamları tarafından yapılmış bir veya birden fazla çalışmadan elde edilen aritmetik ortalama değerleri, *t*-testi değerleri, ki-kare değerleri, yüzde oranı farklılık değerleri veya korelasyon katsayıları kullanılır. Meta analizlerinde, iki farklı değişken veya iki farklı çalışma arasındaki ilişkileri belirlemek için "etki büyüklüğü" (EB) olarak adlandırılan kavram ön plana çıkar. Etki büyüklüğü, test-yeniden test uygulamalarında, uzun süren sıralı ölçümlerde, paralel form uygulamalarında, iki gözlemci arasındaki değerlendirmelerde ve ayrıca cinsiyet, zeka, yaş, meslek gibi bağımsız faktörlerin test / ölçek puanları üzerindeki etkisini belirlemek amacıyla kullanılır. Etki büyüklüğü analizini yapmak isteyen bilim adamı; (a) iki değişkenden *birinin diğeri üzerinde etkili olabileceği,* (b) iki değişken arasındaki ilişkilerin *ara faktörlerin* etkisi nedeniyle yüksek / düşük çıkabileceği veya (c) iki grup ortalaması / oranı arasında fark bulunduğu varsayımlarından hareket eder. Etki büyüklüğü, kullanılan formüle göre *d, r,* r^2, *h,* Δ, *g,* r_λ ve z' gibi simgelerle

gösterilir.[a] Bunlardan *d* simgesi iki ortalama arasındaki farklılığı, *h* iki oran / yüzde arasındaki farklılığı, *r* korelasyona dayalı etki büyüklüğünü ve *z'* iki sürekli veri arasındaki ilişkilere dayalı standardize edilmiş etki büyüklüğünü ifade eder. İki değişken arasındaki ilişkiler "zeka-başarı notu" gibi, sebep-sonuç bağlantısı içinde ele alınıyorsa EB analizinde iki dizi veriden birincisi *bağımsız,* diğeri ise *bağımlı* değişkeni tanımlar. Bazen etki büyüklüğü, ara değişkenlerin etkisinden kaynaklanır.

Örneğin, test-yeniden test uygulamalarında aradan geçen süre içinde test maddelerinde kalibrasyon yapılmışsa birinci ölçüm sonuçları bağımsız ikinci ölçüm sonuçları ise bağımlı değişken olarak kabul edilir. Bu analizde eğer varsa, etki büyüklüğü bağımsız değişkenin kendisinden değil, "kalibrasyon ara faktöründen" kaynaklanmıştır. Yine, müdahale öncesi yapılan ölçümler bağımsız; müdahale sonrası yapılan ölçümler ise bağımlı değişkendir. Ölçümler, eğer deney ve kontrol grupları kapsamında yapılıyorsa kontrol grubu verileri bağımsız değişken, deney grubu verileri ise bağımlı değişken olarak adlandırılır. Bu sonuncusunda EB, deney ve kontrol gruplarının ortalama veya yüzde değerleri arasında fark bulunduğu anlamına gelir. Güvenilirlik analizlerinde etki büyüklünü hesaplama yöntemine başvurma amaçları aşağıdaki gibidir:

1. Geliştirilen ölçek veya testte yaş, zeka, cinsiyet gibi bağımsız faktörlerin etki büyüklüğünü saptamak.
2. Test-yeniden test uygulamalarında, pilot araştırma ve asıl araştırma sonuçları arasında eğer test maddelerinde hiçbir değişiklik, revizyon yapılmamışsa aradan geçen zamanın veya örneklem hacminin etkisini etki büyüklüğü ile ortaya koymak.

---

[a] *Etki büyüklüğü* kavramı aynı zamanda iki değişken arasındaki *ilişkililik derecesi, istatistiksel önem* veya *pratik önem* anlamlarında da kullanılmıştır. İki değişken arasında nedensellik ilişkisi araştırılmıyorsa etki büyüklüğü *ilişkililik derecesi* olarak yorumlanır. Etki büyüklüğü; deney ve kontrol grubu, test-yeniden test gibi farklı çalışmalara aitse elde edilen değer gruplar arasındaki veya *ölçümler arasındaki farklılık* veya *değişkenlik* olarak yorumlanır. *Etki* sözcüğü nedenselliğe işaret eder. İki değişken arasında bir şekilde bir nedensellik faktörü veya ilişkisi yoksa *etki büyüklüğü* ifadesini kullanmamak gerekir. Ancak literatürde nedenselliğe işaret edilmediği halde bu kavramın kullanıldığı görülmektedir. Bu kitapta bizim önerimiz, konu nedensellikle ilgili değilse *etki büyüklüğü* kavramının yanında parantez içinde *ilişkililik derecesi* gibi ek bir açıklamaya gidilmesidir. Böylece okuyucu konunun doğrudan nedensellikle ilgili olmadığı hakkında bir fikir sahibi olacaktır.

3. Test-yeniden test uygulamalarında, pilot araştırma ve asıl araştırma uygulamalarında iki zaman dilimi arasında test maddelerinde değişiklik veya revizyon yapılmışsa bu revizyonun etkisini etki büyüklüğü ile ortaya koymak.
4. Paralel form uygulaması sonucunda elde edilen korelasyon katsayısına bakarak formlar arasındaki ilişkililik derecesini ortaya koymak. (Bu uygulamada EB kavramının anlamı, "etki" değil, "ilişkililik derecesidir" Bu nedenle etki büyüklüğü kavramı yerine, ilişkililik derecesi kavramını kullanmak daha doğrudur.)
5. Bir test aynı ana kütleden seçilen bağımsız örneklemlere test-yeniden test şeklinde uygulandığında elde edilen korelasyon katsayıları arasındaki farkın büyük olup olmadığını belirlemek.
6. Bir test aynı ana kütleden seçilen bağımsız örneklemlere test-yeniden test şeklinde uygulandığında elde edilen güvenilirlik katsayılarını birleştirerek tek bir güvenilirlik katsayısı elde etmek.

Etki büyüklüğü değişik faktörlere dayalı olarak ortaya çıkar. Bilim adamı hangi faktörün etkisinden şüpheleniyorsa onu araştırır. Etki büyüklüğünü yaratan faktörler aşağıdaki gibi sıralanabilir:

1. Şans faktörüne dayalı olarak ortaya çıkabilir.
2. Yapılan farklı ölçümlerdeki uygulama değişiklikleri nedeniyle ortaya çıkabilir.
3. Yaş, zeka, cinsiyet, ırk gibi ara değişkenlerin etkisi nedeniyle ortaya çıkabilir.
4. Cevaplayıcılardaki gelişme ve öğrenme etkisi nedeniyle ortaya çıkabilir.
5. Müdahale etkisi nedeniyle ortaya çıkabilir.
6. Ölçek veya test maddelerindeki kalibrasyon, revizyon çalışmalarının sonucunda ortaya çıkabilir.
7. Zaman içinde tutumlardaki değişme veya becerilerin körelmesi nedeniyle ortaya çıkabilir.
8. Örneklem hacmi değişiklikleri nedeniyle ortaya çıkabilir.

Bilim adamı, etki büyüklüğü değerlerini üç amaçla kullanır; (a) yapılan farklı çalışmalarda bağımsız faktörün veya ara faktörlerin bağımlı değişkendeki etkisini ortaya koymak için, (b) değişik çalışmalarda ortaya çıkan güvenilirliğe ait ölçüm değeri farklılıklarının önemli olup olmadığını görmek için ve (c) farklı etki büyüklüklerini birleştirerek tek bir etki büyüklüğü rakamına ulaşmak için.

Etki büyüklüğünün hesaplanması. Literatürde etki büyüklüğü daha çok üç hesaplama biçimine göre belirlenir. Bunlar aritmetik ortalamalara, korelasyon katsayılarına ve oranlar arasındaki farklılığa dayanan hesaplamalardır.

Ortalamalara dayalı olarak yapılan hesaplama. Aritmetik ortalamalara dayalı olarak yapılan hesaplama türü, eğer ölçümde iki grup varsa ve bağımlı değişken eşit aralıklı veya oranlı ölçek verisi niteliğinde ise uygundur. Standardize edilmiş ortalamalar arasındaki farklılık değerinin dikkate alındığı bu hesaplamada etki büyüklüğü için üç farklı yöntemden yararlanılır. Bunlar Cohen *d*, Hedges *g* ve Glass delta (Δ) formülleridir. Bunlardan Cohen *d* formülü aşağıdaki gibidir.

d = [deney grubu ortalaması]–[kontrol grubu ortalaması]/standart sapma.

Coe'ye (2002) göre, hangisinin deney ve hangisinin kontrol grubu olduğunun bilinmediği durumlarda dahi EB yine hesaplanabilir, ancak böyle bir durumda sadece gruplar arasındaki farklılık ölçülmüş olur. Formüldeki standart sapma aslında ana kütleye ait standart sapma değeridir, fakat bunu bilmek tam olarak mümkün olmadığından ya kontrol grubunun *SS* değeri veya kontrol ve deney gruplarına ait *birleştirilmiş SS*[a] değeri esas alınır.[52] Cohen'e göre ise, grupların varyans değerleri birbirine eşitse deney ve kontrol gruplarından herhangi birisine ait standart sapma değerini kullanmakta bir sakınca yoktur (aktaran Becker, 2003).[53]

Etki büyüklüğü *d*, korelasyon katsayısı, *r* değerlerinden hareket edilerek de hesaplanabilir. Bunun için $d = 2r / \sqrt{(1 - r^2)}$ formülü uygulanır. Hedges

[a] Cohen'e (1988) göre, *birleştirilmiş standart sapma*, iki standart sapma ortalamasının kare köküdür ($\sigma_{birleştirilmiş} = \sqrt{[(\sigma_1^2 + \sigma_2^2) / 2]}$).

g ve t-testi gibi diğer etki büyüklüğü hesaplama yöntemleri ve formülleri için okuyuculara literatüre başvurmalarını öneririz.

Korelasyon değerlerine dayalı olarak hesaplama. Ölçüm çalışması, denkleştirilmiş gruplar arasında yapılıyorsa veya çalışma tekrarlanmış ölçümler şeklinde ise bu kez standardize edilmiş farklılığı ortaya çıkarmak için *korelasyon katsayıları* temel alınır.[54] Etki büyüklüğünün korelasyon katsayılarına bağlı olarak hesaplanması birkaç şekilde olur. Her iki ölçüm verisi de sürekli veri niteliğinde ise Pearson yöntemi uygulanır.[55] Her iki değişken ikili veri yapısına sahipse *phi*, değişkenlerden biri ikili diğeri sürekli veri niteliğinde ise nokta-iki serili r_{pb}, her iki veri büyüklük sırasına sokulmuşsa *rho*, veriler standardize edilmişse Fisher z' yöntemi uygulanır. Etki büyüklüğü; korelasyon katsayılarının *birleştirilmesi*, iki korelasyon katsayısı arasındaki *farklılığının* ortaya konması veya bir *değişkenin diğeri üzerindeki etkisi* şeklinde belirlenebilir.

Etki büyüklüğünün korelasyon katsayıların birleştirilmesi yöntemiyle belirlenmesi. Daha önce Fisher z' yönteminde korelasyon katsayılarının birleştirilmesi yöntemine değinilmiş, ancak konu "etki büyüklüğü" nosyonu çerçevesinde ele alınmamıştı. Etki büyüklüğü nosyonu, analizde ara değişkenlerin etkisinin göz önünde bulundurulmasını gerektirtir. Meta analizinde etki büyüklüğü ölçüsü olarak korelasyon (güvenilirlik) katsayısı r değerleri temel alınmışsa birden fazla ölçüme dayanan ve örneklem büyüklükleri dikkate alınarak belirlenen r_b (birleşik korelasyon katsayısı) ana kütleye ilişkin daha yansız bir güvenilirlik tahmini yapmaya imkan sağlar.[56] Bunun için korelasyon (güvenilirlik) katsayıları basit bir şekilde toplanarak toplam katsayı sayısına bölünür (*bk.*, Eşitlik 6-15).

$$r_b = \frac{\sum r_i}{n} . \qquad (6\text{-}15)$$

Bilim adamı, r_b değerinin ayrıca karesini alarak (r_b^2) "etki eden faktörün ilişkili olduğu diğer faktörde yarattığı değişkenliği" saptayabilir. *Belirlilik katsayısı* olarak da isimlendirilen r_b^2 değeri[a] örneğin, "zekanın başarı

[a] Belirlilik katsayısı, bazı kaynaklarda büyük harf olarak yazılıp R^2 şeklinde diğer bazı kaynaklarda da küçük yazılıp r^2 olarak gösterilmiştir. Bu kitapta küçük yazılış biçimi tercih edilmiştir. Yine bu kitapta büyük *re* harfi (R) *küme içi korelasyon analizi* yöntemiyle hesaplanan güvenilirlik katsayısını gösterir.

notlarını %49 oranında açıkladığı" şeklinde yorumlanır (r =,70 olması halinde). Korelasyon katsayıları, test-yeniden test puanlarına dayanıyorsa, r_b *birleşik güvenilirlik derecesi* olarak isimlendirilir. Cohen'e (1988) göre, birleşik güvenilirlik katsayısı 0,10 ise küçük etki, 0,25 ise orta etki ve yaklaşık 0,40 ise büyük etki olarak yorumlanır.

Birleşik etki büyüklüğünü hesaplamanın ikinci yöntemi ağırlıklı ortalama formülünden yararlanmaktır. Bu uygulamada her bir korelasyon sayısı örneklem büyüklüğü ile çarpılır ve toplam örneklem büyüklüğüne bölünür (*bk.*, Eşitlik 6-16).

$$r_b = \frac{\sum (n_i \cdot r_i)}{\sum n_i} . \qquad (6\text{-}16)$$

Ağırlık verilerek hesaplanan birleşik etki büyüklüğü değeri, eğer küçük örneklemler ve büyük örneklemlerle birlikte hesaplanırsa büyük örneklemden daha fazla etkilenerek büyük çıkma ihtimaline sahiptir.[57]

Araştırmacı etki büyüklüğü konusunda hesaplama yaparken aynı zamanda "düzeltilmemiş" ve "düzeltilmiş" etki büyüklüklerinden hangisini kullanacağına da karar vermelidir. Örnek kütleden elde edilen korelasyon veya güvenilirlik katsayıları *örneklem hata varyansı* nedeniyle büyük ölçüde şişkin çıkar. Daha sağlıklı bir değerlendirme için bu katsayıların standardize edilmesi gerekir. Literatürde standardizasyon işlemi için "düzeltilmiş R^2," Hays ω^2, Fisher z' ve Herzberg R^2 teknikleri önerilmiştir. Etki büyüklüğü ölçüsü olarak d temel alınmışsa düzeltme işlemi için Hedges düzeltme formülü kullanılır. Etki büyüklüğü ölçüsü eğer r ise, bu kez düzeltme işlemi için Fisher z' değeri kullanılır.[58] Bu kitapta sadece Fisher z' üzerinde durmayı uygun gördük.[a] Etki büyüklüğünü tanımlayan r değerleri normal dağılım özelliğine sahip olmadığından Fisher z' bu puanları normalleştirir, puanların normal dağılıma sahip olmasını sağlar. Korelasyon veya güvenilirlik katsayılarının Fisher z' değerlerine dönüştürülmesinden sonra elde edilen rakamların basit bir şekilde ortalaması alınarak veya ağır-

[a] Fisher z' değerlerini SPSS'te hesaplatmak için Transform mönüsü altında Compute düğmesiyle açılan pencereye 0,5 (ln ((1 + var1) / (1 – var1))) formül tanımlaması yapılır. Veri çizelgesine ise, var1 değişkeni olarak korelasyon katsayıları veya güvenilirlik katsayıları girilir.

lık verilerek *birleşik Fisher z´ değerleri* elde edilir. Bazı bilim adamları ağırlıksız ve ağırlıklı Fisher z´değerlerinin her ikisini de hesaplayarak aradaki farkı görmek isterler. Ağırlıklı ve ağırlıksız Fisher z´değerleri bir tür korelasyon katsayısı olmadığından, yorum yapmak için uygun değildir. Bu nedenle ters dönüşüm formülüyle tekrar *yansız r* değerlerine dönüştürülür. Bu kez araştırmacının elinde ağırlıklı ve ağırlıksız olmak üzere yansız iki farklı *r* değeri bulunur: Birincisi ağırlıksız ve yansız r_y ikincisi ise, ağırlıklı ve yansız $r_{\lambda y}$ değeri. Güven aralığı Fisher z´değerlerine göre hesaplanabilir, ancak bu değerler korelasyon katsayısı olmadığından yorum yapma güçlüğüyle karşılaşılır. Bu sorunu aşmak için ya Fisher z´değerlerinin güven aralıkları ayrı ayrı r_y değerlerine dönüştürülür veya dönüştürülmüş *r* değerlerinin güven aralığı yeniden hesaplanır. Güven aralığı konusunda yorum yapabilmek için dönüştürülmüş, yansız *r* değerlerinin temel alınması gerekir. Güven aralıkları, elde edilen değerlerin sıfırdan anlamlı ölçüde farklı olup olmadığını gösterir.

Ters dönüştürülmüş r değerlerinin hesaplanması. Normalleştirilmiş z´ puanları yorum yapmak için uygun olmadığından Eşitlik 6-17'deki ters dönüşüm formülü kullanılarak bu değerler yansız r_y puanlarına dönüştürülür ve etki büyüklüğü r_y puanları cinsinden yorumlanır. Ters dönüştürülmüş *güvenilirlik katsayıları* "yansız etki büyüklüğü tahmin değeri" olarak adlandırılır ve r_{td} veya r_y simgesiyle gösterilir.[a]

$$r_{td} = (\exp(2z') - 1) / (\exp(2z') + 1).\ \text{alt limit}$$
$$r_{td} = (\exp(2z') - 1) / (\exp(2z') + 1).\ \text{üst limit} \qquad (6\text{-}17)$$

$$r = \frac{e^{2z'} - 1}{e^{2z'} + 1}. \qquad (6\text{-}18)$$

exp = 2,718 doğal logaritmanın temeli olan e sabiti.
r_{td} = Ters dönüşüm *r* değeri.
z´ = Fisher z´değeri.

[a] Fisher z´ değerlerini r_y değerlerine SPSS ortamında dönüştürmek için Compute düğmesiyle açılan pencereye ((EXP (z´/ 0.5)) – 1) / (1 + EXP (z´/ 0.5)) formül tanımlaması yapılır. Veri çizelgesine ise, z´ değişkeni olarak Fisher z´ değerleri girilir.

■ Örnek: Bir ölçümde z' değerinin alt limit değeri ,32 ve üst limit değeri ,73 bulunmuş olsun. Hesaplanan Fisher z' *güven aralığı* değerlerinin %95 güven aralığında r_y (r_{td}) cinsinden geriye dönüştürülmüş güven aralığını bulunuz (*bk.*, Eşitlik 6-19, Eşitlik 6-20, Eşitlik 6-21, Eşitlik 6-22, Eşitlik 6-23).

r_{td} = (exp (2*,32) –1) / (exp (2*,32) +1). alt limit
r_{td} = (exp (2*,73) –1) / (exp (2*,73) +1). üst limit (6-19)

r_{td} = (exp (,64) –1) / (exp (,64) +1).
r_{td} = (exp (1,46) –1) / (exp (1,46) +1). (6-20)

r_{td} = (1,89 –1) / (1,89 +1).
r_{td} = (4,30 –1) / (4,30 +1). (6-21)

r_{td} = ,89 / 2,89.
r_{td} = 3,30 / 5,30. (6-22)

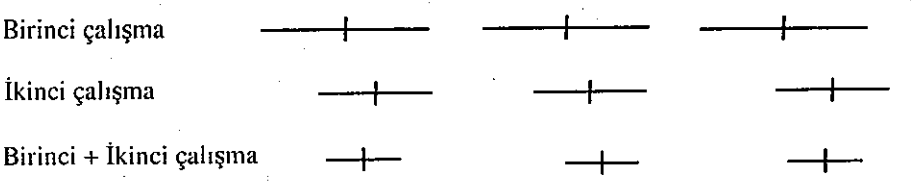
r_{td} = ,30 – ,62. (6-23)

Etki büyüklüğünü inceleme modelleri. Etki büyüklüğü değerleri birleştirilip tek bir etki büyüklüğü rakamına ulaşılmak istenirken iki modelden yararlanılır: sabit etki modeli ve tesadüfî etki modeli. Sabit etki modelinde, araştırmacının kontrolü altında tutulan müdahale türü, araştırma çevresi, araştırma grubu veya bireysel özellikler gibi faktörlerin bağımlı değişkende bir etki büyüklüğü yaratmayacağı varsayımından hareket edilir (*bk.*, Şekil 6-3). Farklı çalışmalardaki etki büyüklüğünün sadece örnekleme hatasından kaynaklandığı düşünülür. Sabit etki büyüklüğü, arka plandaki "gerçek" etki büyüklüğünden etkilenir ve tüm çalışmalardaki ortak etkiyi temsil eder. Buna göre hesaplanan *d* veya *r* etki büyüklükleri sabit faktör olarak ele alınır. Tesadüfî etki modelinde ise, yapılan çalışmadan çalışmaya, *ana kütle etki büyüklüğünün* değişkenlik göstereceği varsayılır. Değişik çalışmaların etki büyüklüğü, süper bir ana kütleye ait *muhtemel etkiler evreninden* alınmış sadece bir örneklem değeridir.[59] Bu iki yöntem arasındaki temel farklılık *birleşik etki* büyüklüğüne ait standart hatanın hesaplanmasında ortaya çıkar.

Sabit etki modelinde, hata terimi olarak sadece *çalışma içindeki değişkenlik* (varyans) dikkate alınırken tesadüfî etki modelinde *çalışma içindeki değişkenliğin* yanında *çalışmalar arasındaki değişkenlik* de göz önünde bulundurulur. Bu kitabın amacıyla doğrudan ilgili olmaması nedeniyle sabit etki ve tesadüfî etki modellerine ilişkin hesaplama yöntemlerine ve kullanılan formüllere ilişkin bilgiler üzerinde ayrıntılı olarak durulmamıştır. Bu konuda daha fazla bilgi edinmek isteyen okuyuculara ilgili literatüre başvurmaları önerilir.

Diğer etki büyüklüklerinin korelasyon katsayılarına dönüştürülmesi. Bilim adamı isterse *d* cinsinden hesaplanan etki büyüklüğünü *r* değerlerine dönüştürmek için $r_\lambda = d / \sqrt{(d^2 + 4)}$ formülünü kullanabilir. Bir diğer hesaplama yöntemi ise, *t*-testi sonucunda elde edilen *t* değerlerinden yararlanmaktır. *t*-Testi sonucunda elde edilen *t* değerinin *r* değerine dönüştürülmesi için $r_\lambda = \sqrt{[t^2 / (t^2 + df)]}$ formülü kullanılır. Etki büyüklüğü *r* değerinin *d* değerine göre daha avantajlı olduğu belirtilmiştir. Çünkü *d* değeri *r* değerine çevrildiğinde bağımsız değişkenin bağımlı değişken üzerindeki etkisi görülmekte veya iki değişken arasındaki ilişkiler belirli bir anlama sahip olmaktadır. Ancak *r* değeri *d* değerine çevrildiğinde bilgi kaybı ortaya çıkmaktadır. Ayrıca *r* değeri ikiden fazla grup için analiz yapmaya uygun iken, *d* sadece iki grupla sınırlı kalmaktadır.[60]

Oranlara veya yüzde değerlerine dayalı olarak yapılan hesaplama. Oranlar (örneğin ,65 gibi) veya yüzde değerleri (örneğin %72 gibi) arasındaki farklılığın ölçümü ($p_1 - p_2$) bağımlı ve bağımsız değişkenlerin her ikisi de ikili veri yapısına sahipse uygundur. İki grubun varyansları eşit olduğunda verilerin normal dağıldığı varsayılır.



Şekil 6-3. Meta analizi sonucunda Fisher z' ve ters dönüştürülmüş r_y puanlarıyla güvenilirlik katsayısının ve bu katsayıya ait güven aralığının yeniden saptanması.

Normal dağılma özelliğinden şüphelenildiği veya bu koşul gerçekleşmediği zaman ölçek verileri yeni ölçek verilerine dönüştürülür. Bunun için üç farklı yöntem vardır: arcsin, logit ve probit yaklaşımları. Bunlardan sık kullanılan *arcsin* (ters sinüs fonksiyonu) yönteminin formülü şu şekildedir: arcsin = (2 arcsin $\sqrt{}$ p) – (2 arcsin $\sqrt{}$ p). Cohen (1977) tarafından geliştirilen ters sinüs fonksiyonu literatürde aynı zamanda *h* istatistiği olarak tanınır. Hesaplanan *h* değerleri daha sonra çubuk grafiği üzerine aktarılarak etki büyüklüğünü belirlemek üzere yüzde farklılıklarını gösteren *h-profili* elde edilir. Logit formülü ise şöyledir: *logit* = ln [p1/(1 – p1)] – ln[p2/(1 – p2)]. Formülde *ln* simgesi doğal logaritmayı gösterir. Probit yöntemi, yığışımlı olasılık için normal dağılım değerini hesaplamayı gerektirir. Bu yaklaşımların içinde arcsin yöntemi en az radikal olanıdır. Probit yöntemi çok daha köklü bir dönüşüm sağlarken logit yöntemi ılımlı bir dönüşüm imkanı sağlar. Bu konuda ayrıntılı bilgi edinmek isteyen okuyuculara D.A. Kenny'nin (1987) *Statistics for the Social and Behavioral Sciences* adlı kitabına başvurmaları önerilir.[61]

Etki büyüklüğünün yorumlanması. Etki büyüklüğü katsayılarının yorumlanmasını Jacop Cohen (1977, 1988, 1992) bir anlamda standart hale getirmiştir. Yorumlarda *d*, *r*, r^2 veya *h* değerleri temel alınabilir. Katsayıların büyüklüğüne ilişkin yorumlar Tablo 6-10'daki gibi yapılır.[62]

Tablo 6-10. Yansız Etki Büyüklüğü Katsayılarının Yorumlanması

d değerleri	r_y değerleri	r^2_y değerleri	Yorumu
> ,80	> ,37	> ,14	Büyük etki/farklılık
,50 – ,80	,24 – ,37	,06 – ,14	Orta derecede etki/farklılık
,20 – ,50	,10 – ,24	,01 – ,06	Küçük etki/farklılık
< ,20	< ,10	< ,01	Önemsiz etki/farklılık

Etki büyüklüğü, standart normal dağılımdaki *z*-puanlarının eşitidir ve bu nedenle *z*-puanları gibi yorumlanır. Örneğin, d = ,60 deney grubundaki bir kişinin ortalama puanının kontrol grubundaki bir kişinin ortalama puanından ,60 standart sapma kadar daha yüksek olduğu anlamına gelir.[63] Eğer d yerine eşiti r değerleri temel alınmışsa, r = ,29 her iki grupta da "başarılı olabileceklerin" oranı ,29'dur şeklinde yorumlanır. Etki büyüklüğünün d değerleriyle ,20'den küçük olması (müdahalenin etkililiğinin incelendiği bir araştırmada) farklılığın önemli olmadığını gösterir. Diğer bir deyişle müdahalenin niteliği ile ortaya çıkan etki büyüklüğü arasındaki ilişki zayıftır. Müdahale, istenen etkiyi tam olarak sağlayamamıştır. Etki büyüklüğünün 0 çıkması deney grubunun ortalamasının kontrol grubunun 50'nci yüzdelik diliminden başladığı anlamına gelir.[64] Etki büyüklüğü değerleri pozitif veya negatif işaretli olabilir. Pozitif işaret *iyileşmeyi* negatif işaret ise *kötüleşmeyi* gösterir. Pilot araştırma-asıl araştırma uygulamasında pozitif işaret güvenilirliğin yükselmiş olduğu biçiminde yorumlanır. Hangi rakamın pozitif değer sayılacağına araştırmacının / bilim adamının kendisi karar verir. Araştırmacı eğer büyük rakamın pozitif olduğunu düşünüyorsa ve bu rakam da kontrol grubunda ortaya çıkmışsa etki büyüklüğü negatif işaretli çıkar. Büyük rakam, bağımlı değişkende veya deney grubunda ortaya çıkmışsa bu kez etki büyüklüğünün işareti pozitiftir. Öte yandan bilim adamı bazen etki büyüklüğünün yüksek ve bazen de etki büyüklüğünün düşük çıkmasını ister. Örneğin, herhangi bir müdahale yapılmadan bir ölçeğin farklı gruplarda test-yeniden test uygulamasıyla sınanmasında etki büyüklüğünün küçük çıkması arzulanırken; pilot araştırma-asıl araştırma uygulamalarında pozitif işaretli büyük EB değerlerine ulaşılmaya çalışılır. Etki büyüklüğü incelenirken eldeki probleme özgü yorumlar yapılmalıdır. Bunun için literatürde benzeri sorunlar için hesaplanan etki büyüklüğünün değerlendirilme kriterlerine bakmak gerekir.[65] Aynı

konuda yapılmış bir araştırma bulunamamışsa benzeri alanlardaki diğer kriterler incelenir, büyük ve küçük değerlerden hangisinin temel alınacağı saptanır.

Etki büyüklüğü türdeşlik (homojenlik) analizi. Etki büyüklüklerinin türdeşlik (eşitlik) analizi, hesaplanan ortalama veya birleşik korelasyon katsayısının ana kütledeki etki büyüklüğünü tahmin etmede kullanılıp kullanılamayacağını belirlemeye yöneliktir. Bilim adamı bunun için sabit etki modelinde "yapılan farklı çalışmalardaki etki büyüklüklerinin birbirine eşit olduğu" sıfır hipotezinden hareket eder. Eşitlik testi ile, etki büyüklüklerinin çalışmadan çalışmaya önemli ölçüde değişip değişmediği gözlemlenir. Etki büyüklükleri eşitse elde edilen değer aynı zamanda ana kütle etki büyüklüğünü gösterir.

H_0: Bütün etki büyüklükleri eşittir ($d_1 = d_2 = d_3$).
H_1: Bütün etki büyüklükleri eşit değildir ($d_1 \neq d_2 \neq d_3$).

Etki büyüklüğü eşitlik analizi kullanılan formüle göre değişik şekillerde yapılabilir. Eğer ortalamalara dayanan d yöntemi kullanılmışsa Hedges, r değerleri temel alınmışsa Fisher z' yöntemi tercih edilir. Oranlara dayalı ölçümlerde ise daha önce sözü edilen arcsin, logit ve probit hesaplama formüllerinden yararlanılır.

Aritmetik ortalamalara dayanan etki büyüklüklerinin türdeş (eşit) olup olmadığını belirlemek için Mantel-Haenszel Q değerinden yararlanılır. Okuyucu bu hesaplama biçimi için literatüre başvurmalıdır. Bu kitapta güvenilirlik katsayılarıyla ilişkisi olması nedeniyle korelasyon katsayılarına dayanan etki büyüklüğü türdeşlik analizi yöntemi ele alınmıştır. Güvenilirlik katsayılarına dayalı olarak Q değerini hesaplamak için, dönüştürülmüş r (veya z') ile dönüştürülmüş r aritmetik ortalama değerinin (veya $\overline{z}'$) arasındaki farkların kareleri alınır. Bu değerler, her bir çalışmaya ait korelasyon katsayısının varyansı ile $(n - 3)$ çarpılarak toplamı alınır ve Q değeri elde edilir. Daha sonra bu değer ki-kare tablo değeriyle karşılaştırılarak güvenilirlik katsayılarının eşit olup olmadığına karar verilir. Fisher z'puanlarının aritmetik ortalaması Eşitlik 6-23'deki formülle ile hesaplanır.

$$\bar{z} = \frac{\sum_{i=1}^{k} (n_i - 3)\, z_i}{\sum (n_i - 3)} . \qquad (6\text{-}23)$$

Bu değer, Eşitlik 6-24'deki formülde yerine konarak Q değeri hesaplanır. Hesaplanan Q değeri ,05 anlamlılık düzeyi ve belirlenen serbestlik derecesinde tablodaki ki-kare değerinden yüksekse farklı çalışmalardan elde edilen güvenilirlik katsayılarının birbirine eşit veya türdeş olmadığına karar verilir.

$$Q = \sum_{i=1}^{k} (n_i - 3)(z'_{r_i} - \bar{z}_r)^2 . \qquad (6\text{-}24)$$

Ki-kare testi için sıfır hipotezi, "ana kütle etki büyüklüğü sıfıra eşittir" veya "etki büyüklükleri arasında fark yoktur" şeklinde belirlenir. Test sonucu p değerine bakılarak da yorumlanabilir. Olasılık değeri eğer anlamlı çıkmışsa bazı etki büyüklüklerinin tutarsızlık gösterdiği ve bu nedenle ortalama etki büyüklüğünün genel etki büyüklüğünü temsil etmediği kararına varılır.[66] Birden fazla korelasyon katsayısının türdeşlik analizini yapmak için Tablo 6-11'deki adımlar atılır.

Tablo 6-11. Fisher z' Değerlerinin Türdeşlik Analizi

Çalışma	r	n	z'	$(n-3)\,z'$	$(n-3)$	$\bar{z}'$	$(z'-\bar{z}')^2$	$(n-3)\,(z'-\bar{z}')^2$
1	0,38	60	,40	22,8	57	,64	,0576	3,28
2	0,62	70	,50	33,5	67	,64	,0196	1,31
3	0,79	80	1,07	82,39	77	,64	,1849	14,23
4	0,43	60	,45	25,65	57	,64	,0361	2,05
Toplam		270	2,42	164,34	258			$Q = 20,87$

Tablo 6-11'de Q değeri 20,87 çıkmıştır. Bu değer ,05 anlamlılık düzeyinde ve 3 serbestlik derecesinde ($ss = k - 1$) tablodaki ki-kare değerinden yüksek çıkması nedeniyle istatistiksel olarak anlamlıdır. Sonuçta etki büyüklüğünü belirleyen güvenilirlik katsayılarının eşit olmadığına karar verilir. Söz konusu güvenilirlik katsayıları ana kütlenin tamamına genellenemez ve bu nedenle tek bir rakam halinde birleştirilmemelidir.

ALINTI YAPILAN KAYNAKLAR

[1] A.J. Broderick, "Testing for Metric Equilance Using Confirmotary Factor Analysis: A Consumer Involvement Study [Teyit Edici Faktör Analizinin Metrik Eşitliğini Test Etme]," <http://research.abs.aston.ac.uk/wpaper/9903.pdf.> (28.01.2003).

[2] D. Grisaffe, "Summary Considerations Supporting 5-Points Rating Scales [Beş Dereceli Ölçekleri Destekleyen Kanıtlar Üzerine Düşünceler]," <http://www.walkerinfo. com/docs/scaling_paper_grisaffe.pdf.> (28.01.2003).

[3] K.G. Jöreskog, "Analysis of Ordinal Variables: Cross-Sectional Data]," <http://www.ssicentral.com/lisrel/ord2.pdf.> (02.02.2003).

[4] D.J. Young, "Characteristics of Effective Rural Schools: A Longitudinal Study of Western Australian Rural High School Students [Kırsal Kesimde Etkili Okulların Özellikleri]," 1998, <http://www.nexus.edu.au/TeachStud/arera/research/Dyoung/DYOUNG.htm> (29.03.2003).

[5] N.J. Birkett, "Selecting the Number of Response Categories for a Likert-Tyle Scale [Likert Tipi Ölçeklerde Yanıt Sayısının Belirlenmesi]," <http://www.amstat.org/sections/SRMS/proceedings/papers/1986_091.pdf> (16.03.2003).

[6] University of Teesside, "Experimental Method [Görgül Yöntem]," <http://sss-studnet.tees.ac.uk/psychology/ modules/year1/research_methods2/Lect2.doc> (02.02.2003).

[7] R.A. Yaffee, "Common Corelation and Reliability Analysis with SPSS for Windows [SPSS İle Korelasyon ve Güvenilirlik Analizleri]," 06 Arl 2000, <http://www.nyu.edu/its/socsci/Docs/correlate.html> (26.01.2003).

[8] Kimi yazarlar bu gibi durumlarda Spearman korelasyon analizinin uygulanmasına karşı çıkmışlar ve daha güçlü olması nedeniyle yine de Pearson korelasyon analizinin kullanılmasını önermişlerdir. Bu konuda *bk.*, D.M. Roberts,"A Note on the Use of the Spearman Rank Order Correlation [Spearman Sıra Korelasyonunun Kullanılması Üzerine Bazı Düşünceler]," <http://roberts.ed.psu.edu/users/droberts/papers/spearman.PDF> (26.01.2003), 13.

[9] Blackwell Science, "Kendall's Tau [Kendall Tau]," *Journal of Clinical Nursing*, 10, 707-715.

[10] L. Crocker ve J. Algina, *Introduction to Classical and Modern Test Theory*, (Chicago: Holt, Reinhart ve Winston, 1986), Aktarılan kaynak, "Phi Coefficient [Phi Katsayısı]," 2003, <http://www.gcd.clemson.edu/Main808/Notes808/PhiCoeff.htm> (15.02.2003).

[11] E.N. Nelson ve E.E. Nelson "Computation of Measures of Association [İlişki Ölçülerinin Hesaplanması]," 15 Ağs 1998, <http://www.csubak.edu/ssric/Modules/COWI/COWIMod/D.htm> (09.02.2003).

[12] IRT Modeling Lab, "Investigating Unidimensionality for Dichotomous Data [İkili Verilerde Tek Boyutluluğun Araştırılması]," <http://work.psych.uiuc.edu/irt/dim_dich1.asp> (14.07.2003).

[13] IRT Modeling Lab, "Investigating."

[14] "Correlations [Korelasyonlar]," <http://www.dal.ca/~houlihan/psy3500/LecturePDF/Correlation_2_8_per_page.pdf> (01.01.2003).

[15] Michigan State University, "Item Analysis [Madde Analizi]," <http://www.msu.edu/dept/soweb/itanhand.html> (03.02.2003).

[16] Johnson, "Meeting 6 [Toplantı 6]," <http://www.ed.sc.edu/edpyrmfn/johnson/meetn06.htm> (04.02.2003).

[17] Fair Test Examiner, "Racial Bias Built into Tests [Testlerde Irk Yanlılığı]," <http://www.fairtest.org/examarts/winter00/Racial_Bias_Built_into_Tests.html> (04.02.2003).

[18] D. Lehmkuhl, "Nonparametric Statistics [Parametrik Olmayan İstatistikler]," <http://www.oandp.org/jpo/library/1996_03_105.asp> (23.09.2002).

[19] University of New Brunswick, "Quantative Methods For Health Research [Sağlık Araştırmalırında Sayısal Yöntemler]," <http://www.unb.ca/courses/mhodgins/n6051/module9b_2002.htm> (23.09.2002).

[20] R.A. Yaffee, "Common Correlation and Reliability Analysis With Spss For Windows [SPSS'te Korelasyon ve Güvenilirlik Analizleri]," 1999, <http://www.nyu.edu/its/socsci/Docs/correlate.html>(23.09.2002).

[21] Symnet, "Statistics Homepage [İstatistik Ana Sayfası]," <http://www.symynet.com/educational_software/teaching_resources/Statistics/alternative_correlations/intro.htm>

[22] P. Barrett, "A Choise Between Enhanced Test Technology [Genişletilmiş Test Teknolojisi Arasında Tercih]," <http://www.liv.ac.uk/~pbarrett/psychom1.pdf.> (28.01.2003).

[23] P. Barett, "Skewness and Correlations [Çarpıklık ve Korelasyonlar]," <http://www.pbarrett.net/statistics_corner.htm> (24.01.2004).

[24] "Reliability [Güvenilirlik]," <http://www.geolog.com/msmnt/mrelobj.htm> (05.01.2003).

[25] Annette M. Green, "Kappa Statistics for Multiple Raters Using Categorical Classifications [Kategorik Sınıflandırmalarda Çoklu Gözlemciler İçin Kappa İstatistiğinin Kullanılması]," <http://www2.sas.com/proceedings/sugi22/POSTERS/PAPER241.PDF> (26.06.2003).

[26] R.A Yaffee, "Common Correlation and Reliability Analysis with SPSS for Windows [SPSS'te Korelasyon ve Güvenilirlik Analizleri]," 1999, <http://www.nyu.edu/its/socsci/Docs/correlate.htm> (30.03.2003).

[27] "Interrater Reliability [Değerlendiriciler Arası Güvenilirlik]," <http://www.uchsc.edu/gcrc/qr4.doc> (26.06.2003).

[28] ACITS, "What is a Good Kappa Coefficient? [İyi Kappa Katsayısı Nedir?]," <http://www.utexas.edu/cc/faqs/stat/general/gen27.html> (26.06.2003).

[29] M. Rudner, "Reliability of Measurement: Kappa [Ölçümlerin Güvenilirliği: Kappa]," 2003, <http://ericae.net/~rudner/educ637/pkappa.htm>

[30] "Alternative Correlation Techniques [Alternatif Korelasyon Teknikleri]," <http://people.brandeis.edu/~vortex/psyc210a/Chapter10.doc> (27.06.2003).

[31] Statistica, "Kendall Tau," <http://www.rrz.uni-hamburg.de/RRZ/Software/Statistica/Handbuch/glosi.html> (19.02.2003).

[32] J. Packard, "Observer Reliability [Gözlemci Güvenilirliği]," <canis.tamu.edu/VFSCcourses/WFSC620/exercises/ExerciG.htm> (24.02.2001).

[33] B. Trochim, "Selecting Statiscis [İstatistik Seçimi]," t.y., <http://trochim.human.cornell.edu/selstat/p8_7.htm> (24.02.2001).

[34] Statistica, "Kendall Tau."

[35] H. Smith, "Statistical Measures [İstatistiksel Ölçümler]," <http://home.wlu.edu/~journalism/J203/statistic.htm> (24.02.2001).

[36] D.C. Howell, "Intraclass Correlation [Küme İçi Korelasyon Analizi]," <http://www.uvm.edu/~dhowell/StatPages/More_Stuff/icc/icc.html> (27.11.2002).

[37] D. Garson, "Reliability [Güvenilirlik]," <http://www2.chass.ncsu.edu/garson/pa765/reliab.htm.> (29.06.2003).

[38] R.A. Yaffee, "Enhancement of Reliability Analysis [Güvenilirlik Analizinin Genişletilmesi]," <http://www.nyu.edu/its/socsci/Docs/intracls.html.> (29.06.2003).

[39] "Intraclass Correlations Coefficients [Küme İçi Korelasyon Katsayıları]," <http://www.powmri.unsw.edu.au/FBRG/The%20last%20word%20on%20ICCs.doc> (27.11.2002).

[40] SPSS, "Technical Support – Statistical Macro Library [Teknik Destek – İstatistiksel Otokod Kütüphanesi]," <http://www.spss.com/tech/stat/macros/lccsf.htm.> (29.06.2003).

[41] "Intraclass Correlations Coefficients [Küme İçi Korelasyon Katsayısı]," t.y., <http://www.powmri.unsw.edu.au/FBRG/The%20last%20word%20on%20ICCs.doc> (08.03.2003).

[42] D. Garson, "Correlation [Korelasyon]," t.y., <http://www2.chass.ncsu.edu/garson/pa765/correl.htm.> (24.02.2001).

[43] "Intraclass Correlations," (08.03.2003).

[44] P. Barlett, "Reliability [Güvenilirlik]," 1999, <http://www.liv.ac.uk/~pbarrett/rater.pdf> (27.11.2002).

[45] F. Gieles, "An Explanation of the Statistics Used in The Meta-Analysis [Meta Analizde Kullanılan İstatistik Yöntemlerin Açıklanması]," <http://www.tegenwicht.org/13_rbt_eng/gieles_explanation.htm#Fisher's-Z> (08.03.2003).

[46] C.H. Yu, "An Introduction to Computing and Interpreting Cronbach Coefficient Alpha in SAS [Cronbach Alfa Katsayısının SAS'ta Hesaplanması ve Yorumlanmasına Giriş]," <http://www2.sas.com/proceedings/sugi26/p246-26.pdf.> (23.03.2003).

[47] R. A. Yaffee, "Common Correlation and Reliability Analysis with SPSS for Windows [SPSS ile Güvenilirlik ve Korelasyon Analizleri]," http://www.nyu.edu/its/socsci/Docs/correlate.html> (08.03.2003).

[48] IFA Services, "Two Correlation Coefficients [İki Korelasyon Katsayısı]," <http://fonsg3.let.uva.nl/Service/Statistics.html> (08.03.2003).

[49] Hunter, "Combining Correlation Coefficients [Korelasyon Katsayılarının Birleştirilmesi]," <http://149.170.199.144/rd/metacorr.htm> (14.10.2002).

[50] A. Field, "A Bluffers Guide to Meta Analysis: Correlations [Meta Analiz İçin Basit Bir Rehber: Korelasyonlar]," <http://www.cogs.susx.ac.uk/users/andyf/research/articles/meta1.pdf> (09.03.2003).

[51] G. Lind, "Formulas For Converting Various Test Statistics into Effect Size Coefficient r (Correlation) [Çeşitli Test İstatistiklerini Etki Büyüklüğü Katsayısına Dönüştürme]," <http://classroom.psy.utexas.edu/HonorsHandouts/ZHandouts/old/comp-based/class12_2/Lind-2002_Formeln%20fuer%20Effe.pdf> (15.03.2003).

[52] R. Coe, "It's the Effect Size, Stupid [Etki Büyüklüğü]," <http://www.leeds.ac.uk/educol/documents/00002182.htm> (03.05.2003).

[53] L. Becker, "Effect Size [Etki Büyüklüğü]," <http://web.uccs.edu/lbecker/Psy590/es.htm#I> (03.05.2003).

[54] B. Thompson, "A Suggested Revision to the Forthcoming 5th Edition of the APA *Publication Manual*[APA Yayın Elkitabının 5. Baskısındaki Düzeltme Önerisi]," 29 May 2000 <http://www.coe.tamu.edu/~bthompson/apaeffec.htm> (02.05.2003).

[55] D.A. Kenny, "Meta Analysis [Meta Analizi], <http://users.rcn.com/dakenny/meta.doc> (04.05.2003).

[56] F. Gieles, "An Explanation of the Statistics Used in the Meta-analysis [Meta Analizlerinde Kullanılan İstatistiklerin Açıklaması], " <http://www.tegenwicht.org/13_rbt_eng/gieles_explanation.htm> (23.03.2003).

[57] Field, "A Bluffers Guide to."

[58] Etki büyüklüğü hesaplamaları için *bk.*, W.G. Hopkins, "On The Fly For Differences Between Means [Ortalamalar Arasındaki Farklılıklar Üzerine]," <http://www.sportsci.org/resource/stats/ssmean.html> ayrıca *bk.*, <http://ericae.net/meta/chap9/chap9.htm> (16.03.2003).

[59] A.P. Field, "Meta-Analysis of Correlations [Korelasyonların Meta Analizi]," *Psychological Methods*, 6 (2), 161-180.

[60] R. Rosenthal ve M R. Dimatteo, "Meta Analysis [Meta Analizi], <http://www.findarticles.com/cf_0/m0961/2001_Annual/73232703/p1/article.jhtml?term=> (09.05.2003).

[61] Kenny, "Meta Analysis."

[62] National Center for Education Statistics, "Statisticel Procedunes [İstatistiksel Prosedürler]," <http://nces.ed.gov/das/epubs/2002209/method_sp3.asp> (28.04.2003).

[63] Coe, "It's the Effect."

[64] Becker, "Effect Size."

[65] Chong-ho Yu, "Meta-analysis and effect size [Meta Analizi ve Etki Büyüklüğü]," <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/es.html> (27.04.2003).

[66] H. Pike, A. Hills ve R. MacLennan, "Personality and Military Leadership [Kişilik ve Askeri Liderlik]," <http://www.rmc.ca/academic/conference/iuscanada/papers/MacLennan_personalitypaper.pdf> (27.04.2003).

VARYANS ANALİZİ VE GÜVENİLİRLİK

Güvenilirliği saptamaya yönelik olarak sık başvurulabilecek istatistikî çözümleme yöntemlerinden bir diğeri varyans analizidir. Bilim adamı, test sonuçlarının güvenilirliğini belirlemek için değişik *varyans analizi* tekniklerinden yararlanabilir. Her bir teknik, bilim adamının tercih ettiği araştırma veya deney tasarımı göz önünde bulundurularak saptanır. Bu nedenle bilim adamı öncelikle araştırma/deney tasarımına uygun varyans analizi yönteminin hangisi olduğunu araştırmalıdır. Varyans analizi, ölçüm sonuçlarında potansiyel olarak değişkenlik yaratma ihtimali bulunan faktörleri belirlemeye hizmet eder. Araştırmacı bu teknikle, eğer farklı gruplarda ölçüm yapmışsa grup sonuçlarının tutarlılık gösterip göstermeğini veya aynı grupta birden fazla ölçüm yapmışsa kişilerin davranışlarının istikrarlı olup olmadığını belirler. Örneğin, psikometrik ölçümlerde kişilere birden fazla test uygulanması halinde t_1, t_2, t_3 zamanında yapılan test sonuçlarının istikrarlılığının saptanması *varyans analiziyle* veya *küme içi korelasyon analiziyle* yapılabilir. Varyans analizi, ölçüm sonuçları arasında önemli sayılabilecek sistematik hata bulunma durumunu araştırmaya yöneliktir. Bu bölümde değişik ölçüm tasarımlarında uygulanabilecek varyans analizi yöntemleri üzerinde durulmuştur.

KULLANIM AMAÇLARI

Varyans analizi, sınıflandırma değişkeni şeklinde belirlenen bir faktörün tepki veya yanıt değişkeni üzerindeki etkisini belirlemeye yöneliktir. Sınıflandırma değişkeni bağımsız, tepki değişkeni ise çoğunlukla bağımlı değişken olarak belirlenir. Seçilen modele göre bağımsız değişken bir, iki veya ikiden fazla olabilir. Aynı şekilde bağımlı değişken sayısı da birden fazla olabilir. Bilim adamları, parametrik nitelikteki ölçek/test genel sonuçları üzerinde çoğunlukla cinsiyet, yaş, kıdem, sınıf, zaman ve sektör gibi bağımsız faktörlerin ne gibi bir etki doğurduğunu görmek isterler. Eğer cinsiyet faktörü ölçüm sonuçları üzerinde herhangi bir etkiye sahip değilse, test sonuçları kadınlar ve erkeklerin her ikisi için de geçerlidir. Kadınlar ve erkekler için ayrı norm değerleri oluşturmaya gerek yoktur. Fakat tam tersine cinsiyet faktörü veya yaş faktörü sonuçlar üzerinde etkili ise o zaman kadınlar ve erkekler için veya değişik yaş

kili ise o zaman kadınlar ve erkekler için veya değişik yaş grupları için ayrı norm değerleri oluşturmak gerekir. Çünkü erkeklerin ölçüm ortalamaları kadınların ölçüm ortalamalarından farklıdır. Erkeklerin ölçüm ortalamaları kadınlar için güvenilir bir şekilde kullanılamaz. Bunun yanında bilim adamı ölçüm sonuçlarını değerlendirirken "yaş ve cinsiyet" veya "eğitim düzeyi ve yaş" gibi bağımsız faktörleri ikili gruplar halinde ele alarak bu grupların etkisini birlikte de görmek isteyebilir. Güvenilirlik konusuyla ilgili olarak varyans analizinden aşağıdaki amaçlarla yararlanabilir.

1. Birbirinden bağımsız, birden fazla grupta yapılan ölçümlere ait toplam veya ortalama puanlarının tutarlılığını saptamak. (Birden fazla grup, tek bir bağımsız değişkenin düzeyleri şeklindedir.)

2. Birden fazla kişi üzerinde sınanan test maddelerinin iç tutarlılık güvenilirliğini saptamak. (Kişi sayısı bağımsız değişken olarak belirlenir ve bağımlı değişkenler ise her bir kişinin örneğin, 10 maddeli bir testten aldığı puanlardır.)

3. Testlerde yapılan kalibrasyon çalışmalarının istenen sonucu verip vermediğini görmek ve ölçüm sonuçlarını karşılaştırarak incelemek. (Bağımsız değişken "kalibrasyon" olarak belirlenir. Veriler; 1= kalibrasyon öncesi ve 2= kalibrasyon sonrası şeklinde kodlanır. Bağımlı değişken ise testlerin toplam veya ortalama puanlarıdır.)

4. Aynı grupta birden fazla tekrarlanan (test-yeniden test şeklinde) ölçümlere ait toplam puanların tutarlılığını veya güvenilirliğini saptamak. (Bağımsız değişken ölçüm sayısı olarak kodlanır. Bağımlı değişken ise testlerin toplam veya ortalama puanlarıdır.)

5. Birden fazla değerlendiricinin/gözlemcinin puan toplamları/ortalamaları arasında önemli bir farklılık bulunup bulunmadığını saptamak. (Bağımsız değişken "gözlemci no" olarak kodlanır.)

6. Değerlendiriciler, katılımcılar, zaman, koşullar ve uygulama yöntemi gibi değişik faktörlerin/yüzeylerin ölçüm sonuçları üzerindeki etkisini ortaya koymak. (Varyans bileşenleri analizi yöntemi uygulanır.)

7. Bir grupta yapılan ölçüm sonunda test puanlarını yarıya bölerek puanların iki yarısı arasında tutarlılık olup olmadığını belirlemek. Bu tür uygulamalardan sonra çıkan güvenilirlik katsayısının Spearman-Brown formülüyle düzeltilmesi gerekir. (Bağımsız değişken "grup no" olarak kodlanır.)

Görüldüğü gibi varyans analizi, testlerin esas olarak iç tutarlılığını değil, gruplar arasındaki istikrarlılığını, test sonuçlarını etkileyen yüzeylerdeki değişkenliği ve dereceleme / puanlama çalışmalarında ise, gözlemciler arasında tutarlılık bulunma durumunu tespit eder.

VARSAYIMLARI

Varyans analizinin uygulanabilmesi için, diğer testlerde olduğu gibi belirli ön koşulların sağlanması gerekir. Bu ön koşullar varyans analizinin türüne göre değişebileceğinden bilim adamının bu konuda bir ön araştırma yapmasında yarar vardır. Bu varsayımların en önemlileri *normallik* ve *varyansların türdeşliği* konusuyla ilgilidir. Bununla birlikte, söz konusu varsayımları "olmazsa olmaz" gibi kesin bir kural olarak da değerlendirilmemek gerekir. Bu varsayımların gerekleri belirli ölçüde karşılanamasa veya verilerin dağılımı asimetrik bir özelliğe sahip olsa bile eğer;

1. karşılaştırılan örneklem gruplarının büyüklükleri birbirine eşitse veya,
2. normallikten sapma önemli ölçüde büyük değilse

varyans analizi yine de güçlü bir teknik olarak değerlendirilir. Dağılımın sivri veya basık olması *normallik* koşulunu büyük ölçüde etkilemez. Varyans analizi, bağımsız ölçümlere veya bağımlı ölçümlere dayalı olarak gerçekleştirilir. *Bağımsız ölçümler varyans analizinin* uygulamasına ilişkin ön koşullar veya varsayımlar dört tanedir. Bağımlı, tekrarlanan ölçümlerde ise bunlara iki yeni ön koşul / ön kabul daha eklenir: *bileşik simetri* özelliği ve *küresellik* özelliği. Aşağıdaki paragraflarda varyans analizi yapmadan önce söz konusu ön koşulların sağlanmasıyla ilgili bilgiler üzerinde durulmuştur.

Normallik. Varyans analizinde, *t*-testinde olduğu gibi test sonuçlarını gösteren bağımlı değişkene ait verilerin (toplam puan veya aritmetik ortalama) yaklaşık olarak normal dağılım özelliğine sahip olması gerekir. Bu nedenle varyans analizi test ve ölçeklerin maddeleri için değil, toplam veya ortalama puanları için uygulanır. Örneklem gruplarının her birinde verilerin dağılımı simetrikse varyans analizi ön koşul ihlallerine karşı güçlüdür. Söz konusu güçlü olma hali, örneklem hacmi büyüdükçe daha da artar. Eğer örneklem hacmi küçükse ve verilerin dağılımı yatık/basık bir görünüme sahipse test güç kaybeder. Tam tersine örneklem hacmi küçük fakat, veriler sivri bir dağılıma sahipse test güç kazanır.[1] Yaygın kullanılan tek yönlü varyans analizlerinde (TYVA) normal dağılımdan hafif derecedeki bir sapmanın elde edilen sonuçları önemli ölçüde etkilemediği

bulunmuştur (Kirk, 1982, aktaran Helberg).[2] Ancak bu bulgu veya değerlendirme bilim adamını normallik testi yapmaktan alıkoymamalıdır. Önemli ölçüde sağa veya sola çarpık veri yapılarında varyans analizi yöntemini uygulamak doğru değildir. Varyans analizinin çarpıklıklara karşı *güçlü* olması, normallik varsayımının göz ardı edilebileceği anlamına gelmez.

Verilerde büyük ölçüde çarpıklık saptanmışsa, örneklemin genişletilmesi yöntemine başvurulur veya uygun olduğu durumda verilerin logaritma, karekök ve ters logaritma yöntemleriyle standart puanlara dönüştürülmesi düşünülür. Normallik analizi SPSS'te Explore mönüsü altında Normality plots seçeneği ve Kolmogorov - Smirnov - Shapiro-Wilks testleri ile yapılır.

Varyansların türdeşliği. Varyans analiziyle ilgili ikinci ön koşul, ölçüm yapılan gruplarda varyans değerlerinin türdeş/eşit olmasıdır. Muhtemel ölçüm çiftlerinin varyansları birbirine eşitse varyansların türdeşliği sağlanmıştır. Aslında grup varyanslarının eşit olması, temsil ettikleri ana kütle varyanslarının da eşit olduğu anlamına gelir. Her bir ölçüm grubundaki ham değerlerin standart sapmalarının (varyanslarının) birbirine benzer çıkması veri dizilerinin türdeş olduğunu gösterir. İki, üç veya üçten fazla gruba ait muhtemel bütün çiftlerin varyans ve kovaryans değerlerinin eşitliğini tanımlamak üzere, bu olguya aynı zamanda *türdeşsellik* veya *sabit varyans* (homoscedasticity) adı verilmiştir. Varyans analizi, grup değerlerinin varyansları yaklaşık olarak birbirine eşit olduğu zaman kullanılabilir. Grupların varyansları birbirine eşit değilse bu duruma "ayrısallık" veya değişken varyans (heterosedasticity) adı verilir. Ölçüm gruplarına ait varyansların eşit olup olmadığını saptamak için değişik hesaplama yöntemlerine başvurulur. Bunlardan sık kullanılan testler aşağıdaki gibidir:

(a) Levene testi, (b) Bartlett testi, (c) F_{max} testi.

İstatistiksel analiz programı SPSS'te Levene testi Explore veya GLM mönüsüyle hesaplanabilir. Varyansların türdeş olup olmadığını grafik üzerinde görmek isteyen araştırmacılar Boxplots grafiğini kullanabilirler. Grafik yönteminde, Boxplots mönüsünden hareket edilerek Plots düğmesindeki *Boxplots: Factor levels together* seçeneği işaretli hale getirilir. Grafikteki bıyıkların uzunluğu ve kutunun medyana göre konumu, verilerin dağılım biçimi hakkında okuyucuya fikir verir. Dağılımda büyük ölçüde dengesizlik varsa gruplar arasındaki değişkenliğin anlamlı olduğuna karar verilir.

Levene testi, Explore mönüsünde Plots düğmesi altında *Spread vs Level with Levene Test* başlığında yer alan *None, Power estimation, Transformed, Untransformed* seçeneklerinden biri işaretli hale getirilerek hesaplatılır.[3] None, şıkkı türdeşsellik testini ön tanımlı olarak iptal eder. "Güçlü tahmin" yöntemi (power estimation), kartiller arasındaki dağılımı gösteren doğal logaritma grafiğini çi-

zer. Bu grafik, varyansları daha homojen hale getirerek daha güçlü bir tahmin yapılmasına olanak verir. Dönüştürülmüş (transformed) yöntemde, veriler grafik çizilmeden önce dönüştürülür. Bu yöntemi kullanmayı düşünen bilim adamları Power kutusunun yanındaki dönüştürme şıklarından birini seçmeleri gerekir. Dönüştürülmemiş (untransformed) yöntemde ham veriler kullanılarak kartiller arasındaki dağılımın grafiği verilir. Bu tahmin yönteminde "güç" şıkkı yoktur.

Varyansların eşitliği analizi, ikinci bir yöntem olarak Options düğmesi altında Statistics kartında Homogeneity Variances şıkkı seçilerek hesaplatılabilir. Levene istatistiği anlamlılık değeri ,05'ten küçük ise iki varyansın önemli ölçüde birbirinden farklı; ,05'ten büyük ise varyansların eşit olduğuna karar verilir. Bilim adamının amacı anlamlılık değerinin ,05'ten büyük çıkmasını sağlamaktır. Levene testi aşırı ölçüde duyarlı bir test olarak değerlendirildiğinden veriler eğer normal dağılım özelliği göstermiyorsa Levene testinin duyarlılığını azaltmak için alfa değerinin ,001'e düşürülmesi önerilmiştir (Tabachnick ve Fidell, 1996, aktaran Triggs ve Moss).[4]

Ölçüm yapılan örneklem büyüklükleri yaklaşık olarak birbirlerine eşitse varyans analizi türdeşsellik koşulunun sağlanamamasına karşı oldukça güçlüdür. Ancak Nagy' göre (2003), belirli koşullarda bu test istikrarsız sonuçlar verir. Bu koşullardan birincisi, örneklem büyüklükleri eşit değilse ve büyük varyans değerleri büyük örneklem gruplarıyla ilgili ise böyle bir durumda test sonucu tutucu çıkar. Diğer bir deyişle Tip I hatası yapma olasılığı belirlenen α düzeyinden düşüktür. İkincisi, örneklem büyüklükleri yine eşit değilse fakat, büyük varyans değerleri küçük örneklem gruplarıyla ilgili ise böyle bir durumda Tip I hatası yapma olasılığı belirlenen α düzeyinden yüksektir.[5]

Bartlett testini kullanmak isteyen araştırmacılar, SPSS'te Analyze, Data reduction, Factor ve Descriptives mönüleriyle bu hesaplamayı yaptırabilirler. Analiz, verilerin çok değişkenli normal dağılım özelliğine sahip olma durumunu ortaya koyar. Bartlett test sonucu ,001'den büyük çıkmışsa birim matrisinden farklı olarak değişkenler arasında ilişki olduğuna karar verilir. Bartlett ve Levene testlerinden hangisinin kullanılması gerektiği konusunda tereddüt eden araştırmacılara, daha güçlü olması ve SPSS gibi yazılımlarda ön tanımlı olarak seçilmiş olması nedeniyle Levene testinin kullanılması önerilmiştir.[6]

Veri yapısı. Varyans analizinde, "bağımlı değişken" eşit aralıklı veya oranlı ölçek verisi niteliğinde olmalıdır. Analiz, eğer test maddelerinin iç tutarlılığını belirlemek için yapılıyorsa Likert ölçeği gibi tutum ölçeklerinde derecelerin eşit aralıklı olduğu varsayılır. Özellikle 7 dereceli ölçeklerde ve büyük örneklem büyüklüğüne sahip çalışmalarda bu konuda ciddî bir sorunla karşılaşılmaz. Tutum ölçeklerinin toplam veya ortalama puanları ise eşit aralıklı ölçek verisi olarak

nitelendirilir. Gözlemcilerin verdikleri puanların da en azından yaklaşık olarak eşit aralıklı ölçek verisi niteliğinde olması gerekir.

Tesadüfî olarak belirlenmiş olma. Veriler ana kütleden tesadüfî olarak seçilmiş örnek kütleye ait olmalıdır. İradî örnekleme, kolayda örnekleme ve kota örneklemesi yöntemiyle elde edilen verilere dayalı sonuçlar yanlıdır.

Gözlemlerin bağımsızlığı. Varyans analizinde önemli ön kabullerden bir diğeri, ölçüm yapılan her bir örneklem veya gruptaki gözlem değerinin birbirinden bağımsız olduğudur. Ancak dikkatli yapılmayan bir ölçümde gözlem değerleri tam bağımsız olmayabilir. Aynı sınıfta yer alan öğrenciler, aynı mağazadan alış veriş yapan tüketicilerin birbirlerinden tam olarak bağımsız oldukları söylenemez. Bu tür ölçümlerde belli ölçüde yanlılık vardır.[7] Böyle bir durumda analiz birimi olarak tek tek "öğrenciler" değil, bir bütün olarak "sınıf" dikkate alınır.

Grupların bağımsızlığı. Bu varsayım özellikle "bağımsız ölçümler TYVA" için geçerlidir. Ölçüm yapılan her bir grup/örneklem diğerinden bağımsızdır. Çalışmada k sayıda gruptan yararlanılmış olabilir. Bu gruplardaki kişilerin her birinde farklı ölçümler yapılmıştır. Tekrarlanmış ölçümler şeklinde gerçekleştirilen uygulamalarda bu ön koşul aranmaz.

Örtüşen simetri (compound symetri). Bağımsız ölçümler TYVA'da gruplara ait varyans değerlerinin kabaca birbirine eşit olduğu varsayılmıştı. Benzer fakat biraz daha karmaşık bir ön kabul, "tekrar eden ölçümler için" söz konusudur. Tekrar eden ölçümlerde, ölçüm verilerinin varyanslarının eşitliğine ilave olarak ölçüm çiftlerine ait kovaryans (korelasyon) değerlerinin de eşit çıkması istenir. Bu olgu, kovaryans matrisinin simetrik özellik göstermesi nedeniyle "örtüşen simetri" terimiyle ifade edilmiştir. Örtüşen simetride, ölçümler arasında hem varyans ve hem de kovaryans (korelasyon) değerleri birbirine eşittir. Örtüşen simetride, muhtemel bütün çiftlere ait korelasyon katsayıları kabaca birbirine benzer.[8] Aynı örneklem grubunda t_1, t_2, t_3 ve t_4 gibi farklı zamanlarda yapılan ölçümler sonucunda t_1 ile t_2 ve t_1 ile t_3 gibi farklı ölçüm çiftleri arasındaki korelasyon katsayıları yaklaşık olarak birbirine eşit çıkmışsa "örtüşen simetri özelliği sağlanmıştır" denilir. Verilerde örtüşen simetri özelliği varsa küresellik özelliği de vardır. Fakat, veriler örtüşen simetri özelliğine sahip olmadığı halde küresellik ön şartını sağlamış olabilir.[9]

Küresellik (sphericity). Küresellik, bağımlı örneklemlerde tekrarlanan çoklu ölçümlere dayalı TYVA tasarımında kullanılan matematiksel bir ön kabuldür.[10]

Küresellik, kovaryans matrisindeki[a] verilerin dairesel/küresel bir biçimde yığılmasını temsil eder.[11] İki değişkenli veya iki serili ölçümlerde sadece tek bir korelasyon katsayısı olacağından küresellik şartı aranmaz. Küresellik, bağımsız örneklemlerdeki "varyansların eşitliği" testine benzer ve bir anlamda da bu testin bağımlı örneklemlerdeki uzantısı niteliğindedir. Küresellik, istatistiksel metinlerde ε veya *e* simgesiyle gösterilir ve bazen *dairesellik* terimiyle ifade edilir.

Aynı kişiler üzerinde yapılan çoklu ölçümlerde, "ikili karşılaştırmalar şeklinde yapılan ölçüm farklarına ait serilerin varyans değerlerinin eşit veya aynı çıkması" küresellik olarak tanımlanmıştır. Olguyu *matris cebri* açısından da inceleyebiliriz. Kovaryans matrisinin çaprazı üzerinde ve çaprazın dışında kalan değerlerin birbirine eşit olması daha önce *örtüşen simetri* olarak isimlendirilmişti. Ölçüm verilerinde *örtüşen simetri* özelliği sağlanmışsa *küresellik* koşulu da sağlanmış demektir. Kovaryans matrisini incelediğimizde kovaryans değerleri birbirine eşitse ve bunun yanında varyans değerleri de birbirine benzer özellik gösteriyorsa küresellik varsayımının karşılanması konusunda bir sorun yok demektir.[12] Ancak sosyal bilimciler matrislerle çalışmadıklarından Tablo 7-1'de ölçüm farklarının varyansları arasındaki benzerliğin ne şekilde hesaplandığına ilişkin daha somut bir örnek verilmiştir.

Tablo 7-1. Hayalî Veriler Üzerinde Küreselliğin Hesaplanması

A t_1 zamanında yapılan ölçüm	B t_2 zamanında yapılan ölçüm	C t_3 zamanında yapılan ölçüm	A – B $t_1 - t_2$	A – C $t_1 - t_3$	B – C $t_2 - t_3$
5	6	12	–1	–7	–6
8	10	14	–2	–6	–4
10	8	12	2	–2	–4
12	10	12	2	0	2
14	11	13	3	–1	2
9	10	15	–1	–6	5
Varyans			4,3	9,0	19,4

[a] Veri matrisinin köşegenindeki hücrelerde yer alan rakamlar grupların varyans değerlerini gösterirken köşegenin dışındaki rakamlar kovaryans değerleridir ve bu değerlerin hepsine birlikte kovaryans matrisi adı verilir.

Tablo 7-1'de, ikiden fazla bağımlı örnek kütle arasında yapılan ikili karşılaştırmaların sonuçları verilmiştir. Bu sonuçlara göre, grupların toplam puanları arasındaki farkların varyansları birbirinden önemli ölçüde farklı olduğundan küresellik sağlanamamıştır. Küreselliğin sağlanabilmesi için varyans değerlerinin aynı veya yaklaşık olarak birbirine eşit çıkması gerekir.

- Varyansların eşitliği.

$$S^2_{A-B} \approx S^2_{A-C} \approx S^2_{B-C} \cdot$$

Üç ölçümden en azından ikisinde varyans benzerliği yakalanmışsa buna "lokal küresellik" adı verilir ve genelde küresellik şartının karşılanmış olduğuna karar verilir.

Küresellikten ayrılmanın şiddetini değerlendirme. Bir ölçümde örtüşen simetrinin sağlanamamış olması küreselliğin bulunmadığı anlamına gelmez. Küresellik, örtüşen simetriden bağımsız olarak elde edilebilir. Küresellik şartının sağlanıp sağlanmadığını veya sağlanamamışsa sapmanın şiddetini veya derecesini belirlemek için SPSS'te *Mauchley testi* uygulanır. Bu test ile *farklılık varyanslarının eşit olduğu* hipotezi test edilir. Mauchly testi, ki-kare testini kullanarak *p* anlamlılık değerini verir. Test sonucunda $p < ,05$ ise küresellik koşulunun sağlanamadığına, $p > ,05$ ise küresellik koşulunun sağlandığına, diğer bir deyişle tekrarlanan ölçümlerin varyans değerleri arasında önemli bir farklılık bulunmadığına karar verilir. Mauchly testi sonucu anlamlı çıkmışsa ($p < ,05$ ise) bu kez tekrarlanan ölçüm sonuçlarının kovaryans değerlerinin türdeş olmadığına karar verilir. Ölçümlerin arka planındaki dağılımın homojen olmadığı ve çok değişkenli dağılımın normallik özelliği göstermediği anlaşılır. Böyle bir durumda SPSS çıktılarındaki düzeltilmiş Greenhouse-Geisser ve Huyn-Feldt değerleri incelenerek bunlardan hangisinin temel alınacağı belirlenmeye çalışılır. Düzeltilmiş değerlere başvurulmasının amacı Tip I hatası yapma olasılığını azaltmaktır. Bazı bilim adamları istatistiksel bir güce sahip olmaması nedeniyle *Mauchley* testinin uygulanmasına itiraz etmişler ve testin küçük hacimli örneklemlerde Tip II hatasına yol açtığını bildirmişlerdir.

Küresellikten sapmanın düzeltilmesi. Bağımlı örneklemlerde yapılan ölçümlerde küresellik koşulu sağlanamamışsa geçerli bir *F*-oranı elde etmek için düzeltme yoluna başvurulur veya başka bir varyans testi seçilir. Düzeltme yönteminde üç farklı yaklaşımdan yararlanılır:

1. Greenhouse ve Geisser (G-G) yaklaşımı.
2. Huyunh ve Feldt (H-F) yaklaşımı.
3. Alt sınır tahmin değeri yaklaşımı (lower bound estimate).

Bu yöntemlerin her birinde küresellik etkisini düzeltmek için elde edilen tahmin değerleri serbestlik dereceleriyle çarpılır.

Greenhouse ve Geisser (G-G) yaklaşımı. Greenhouse ve Geisser (1958) tarafından geliştirilen yöntem diğerlerinin içinde daha tutucu sonuçlar verir. (daha küçük değerler elde edilir).[13] G-G yönteminde (çoğunlukla $\hat{\varepsilon}$ simgesiyle gösterilir) serbestlik derecesi yükseltilerek veya düşürülerek daha anlamlı bir p değeri elde edilir.

Huyunh ve Feldt yaklaşımı. Huyunh ve Feldt (1976) tarafından önerilen bu yaklaşımından daha liberal değerler (daha yüksek) elde edilir.[14] Yöntemde Epsilon olarak isimlendirilen düzeltme faktörü bir ölçüde daha yüksek elde edilir. Düzeltme, *alt sınır tahmin değeri* kadar katı veya cimri değildir.

Alt sınır tahmin değeri. Alt sınır düzeltmesinde, Tip I hatası olasılığını azaltmak için en yüksek F kritik değeri kullanılır.[15] SPSS'in dışındaki dünyada ise, alt sınır tahmin değeri olarak Greenhouse ve Geisser yaklaşımı kullanılır.

Çok yönlü varyans analizi. Küresellik koşulu sağlanamamışsa başvurulacak bir diğer yöntem çok yönlü varyans analizi – ÇYVA (multiple analysis of variance – MANOVA) yöntemini uygulamaktır, çünkü bu testte örneklem hacmi 20'den küçük değilse küresellik ön kabulü veya şartı aranmaz. Ancak, her koşulda çok yönlü varyans analizinin kullanılması doğru olmayabilir. Maxwell ve Delaney (1990) tek yönlü varyans analizinin çok yönlü varyans analizinden daha güçlü olduğunu bildirmiş ve n, eğer $a + 10$'dan küçükse çok yönlü varyans analizinin kullanılmamasını önermiştir. Burada a simgesi tekrarlanan ölçümlerdeki düzey sayısını gösterir. Öyle anlaşılmaktadır ki genel bir kural olarak, küresellik ihlalinin büyük ($e < ,70$) ve örneklem hacminin $a + 10$ olduğu durumda çok yönlü varyans analizi daha güçlüdür (aktaran Field, 2003).[16]

Küreselliğin SPSS'te analizi. Küreselliği SPSS'te analiz etmek için GLM Repeated Measures mönüsü kullanılır. Açılan diyalog kutusuna bağımlı ve bağımsız değişkenler tanıtılır. Bağımsız değişkenler ölçüm sayıları, demografik değişkenler veya farklı gruplar olabilir. Hesaplama sonucunda Mauchly test sonucu anlamlı çıkmamışsa Epsilon düzeltme faktörü hesaplanmaz. Mauchly test sonucu anlamlı çıkmışsa ve söz konusu anlamlılık alt sınır değeriyle elde edilmişse farklılığı incelemek için yeterince güç var demektir. Diğer bir deyişle bü-

tün testlerden aynı sonuçlar çıkarılabiliyorsa küreselliğin ihlal edilip edilmediği konusunda testler arasında farklılık aranmaz. Eğer çeşitli düzeltme testleri farklı sonuçlar veriyorsa böyle bir durumda Huyunh ve Feldt test sonucu kullanılır, çünkü bu test sonucu daha az katıdır.[17]

Tekrar eden ölçümlerde aynı katılımcıların kullanılıyor olması daha sonraki ölçüm sonuçlarının önceki ölçüm sonuçlarına bağımlı olması gibi bir durum yaratır. Birinci ölçümde yüksek puan alan kişiler sonraki ölçümlerde de yüksek puan alırlar. Gözlemlerdeki bağımlılık, ölçüme katılan bazı kişilerin diğerlerinden daha iyi puanlar alması sebebiyle ortaya çıkar.

Grup büyüklüklerinin eşitliği. Varyans analizi yapılacak ölçümlerin / grupların / örneklemlerin büyüklükleri de yaklaşık olarak birbirine eşit olmalıdır.

TEMEL KAVRAMLAR

Araştırmacı, varyans analizi sonuçlarını analizle ilgili temel kavramlara hakim olursa etkili bir şekilde yorumlayabilir. Bu başlıkta söz konusu kavramlar ve bu kavramların anlamları üzerinde durulmuştur.

Faktör

Faktör; bilim adamının bağımlı değişken üzerindeki etkisini görmek için kendi kontrolü altında tuttuğu bağımsız değişkendir. Ölçümde birden fazla bağımsız değişken varsa öyle bir durumda "faktörlerden" söz edilir. Varyans analizinde en az bir bağımsız değişken olmalıdır. Ölçümü yapan kişi araştırmasında birden fazla bağımsız değişkenin/faktörün etkisini ortaya çıkarmak isteyebilir. Herhangi bir faktör iki veya daha fazla düzeyden oluşur. Bağımsız değişkene ait "şıklar" veya "seçenekler" faktörün düzeyleridir. Bu düzeyler aynı zamanda duruma göre bağımsız gruplar veya bağımlı gruplar olabilir. Normal koşullarda bir faktörde 10'dan fazla düzey bulunmaz. Faktörlerin eşit aralıklı ölçek niteliğinde olması gerekmez. Faktörler genellikle nominal veya sıralı ölçek verisi niteliğindedir. Örneğin; cinsiyet, yaş grubu, kontrol grubu, ölçüm zamanları, coğrafî bölge değişkenleri genellikle faktör olarak belirlenir. Faktörler koyu siyah yazılmış büyük harflerle gösterilirler (ör., **A**, **B**, **C**, ... , **Z**).[18] Faktörlerin indisleri ise düzeyleri gösterir. Rakam veya harf olarak yazılabilen indisler koyu siyah biçimle değil, romen tarzıyla gösterilir (*ör.*, . $\mathbf{A}_2$, $\mathbf{A}_3$). Bazı karakterler her zaman belirli faktörleri göstermek üzere kullanılır. Örneğin, **S** kişi faktörünü, **E** hata faktörünü ve **G** grup faktörünü göstermek üzere kullanılır.

Bağımsız faktörler değişik şekillerde sınıflandırılmıştır. Bunlardan en sık kullanılanları *sabit faktör* (fixed) ve *tesadüfî faktör* (random) kavramlarıdır. Pek çok araştırmada bir faktörün sabit veya tesadüfî nitelikte olup olmadığını saptamak

zor olabilir. Genelde süjeler (S) ve gruplara (G) ait faktörlerin ana kütleden rasgele seçilmeleri nedeniyle tesadüfî olduğu varsayılır. Tesadüfî faktörlerde araştırmacı bulgularını daha büyük ana kütleye genellemek ister. Sabit faktörlerde ise araştırmacı ana kütleye genelleme hedefi gütmeksizin sadece mevcut örneklemi dikkate alarak faktörün düzeyleriyle ilgilenir. Bir faktörün sabit mi yoksa tesadüfî mi olduğunu belirlemek için şu soruyu sorarız: "Eğer ölçümü tekrarlamış olsaydık bu işlemi hangi koşullarda yapardık?" Eğer aynı kişiler üzerinde ölçüm yapmayı düşünmüşsek sabit faktör koşulları geçerli olurdu. Tam tersine ana kütleden çok farklı kişilerin seçilebileceğini düşünüyorsak o zaman tesadüfî faktör koşulları geçerlidir.[19] İstatistiksel analiz yazılımına tesadüfî faktör birinci deneme, ikinci deneme üçüncü deneme gibi yine kategorik bir değişken olarak tanıtılır. Ancak ana kütleye genelleme hedefi güdülüyorsa deneme sayısı değişkeni tesadüfî faktör olarak atanır. Bir değişken tesadüfî faktör olarak belirlenmişse bu değişken ana kütleden tesadüfî olarak seçilmiş varsayılır. Sonuçlar bağımsız değişkenle temsil edilen örneklem verilerine değil, ana kütleye genellenir.

Faktöriyel

İki yönlü varyans analizinde kullanılan bir terimdir. Literatürde iki yönlü varyans analizine aynı zamanda "faktöriyel tasarım" adı da verilir. İki yönlü varyans analizinde iki bağımsız değişken vardır. Bir bağımsız değişkenin bütün düzeylerinin diğer bağımsız değişkenin tüm düzeyleriyle ilişki içinde olması ve sonuçta bu ilişkiye dayalı olarak birden fazla ilişki kombinasyonun ortaya çıkması faktöriyel tasarım olarak isimlendirilmiştir. Örneğin, 2x2 faktöriyel tasarımında her bir bağımsız değişken iki düzeye sahiptir ve sonuçta rasgele seçilmiş katılımcılar bu faktöriyel tasarımda dört farklı grup içinde sınıflandırılabilir. Bir başka tasarım, 3x3 faktöriyel yapısında ise dokuz farklı grup arasında karşılaştırma yapılır.[20]

Bağımlı Değişken

Bilim adamının yaptığı araştırma/ölçüm sonucunda belirli bir davranış veya tutuma ilişkin olarak elde ettiği puanlardır. Tek yönlü ve iki yönlü varyans analizinde tek bir bağımlı değişken varıdır. Çok yönlü varyans analizinde (ÇYVA) ise birden fazla davranışın / tutumun ölçümü söz konusudur ve bu nedenle birden fazla bağımlı değişkenle çalışılır. Bağımlı değişken büyük *X* harfiyle gösterilir.

Faktörler Arasındaki İlişkiler

Varyans analizinde bilim adamı ölçüm değişkenleri arasındaki ilişkileri önceden belirleyerek istatistiksel yazılım programının veri matrisine buna uygun bir şekilde girmek zorundadır. Veri matrislerinin türleri aşağıdaki gibidir:

Çapraz veri matrisi. İki faktör veya iki bağımsız değişkenin bir ölçüm değişkeni karşısındaki durumu görülmek istendiğinde yapılan düzenlemeye çapraz tasarım adı verilir. Çapraz tasarım *x* harfiyle gösterilir. Örneğin, A ve B gibi iki faktör bulunsun. Ölçüm değişkenini ise C olarak belirleyelim. A cinsiyet faktörünü, B medenî durum faktörünü ve C ise stres puanlarını göstersin. Cinsiyet faktörü medeni durum faktörüyle karşılaştırıldığı zaman (AxB) evli ve bekar olan bütün erkek ve kadınlar analize alınacak demektir. Stres düzeyini belirleyen analiz sonucunda dört farklı gruba ait sonuçlar alınır: Evli erkekler, evli bayanlar, bekar erkekler, bekar bayanlar.

Yuvalanmış tasarım. Eğer B faktörünün düzeyleri A faktörünün sadece tek bir düzeyiyle ilintili olarak analiz edilmek isteniyorsa bu düzenlemeye yuvalanmış tasarım adı verilir. Düzenleme B(A) simgesiyle gösterilir ve bu durumu tanımlamak için "A içinde yuvalanmış B faktörü" ifadesi kullanılır. Bu tür ilişki düzenlemesi literatürde aynı zamanda "hiyerarşik tasarım" biçiminde adlandırılmıştır. Toplumsal yaşamda ve örgütlerde pek çok faktör hiyerarşik bir şekilde sıralanmıştır. Örneğin, bilim dalları ana bilim dalları içinde, ana bilim dalları ise bölümler içinde yuvalanmıştır.

Genel Ortalama

Değişik gruplarda yapılan tüm ölçüm puanlarının toplanması ve toplam örneklem büyüklüğü sayısına bölünmesiyle elde edilen değerdir. Bir diğer tanımı, grup ortalamalarının ortalamasıdır. Grup ortalamalarının toplanarak grup sayısına bölünmesiyle bulunur.

Toplam, Grup İçi ve Gruplar Arası Değişkenlik

Varyans analizinde üç tür değişkenlik vardır: toplam, grup içi ve gruplar arası değişkenlik. Toplam değişkenlik, bağımlı değişken puanlarının genel olarak dağılımıdır. Varyans analizinde bağımlı değişken puanlarındaki toplam değişkenliğin ne kadarının bağımsız değişkenle açıklandığını ve ne kadarının ise açıklanamadığını anlamaya çalışırız. Gruplar arası değişkenlik, bağımsız değişkenin etkisini gösterir. Varyans analizinde toplam değişkenlik, varyans değildir; grup içi ve gruplar arası değişkenliğin toplamından oluşan bir değerdir. Gruplar arası değişkenlik, bir gruba ait aritmetik ortalamanın genel ortalamayla karşılaştırıldığında ortaya çıkan sapmadır. Gruplar arası değişkenlik açıklama getirilebilen bir varyansa sahiptir. Grup içi değişkenlik ise, bir kişiye ait bir puan grup ortalamasıyla karşılaştırıldığında ortaya çıkan sapmadır. Grup içi değişkenlik açıklanamayan varyansa sahiptir. Grup içi değişkenlikte ölçüm hataları ve bireyden kaynaklanan ölçüm sorunları söz konusudur. Toplam varyans, açıklanan varyans +

açıklanamayan varyansa eşittir. Grup içi değişkenlik, bağımsız değişkenin etkisi çıkarıldıktan sonra kalan değerdir ve bu nedenle aynı zamanda "artık değer" olarak isimlendirilir.

İstatistiksel analizde *değişkenlik*, "karelerin toplamı - KT" terimiyle ifade edilir (sums of squares - SS). Toplam değişkenlik, TKT kısaltmasıyla (Total SS - TSS), gruplar arası değişkenlik, GAKT kısaltmasıyla (Between groups SS -BSS) ve grup içi değişkenlik ise GİKT kısaltmasıyla (Within Groups SS - WSS) gösterilir.

Varyans analizinin temelinde yatan fikir, gruplar arasındaki değişkenliği grup içindeki değişkenlikle karşılaştırarak bir sonuca varmaktır. Analizde, örneklemler / gruplar arasındaki değişkenlik örneklem veya grup içindeki değişkenlikten daha yüksekse *örneklem ortalamaları aynı değildir* sonucuna varılır. Böyle bir durumda her bir grup diğerinden bağımsızdır ve gruplar arasında etkileşim yoktur.

F İstatistik Değeri

Varyans analizinde *F* oranı *gruplar arası* varyansın *grup içi* varyansa bölünmesiyle bulunur. Güvenilirliği belirlemeye yönelik olarak yapılacak varyans analizinde *F* değerinin anlamlı çıkmaması gerekir. Diğer bir deyişle sıfır hipotezi kabul edilerek grup ortalamaları arasında anlamlı bir farklılık olmadığı belirtilmelidir. Eğer anlamlı bir farklılık çıkmışsa bu farklılığı yaratan grup araştırmadan çıkarılarak başka bir grup üzerinde yeniden araştırma yapılır. Araştırmaya veya farklı gruplar seçme işlemine *F* değeri anlamlı çıkmayıncaya kadar devam edilir. Sıfır hipotezi kabul edildiği durumda güvenilirlik katsayısını hesaplamak için Eşitlik 7-1'deki formül kullanılır.[21]

$$R = (MS_B - MS_E) / MS_B. \quad \text{(İngilizce simgeler)}$$
$$R = (GAKO - HDKO) / GAKO. \quad \text{(Türkçe simgeler)} \qquad (7\text{-}1)$$

Formüldeki kısaltmalar ve anlamları aşağıdaki gibidir:[b]

MS_B = Mean square between.
GAKO = Gruplar arası kareler ortalaması.

---

[b] İstatistik analiz programı SPSS'in dökümleri yabancı dilde veriyor olması nedeniyle İngilizce ve Türkçe kısaltmaların her ikisi de gösterilmiştir.

MS_E = Mean square error. = $(SS_w + SS_i) / df_w + df_i$.
HDKO = Hata değerlerinin kareler ortalaması.

SS_w = Sum of squares within.
GİKT = Grup içi değerlerin karelerinin toplamı.

SS_i = Sum of squares interaction.
EDKT = Etkileşim değerlerinin karelerinin toplamı.

df_w = Degrees of freedom within.
sd_{gi} = Grup içi değerlerin serbestlik derecesi.

Df_i = Degrees of freedom interaction.
sd_e = Etkileşim değerlerinin serbestlik derecesi.

VARYANS ANALİZLERİNİN TÜRLERİ

Literatürde varyans analizleri değişik gruplar altında sınıflandırılmıştır. Bu bölümde ise sadece güvenilirlik analizlerinde uygulanabilecek temel teknikler ele alınmıştır. Bu teknikler; tek yönlü varyans analizi, iki yönlü varyans analizi, çok yönlü varyans analizi ve varyans bileşenleri analizidir.

Tek Yönlü Varyans Analizi

Tek yönlü varyans analizi (TYVA), iki veya daha fazla örnek kütleye veya gruba ait puan ortalamaları/toplam puanları arasında önemli ölçüde farklılık bulunup bulunmadığını belirlemek için yapılır. Tek yönlü varyans analizi bağımsız örnek kütlelerde veya bağımlı örnek kütlelerde uygulanabilir.

Bağımsız örnek kütlelerde tek yönlü varyans analizi. Bu yaklaşım, aynı zamanda "gruplar arası TYVA tasarımı" (between subjects factor) olarak adlandırılır. Bazı araştırmacılara göre TYVA "bağımsız iki örneklem *t*-testinin kuzenidir."[22] Bağımsız örneklemler TYVA'nın *t*-testinden farkı, ikiden fazla grup / örneklem verileri arasındaki tutarlılığı belirlemesidir.

Bağımsız ölçümler TYVA'da, bilim adamı ikiden fazla bağımsız gruptan elde ettiği veriler arasındaki tutarlılığı araştırır. İkiden fazla bağımsız grup aynı ana kütleye veya farklı ana kütlelere ait olabilir. Genelleme ait olduğu ana kütlelere ilişkin olarak yapılır. Bağımsız ölçümler varyans analizinde bir gruplama değişkeni (bağımsız) ve bir bağımlı değişken vardır. Bu uygulama literatürde "tek faktör tasarımı" ve "katılımcı grupları arasındaki tasarım" adları ile de anılır. Tek yönlü varyans analizinde bağımsız değişken iki, üç veya daha fazla sayıda

kategoriden / düzeyden oluşur. Tek yönlü varyans analizinin düzenleme biçimleri aşağıdaki şekillerde olabilir:

1. *Aynı ana kütleden bağımsız örneklemler seçme.* Birinci örneklem, ikinci örneklem, üçüncü örneklem.
2. *Farklı ana kütlelerden örneklemler seçme.* İşletmeci, asker, mühendis, öğretmen vb. gibi.
3. *Aynı ana kütleden değişik yaş grupları ile çalışma.* Gençler, orta yaşlılar, yaşlılar gruplarında olduğu gibi.

Bağımlı değişken, en az üç farklı gruptan veya örneklemden elde edilmiş olan ölçek veya testin toplam/ortalama puanıdır. Bağımsız örneklemler tek yönlü varyans analizinde, bilim adamı faktör düzeylerine göre bağımlı değişken puanlarının önemli ölçüde değişiklik gösterip göstermediğini anlamaya çalışır. Güvenilirlik açısından, örneklem grupları eğer aynı ana kütleden seçilmişse toplam veya ortalama puanları arasında önemli bir farklılığın çıkmaması gerekir. Grup ortalamaları önemli ölçüde farklı ise, bağımsız değişkenin bağımlı değişken üzerinde bir etkisi olduğuna karar verilir ve test/ölçek sonuçlarının bağımsız değişken grubu ve bu gruba ait kategoriler açısından güvenilir olmadığı sonucu ortaya çıkar. Eğer çalışma, "gözlemciler arasındaki puanlama tutarlılığını" belirlemeye yönelik ise, bu kez gözlemciler bağımsız değişken olarak atanır.

Varyans analizi, değişik gruplara uygulanan test/ölçeğin duruma göre, madde puanları, toplam puanları veya ortalama puanları arasındaki farklılığı saptar. Bununla birlikte testin amacı, her bir ölçüme/gruba ait *varyanslar arasındaki farklılığı* ortaya çıkarmak değildir. Testin adında *varyans* sözcüğünün geçmesi, hesaplamada varyans prosedürünün kullanılması sebebiyledir. Bu model için daha iyi bir isimlendirme "ölçüme ait madde/toplam/ortalama puanları arasındaki farklılıkların analizi" olabilirdi.[23]

■ TYVA'da *F- Oranı* formülü.

$$F\text{- Oranı} = \frac{\text{Gruplar arasındaki değişkenlik}}{\text{Grup içindeki değişkenlik}} . \qquad (7\text{-}2)$$

Bu analizde, önce birinci ölçüme/gruba ait değerler içindeki değişkenlik incelenir ve daha sonra gruplar veya farklı ölçümler arasındaki değişkenlik ele alınır.

Varyans analizinin güvenilirlik analizi olarak kullanılması tartışmalı bir konudur. Bilim adamlarının bir kısmı ölçüm sonuçlarının tutarlılığını belirlemek

için bu yöntemi kullanırlarken diğerleri bu uygulamaya karşı çıkmışlardır. Karşı çıkan bilim adamlarına göre TYVA, tekrar eden ölçümlere dayalı olarak yapılan test-yeniden test korelasyon katsayısını verir. Söz konusu katsayı, farklı ölçümlerde farklı örneklem büyüklükleri kullanılmışsa veya örneklem büyüklüğü küçükse "yanlı" çıkar. Analizde "grup içi varyans" *toplam hata* ile aynıdır. Ölçüm uygulamaları arasında sistematik değişiklikler varsa bu hata daha yüksek çıkar. Bu nedenle Hopkins (2003) uzun süren *çok aşamalı araştırmalarda* (longtiudinal) bu yöntemin kullanılmamasını önermiştir.[24]

Hipotez testi. Tek yönlü varyans analizinde sıfır hipotezi aşağıdaki gibi belirlenir:

H_0: Grup ortalamaları birbirine eşittir. Bağımsız değişken olarak belirlenen faktörün ölçüm sonuçları üzerinde herhangi bir etkisi yoktur. $H_0: \mu_1 = \mu_2 = ... = \mu_n$

H_1: Grup ortalamalarından en az bir tanesi diğerlerinden farklıdır. Bağımsız değişkenin veya faktörün düzeylerinden en az biri ölçüm sonuçlarının eşit çıkmasını engellemektedir. $H_0: \mu_1 = \mu_2 = ... = \mu_n$

Karar kriteri. Varyans analizinde karar kriteri F değerine bakılarak belirlenir. Eğer F değeri anlamlı ise grup ölçümleri arasında önemli bir farklılığın olduğuna karar verilir. Böyle bir durumda post-hoc testleri yapılarak güvenilirliği bozan ölçümün hangi gruba ait olduğu saptanır. Post-hoc testler, aynı zamanda "çoklu karşılaştırma prosedürü" olarak da isimlendirilir.

■ Varyans analizinde serbestlik dereceleri.

$sd_{\text{süt.sayısı}}$ = sütun (grup) sayısı, $k-1$, $5-1$ $= 4$.
$sd_{\text{kat. sayısı}}$ = katılımcı sayısı, $N-1$, $10-1$ $= 9$.
sd_E = $(k-1)(N-1)$, $(5-1)(10-1)$ $= 36$.

Hesaplanan F değeri, $k-1$ pay ve $N-k$ payda serbestlik dereceleri dikkate alınarak önceden belirlenen tablo değerinden büyükse sıfır hipotezi ret edilir. Sıfır hipotezinin ret edilmesi en azından bir gruba ait örneklem ortalamasının farklı olduğu anlamına gelir. Fakat TYVA testi bu farklılığın hangi gruptan kaynaklandığını göstermez. Hangi gruptan kaynaklandığını görmek için verilere Scheffe veya Tukey testi uygulanır.

Bağımlı örnek kütlelerde tek yönlü varyans analizi. Bu analiz bağımlı örneklemler t-testinin bir uzantısı niteliğindedir. Bu uygulamada bağımsız değişken / faktör "ölçüm koşullarını" veya "zamanı" temsil eder. Ölçüm koşulları,

yönetici, eşleri ve çocukları gibi birbirine bağlı olarak yapılan ölçüm düzeylerini veya aynı örnek kütlede değişik gruplara karşı takınılan tutumları belirlemek için yapılan birden fazla ölçümü ifade eder. Zaman ise, aynı örnek kütlede farklı günlerde, aylarda veya yıllarda tekrarlanan ölçümlerle ilgilidir. Örneğin, eğitim öncesinde, eğitim sırasında ve eğitim sonrasında yapılan ölçümler zaman faktörünün düzeyleri olarak belirlenir.

Düzenleme özelliği açısından bu yaklaşım literatürde "grup içi tasarım" ve "tekrarlanmış ölçümler TYVA analizi" adlarıyla tanımlanmıştır. Bağımlı örnek kütlelerde tek yönlü varyans analizi "Tekrarlanmış Ölçümler Tasarımı" başlığı altında ayrıntılı olarak ele alınmıştır.

Bağımlı örnek kütlelerde varyans analizi yapmak isteyen bilim adamlarının, bağımsız örneklemler varyans analizi için belirlenen temel ön koşullara ek olarak örtüşen simetri ve küresellik koşulunun yerine getirilme durumunu araştırmaları gerekir.

SPSS'te Tek Yönlü Varyans Analizi. Tek faktörlü tasarımların tamamında tek bir bağımsız değişken vardır. İstatistik analiz programı SPSS'te bu analizler One-Way ANOVA ve GLM mönüleriyle hesaplanır. Programda varyans analizini yapmadan önce normallik testleri yapılarak verilerin normal dağılım özelliğine sahip olup olmadığı araştırılmalıdır.

Post hoc testler. Gruplar arasında anlamlı ölçüde farklılık varsa bu farklılığın hangi gruptan kaynaklandığını görmek için post hoc testler yapılır. Her bir grupta eşit sayıda vak'a varsa bunun için Tukey testi, gruplardaki vak'a sayısı farklı ise bu kez Bonferroni testi seçilir. SPSS anlamlı ölçüdeki farklılıkları yıldız imiyle işaretler. Bu işaretlere bakarak hangi grubun diğerlerinden anlamlı ölçüde farklı olduğuna karar verilir.

Çok Yönlü Varyans Analizi

Çok yönlü varyans analizinde, tek yönlü varyans analizinden farklı olarak iki veya daha fazla bağımsız değişken / faktör vardır. Bu bağımsız değişkenlerden biri kontrol değişkeni olabilir. Örneğin, sayısal yetenek testi puanlarının birinci sınıf, ikinci sınıf ve üçüncü sınıf öğrencileri arasında farklılık gösterip göstermediği, öğrencilerin kolej kökenli olup olmadıklarına göre test edilebilir. Burada kolej kökenlilik kontrol değişkeni olarak belirlenmiştir. "Çok yönlü varyans analizi türdeşsellik (sabit varyans) varsayımındaki ihlallere karşı daha az duyarlıdır."[25] Uygulamada araştırmacıların genellikle *iki yönlü varyans analizi* terimini kullanmaları nedeniyle bundan sonraki paragraflarda "çok yönlü" ifadesi yerine "iki yönlü varyans" teriminin kullanılması tercih edilmiştir.

İki yönlü faktör analizinde bilim adamı iki bağımsız değişkenin etkileşim içinde bağımlı değişkeni etkilediğini veya iki bağımsız değişkenin birbirinden bağımsız olarak bağımlı değişken üzerinde etkili olduğunu düşünebilir. Bağımsız olduğunu düşünüyorsa bu tür iki yönlü varyans analizleri "birikimsel model" (additive model) olarak isimlendirilmiştir. Birikimsel modelde bağımsız değişkenler arasındaki ilişkiler dikkate alınmaz. Etkileşim modelinde ise aynı zamanda iki bağımsız değişkenin ne ölçüde etkileşim etkisi yarattığına bakılır.[26] İki yönlü varyans analizi grup içi, gruplar arası veya karma bir tasarım içinde yapılmış olabilir. Birincisinde her iki faktör de *grup içi tasarım* özelliğine (tekrarlanmış ölçümler), ikincisinde her iki faktör de gruplar arası tasarım özelliğine ve üçüncüsünde ise, bir faktör tekrarlanmış ölçümler tasarımına diğeri ise gruplar arası tasarım özelliğine sahiptir. Üçüncüsüne karma tasarım (split-plot) adı verilir.

Analiz sonucunda anlamlı bir *F* değeri elde edilmişse anlamlı bir ana etken olduğundan söz edilir. Etkileşim ise şu şekilde yorumlanır: "Bir faktörün etkisi diğer faktörün düzeylerine dayanıyorsa iki bağımsız faktör arasında etkileşim etkisi vardır." Etkileşim etkisi grafikle daha iyi anlaşılır. Eğer iki faktöre ait grafik çizgileri yaklaşık olarak paralel bir seyir izliyorsa önemli bir etkileşimin olmadığı söylenir. Önemli bir etkileşim varsa çizgiler paralel seyretmez.

Çok yönlü faktör analizlerinde bilim adamları öncelikle etkileşim etkisinin yorumlanmasını önermişlerdir. Eğer etkileşim etkisi önemli değilse o zaman ana etkenin yorumlanması yapılır. Eğer etkileşim anlamlı çıkmışsa ana etkenler çok fazla bilgi vermez.[27]

Varsayımları. İki yönlü varyans analizinin varsayımları tek yönlü varyans analizinin varsayımlarıyla aynıdır. Buna göre (a) kişilerin ölçüm gruplarına rasgele ve bağımsız bir şekilde atandıkları, (b) örneklemlerin birbirinden bağımsız oldukları (c) bağımlı değişkenin normal dağılım özelliğine sahip olduğu ve (ç) ana kütle varyanslarının eşit olduğu varsayılır. İki yönlü varyans analizinde üzerinde ölçüm yapılan örneklem gruplarının büyüklükleri eşit olmalıdır.

Hipotez. İki yönlü varyans analizinde üç sıfır hipotezi ve üç alternatif hipotez vardır. Birinci hipotezde birinci faktörün düzeylerine ait ana kütle ortalamalarının eşit olduğu varsayılır. İkinci hipotezde ikinci faktörün düzeylerine ilişkin ana kütle ortalamalarının eşit olduğu varsayımından hareket edilir. Üçüncü sıfır hipotezinde ise iki faktörün düzeyleri arasında etkileşim etkisinin bulunmadığı varsayımı test edilir. Etkileşim etkisi AxB simgesiyle gösterilir. Etkileşim etkisi, karma veya kombine bir etkiyi gösterir. Bilim adamı öncelikle etkileşim etkisi hipotez testi sonuçlarını yorumlamalıdır. Eğer etkileşim etkisi anlamlı değilse o zaman birinci ve ikinci faktörün düzeylerine ait ana kütle ortalamalarının eşit olduğu hipotez test sonuçları araştırılır. Literatürde birinci ve ikinci faktörün düzeyle-

rine ait ana kütle ortalamalarının eşit olduğu hipotez test sonuçları anlamlı çıkmışsa bu durum "ana etki" terimiyle ifade edilmiştir. Ana etki, bir faktörün düzeyleri arasında süreklilik gösteren farklılığı simgeler. Bağımsız değişkenin/faktörün düzeyleri ikiden fazla ise araştırmacı düzey ortalamalarını karşılaştırmayı düşünebilir. Bu işlem sonucunda farklılığın hangi düzeyden kaynaklandığı görülür. Gerek birinci ve gerekse ikinci faktörün düzeleri arasında anlamlı bir farklılık yoksa ana etkinin anlamlı olmadığı söylenir. Güvenilirlik analizinde ana etkilerin anlamlı olmaması test sonuçlarının örneğin, cinsiyet ve belirlenen yaş grupları için istikrarlı sonuçlar verdiği anlamına gelir. İki yönlü varyans analizinde sıfır ve alternatif hipotezleri aşağıdaki gibi belirlenir.

H_0: A faktörü için ana etki yoktur.
H_1: A faktörü için ana etki vardır.

H_0: B faktörü için ana etki yoktur.
H_1: B faktörü için ana etki vardır.

H_0: Etkileşim etkisi yoktur.
H_1: Etkileşim etkisi vardır.

Araştırmacı analiz sonucunda anlamlı etkileşim etkisi bulurken hiçbir ana etki bulamayabileceği gibi, bir faktör için anlamlı ana etki bulurken bunun yanında etkileşim etkisi bulamayabilir. Bazen de her iki faktör için anlamlı ana etki bulunurken etkileşim etkisi ortaya çıkmayabilir.

SPSS'te çok yönlü varyans analizi. İstatistiksel analiz yazılımı SPSS'te iki yönlü varyans analizi yapabilmek için iki bağımsız ve bir bağımlı değişkenin veri matrisi penceresine yüklenmiş olması gerekir. Analiz yazılımın GLM mönüsünden Univariate şıkkı seçilerek bağımsız değişkenler Fixed factors veya Random factors penceresine alınır. Etkileşim etkisini görmek için Plots seçeneği açılarak Add tuşuyla analiz penceresine alınması sağlanır. Faktörlerden herhangi biri eğer iki düzeyden daha fazla ise post hoc testleri yapılır.

Çok Değişkenli Varyans Analizi

Çok değişkenli varyans analizi – ÇODVA, (Multivariate Analysis of Variance – MANOVA) tek bağımlı değişken yerine birden fazla bağımlı değişken olması halinde uygulanan bir istatistik tekniktir. ÇODVA, tek değişkenli varyans analizine göre daha güçlüdür ve bu teknik araştırmacıyı Tip I hatası yapmaya karşı korur. Eğer bir dizi bağımlı değişken ve birden fazla bağımsız değişkenler arasındaki ilişkiler araştırılıyorsa ÇODVA tekniği uygulanır. Bu tekniğin güvenilir-

lik analizleriyle ilgisi, cinsiyet ve yaş gibi iki bağımsız değişkenin bir ölçüm aracının (örneğin, stres ölçeği veya yetenek testi gibi) üç farklı zaman diliminde yapılan ölçüm sonuçlarını karşılaştırmak için kullanılabilmesidir. Farklı zamanlarda yapılan ölçüm sonuçları arasında cinsiyet ve yaş faktörüne göre değişiklik olup olmadığını görmek için kullanılır. Ölçümde bağımlı değişkenlerin ortalaması veya bileşkesi vektör olarak isimlendirilir. Bu analizde, bağımlı değişkenlerin doğrusal kombinasyonundan oluşan yeni bir değişken üretilir. Bu yeni değişken grup farklılıklarını maksimize edecek şekilde oluşturulur. Buna göre bağımlı değişkenler bileşke olarak bir üst değişken haline gelir. Güvenilirlik analizlerinde ÇODVA tekniğini kullanmak isteyen bilim adamlarının amacı, iki veya daha fazla bağımlı değişken ortalamasını temsil eden bir vektörün (bileşik değişken Y'nin) faktörler baz alındığında aynı ana kütleden gelip gelmediğini belirlemektir. Sıfır hipotezi, vektör ortalamasının bağımsız gruplar arasında farklılık göstermediği, vektör ortalamasının tüm gruplarda aynı olduğu şeklinde belirlenir. Test, iki veya daha fazla faktörün aynı şapka altında birleşme olasılığını verir.[28]

ÇODVA testinin amacı, bağımlı değişkenlerin bağımsız değişkenler karşısındaki konumunu görmektir. Araştırmacı bağımsız değişkenlerden birinde herhangi bir değişikliğe gittiğinde bağımlı değişkenlerin bu değişiklik karşısında nasıl bir davranış gösterdiği ÇODVA testi ile incelenir. Bilim adamı ÇODVA testi ile aşağıdaki sorulara cevap bulmaya çalışır:[29]

Bağımsız değişkenlerden hangisi (hangileri) ana etkiye sahiptirler?
Bağımsız değişkenler arasındaki etkileşim etkisi nedir?
Bağımlı değişkenlerin önemi nedir?
Bağımlı değişkenler arasındaki ilişkinin gücü nedir?
Araya giren değişkenlerin etkisi nedir?

Çok değişkenli varyans analizinde iki konu araştırılır. Birincisinde bağımlı değişkenlerin birbirleriyle ne ölçüde ilişkili olduğu incelenir. Buna aynı zamanda bağımlı (ilişkili) örneklemler ÇODVA testi denir. Vektörlerin her zaman birbirleriyle ilişkili değişkenlere dayanması gerekmez. Tek yönlü varyans analizinde olduğu gibi bazen bağımlı değişkeni temsil eden değişkenler birbirinden bağımsız da olabilir. Örneğin, bir ölçeğin güvenilirliğinin farklı yüksek lisans programlarına devam eden öğrencilerde sınanmak istenmesi durumunda her bir programdaki öğrenciler bağımsız grupları oluştururlar.

İkincisinde ise, bağımsız değişkenlerin çok sayıda bağımlı değişkeni ne şekilde etkilediği araştırılır. Bağımsız değişkenler, sonuçlar üzerinde etkili olabilecek olan cinsiyet, yaş, kıdem, eğitim, meslek ve deneyim gibi faktörlerdir.

Çok değişkenli varyans analizini türleri. ÇODVA'nın üç farklı türü vardır: Hotelling T, tek yönlü ÇODVA ve faktöriyel ÇODVA. Hotelling T yönteminde iki düzeyli bir bağımsız değişken ve çoklu bağımlı değişken var iken, tek yönlü ÇODVA yönteminde çok düzeyli bir bağımsız değişken ve çoklu bağımlı değişken bulunur. Faktöriyel ÇODVA ise, faktöriyel TYVA'ya benzer. Diğer bir deyişle çok düzeyli çok sayıda bağımsız değişken ve çoklu bağımlı değişken istatistiksel analize alınır.[30]

Varsayımları. Çok değişkenli varyans analizinin varsayımları tek değişkenli varyans analizinin varsayımlarına benzer. Test varsayımların ihlallerine karşı oldukça güçlü olmakla birlikte varsayımların karşılanma durumu yine de araştırılmalıdır. Çok değişkenli varyans analizinin varsayımları aşağıdaki gibidir.

Bağımsızlık. Tesadüfî olarak seçilen örneklemlerin birbirinden bağımsız olmasıdır. Katılımcıların her bir bağımlı değişkendeki puanları istatistiksel olarak diğer katılımcıların o bağımlı değişkendeki puanlarından bağımsızdır. Hatalar da birbirinden bağımsızdır. Bağımsızlık varsayımının karşılanıp karşılanmadığını belirlemek için "artık değerler grafiği" (residual plots) incelenir.

Çok değişkenli normallik. Her bir bağımlı değişkendeki veriler çoklu normal dağılım özelliğine sahiptir. Bağımlı değişkenlerin doğrusal kombinasyonları da çoklu normal dağılım özelliği gösterir. Tüm alt düzey değişkenlerin hepsi normal dağılım özelliği göstermelidir. Çok değişkenli normallik koşulunun karşılanma durumunu belirlemek için ya verilerin elips biçiminde dağılım özelliğine sahip olup olmadığına bakılır veya "artık değerlerin" dağılımı incelenir. Verilerin *çok değişkenli normal dağılım* özelliği göstermesi gerekir. Ayrıca bağımlı değişkenlerin dağılım grafikleri incelenerek normal dağılım özelliği gösterip göstermediğine bakılır.

Türdeşlik. Her bir bağımlı değişkenin gruplar/koşullar karşısındaki varyansı aynı olmalıdır. Bunun yanında, bağımlı değişken çiftlerinin kovaryansları da aynı olmalıdır. Buna "kovaryans matrisinin türdeşliği" adı verilir. Varyansların türdeşliği Barlett'in ki-kare testi ile veya Box'un M testi ile sınanır. Eğer türdeşlik varsayımı karşılanamamışsa ve örneklem büyüklükleri eşitse veya yaklaşık olarak eşitse test bir ölçüde güç kaybına uğramıştır. Örneklem büyüklükleri önemli ölçüde birbirinden farklı ise, Tip I hata oranı yüksek veya düşük çıkar.

Doğrusallık. Bağımlı değişkenler arasındaki ilişkilerin doğrusal olmasıdır.

Testin aşamaları. ÇODVA testinin gerçekleştirilmesi belirli aşamalarda gerçekleştirilir. Güvenilirlik analizi çerçevesinde bu teknikten yararlanılırken araştırmacı esas olarak ÇODVA sonucunun anlamlı çıkmamasını araştırır. Sonuç anlamlı değilse, bağımsız faktörlerin hepsi veya tüm düzeyleri için sonuçlar geçerlidir demektir. ÇODVA sonucu anlamlı çıkmışsa bu kez TEYVA yapılır ve sonucun bağımsız değişkenler açısından anlamlı olup olmadığına bakılır. Sonuç anlamlı çıkmışsa üçüncü aşamada Post Hoc testi yapılarak farklılaştıran grubun hangisi olduğu araştırılır.

Hipotezleri. Çok değişkenli varyans analizinin hipotezleri bağımlı değişken sayısı kadardır. Her bir bağımlı değişken için sıfır hipotezi ayrı ayrı belirlenir. Hipotezin istatistiksel simgelerle gösterilmesi aşağıdaki gibidir:[31]

$BD_1, H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$ ve
$BD_2, H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$ ve
...
$BD_m, H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$

ÇODVA, ortalamaların vektörlerini karşılaştırır. Çok değişkenli varyans analizinde alternatif hipotez, bağımlı değişkenlere ait gruplardan en az birinde farklılık olduğu şeklinde belirlenir.

SPSS ve çok değişkenli varyans analizi. İstatistiksel analiz programı SPSS'te çok değişkenli varyans analizi, "GLM" (General Linear Model) başlığı altında tanımlanmıştır. Bu bölümde GLM-Multivariate şıkkı seçilerek değişkenler yazılıma tanıtılır. Ölçüm sonuçları bağımlı değişken; cinsiyet ve yaş değişkenleri ise sabit faktörler (Fixed factors) olarak girilir. Daha sonra diyalog kutusundaki model düğmesiyle ikinci bir karta geçilir. Burada Custom düğmesi seçili hale getirilir. Cinsiyet ve yaş değişkenleri seçilerek bu değişkenler model alanına veya kutusuna alınır. Açılır listeden Main effect şıkkı seçilir ve GLM-Multivariate box diyalog kartına dönülür. Bu kez Contrasts düğmesine basılır. Bu bölümde, açılır listeden Simple seçeneği seçilir. Referans kategorisinde First seçeneği işaretlenir. Change düğmesine basılarak bağımsız değişkenlerin pencere içine gelmesi sağlanır ve GLM-Multivariate box diyalog kartına dönülür. Üçüncü aşamada Plots düğmesine basılır. Açılan diyalog kutusunda Cinsiyet faktörü yatay eksene alınır. Daha sonra Add düğmesine basılarak Cinsiyet faktörü alt mönüye alınır ve GLM-Multivariate box diyalog kartına dönülür. Dördüncü aşamada Post Hoc düğmesine basılır. Açılan diyalog kutusunda Cinsiyet ve Yaş değişkenleri Post Hoc analiz alanına alınır. Daha sonra Benferroni şıkkı seçili hale getirtilir. Beşinci aşamada Save düğmesine basılır. Açılan diyalog kutusun-

da Tahmin Değerleri bölümünde Unstandardized ve Standard Error kutucukları seçili hale getirilir. Teşhis bölümünde Cook's Distance ve Leverage Values kutucukları seçilir. Artık değerler bölümünde ise Unstandardized ve Standardized kutucukları işaretlenir. Daha sonra GLM-Multivariate kartına dönülür. Altıncı aşamada Options diyalog kartı açılır. Bu bölümde tanımlayıcı istatistik, etki büyüklüğü tahmini, gözlemlenen güç, Residual SSPC, türdeşlik testi, ve artık değerler grafiği kutucukları seçili hale getirilir. Ayrıca *ana etkileri karşılaştır* kutusu işaretlenir. Bundan sonra Continue tuşuna basılarak işlem yaptırılır.[32] ÇODVA için veriler bilgisayara Tablo 7-2'de görüldüğü gibi girilir.

Tablo 7-2. Çok Değişkenli Varyans Analizi İçin Veri Matrisi

Cinsiyet	Yaş	1. ölçüm	2. ölçüm	3. ölçüm
1	2	12	14	11
2	1	10	14	13
1	3	11	13	12

Çok değişkenli varyans analizi çıktıları. Çok değişkenli varyans analizi çıktılarından öncelikle tanımlayıcı analiz tablosu elde edilir. Bu tabloda her gruba ait ortalama ve standart sapma değerleri vardır.

Çok değişkenli varyans analizinde ayrıca her bir grubun kovaryans matrisinin diğer grupların kovaryans matrislerine eşit olduğu varsayılır. Bu varsayım SPSS'te Box'un M testi ile analiz edilir. Box'un M testi anlamlılık değeri ,05'ten büyükse homojenlik (türdeşlik) ön kabulünün/varsayımının ihlal edilmediğine karar verilir.

Bundan sonraki aşamada Bartlett küresellik testi sonuçları incelenir. Bu testin yapılma amacı bağımlı değişkenlerin gerçekten bağımlı olup olmadığını görmektir. Bartlett test sonucunun ,00 çıkması halinde değişkenlerin bağımsız olduğu hipotezi ret edilir ve değişkenler arasında anlamlı bir korelasyon olduğuna karar verilir. Değişkenlerin bağımlılığının istatistiksel olarak kanıtlanmasıyla ÇODVA analizini yapmak için sağlam bir dayanak elde edilmiş olur.

Daha sonra çok değişkenli varyans analizinin ana çıktısı olarak "çok değişkenli *F* istatistik" değerleri (Multivariate Tests) tablosu elde edilir. Bu tablo, bağımsız değişkenlerin her birinin bağımlı değişkenler üzerindeki etkisini ortaya koyar. Sonuçlar dört anlamlılık değeri ölçüsüne göre değerlendirilir. Bu anlamlılık değerlerinin hepsi farklı bakış açısı ve değerlendirme yöntemleri dikkate alınarak aslında aynı değeri verir.

1. Pillai set değeri.[c]
2. Wilks'in Lambda değeri.
3. Hotelling'in set değeri.
4. Roy'un en büyük kök değeri.

Wilks'in Lambda değeri, bağımlı değişkenlerde bağımsız değişkenlerden kaynaklanmayan varyansın yüzdesi olarak yorumlanır. Wilks'in Lambda değeri üç veya üçten fazla bağımlı değişkenden yararlandığı durumlarda kullanılır. Hotelling'in set değeri ise iki bağımlı değişkenin bulunduğu tasarımlarda kullanılır. Literatürde Pillai'nin set değeri ile Roy'un en büyük kök değerinden daha az yararlanılmıştır. Hesaplanan *F* değerinin anlamlılık değeri, bağımsız değişkenin yarattığı etkinin anlamlı olup olmadığını gösterir. Olasılık değeri (sig.) eğer ,05'ten küçük çıkmışsa bağımsız değişkene ait bütün grupların etkinlerinin anlamlı olduğu söylenir. Örneğin, cinsiyet faktörünü temel alırsak üç farklı ölçümü temsil eden stresle ilgili vektör puanlarının kadınlar ve erkeklerde birbirinden önemli ölçüde farklı olduğu sonucu ortaya çıkar. Tablodaki Eta Squared değerleri bağımlı değişkende bağımsız değişkenin sorumlu olduğu değişkenliği açıklar. Sonuçların anlamlı çıkması aynı zamanda Tip I hatası yapma olasılığını da içinde barındırırken; tablonun son sütununda yer alan Power değerleri Tip II hatası yapma şansı hakkında bilgi verir. Tip II hatası yapmamak için bu değerlerin ,90'ın üzerinde olması gerekir.[33]

Posthoc testler. Analiz sonucunda, bağımlı değişken ortalamalarına ait vektörün (bileşenin), gruplar arasında anlamlı bir farklılık gösterdiğini ortaya çıkarmışsa bu farklılığın hangi gruptan kaynaklandığını görmek için posthoc testi yapılır. Posthoc testi ile hangi grubun diğerinden anlamlı ölçüde farklı olduğu belirlenmeye çalışılır. Gruplar ikişerli olarak birbirleriyle karşılaştırılarak benzerlik ve farklılıkları ortaya konur. Posthoc testi için Tukey testi tercih edilir.

Sonuçların sunumu. Çok değişkenli varyans analizi sonuçlarına ilişkin tablolar metinde nadir olarak verilir. Daha çok ortalamalar ve *F* değerleri metin içinde gösterilir.

■ Çok değişkenli varyans analizi sonuçlarının yorumu.

Test-yeniden test değerleri, cinsiyet ve yaş faktörleri ile birlikte çok değişkenli varyans analizi yöntemi uygulanarak (test-yeniden test x cinsiyet x yaş) incelenmiş-

---

[c] Set değeri (Trace). Bir matrisin köşegeninde yer alan değerlerin toplamı.

tir. Analiz sonucunda cinsiyet $F(2, 103) = 68,89$, $p < ,001$ ve yaş faktörlerinde $F(1, 103) = 148,03$, $p < ,001$ ana etki faktörünün var olduğu saptanmıştır.

Tekrarlanmış Ölçümler Varyans Analizi

Tekrarlanmış ölçümler varyans analizi, (a) tek yönlü ve (b) çift yönlü olmak üzere iki şekilde uygulanır. Çift yönlü tekrarlanmış varyans analizi de kendi içinde tekrar ikiye ayrılır: *tekrarlanmış ölçümler çift yönlü varyans analizi* ve *iki yönlü karma varyans analizi* (2-way mixed ANOVA). Aşağıdaki bölümde önce tek yönlü tekrarlanmış ölçümler varyans analizi tekniği üzerinde durulmuş ve daha sonra çift yönlü tekrarlanmış ölçümler varyans analizi tekniklerine değinilmiştir.

Tek yönlü tekrarlanmış ölçümler varyans analizi. Bu teknik, tek yönlü varyans analizinin bağımlı gruplarda uygulanan şeklidir. Yöntemde, aynı kişiler üzerinde yapılmış ölçümler sonucunda elde edilen ikiden fazla ölçüme ait ortalama değerleri veya toplam puanları vardır. Aynı kişiler üzerinde yapılan ölçümler değişik şekillerde tasarlanabilir.

1. Yeni geliştirilen bir genel yetenek testi aynı kişilere üçer ay arayla dört defa uygulanabilir. Bu uygulamada zamana dayalı gerçek bir tekrarlama vardır.
2. Aynı özelliği ölçtüğü düşünülen üç veya dört paralel test formu aynı kişilere bir kerede uygulanarak bu formların ne ölçüde birbirine benzer olduğu araştırılabilir. Bu uygulamada farklı zamanlarda ölçüm yapma durumu söz konusu değildir, ancak kısa arlıklarla da olsa tekrarlanmış bir ölçüm söz konusudur.

Tekrarlanmış ölçümler varyans analizine, daha çok uzun süren çok aşamalı araştırmalardan veya ölçümlerden elde edilen verilerin güvenilirliği için başvurulur. Birbiriyle "ilişkili" veri yapıları arasındaki tutarlılığı belirlemek için "bağımsız tek yönlü varyans analizi" yerine "tekrarlanmış ölçümler varyans analizi" yönteminin kullanılması gerekir. Örneğin, ölçümler test-yeniden test şeklinde yapılmışsa; ölçümler müdahale öncesinde, müdahale sırasında ve müdahale sonrasında yapılmışsa veya ölçümlerle paralel formların benzerliği araştırılıyorsa "bağımsız gruplar arası tasarım" yerine "tekrarlanmış ölçümler TYVA" tasarımı kullanılır.

Tekrarlanmış ölçümlerde test veya ölçek (duruma göre ölçekler) aynı kişilere ikiden fazla sayıda uygulandığından bu teknik aynı zamanda *katılımcılar içi faktör analizi* (within subjects factor analysis) yöntemi olarak da isimlendirilir. Ancak burada, klasik "faktör analizi" yönteminde gördüğümüz gibi herhangi bir

gruplama değişkenin araştırılması söz konusu değildir. Tekrarlanmış ölçümler varyans analizinde tek bir gruba ait çeşitli ölçüm ortalamalarının birbirinden önemli ölçüde farklı olup olmadığı incelenir. Sıfır hipotezi "ölçüm ortalamaları birbirine eşittir" şeklinde belirlenir. Yöntemde gruplama değişkeni kullanılmadan analiz yapılır. Tekrarlanan ölçümler arasındaki güvenilirlik Eşitlik 7-4'deki formüle göre saptanır (Fisher, 1946; Horst, 1949 aktaran, C.H. Yu).[34]

$$r = \frac{MS_{ölçümler\ arası} - MS_{kalan}}{MS_{ölçümler\ arası} + (sd_{kisilerin\ \ kendi\ içinde} \times MS_{kalan})} \qquad (7\text{-}4)$$

Tekrarlanmış ölçümler ve SPSS. Tekrarlanan ölçüm verileri SPSS veri penceresine kullanılacak analiz modülüne göre iki farklı şekilde tanıtılabilir. Birincisi tek değişkenli tanıtım biçimi ve ikincisi ise çok değişkenli tanıtım biçimidir. Tek değişkenli tanıtım biçimi One-Way ANOVA modülü kullanılarak yapılır. Tek değişkenli tanıtım biçiminde tekrarlanan ölçümler bağımsız değişken olarak belirlenir. Çok değişkenli tanıtım biçiminde ise yazılımın GLM modülünden yararlanılır. Tekrarlanmış ölçümler için araştırmacılara çok değişkenli tanıtım biçimini kullanmaları önerilir.

Tek değişkenli tanıtım biçimi. Burada "tek değişkenli" sözcüğünden kasıt bağımlı değişkendir. Ölçüm sayısı ve kişiler veri penceresine bağımsız değişken olarak tanıtılır.

■ Tek değişkenli veri matrisi.

Kişiler	Zaman / Ölçüm	Stres puanı
1	1	67
1	2	87
1	3	57
2	1	59
2	3	62

Çok değişkenli tanıtım biçimi. Bu uygulamada farklı zamanlarda yapılan her bir ölçüm ayrı bir değişken olarak tanıtılır. Paralel formlar yöntemi kullanılmışsa her bir formdan elde edilen değerler ayrı bir sütuna yazılır.

■ Çok değişkenli veri matrisi.

Kişiler	t_1 zamanında yapılan ölçüm	t_2 zamanında yapılan ölçüm	t_3 zamanında yapılan ölçüm
1	67	87	57
2	59	64	62

Çok değişkenli tanıtım biçimi uygulanmışsa veriler istatistiksel analiz programı SPSS'te Analyze / General Linear Model / Repeated Measures mönüleriyle hesaplatılır. Bu bölümde Within-Subject Factor Name kutusuna ölçümü yapılan değişkenin adı yazılır. Bu değişken "zaman" veya "form sayısı" olabilir. Bu bölümde yer alan Numbers of Levels kutusuna ise tekrarlanan ölçüm sayısı belirtilir. Daha sonra açılan diyalog kutusuna faktörün düzeylerinin adları tanıtılır. Diyalog kutusunun alt bölümünde yer alan Contrast düğmesine basılarak post hoc analizi yapılacak değişkenler belirlenir. Açılan diyalog kutusunda Repeated şıkkı seçili hale getirilir. Change düğmesine basılarak gerekli değişiklik yapılır ve işleme devam edilir. Daha sonra Plot düğmesi tıklanarak grafik seçeneği seçili hale getirilir. Bundan sonra Options düğmesine basılarak tanımlayıcı istatistik analiz kutuları (descriptive istatics ve estimates of effect size) seçilir.[35]

Tekrarlanmış ölçümler tek yönlü varyans analizinin çıktıları. Tekrarlanmış ölçümler tek yönlü varyans analizi sonucunda dört tablo elde edilir. Birinci, tablo tanımlayıcı istatistik değerlerini verir. İkinci tablo *within-subject effects* başlığını taşır ve araştırmacının esas görmek istediği bilgileri içerir. Bu tabloda tekrarlanan ölçümler arasında önemli bir farklılık olup olmadığını anlamak için anlamlılık değerlerine (sig.) bakılır. Olasılık değeri ,05'ten küçük çıkmışsa ölçüm sonuçlarının anlamlı ölçüde birbirinden farklı olduğuna karar verilir. Güvenilirlik için sonuçların ,05'ten büyük çıkması gerekir.

İki yönlü tekrarlanmış ölçümler varyans analizi. İki yönlü tekrarlanmış ölçümler varyans analizi, daha önce belirtildiği gibi iki yönlü ve karma olmak üzere iki farklı şekilde uygulanır.

Tekrarlanmış ölçümler iki yönlü varyans analizi. Bu yaklaşımda iki bağımsız değişken vardır ve söz konusu bağımsız değişkenin düzeylerinde yer alan tüm kişiler üzerinde tekrarlanan ölçümler yapılmıştır. Belirli bir düzeyde yer alan kişilerin ölçüme katılmaması gibi bir durum söz konusu değildir. Tek yönlü tekrarlanan ölçümlerde, örnekleme giren kişiler arasında herhangi bir sınıflandırma yapılmazken iki yönlü tekrarlanan ölçümlerde örnekleme giren kişiler cinsiyet ve yaş; meslek ve kıdem gibi faktörlere dayalı olarak sınıflandırılabilir. Bu sınıflan-

dırma değişkenleri "Between Subject Factor" (Kişiler Arası Faktör)[d] olarak nitelendirilir. Diyelim ki cinsiyet ve yaş faktörlerini temel aldık. Böyle bir durumda cinsiyet faktörünün düzeyleri ile yaş faktörünün düzeylerinin bağımlı ölçümler üzerindeki etkisi araştırılacaktır. Bu etki, ana etki veya etkileşim etkisi olarak soruşturulur. Güvenilirlik açısından araştırmacı ana etki ve etkileşim etkisinin her ikisinin de olmasını istemeyebilir veya tam tersine norm değerlerini belirli bir grup için özel olarak oluşturmak istiyorsa böyle bir etkinin ortaya çıkmasını arzu edebilir.

Tekrarlanmış ölçümler iki yönlü varyans analizi yöntemi SPSS'te hesaplanırken değişken tanımlama diyalog kutusunda Between Subject Factor bölümüne bağımsız değişkenler (cinsiyet ve yaş gibi faktörler) tanıtılır. Analiz sonucunda araştırmacıyı esas olarak ilgilendiren iki tablo elde edilir. Bunlardan birincisi Tests of within-subjects effects isimli olanıdır. Bu tabloya bakılarak üç veya dört farklı ölçüm uygulaması sonuçları arasında anlamlı bir farklılık olup olmadığına karar verilir. İkinci tablo ise, Tests of between-subjects effects adını taşır. Bu tablodaki veriler ise bağımsız değişkenlere ait düzeylerin sonuçlar üzerindeki etkisini ortaya koyar. Anlamlılık değeri eğer ,05'ten büyük çıkmışsa bağımsız değişkenlere ait düzeylerin sonuçlar üzerinde hiçbir etkisinin olmadığı söylenir. Ana etki veya etkileşim etkisi yoksa böyle bir durumda sonuçlar örneğin, erkek ve kadınların her ikisine ve ayrıca belirlenen erimdeki tüm yaş gruplarına rahatlıkla genellenebilir.

Tek bir faktör üzerine dayanan tekrarlanmış ölçümler varyans analizi. Bu tekniğe aynı zamanda "iki yönlü karma TYVA" adı verilir. Bu yöntemde de yine iki adet bağımsız değişken vardır (örneğin, cinsiyet ve yaş gibi). Ancak ölçümler farklı bir tasarım içinde yapılır. Bir veya daha fazla değişkende aynı kişiler kullanılırken, bir veya daha fazla değişkenin tekrarlanan ölçümlerinde farklı kişilerden yararlanılır. Örneğin, cinsiyet faktörü temel alındığında üç ölçüm yapılmışsa her üç ölçümde de aynı erkek ve kadınlar kullanılırken; yaş faktörü temel alındığında birinci ölçüme alınan kişiler ile ikinci ölçüme alınan kişiler ve üçüncü ölçüme alınan kişiler aynı olmayabilir. İkinci bağımsız değişkende farklı örneklemlerden yararlanılmışsa bu uygulama karma TYVA tasarımıdır. Daha sonra karma niteliğe sahip söz konusu örneklem sonuçları bir araya getirilerek karma TYVA analizi yapılır.

---

[d] Kişiler Arası Faktör, kişilerin belirli kriterlere göre gruplandırılması anlamına gelir.

Varyans Bileşenleri Analizi

Varyans bileşenleri analizi, Cronbach ve arkadaşları tarafından geliştirilen (1963, 1972) *genellenebilirlik kuramı* çerçevesinde yapılan gözlemlerin ve ölçümlerin güvenilirliğini test etmek için kullanılan bir tekniktir. Varyans bileşenleri analizinin temel amacı, ölçüm yapılan örneklemde bağımlı değişken ile tesadüfî faktörler (bağımsız değişkenler) arasındaki ilişkilerden hareket ederek bağımsız değişkenlerin ana kütledeki ortak varyansını (ortak değişkenliğini) tahmin etmektir. Diğer bir deyişle "bağımlı değişken – tesadüfî faktörler" etkileşiminin ana kütlede ne gibi bir değişkenlik gösterebileceğini ortaya koymak, böylece gözlem puanlarıyla evren puanı (gerçek puanı) arasındaki farkı, hatayı belirlemektir.

Genellenebilirlik kuramına göre bir ölçümün güvenilirliğini saptamak için birden fazla *yüzeydeki* hata oranını tespit etmek gerekir. Bunun için *varyans bileşenleri analiz tekniği* kullanılarak potansiyel hata kaynağı olan değişik yüzeylerdeki varyansın oranı saptanır. Ayrıca hesaplanan varyans değerlerine bağlı olarak *genellenebilirlik katsayısı* ile *dayanıklılık indeks* değerleri ortaya konarak ölçüm sonuçlarının genelleme yapılmak istenen evren için ne ölçüde güvenilir ve geçerli olduğu belirlenir. Varyans bileşenleri analizinde bilim adamı, bağımlı değişkendeki değişkenliğin hangi oranda belirlenen yüzeylere[c] atfedilebileceğini saptamaya çalışır. Bir yüzeydeki varyans düşük çıkmışsa bağımlı değişken, bu yüzeye daha fazla genelleme yapma özelliğine sahip olacak demektir. Bu analiz sonucunda, incelemeye alınan tesadüfî değişkenlerle evrene ne ölçüde genelleme yapılabileceği belirlenmiş olur. Varyans bileşenleri analizi belirli aşamalarda gerçekleşir. Birinci aşama yüzeylerin belirlenmesidir. Araştırmacı bir, iki, üç veya en çok dört farklı yüzeyde değişkenliği araştırabilir. İkinci aşama ölçüm tasarımının belirlenmesidir. Tasarım; gözlem verilerinin tesadüfî olarak veya araştırmacının kontrolü altında toplanmasına göre *tesadüfî* veya *sabit* bir niteliğe sahip olabilir. Tasarım, ayrıca yüzeylerin birbirleriyle ilişkisi açısından da net bir şekilde ortaya konmalıdır. Bütün yüzeyler tüm düzeyleriyle birlikte diğer yüzeylerle ilgili ise çapraz tasarım; diğer yüzeylerin bazı düzeyleriyle ilgili ise bu kez yuvalanmış tasarım söz konusudur. Üçüncü aşamada, belirlenen tasarıma uygun bir biçimde veriler toplanır ve varyans bileşenleri istatistiksel analiz yöntemi uygulanır. Aşağıdaki bölümde önce bu aşamalar üzerinde durulmuş, daha sonra istatistiksel analiz programı SPSS'te analizin nasıl yapılacağı ve çıktılarının nasıl yorumlanacağı konularına değinilmiştir.

---

[c] TYVA (ANOVA) analizindeki "faktörler" ile genellenebilirlik kuramındaki "yüzeyler" aynı anlamdadır. Yüzey kavramı, potansiyel değişkenlik (varyans–hata) kaynağı anlamında kullanılmıştır.

Evren ve yüzeyler. En geniş anlamıyla evren, ölçüm alanıdır. Evren, ölçüm yapan ve karar vermek isteyen bilim adamının tanımlamasına göre basit veya karmaşık; homojen veya farklılaşmış; küçük veya büyük olabilir.[36] Genellenebilirlik kuramı, ölçüm sonunda elde edilen kişilere ait puanların "kabul edilebilir gözlemler evrenine"[f] veya "gizli yapı hakkında doğru çıkarımlar yapmaya" ne ölçüde elverişli olduğuyla ilgilidir. Genellenebilirlik kuramında gözlem puanlarının kişilerin davranışlarını ne ölçüde yansıttığı ve bu puanların ne ölçüde onların daha sonraki davranışlarına genellenebileceği konusu üzerinde durulur. Kabul edilebilir gözlemler evreni, yüzeylerle tanımlanır. Yüzeyler, potansiyel hata kaynağı olan faktörlerdir. Evrende ölçümü etkileyebilecek pek çok yüzey söz konusu olabilir. Hata kaynağı olan yüzey sayısını artırmak ölçümün hassasiyetini artırmaz, tam tersine ölçüm sonuçlarını karmaşıklaştırır. Bu nedenle araştırmacılar az sayıda yüzey ile çalışmayı tercih ederler. Psikolojik test uygulamalarında genelleme yapılan yüzeyler aşağıdaki başlıklarda ele alınmıştır:

1. Test uygulama zamanı.
2. Test yeri.
3. Testi veren değerlendiriciler.
4. Test maddeleri.
5. Test uygulama yöntemi.
6. Test boyutları.

Yüzeyler, simgesel olarak üçgen prizma, bir küp veya sekizgen prizma şeklinde gösterilebilir. Bir küpün üç temel boyutu vardır: en, boy ve yükseklik. Yüzeyler bu üç boyutta yer alır. Her bir yüzeyin kendi içinde değişkenliği söz konusu olabileceği gibi bütün yüzeylerin ortaklaşa etkileşiminden kaynaklanan *ortak yüzey değişkenliğinden* de söz edebiliriz.

Ölçüm puanlarında; test maddelerinin farklı olmasından, testin farklı iki zamanda uygulanmasından ve (farklı gözlemcilerden yararlanılmışsa) değerlendiricilerin kendilerinden kaynaklanabilen değişkenlikler ve hatalar söz konusudur. Ölçümü etkileyebilecek potansiyel hata kaynakları "yüzey" olarak isimlendirilir.

---

[f] "Kabul edilebilir gözlemler evreni" (universe of admissible observations), genellenebilirlik kuramında kullanılan özel bir terimdir. Bu terimin anlamı, mümkün olduğu kadar geniş bir çerçevede (yüzeyde) yapılabilecek tüm gözlemlerin aynı sonucu vermesi ve aynı gizli yapıyı ortaya çıkarmasıdır. Çok boyutlu uzayda sonuç çıkarma alanları, kabul edilebilir gözlemler evrenini belirler. Araştırmacının elinde gözlem yapmak için kullanabileceği birden fazla alan varsa ve bu alanlar da birbirinin yerine geçme / ikame etme özelliğine sahipse söz konusu alanların tümüne birden "kabul edilebilir gözlemler evreni" adı verilir.

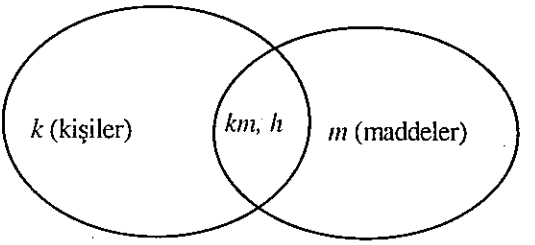
Hata yüzeylerinin düzeylerine ise "koşullar" adı verilir. Bir yüzeye ait düzey/koşul sayısı sonsuz sayıda olabilir. Örneğin, "maddeler" yüzey olarak belirlenmişse madde 1, madde 2 ... madde *n*, koşullardır. "Gözlemciler" yüzey olarak belirlenmişse her bir gözlemci, bu yüzeye ait "koşulları" tanımlar.

Yüzeyler arasındaki ilişkiler. Bilim adamı genellenebilirlik çalışmasında başlıca potansiyel hata kaynaklarını (yüzeyleri) belirledikten sonra bu yüzeyler arasındaki ilişkilerin niteliğini ortaya koyar. Yüzeyler arasındaki ilişkiler araştırmanın tasarımıyla ilgilidir. Araştırma tasarımı tek, iki, üç veya dört yüzeyli olabilir. Daha fazla yüzeyli araştırma tasarımları teorik olarak mümkün olmakla birlikte pratik bir değeri yoktur. Yüzeylerin sayısı, bilim adamının karar verirken kullanacağı ve genelleme yapacağı alana bağlı olarak belirlenir.

Tek yüzeyli tasarım. Genellenebilirlik kuramında en az bir yüzey, potansiyel değişkenlik kaynağı olarak belirlenmelidir. Söz konusu yüzey, aynı zamanda araştırmacının ölçüm sonuçlarını genellemek istediği alandır. Buna, *tek yüzeyli evren*[8] veya *tek yüzeyli tasarım* adı verilir. Diğer bir deyişle analize alınan değişkenlerden en az biri tesadüfî faktörü tanımlamalıdır. Örneğin, sadece test maddeleri arasındaki varyansın veya maddeler arasındaki iç tutarlılığın araştırıldığı bir ölçüm, tek yüzeyli bir tasarımdır. Tek yüzeyli tasarım, *k* x *m* simgesiyle gösterilir. Terimdeki *k* kişileri, *m* ise maddeleri temsil eder. Tasarımdaki *k* simgesiyle gösterilen "kişiler" yüzey olarak değerlendirilmez. Kişiler ölçüm objesidirler. Yüzey ise, bu puanları etkileyen faktörlerdir. Tek yüzeyli *k* x *m* tasarımında *kişilerin puanlarına ait varyans* σ_k^2 simgesi ile, *maddelere ait varyans* σ_m^2 simgesi ile, *kişi-madde etkileşimi*, σ_{km}^2 simgesiyle ve "artık hatalar" ise *h* harfi ile ifade edilir. Artık hatalar, araştırmaya dahil edilmeyen veya edilemeyen diğer yüzeylere ilişkindir. Bir *k* x *m* tasarımında öngörülen tesadüfî faktörlerin dışında artık hatalar da söz konusu olabileceğinden kişi – madde etkileşimine artık hatalar da dahil edilir ($\sigma_{km,h}^2$). Sonunda tek yüzeyli bir tasarımda değişkenlik yaratabilecek üç kaynak (kişi, madde, artık hata) ve her biriyle ilgili üç varyans etkeni söz konusu olur.[37] Tek yüzeyli bir uygulamada *k* x *m* tasarımının ortak değişkenlik ve etkileşim alanı "kovaryans" olarak tanımlanır (*bk.*, Şekil 1).

---

[8] *Evren* sözcüğü araştırmacının genelleme yapmak istediği ana kütleyi tanımlar.



Şekil 7-1. Tek yüzeyli bir tasarımda ortak etkileşim alanı (kovaryans).

Tek yüzeyli bir tasarımda bir kişinin bir maddeden aldığı puan (X_{km}) Eşitlik 7-4'teki formülle gösterilir.

$X_{km} = \mu$ [Genel ortalama, kişilerin evren puanlarının ortalaması]

$+ \mu_k - \mu$ [Kişi etkisi, kişiye özgü özelliklerin genel ortalamadan çıkarılması]

$+ \mu_m - \mu$ [Madde etkisi, maddeye özgü özelliklerin gen. ort. çıkarılması]

$+ \mu_{km} - \mu_k - \mu_m - \mu$. [Artık değer]

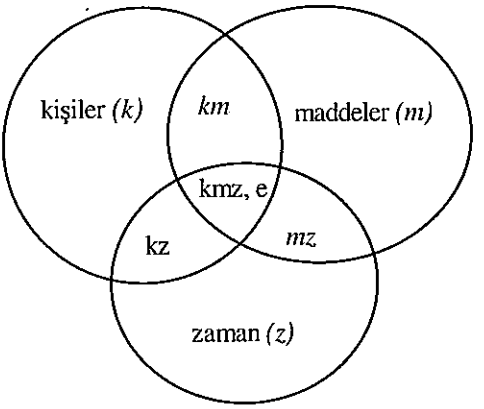
$$X_{km} = \mu + (\mu_k - \mu) + (\mu_m - \mu) + (\mu_{km} - \mu_k - \mu_m - \mu) . \quad (7\text{-}4)$$

Kişiler ve maddeler arasındaki etkileşime ait gözlem puanlarının varyansı Eşitlik 7-5 ile belirlenir:

$$\hat{\sigma}^2(X_{km}) = \hat{\sigma}^2_k + \hat{\sigma}^2_m + \hat{\sigma}^2_{km,h} . \quad (7\text{-}5)$$

Tek yüzeyli bir tasarım için varyans bileşenleri analizini uygulamaya gerek yoktur. Basit varyans analizi, tek yüzeyli tasarımda hata varyansını belirlemek için yeterlidir.

İki yüzeyli tasarım. İki yüzeyli tasarım veya bazı kaynaklarda belirtilen şekliyle *iki yüzeyli tesadüfî etki modeli,* bağımlı değişkenin dışında iki bağımsız değişkenin bulunmasını gerektirir. Bu tasarımda ölçüm yapılan kişilerin puanlarını etkileyen iki bağımsız değişken / tesadüfî faktör / yüzey vardır. Bu yüzeyler örneğin, ölçeğin *maddeleri* ve *zaman* olabilir. Model simgesel olarak, *k* x *m x z, h* şeklinde gösterilir (*bk.*, Şekil 7-2). Modelin anlamı; belirli bir *k* bireyinin, *m* maddesinden *z* zamanında aldığı ve modele alınmayan diğer hataları da içeren puandır.



Şekil 7-2. İki yüzeyli bir tasarımda ortak etkileşim alanı.

Kaynak. B. Amick, "Generalizability Theory [Genellenebilirlik Kuramı]," <http://www.sph.uth.tmc.edu:8053/behsci/lmasse/ph1130/GT_CLASSnotes6.rtf>(28.11.2002).

İki yüzeyli bir tasarımda (madde ve zaman) bir kişinin aldığı puanlar (X_{kmz}) Eşitlik 7-6'daki formülle ifade edilir.

$$
\begin{aligned}
X_{kmz} &= \mu && \text{[Genel ortalama, Evren puanı]}\\
&+ \mu_k - \mu && \text{[Kişi etkisi]}\\
&+ \mu_m - \mu && \text{[Madde etkisi]}\\
&+ \mu_z - \mu && \text{[Zaman etkisi]}\\
&+ \mu_{kz} - \mu_k - \mu_z + \mu && \text{[Kişi - zaman etkisi]}\\
&+ \mu_{km} - \mu_k - \mu_m + \mu && \text{[Kişi - madde etkisi]}\\
&+ \mu_{mz} - \mu_m - \mu_z + \mu && \text{[Madde - zaman etkisi]}\\
&+ \mu_{kmz} - \mu_{km} - \mu_{kz} - \mu_{mz} + \mu_k + \mu_m + \mu_z - \mu && \text{[Artık değer]}
\end{aligned}
\qquad (7\text{-}6)
$$

Kişiler ve maddeler arasındaki etkileşime ait gözlem puanlarının varyansı Eşitlik 7-7 ile belirlenir.

$$\hat{\sigma}^2(X_{kmz}) = \hat{\sigma}^2_k + \hat{\sigma}^2_m + \hat{\sigma}^2_z + \hat{\sigma}^2_{kz} + \hat{\sigma}^2_{km} + \hat{\sigma}^2_{mz} + \hat{\sigma}^2_{kmz,h} \; . \qquad (7\text{-}7)$$

Genellenebilirlik kuramında en fazla iki yüzeyli tasarımların kullanılmış olması nedeniyle üç yüzeyli tasarım üzerinde durulmamıştır. Bu konuya ilgi duyan okurların ilgili literatüre başvurmaları önerilir.

Tesadüfî etki – sabit etki tasarımı. Herhangi bir araştırmaya dayalı olarak yapılacak test ve ölçümlerde araştırmacının kontrolü dışında pek çok faktör tesadüfî olarak rasgele bir etkiye sahip olur. Bir gurup veriye ait kontrol altına alınamayan yorgunluk, stres, kaygı, dikkatsizlik gibi faktörler tesadüfî hata nedenidir. Kişilerin demografik özellikler açısından rasgele seçilmeleri de tesadüfî hata faktörü olarak ortaya çıkar. Tesadüfî hataları içeren veri dizisine *tesadüfî değişken* denir. Genellenebilirlik kuramında yüzeyler tesadüfî etkiye sahiptirler, çünkü tesadüfî örneklem koşullarında uygulanırlar.

Bunun yanında, bilim adamının araştırma sürecine kendisinin yaptığı müdahaleler nedeniyle de bazı değişkenlikler söz konusu olabilir. Bu tür değişkenlikleri içeren verilere ise *sabit değişken* adı verilir. Eğer araştırmacının müdahalesi de söz konusu ise böyle bir durumda hata tanımını yeniden yapmak gerekir. Bu tür araştırmalarda *sabit faktör hatası + tesadüfî faktör hatası* birlikte vardır. Araştırmacı geliştirdiği modelde, hata yüzeyini tek bir sabit değişken olarak tanımlamışsa sadece bu yüzeydeki değişkenlikle ilgilenir. Fakat, potansiyel hata içeren yüzeyleri tesadüfî faktörler olarak belirlemişse elde ettiği sonuçları bu kez tesadüfî faktörleri içeren örneklemin seçildiği ana kütleye geneller.

Çapraz – yuvalanmış veri tasarımı. Genellenebilirlik kuramında veriler iki şekilde toplanır ve bilgisayara da buna uygun bir şekilde girilir. Model eğer eşit düzeyli olmayan faktörlerden (bir yüzeydeki koşullar diğer yüzeydeki bütün koşullar için eşit değilse) oluşmuşsa yuvalanmış tasarım, tam tersine eşit n x k faktöründen oluşmuşsa bu kez çapraz tasarım tercih edilir.

Çapraz tasarım. Bu kodlamada bir yüzeydeki bütün koşullar aynı zamanda diğer yüzeydeki bütün koşullar için de geçerlidir. Çapraz tasarımda her bir kişi için birden fazla puan vardır. Bir kişiye yönelik olarak ya birden fazla zamanda ölçüm yapılmıştır, ya birden fazla madde vardır veya kişi birden fazla gözlemci tarafından değerlendirilmiştir. Örneğin, iki öğretim üyesinin yüksek lisans sınavına giren ve 10 klasik sorudan oluşan bütün öğrencilerin kağıtlarını farklı iki zamanda değerlendirmeleri tipik bir çapraz tasarım uygulamasıdır (k x z x m). Genellenebilirlik kuramında, bağımlı değişken olarak iki şıklı maddelerin toplam puanı değil ortalaması alınır. Cevapların doğru veya yanlış olmasına göre 0 ve 1 şeklinde kodlanan tek yüzeyli bir çapraz tasarım örneği Tablo 7-3'deki gibidir.

Tablo 7-3. Çapraz Tasarım Örneği

Kişiler	Madde 1	Madde 2	Madde 3	Madde 4	Ortalama[a]
1	1	0	1	0	,50
2	1	1	1	1	1,00
3	0	1	1	0	,50
4	1	0	0	0	,25
5	0	1	0	0	,25
6	0	0	0	1	,25
7	0	0	1	0	,25
8	1	1	1	1	1,00
9	0	0	0	0	,00
10	0	0	1	0	,25
11	1	1	1	0	,75
12	0	1	0	0	,25
13	1	0	0	0	,25
14	0	0	0	0	,00
15	0	1	0	0	,25
Madde ort.	,55	,50	,50	,20	,25

[a] Kişilerin ortalaması, bireyin evren puanına işaret eder. Aslında evren puan, kabul edilebilir tüm yüzeylerden bir kişinin alabileceği ortalama değerdir. Ancak tüm yüzeylerde ölçüm yapılamadığından ortalama gözlem puanı evren puan gibi kabul edilir.

Yuvalanmış Tasarım. Kişiler farklı koşullar altında ölçüme tabi tutulmuşlarsa bu tasarım "yuvalanmış" olarak isimlendirilir.[38] Tek yüzeyli tasarım, her bir ölçümde farklı değerler alan tek bir faktörün veya yüzeyin olduğunu belirler. Psikolojik araştırmalarda üç tür yuvalanmış tasarım kullanılır: hiyerarşik, bölünmüş fidelik (split plot)[h] ve tekrarlanan ölçümler. Bu tasarımların esas amacı, ölçüm

[h] Bölünmüş fidelik (split plot). Tarım kökenli bu kavram, çok sayıda bloktan oluşan bir tarım alanında bir bloğa ait fideliğin veya yatağın tekrar bölünmesi anlamına gelir. İstatistikte ise, çok değişkenli faktöriyel tasarımlarda en az iki faktörün kendi içinde ana ve daha sonra alt işlevlere ayrılmasıdır. Örneğin üç ana blok, ikişer fideliğe ve her bir fidelik de tekrar ikişer alt fidelik bölümlerine ayrılabilir. Bir okulda öğrenciler sınıflara (blok), sınıflar gündüz ve gece eğitimi gruplarına (fidelik) ve her bir gruptaki öğrenciler de erkek ve kız öğrenciler alt gruplarına (alt fidelik) ayrılabilir.

yüzeylerindeki kontrol altına alınamayan değişkenliği ortadan kaldırmaktır. Yuvalanmış tasarımlar, bir faktöre ait yüzeyler diğer faktörün tüm yüzeylerinde temsil edilmediği zaman kullanılır.[39] Örneğin, bir araştırmacı öğrencilerin puanlarını cinsiyet ve sınıf bazında değerlendirmek isteyebilir. Cinsiyet faktörü öne çıkarıldığında kişilerle ilgili farklı iki ölçüm yapılmış olduğundan *yuvalanmış tasarım* şartları geçerlidir. Araştırmacı, sınıflara göre cinsiyet faktörünün sonuçlarında ortaya çıkabilecek değişkenliği görmek isteyecektir. Yuvalanmış tasarım simgelerle aşağıdaki gibi gösterilir:

$S\ (c)\ m \times k$	Parantezli gösterim biçimi (s = sınıf, c = cinsiyet, m = maddeler).
$s{:}c \times m \times k$	İki nokta işareti kullanılarak yapılan gösterim biçimi.

Tek yüzeyli, yuvalanmış tasarımda birinci değişken gruplama değişkeni olarak tanımlanır. Bu bir bağımsız değişkendir. Gruplama değişkenlerinden sonra gelen değişken ise kişilerin ölçüm puanlarını gösterir. Bu puanlar çoğunlukla bağımlı değişkeni tanımlar. İki yüzeyli yuvalanmış tasarımda ise iki tane bağımsız değişken vardır. Tablo 7-4'te yuvalanmış değişkenlerin SPSS'e tanımlanmasına ilişkin bir örnek verilmiştir:

Tablo 7-4. İki yüzeyli Yuvalanmış Tasarım

Kişi no	Sınıfı	Kızlar/erkekler	Test puanları
1	1	1	4
2	1	1	6
3	2	1	3
4	2	1	4
9	3	1	8
10	3	1	7
11	4	1	6
14	4	1	9
18	1	2	1
19	1	2	2
23	2	2	2
25	2	2	4
27	3	2	5
30	3	2	6
31	4	2	9
32	4	2	10

Analiz. Genellenebilirlik kuramında *evren puanı*, ölçüm nesnesinin (kişinin μ_k) bütün yüzeylere ait koşulların kombinasyonundan alabileceği ortalama gerçek puandır. Evren puanı varsayımsal bir değerdir, gerçekte böyle bir puan yoktur. Ölçümlerden sadece gözlem/test puanları elde edilir. Bu test puanları değişik yüzeylere ait koşulların ortalama değeridir ve evren puanı tahmin etmek için kullanılır.

Genellenebilirlik kuramına göre güvenilirlik analizinin yapılması için faktör temelli, TYVA Bileşenleri (ANOVA Components) tekniği uygulanır. Varyans Bileşenleri hata yüzeylerindeki varyans değişkenliğini tahmin etmek için kullanılan bir teknik olarak tanımlanmıştır.[40] Bu teknikte, bir bağımsız değişkenin bağımlı değişken üzerindeki etkisinin diğer yüzeylerdeki faktörlerden de (veya diğer bağımsız değişkenlerden de) etkilenerek ortaya çıkacağı varsayımından hareket edilir. Varyans bileşenleri yönteminin uygulanabilmesi için istatistiksel analiz programına öncelikle bağımlı ve bağımsız değişkenlerin (yüzeylerin) tanıtılması gerekir. Yüzeyler kategorik değişkenlerdir; gözlemcilere, zamana veya kıdeme ait sınıflandırma gruplarını tanımlar.

Genellenebilirlik analizi, istatistik analiz programı SPSS'te Analyze mönüsü altında General Linear Models, düğmesinden alt program Variance Components şıkkı seçilerek yapılır. Açılan pencerede önce bağımlı değişken tanıtılır. Bağımlı değişken tek bir tanedir ve ölçümü yapılmak istenen hedefi belirler. Bu hedef, bir tutum ölçeğinin toplam puanı, psikometrik testin toplam puanı, çoktan seçmeli bir bilgi testinin ortalama puanı veya bir konuya ilişkin indeks değeri olabilir. Daha sonra bağımlı değişkene etki eden / etkileyen sabit faktörler ve tesadüfî faktörler tanıtılır. Sabit değişken yoksa sadece tesadüfî faktörler tanıtılır.

Sabit faktörler; bir müdahale, tedavi veya etki ile araştırmacının kontrolü altında bulunan ve araştırmacının etkilediği değişkenlerdir (zaman, yorgunluk, ışık, açlık, tokluk, mekan vb. gibi). Araştırmacı geliştirmiş olduğu modele ve araştırma hipotezine göre gerektiğinde birden fazla sabit faktör belirleyebilir. Öte yandan, tek tesadüfî hata yüzeyine sahip bir tasarımda model, sabit değişken içermez. Örneğin, sadece maddeler arasındaki iç tutarlılığı belirlerken sabit faktör kullanılmaz.

İstatistik yazılımı SPSS'te sabit faktörlerin hemen arkasından tesadüfî faktörler tanıtılır. Tesadüfî faktörler araştırmacının kontrolü altında olmayan araştırma sonucunda tesadüfî olarak değişik değerler alabilen değişkenlerdir. Sabit faktör olmadan hesaplama yapılabilirken, tesadüfî değişken kullanılmadan hesaplama yapılamaz. Örneğin, araştırmaya katılan kişilerin özelliklerine ve araştırmanın yapılış biçimine ilişkin (cinsiyet, yaş, eğitim durumu, memleketi, sektörü, uygulanan zaman gibi) değişkenler, tesadüfî faktörleri tanımlar. İstatistik yazılımı SPSS'te tesadüfî değişkenlerin sayısı belirli bir miktarı aştığında bilgisayarda

bellek yetersizliği nedeniyle hesaplama güçlüğü ile karşılaşılabilir. Uzayda çok sayıda *potansiyel değişkenlik yüzeyi* vardır ve bu nedenle çok sayıda tesadüfî değişken belirlenebilir, ancak araştırmacı bunların içinden genelleme yapmak istediği alanla ilgili en *önemlilerini* analize almalıdır. Bu anlamda tesadüfî değişken sayısı genelleme yapılmak istenen birkaç yüzey sayısıyla sınırlandırılır.

Araştırmacı geliştirmiş olduğu modelde bağımlı değişkenleri etkileyen faktörlerin tamamının tesadüfî faktörlerden kaynaklandığını düşünüyorsa *tesadüfî etki modelini* kullanır. Bağımlı değişkeni etkileyen faktörlerin bir kısmı tesadüfî ve diğer kısmı ise sabit faktörlerden oluşmuşsa böyle bir durumda *karma etki modelinden* yararlanır.[41] Değişkenlerin tanımlanmasından sonra istatistik yazılımı SPSS'te Options düğmesine basılır. Bu düğmenin altında dört farklı seçenek vardır ve bunlar aşağıdaki gibidir:

1. MINQUE.
2. ANOVA.
3. Maximum likelihood.
4. Restricted maximum likelihood.

Minque. İngilizce *Minimum Variance Quadratic Unbiased Estimators* ifadesindeki kelimelerinin baş harflerinden meydana gelmiştir. Bu ifade "Yansız En Küçük Hata Tahmin Değeri" anlamındadır. Birleşik verilerin normal dağılım özelliği gösterip göstermediğini belirlemek için kullanılan bir araçtır. Yansız bir hata tahmini yapmak için en küçük değerin ne olabileceğini gösterir. Güvenilirlik veya genellenebilirlik analizinde bu model kullanılmaz.

Maximum likelihood (en yüksek olasılık)[i]. Gözlem değerlerinde, çıkabilecek *muhtemel en yüksek olasılık* durumunu saptamaya çalışan bir hesaplama yöntemidir. Tekrarlanarak yapılan işlemler sonucunda belirlenir. Bu yöntem güvenilirlik analizlerinde kullanılmaz.

Restricted maximum likelihood (sınırlandırılmış en yüksek olasılık). Önceki yöntemde herhangi bir kısıtlamaya gidilmeden hesaplama yapılırken bu yöntemde belirli kısıtlamalar altında parametrenin gerçekleşme olasılığı tahmin edilmeye çalışılır. Yöntem, tesadüfî etkiye bağlı varyans bileşenleri tahmininin sabit etki modeline göre hesaplanmış artık değerlere dayandırılmasını esas alır. Bu

[i] Ana kütle parametrelerini tahmin etme yöntemidir. Bazı kaynaklarda "doğrusal en çok olabilirlik yöntemi" olarak çevrilmiştir.

yöntem güvenilirlik analizlerinde kullanılmaz. (Modelin oldukça ayrıntılı matematiksel formülleri için ilgili literatüre bakınız.)

ANOVA. Genellenebilirlik analizinde ANOVA düğmesinden yararlanılır. Bu düğme seçili hale getirildiğinde kareler toplamı (sum of squares) bölümü altında Type I ve Type III seçenekleri görülür. Bunlardan Type I araştırma modeli hiyerarşik modele uygunsa seçili hale getirilir. Veriler çok düzeyli veya Hiyerarşik Doğrusal Modele (HDM)[j] uygunsa tercih edilir. Tip I modelinde, kareler toplamındaki artık değerlerin düzeltilmesi amaçlanır. ANOVA Tip I (hiyerarşik kompozisyon) bütün etkilerin tesadüfî ve örneklem büyüklüklerinin eşit olduğunu varsayar. Tasarım dengeli ise ve *nk* hücrelerinde hiç bir eksik veri yoksa bu model uygulanır.[42] ANOVA Tip III ise, karma etki modelini tanımlar. Hücrelerinde eksik veri bulunmamakla birlikte veri yapıları dengesiz olarak yuvalanmış tasarımlarda kullanılır. İstatistik analiz programı SPSS'te Genel Sayısal Model (GSM) için Tip III seçeneği, kullanılan sürüm numarasına göre değişmekle birlikte önceden tanımlanmış olabilir. Bu modelde ilgilenilmesi gereken artık değer bulunmadığı varsayılır.[43]

İstatistiksel analiz programı SPSS'te değişkenler tanımlandıktan sonra Options düğmesinden ANOVA şıkkı seçili hale getirilir. Ayrıca Display bölümünden Kareler Toplamı ve Beklenen Kareler Ortalaması şıkları işaretlenir. Araştırmacı bundan sonra Model kartına girerek buradan Full Faktoriyel veya Model şıklarından birisini seçer. Model şıkkı seçildiğinde bağımlı değişkeni etkileyen faktörlerden hangilerini analize almak istiyorsa bunları analiz penceresine aktarır. Bundan sonraki aşamada, OK düğmesiyle hesaplamayı yaptırır.

Analizin yapılabilmesi için tüm gözlemlere ilişkin bağımlı değişkenin aritmetik ortalama değerlerinin olması gerekir. Ayrıca ideal bir ölçümde tüm test / ölçek maddelerinin güçlük derecelerinin eşit olması arzulanır. Maddelerin zorluk düzeyleri daha geniş bir dağılma sahipse bu dağılımın etkisiyle değişkenlik daha fazla olacak ve genellenebilirlik azalacaktır. Bu durum "madde koşulları varyansı" olarak isimlendirilir.[44] Hata yüzeylerinin etkileşimi sonucunda değişkenlik artarken genellenebilirlik azalır. Etkileşim etkisi ayrıca "kalanlar

---

[j] Çok düzeyli analiz modeli literatürde değişik isimlerle anılmıştır: hierarchical linear model (HLM) analysis, random coefficient model, empirical Bayes, growth curve modeling terimleri bunlar arasındadır. Bu uygulamada ölçümler birden fazla ve farklı düzeylerde yapılır. Örneğin ölçümler önce okul, daha sonra sınıf ve son olarak şube bazında yapılır. Ölçümler öğrencilerin özelliklerinin şubeye, sınıfa ve okula göre saptanması şeklinde gerçekleştirilebilir. Örneğin öğrencilerin başarıları ölçülüyorsa analiz şube, sınıf ve okul bazında yapılır. Çok düzeyli analizde verilir bilgisayara O1S1, O1S2, O1,S3, O2S1, O2S2, O2S3 şeklinde tanıtılır.

varyansı"[k] olarak rapor edilir. Verilerle modelin uyuşmaması halinde "artık hata terimi" daha yüksek çıkar ve daha düşük genellenebilirlik durumu söz konusudur. Hesaplama sonuçlarına bakarak karar verilirken nötr varyans ihmal edilerek *nispî* bir karar veya tüm varyanslar dikkate alınarak *mutlak* bir karar verilir.

Çıktılar. Hesaplama sonucunda dört tablo elde edilir. Bunlar: Factor level information, ANOVA, Expected mean squares ve Variance estimates tablolarıdır. Analiz çıktısına ilişkin yorumlar Expected mean squares tablosuna göre yapılır. Bu tablodaki Intercept tesadüfî faktörlerini varyanslarının sıfır olması halinde beklenen kareler ortalamasının ne olacağını gösterir. Expected mean squares tablosundaki Varience Compenent başlığı altında ise tesadüfî faktörün veya duruma göre tesadüfî faktörlerin varyans değeri (değerleri) gözükür. Ayrıca analizde eğer tesadüfî faktörle birlikte sabit faktör de bulunuyorsa bu iki faktör arasındaki etkileşimin varyans değerlerini de incelemek mümkündür. Varyans analizi çıktılarından güvenilirlik analizleri için esas olarak Variance Estimates başlıklı tablo kullanılır. Bu tabloda tesadüfî değişkenin varyansı, tesadüfî değişkenle sabit değişkenin etkileşimine ilişkin varyans ve hata varyansı vardır.

G ve K çalışması. Genellenebilirlik kuramında iki tür çalışma yapılır: Birincisi *genellenebilirlik* çalışması (G çalışması) ikincisi ise *karar* çalışmasıdır (K çalışması). Genellenebilirlik çalışmasında analize alınan hata yüzeyleriyle ilgili olarak belirli bir ölçüm prosedürü sonunda varyans bileşenleri tahmin edilmeye ve bu yüzeylerin birbirleriyle olan etkileşimleri hakkında bilgi sağlanmaya çalışılır. Ölçümü etkileyen varyans bileşenlerini tahmin edebilmek için çok sayıda hata faktörü üzerinde durulur. Bir anlamada, G çalışması bir ölçüm sürecinin pilot araştırması gibi düşünülür. Genellenebilirlik, çok sayıda araştırma sonucunda daha rafine hale gelir. G çalışmaları değişik yüzeylerdeki etkilerin birbirine karışmasını önlemek için genellikle çapraz tasarım şeklinde gerçekleştirilir.

Karar çalışmaları bölümünde ise, ölçüm nesnesiyle ilgili olarak verilmesi gereken kararlar üzerinde durulur. Araştırmacı, bu konuda genelleme yapmak istediği evreni tanımlar. Gözlem verilerine dayalı olarak sonuçları ya bütün yüzeylere veya sadece belirli yüzeylere genelleştirir. *K* çalışmasında kararlar tek bir gözlemin sonuçlarına bağlı olarak değil, çok sayıda gözlem sonuçlarının ortalamasına dayalı olarak verilir. Örneğin, *n* sayıdaki kişinin madde ve zaman ortalamaları alınır ve bu durum simgesel olarak X_{MZ} şeklinde gösterilir. G kuramına göre davranışsal ölçümlerde araştırmacı iki tür karar verir: Nispî karar ve mutlak

---

[k] Kalanlar varyansı (residual variance).

karar. Nispî kararlar, kişilerin birbirleriyle karşılaştırılmasına dayalı olarak verilir. Norm referanslı testlerde *nispî kararlar* kullanılır. Araştırmacı karar verirken normları göz önünde bulundurur. Mutlak kararlarda ise, bizzat deneklerin performansı göz önünde bulundurulur ve bu performans *kriter değerle* karşılaştırılır, diğer deneklerin puanlarına veya ne yaptıklarına bakılmaz. Kriter referanslı testlerde *mutlak karar* biçimi kullanılır.

Güvenilirlik katsayıları. Genellenebilirlik kuramının amacı, varyans bileşenleri ile ölçüm hatalarının önemini vurgulamak olmakla birlikte, klasik test kuramındaki güvenilirlik katsayısına benzer şekilde bu kuramda da özet katsayılar elde edilir. Kuramda belirlenmiş olan formüller çerçevesinde, nispî kararlar için *genellenebilirlik katsayısı* ve mutlak kararlar için ise, *dayanıklılık indeks* değerleri elde edilir.[45]

Genellenebilirlik katsayısı. Genellenebilirlik katsayısı (ρ) klasik test kuramındaki güvenilirlik katsayısına benzer ve her bir yüzey için ayrı ayrı hesaplanabilir. Ancak biz ölçümlerde *kişilere ait katsayıyla* ilgileniriz. Bu katsayı, kişilerin puanlarındaki sistematik hatayı veya değişkenliği gösterir. Genellenebilirlik katsayısı Eşitlik 7-8'deki formülle hesaplanır:

Genellenebilirlik katsayısı = Bağımlı değişkenin varyansı / (bağımlı değişkenin varyansı + etkileşimin varyansı).

$$G_{GK} = \sigma^2_s / (\sigma^2_s + \sigma^2_{si,e}) . \qquad (7\text{-}8)$$

G_{GK} = Güvenilirlik katsayısı.
σ^2_s = Bağımlı değişkenin varyansı.
$\sigma^2_{si,e}$ = Etkileşime katılan hata yüzeylerinin varyansı.

Dayanıklılık indeksi. Dayanıklılık indeksi de *G* katsayısına benzer ve Eşitlik 7-9'daki formülle hesaplanır.

■ Dayanıklılık indeksi = Bağımlı değişkenin varyansı / (bağımlı değişkenin varyansı + bağımsız değişkenlerin varyansı + etkileşimin varyansı).

$$\Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Lambda} . \qquad (7\text{-}9)$$

Φ = Her hangi bir tasarım veya evreni tanımlayan genel simge.
$\sigma^2_\Lambda = \sigma 2_i + \sigma 2_{si,e}$ (bağımsız değişken ve etkileşimin varyansı).

Shavelson and Webb (1991) genellenebilirlik katsayısının en az ,80 olması gerektiği belirtilmişlerdir (aktaran Titus).[46] Öte yandan ANOVA'nın varyans bileşenleri analizinin genelde yanlı olduğu ve bazen sonucun (varyans, tanımı gereği sıfır veya pozitif olması gerektiği halde) negatif çıkabildiği bildirilmiştir.

Analiz sonuçlarının güvenilirliği. Genellenebilirlik kuramı çerçevesinde yapılan analiz sonuçları ile elde edilen dayanıklılık indeksi ve güvenilirlik katsayısı başka örneklemler ve başka çalışmalarda aynı sonuçları vermeyebilir. Neden – sonuç ilişkilerinin incelendiği araştırmalarda, "kesinlik" daha önemlidir. Bilim adamı araştırma sonuçlarına dayalı olarak genelleme yapmak istediği zaman kesinlikten bir ölçüde uzaklaşır. Kesinliğe önem verilip kesinlik sağlandığında ise bu kez genelleme yapmak zorlaşır. Kesinlik ve genellenebilirliği eş zamanlı olarak sağlamak çok zordur, çünkü birinin üzerinde odaklanma diğerinin güvenilirliğini azaltır. Bu güçlüğü yenmenin bir yöntemi, çok sayıda gözlem yapmak ve birden fazla örneklem üzerinde çalışmaktır.[47]

ALINTI YAPILAN KAYNAKLAR

[1] J. Nagy, "Assumptions of One-Way Analysis of Variance [Tek Yönlü Varyans Analizinin Varsayımları]," <http://lsvl.la.asu.edu/bio415/jnagy/documents/Notes/PDF/Lecture%2016.pdf> (09.07.2003).

[2] C. Helberg, "Pitfalls of Data Analysis [Veri Analizinde Yapılan Yanlışlıklar]," <http://my.execpc.com/~helberg/pitfalls/> (14.03.2003).

[3] L.A. Becker, "One-Way ANOVA [Tek Yönlü Varyans Analizi]," <http://web.uccs.edu/lbecker/ SPSS/ glm_1way.htm> (14.03.2003).

[4] Tom Triggs ve Simon Moss, "Homogenitiy of Variances and Covariance [Varyansların ve Kovaryansların Türdeşliği]," <http://www.med.monash.edu.au/psych/research/rda/Homogeneity%20of%20covariance.htm.> (09.07.2003).

[5] Nagy, "Assumptions of One-Way ."

[6] D. Garson, "Student's t-Test of Difference of Means [Ortalamalar Arasındaki Farklılıklar İçin Student t-Testi]," <http://www2.chass.ncsu.edu/garson/pa765/ttest.htm.> (09.07.2003).

[7] Helberg, "Pitfalls of Data."

[8] L. Sherry, "Repeated Measures and Expected Mean Squares Tekrarlanan Ölçümler ve Beklenen Kareler Ortalaması]," <http://carbon.cudenver.edu/~lsherry/rem/ems.html> (03.04.2003).

[9] D. Howell, "Repeated Measures Analysis of Variances [Tekrarlanan Ölçümler Varyans Analizi]," <

[10] T. Baguley, "An Introduction to Sphericity [Küreselliğe Giriş]," <http://www-staff.lboro.ac.uk/~hutsb/Spheric.htm> (15.05.2004).

[11] D. Garson, "GLM: MANOVA and MANCOVA [Tek ve Çok Değişkenli Varyans Analizi]," <http://www2.chass.ncsu.edu/garson/pa765/manova.htm> (16.05.2004).

[12] T. Baguley, "An Introduction to Sphericity [Küreselliğe Giriş," <http://www-staff.lboro.ac.uk/~hutsb/Spheric.htm.>

[13] A. Field, "Sphericity [Küresellik]," <http://www.cogs.susx.ac.uk/users/andyf/teaching/pg/glm3/sld006.htm> (07.07.2003).

[14] A. Field, "Sphericity [Küresellik]," <http://www.cogs.susx.ac.uk/users/andyf/teaching/pg/glm3/sld006.htm> (07.07.2003).

[15] J. Newsom, "Some Comments and Definitions Related to the Assumptions in Within-subjects ANOVA [ANOVA Testinin Ön Kabullerine İlişikin Yorumlar]," <http://www.ioa.pdx.edu/newsom/da1/ho_wsassump.doc.> (07.07.2003).

[16] A. Field, "A Bluffer's Guide to ... Sphericity [Küresellik İçin Rehber]," <http://www.cogs.susx.ac.uk/users/andyf/research/articles/sphericity.pdf> (08.07.2003).

[17] Newsom, "Some Comments."

[18] D.W. Stockburger, "Linear Models and Analysis of Variance: Concepts, Models, And Applications [Doğrusal Modeller ve Varyans Analizi: Kavramlar, Modeller ve Uygulamalar]," <http://www.psychstat.smsu.edu/multibook/multi2.htm> (16.07.2003).

[19] Gerard E. Dallal, "Fixed and Random Factors [Sabit ve Tesadüfi Faktörler]," <http://www.tufts.edu/~gdallal/fixran.htm> (19.05.2004).

[20] D. Page, "Two-way Analysis of Variance [İki Yönlü Varyans Analizi]," <http://www-pub.naz.edu:9000/~dapage/two-wayANOVA_copy.html> (22.07.2003).

[21] B.W. Bergemann, "Measuring Research Variables [Araştırma Değişkenlerinin Ölçümü]," <http://www.campbell.edu/faculty/bergemann/res10.html> (26.03.2003).

[22] "Comparison of Multiple Groups: One-way ANOVA [Çoklu Grupların Karşılaştırılması]," <http://psych.boisestate.edu/rturrisi/class_info/Advanced%20Stats/Chapter%2008%20Comparison%20of%20multiple%20groups.htm> (24.03.2003).

[23] W.G. Hopkins, "t-Test and One Way ANOVA [t-Testi ve Tek Yönlü Varyans Analizi]," <http://www.sportsci.org/resource/stats/ttest.html> (14.03.2003).

[24] W.G. Hopkins, "New Wievs of Statistics [İstatistik Hakkında Yeni Görüşler]," <http://www.sportsci.org/resource/stats/relycalc.html> (14.03.2003).

[25] D. Garson "ANOVA [Tek Yönlü Varyans Analizi]," <http://www2.chass.ncsu.edu/garson/pa765/anova.htm> (28.03.2003)

[26] J. Lethen, "The Additive Model [Birikimsel Model]," <http://stat.tamu.edu/stat30x/notes/node142.html#SECTION00111100000000000000> (22.07.2003)

[27] I. Price, "Two-Way ANOVA [İki Yönlü Varyans Analizi]," <http://www.une.edu.au/WebStat/unit_materials/c7_anova/twoway_anova.htm> (22.07.2003).

[28] G. Carey, "Multivariate Analysis of Variance: I. Theory [Çok Değişkenli Varyans Analizi: I. Kuram],"<http://ibgwww.colorado.edu/~carey/p7291dir/handouts/manova1.pdf.> (28.07.2003).

[29] A. French, J. Poulsen, ve A. Yu, "Multivariate Analysis of Variance [Çok Değişkenli Varyans Analizi]," <http://userwww.sfsu.edu/~efc/classes/biol710/manova/manovanew.htm> (18.06.2004).

[30] C.A. Wendorf, "Multivariate Analysis of Variance [Çok Değişkenli Varyans Analizi],"<http://www.uwsp.edu/psych/cw/statmanual/manovauses.html> (28.07.2003).

[31] E.L. Zurbriggen, "MANOVA: Multivariate Analysis of Variance [ÇODVA: Çok Değişkenli Varyans Analizi]," <http://psych.ucsc.edu/faculty/zurbrigg/psy214a/03MV9b.pdf> (28.07.2003).

[32] R. Gebotys, "MANOVA [Çok Değişkenli Varyans Analizi]," <http://www.wlu.ca/~wwwpsych/gebotys/manova.pdf> (18.06.2004).

[33] D. Garson, "General Linear Model (Genel Doğrusal Model)," <http://www2.chass.ncsu.edu/garson/pa765/mancspss.htm> (18.06.2004).

[34] C.H. Yu, "Reliability of Self-report Data [Kişisel Bildirimli Verilerin Güvenilirliği]," <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/memory.html> (28.03.2003).

[35] Allegheny Psychology Department, "One-Way Within Subjects ANOVA [Tek Yönlü Katılımcılar İçindeki TYVA]," <http://www.google.com.tr/search?q=cache:XNXpLJ9GwaMJ:webpub.alleg.edu/dept/psych/SPSS/SPSS1wANOVA.html+within-subjects+ANOVA&hl=tr&ie=UTF-8> (22.07.2003)

[36] G.E Matt, "Generalizabitiy Theory [Genellenebilirlik Kuramı]," <http://www.psychology.sdsu.edu/faculty/matt/Pubs/GThtml/GTheory_GEMatt.html> (20.01.2003).

[37] D.A. Tober ve Diğerleri, "Use Of Generalızability Theory In Examining The Dependability [Güvenilirliğin İncelenmesinde Genellenebilirlik Kuramının Kullanımı]," Measurement in Physical Education & Exercise Science, 3 (1999).

[38] M.V. Ohland, "Comparing the Reliability two Peer Evaluation Instruments [İki Meslektaş Değerleme Anketinin Güvenilrliğinin Karşılaştırılması]," <http://216.239.39.100/search?q=cache:WtFdYRCGRNUC:www.succeed.ufl.edu/papers/00/00072.pdf+generalizability+facet+two&hl=tr&ie=UTF-8> (20.01.2003).

[39] E.F. Connor, "Nested and Repeated Measures of ANOVA [Yuvalanmış ve Tekrarlanan ANOVA Övlçümleri]," <http://online.sfsu.edu/~efc/classes/biol458/labs/lab8/lab8.htm> (27.11.2002).

[40] R.J. Shevelson, ve N.M. Webb, "Generalizability Theory [Genellenebilirlik Kuramı]," <http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/Published%20paper%20PDF/Generalizability%20Theory_ESM_Final.pdf> (28.11.2002).

[41] Stat Soft., "Variance Compenents [Varyan Bileşenleri]," <http://www.statsoftinc.com/textbook/stvarc om.html> (11.10.2002).

[42] L. Becker, "GLM: Unequal n Design [Genel Doğrusal Model: Eşit Olmayan Tasarım]," <http://web.uccs.edu/ lbecker/spss80/glm_uneqn_80.htl> (11.12.2002).

[43] D. Nichols. "Type I, II, III Sum of squares [Kareler Toplamı, Tip I, II, III]," <http://www.math. yorku. ca/ Who/Faculty/Monette/Ed-stat/0376.html> (27.11.2002).

[44] Institute for Objective Measurement, "Generalizability Theory and Rasch Measurement [Genellenebilirlik Kuramı ve Rasch Ölçümleri]," <http://www.rasch.org/rmt/rmt151s.htm> (11.10.2002).

[45] R.J. Shevelson, "Generalizability Theory [Genellenebilirlik Kuramı]," <http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/Published%20paper%20PDF/Generalizability%20Theory_ESM_Final.pdf> (28.11.2002).

[46] J.Titus, Generalizability Theory [Genellenebilirlik Kuramı], <http://www.cquest.utoronto.ca/ env/aera/ aera-lists/aera-d/97-11/0039.html> (11.10.2002).

[47] Colorado State Un., "Generalizabilitiy: Potential Limitations [Genellenebilirlik: Potansiyel Kısıtlar]," <http://writing.colostate.edu/references/research/gentrans/com2b3.cfm> (21.01.2003).

FAKTÖR ANALİZİ VE GÜVENİLİRLİK

Faktör analizi, güvenilirlik ve geçerlilik çalışmalarında yararlanılan temel istatistikî analiz araçlarından biridir. Bu analiz sonucunda "güvenilirlik katsayısı" veya "güvenilirlik indeksi" gibi değerler elde edilmez. Tam tersine faktör analizi, ölçüm aracının veya ölçüm verilerinin faktöriyel yapısını ortaya çıkararak aştırmacıya güvenilirlik ve geçerliliğin hangi boyutlarda araştırılması gerektiğine ilişkin bir yol haritası sunar. Bu bölümde faktör analizi, genel açıklamalardan sonra iki temel başlıkta ele alınmıştır. Birinci başlıkta, ölçüm aracıyla toplanan verilerin hangi faktörleri veya bileşenleri içerdiğine ilişkin olarak yapılan keşfedici faktör analizi üzerinde durulmuş, ikinci başlıkta ise geliştirilen modelin güvenilirliğini test eden teyit edici faktör analizi konusuna değinilmiştir.

GENEL

Kullanım Amacı

Faktör analizi, çok genel bir sınıflamayla ele alınırsa; (a) bir testteki maddelerin hangi temel bileşenlere işaret ettiğini belirlemek, (b) test maddelerini etkileyen arka plandaki gizli yapıyı veya gizli değişkenleri ortaya çıkarmak, veya (c) faktörler ve değişkenler arasındaki ilişkilerin niteliğini saptamak için kullanılır. Bilim adamı eğer gizli değişkenleri belirlemeye yönelik bir çalışma yapmışsa, araştırmasını bir adım daha ileriye götürüp saptadığı veya öngördüğü gizli değişkenler arasındaki ilişkileri neden-sonuç hipotezleriyle test etmeyi düşünebilir. Faktör analizi yöntemi yaklaşık 100 yıl önce, Charles Spearman (1904) tarafından keşfedilmiştir. Spearman faktör analizini zekayı açıklamak üzere *Genel Yetenek* adını verdiği *G* Faktörünü ortaya çıkarmak için kullanmıştır. Onun *çift faktör kuramı* adını verdiği modelde zeka başlıca iki faktörle açıklanıyordu: *G* faktörü ve *S* faktörü. Bu modelde *G* Faktörü "genel zihinsel enerji" anlamında kullanılırken, *S* Faktörü, zekayı ölçmek için kullanılan test bataryasındaki her bir testin *ortak faktörün* yanında *spesifik* bir özelliği de birlikte ölçmesi anlamına geliyordu. Spearman zekayı açıklamak için özellikle *G Faktörü* üzerinde durmuştu. Burt ve

Spearman; *G* faktörünün kalıtsal olduğunu, değiştirilemeyeceğini, fakat *S* faktörünün eğitimle değiştirilebileceğini kabul etmişlerdir (aktaran, Weinberg, 2004).[1] Daha sonraki yıllarda yapılan araştırmalar *tek faktör kuramının* zihinsel yetkinliği açıklamada yeterli olamayacağını ortaya koymuş, araştırmacılar tek faktör yaklaşımının yanında zekayı daha iyi açıklamaya yardım edecek *grup faktörleri* olgusunu da dikkate almaya başlamışlardır.

Cyril Burt, Godfrey Thompson ve Spearman'ın kendisi bir süre sonra çift faktör kuramı yerine başka bir öneride bulunmuşlar ve *G* Faktörünün yanında ilgili testlerin birleşerek bir grup oluşturmasıyla ortaya çıkan "grup faktörlerinden" söz etmişlerdir.[2] Yine bu dönemde, Philip Vernon ve Burt "sözel-eğitsel" ve "düzlemsel-mekanik" yeteneklerin önemli iki *grup faktör* olduğunu belirtmişlerdir.[3] Sözel-eğitsel grup faktörü; okuma, doğru yazma, kelime hatırlama ve kelime hazinesinin zengin olması gibi yetenekleri kapsarken düzlemsel-mekanik yetenek grup faktörü; pratik beceri, şekil algısı, mekanik şekil algısı ve bazı matematik yetenekleri içine almaktadır. Burt ve Vernon "grup faktörleri"nin eğitimin fonksiyonu olduğuna inandıklarını ifade etmişlerdir. Bu yaklaşımda sözel/eğitsel ve düzlemsel/mekanik yetenekleri ölçen bileşenler *major grup faktörleri* şeklinde adlandırılırken diğer yetenekleri ölçen bileşenlere ise *minör grup faktörleri* adı verilmiştir.[4] Burt ve Vernon'un yaklaşımı yeteneklerin hiyerarşik bir şekilde sıralanmasına dayanır:

1. Genel yetenek (*g*).
2. Majör grup faktörleri.
3. Minör grup faktörleri.
4. Spesifik faktörler (*s*).

Faktör analizinin gelişim sürecinde daha sonraları E. L. Thorndike zekayı açıklamak için kullanılan "genel faktör" kuramını bütünüyle ret etmiş ve kendisine ait *çoklu faktör kuramını* formüle etmiştir. Bu kurama göre zeka birbirinden bağımsız çok sayıda ögeden meydana gelir. Söz konusu ögeler bir araya gelerek hep birlikte zekayı meydana getirir. Buna göre kişi bir faktörde çok güçlü olabilirken diğer bir faktörde oldukça zayıf olabilir, ancak söz konusu kişi yine de bazı yetenekleri rezerv edilmiş diğer kişiler kadar zekidir.[5]

Thurstone (1887-1955) 1930'lu yıllarda olguya farklı bir bakış açısından yaklaşmış ve *temel grup faktörleri* adını verebileceğimiz kendi özel yaklaşımını geliştirmiştir.[6] Bu yaklaşımda zekayı açıklayan faktörler birbirinden ilişkisiz olarak *temel zihinsel yetenekler* adı altında yedi başlıkta toplanmış-

tır.[a] Thurstone başlangıçta bu faktörlerin ilişkisiz olduğunu düşünmekle birlikte meslekî kariyerinin son yıllarında (1946) tam tersi bir görüş ileri sürerek bu faktörlerin ilişkili olduğunu ve hep birlikte Spearman'ın *G* faktörünü açıkladığını söylemiştir.[7] Günümüzde bilim adamları zekanın ölçülmesinde hangi faktörlerin temel alınması gerektiği konusunda belirli bir mutabakat içinde olmasalar da genelde çoklu faktör kuramını benimsemişlerdir.

Faktör analizinde seçilen modele göre, ya gizli değişkenlerin gözlem değişkenlerini etkilediği veya gözlem değişkenlerinin gruplaşarak temel bileşenleri ortaya çıkardığı varsayılır. Literatürde faktör analizinden değişik amaçlarla yararlanılabileceği kapsamlı açıklamalarla ele alınmıştır. Bu amaçların sadece başlıklarını vererek şu şekilde sıralayabiliriz:[8]

1. Birbiriyle ilişkili olan veri yapılarının ortaya çıkarılması.
2. Ortak faktörün / faktörlerin ortaya çıkarılması.
3. Ampirik sınıflandırma modelleri geliştirme.
4. Bağımsız faktörlere dayanan alt ölçekleri geliştirme.
5. Hipotez test etme.
6. Veri dönüştürme.
7. Olgular arasındaki ilişkileri açıklama.
8. Faktörler ve değişkenler arasında anlamlı bir görünüme sahip ilişkilerin yol haritasını çıkarma.
9. Kuram geliştirme.

Faktör analizi, ölçüm verilerinin güvenilirliğinin saptanmasıyla ilgili işlemlerde sık başvurulan bir yöntemdir. Bilim adamları güvenilirlik değerlendirmesi sürecinde faktör analizini iki şekilde kullanabilirler. Birinci yaklaşımda geliştirilen *testin/ölçeğin verilerine ait güvenilirlik* ön plandadır. Bu yaklaşıma kısaca "testin güvenilirliği" adını verebiliriz. İkinci yaklaşımda ise geliştirilen "ölçüm modelinin güvenilirliği" soruşturulur. Testin güvenilirliğini belirlemeye yönelik çalışmalar da kendi içinde tekrar iki grupta ele alınabilir. Bilim adamı birinci yaklaşımda önce alfa güvenilirlik analizi yaparak ölçeğe/teste ait sonuçların güvenilir olup olmadığını, maddeler arasın-

---

[a] Bu yedi yetenek şunlardır: (1) sözel kavrayış (kelimeleri, kavramları ve terimleri tanımlama ve anlama), (2) sözel akıcılık (kelime hazinesi, belirli konudaki kelimeleri hatırlayarak hızla yazma), (3) tümevarımsal muhakeme (sentezsel akıl yürütme, bir dizi gözlemi tanımlayan kuralları bulma yeteneği – inductive reasoning), (4) düzlemsel görselleştirme (üç boyutlu uzay ilişkilerini görme yeteneği), (5) sayısal yetenek (matematik problemlerini çözme), (6) bellek (ezberleme ve hatırlama yeteneği), (7) algılama hızı (şekilsel nesneler arasındaki benzerlik ve farklılıkları görme).

da negatif işaretli korelasyon katsayısının bulunup bulunmadığını belirler. Güvenilirlik yüksek çıkmışsa bir sonraki aşamada faktör analizini ölçeğin kaç faktör içerdiğini görmek için kullanır. İkinci yaklaşımda ise çok sayıda madden oluşan ölçeğe önce faktör analizi yapılarak ölçek veya testin tek boyutlu olup olmadığı belirlenir. Çıkan sonuca göre tek boyutu temsil eden faktöre dayalı olarak veya eğer çok boyutlu ise her bir boyuta göre Cronbach alfa değerleri ayrı ayrı hesaplatılır. Bu bölümde faktör analizinin önce *testin güvenilirliği* için ve daha sonra ise, *modelin güvenilirliği* için kullanılması konuları üzerinde durulmuş ve faktör analizi türleriyle güvenilirlik arasındaki ilişkiler incelenmiştir. *Testin güvenilirliği* keşfedici faktör analizi, *modelin güvenilirliği* ise teyit edici faktör analizi ve yapısal eşitlik modelleri başlıkları altında ele alınmıştır.

Herüstik Düşünme

Faktör analizi, *mutlak düşünme* tarzından uzak olarak bilim adamının ölçüm yapılan olguyu değişik bakış açılarından değerlendirmesine imkan sağlayan bir tekniktir. Mutlak düşünme; kesin ve belirli olmayı, netleşmeyi gerektirirken *herüstik düşünme* farklı bakış açılarından değişik görüşlerin de uygun olabileceğinin kabul edilmesi anlamına gelir. Bu bakış açısında bilim adamı kendisini belirli bir yerde konumlandırmaz. Olaylara, duruma ve koşullara göre pozisyonunu yeniden belirler. Spearman'ın *G* kuramı, faktör analizinde mutlak düşünme anlamında kesin bir teori iken, örneğin Rubenstein'in "merak" konusuyla ilgili olarak geliştirmiş olduğu yedi faktör kuramı, sadece "belirli bir bakış açısından" önemli görülen bileşenleri açıklar. Merak olgusunun ögeleri, söz konusu yedi faktörden ibaret değildir veya olgu yedi faktör yerine pekâlâ dört faktörle de açıklanabilirdi.[9] Belirli bir olgunun ölçüm verilerine veya belirli bir kavramsal yapıyı ölçmek üzere oluşturulan çok sayıda maddeye faktör analizi uygulandığında bir örneklem grubunda üç faktör çıkmışken başka bir örneklem grubunda beş faktör çıkabilir. Froman'a göre (2001) keşfedici faktör analizi kavramsal yapının arka plandaki gerçek boyutlarını tam olarak ortaya çıkarmaz.[10] Özellikle faktör analizi yapılacak maddeler iyi incelenmiş bir kuramsal temele dayanmıyorsa ortaya çıkan boyutlar gerçeği temsil etmekten uzaktır. Bu olgu Türkiye'nin şehirlerinin gruplandırılmasıyla da açıklanabilir. Şehirler, bölge temelinde 7 faktörlü olarak gruplandırılabileceği gibi Anadolu ve Rumeli olmak üzere 2 faktörlü olarak da bölünebilir. Faktör sayısı azaldıkça kuramsal açıklamalar, basitleşir ve resim daha net görülür, fakat bu yaklaşım gerçeğe çok uygun değildir. Faktör sayısı arttıkça teoriyle veriler arasında daha iyi bir uyuşum elde edilir, fakat faktör sayısı belirli bir rakamın üstüne çıktığında, örneğin 20 veya 30 faktör söz konusu olduğunda, incelenen olgunun kavranması

zorlaşır. Bu nedenle faktöriyel yapının optimum büyüklük içinde belirlenmesine çalışılır.

Herüstik yaklaşım, faktör yapılarının hiçbir zaman tam olarak ortaya çıkarılmayacağı anlamına da gelmez. Belirli bir ana kütleden seçilen örneklemlerde etnik köken, cinsiyet, yaş, meslek ve eğitim gibi ölçümü etkileyebilecek diğer ara değişkenler kontrol altına alınmışsa söz konusu farklı örneklemlerden elde edilen faktör yapılarının küçük farklılıklarla büyük ölçüde birbirine benzer çıkması gerekir.

Uygulama Sırası

Faktör analizinin klâsik test kuramına göre bir testin güvenilirliği için kullanılmasında literatürde iki farklı yaklaşım vardır. Bazı bilim adamları faktör analizinin önce yapılmasını savunurlarken diğerleri önce alfa güvenilirlik analizinin yapılmasından yana görüş bildirmişlerdir. Her iki yöntem de belirli koşullarda geçerli veya gerekli olabilir.

Önce güvenilirlik analizinin yapılması. Tek boyutlu bir ölçek geliştirmeyi hedefleyen araştırmalarda bu yönteme başvurulur. Bu tür çalışmalarda bilim adamı araştırılan faktörle veya gizli yapıyla ilgili olduğunu düşündüğü az sayıda madden oluşan (3-12 gibi) bir test/ölçek geliştirir. Kavramsal yapıyla ilgisi olduğu düşünülen veya varsayılan bu değişkenlere/maddelere "öncü değişkenler" (marker variables) adı verilir. Testin geçerlilik analizlerinden sonra bilim adamı öncü değişkenler üzerinde güvenilirlik analizleri için alfa katsayısını hesaplar. Madde sayısının az ve maddelerin homojen nitelikte olması nedeniyle güvenilirliğin normal olarak yüksek çıkması gerekir. Alfa katsayısı eğer yüksek çıkmışsa ölçek veya testin türdeş olduğuna karar verilir. Bilim adamı böyle bir durumda ayrıca faktör analizi yapmaya gereksinim duymayabilir. Bununla birlikte, yüksek alfa değerine karşın bir test veya ölçek büyük bir ihtimalle birden fazla faktöre (bileşene) sahiptir. Çünkü türdeşlik ile tek boyutluluk farklı olgulardır. Bilim adamı türdeşliğin dışında boyutsallığı da görmek istiyorsa bu kez faktör analizi yöntemine başvurur. Çünkü yüksek güvenilirlik katsayısı, tek boyutluluğun garantisi değildir. Güvenilirlik analizinden sonra yapılan faktör analizi sonucunda baskın bir faktörle karşılaşılmışsa ölçeğin yine tek boyutlu olduğuna karar verilir. Baskın faktör tanımlaması izafîdir. Bazı bilim adamları göstergelerin tek bir faktörle asgarî ,30 - ,35 faktör yüklerine sahip olması ve göstergelerin diğer faktörlere ait faktör yüklerinin ise sıfıra yakın düşük değerler olması halinde "baskın faktör" tanısı yapılabileceğini belirtmişlerdir (Norman ve Streiner, 1994, aktaran Garson).[1] Bu kitapta kavram ve rakam karmaşasına neden olmamak için bir faktörde ,40 ve diğerlerinde sıfıra yakın faktör yükü

ile karşılaşılması halinde bu oran baskın faktör için eşik değeri olarak kabul edilmiştir.

Önce faktör analizinin yapılması. Araştırmacı geliştirdiği ölçeğin çok boyutlu / faktörlü olduğunu literatürdeki kuramsal bilgilerden biliyorsa, amacı *karmaşık kavramsal yapıları* ölçen çok boyutlu bir ölçek geliştirmek ve gizli yapıları ortaya çıkarmak ise veya amacı tek boyutlu bir ölçek geliştirmek olmakla birlikte sezgisel olarak testin birden fazla ve önemli ağırlıklara sahip boyutlar içerdiğini düşünüyorsa, boyut sayısı hakkında hiçbir fikri yoksa, korelasyon matrisi verilerinin ortalaması düşükse, maddeler arası korelasyon katsayılarının bazıları negatif işaretli ise o zaman, bir dizi işlemi gerektiren güvenilirlik analizi değerlendirmelerinden önce faktör analizi yöntemini uygulamalıdır. Güvenilirlik analizleri, faktör analizi sonuçlarının alınmasından sonra belirlenen faktör yapıları dikkate alınarak her bir boyut için ayrı ayrı yapılır. Bazı araştırmacılar, alt boyutların/faktörlerin tek bir kavramsal yapıya işaret ettiği durumlarda faktörleri dikkate almadan Cronbach alfa değerini tüm maddeler için hesaplama yoluna başvurmuşlardır. Fakat, böyle bir durumda alfa değeri eğer ,90'ın üzerinde çıkmışsa bu rakam ölçeğin güvenilir olduğunu değil, kuşkulu olduğunu gösterir. Çünkü birleşik ölçümlerde alfa güvenilirlik değeri gerçek değeri olduğundan daha düşük gösterir. Düşük değer ,90 ise gerçek değer çok daha fazla çıkacak demektir. Birleşik ölçümlerde alfa değerinin bu şekilde yüksek çıkması maddelerdeki fazlalığa işaret eder. Bunun anlamı ölçeğin yapıyı geniş bir yelpazede ölçmediğidir. Araştırmacı böyle bir sonuçla karşılaşmışsa belirlenen faktörlerin tek bir yapıyı ölçtüğü iddiasından uzak olarak güvenilirliği alt faktörler bazında analiz etmelidir.[12]

Bilim adamı eğer karmaşık kavramsal yapılara dayanan çok boyutlu ölçeklerin güvenilirlik analizini birleşik olarak yapmak istiyorsa bu amaçla hazırlanmış MSP, LISREL gibi özel yazılımlardan yararlanabilir veya genel bir fikir edinmek için *birleşik ölçümlerin güvenilirlik katsayısı* formülünü uygulayabilir.

TESTİN GÜVENİLİRLİĞİ İÇİN KEŞFEDİCİ FAKTÖR ANALİZİ

Testin güvenilirliği için yapılan faktör analizinde ölçüm maddelerinin belirli faktörleri veya kavramsal yapıları temsil etme güvenilirliği konusu üzerinde odaklanılır. Bu nedenle yaklaşım "keşfedici faktör analizi" olarak isimlendirilmiştir. Bilim adamının ilgisi faktörlere dayalı "bir ölçüm modelinin" güvenilirliğini test etmek değildir. Tam tersine ölçüm değişkenlerinin ne şekil-

de gruplaştığını veya bu maddelerin arka planında hangi faktörlerin bulunduğunu görmektir.

Bir batarya[a], bir test, belirli bir madde kümesi veya bir ölçeğin kaç faktör içerdiği önceden bilinmiyorsa, bu konudaki kuramsal bilgiler yetersizse böyle bir durumda varlığından şüphe edilen "gizli değişkenler" veya ortaya çıkarılmaya çalışılan "temel boyutlar/bileşenler" keşfedici faktör analizi yöntemi ile araştırılır. Analize, "keşfedici" denmesinin nedeni literatürde konuyla ilgili kuramsal bilgilerin bulunmaması, ölçüm yapılan konunun kaç faktörden oluştuğunun önceden bilinmemesi sebebiyledir. Araştırmacı böyle bir durumda faktör sayısını *keşfetmeye* çalışacaktır. Keşfedici faktör analizi, önceki kuramsal bilgilerin teyidi amacıyla yapılmaz, tersine kuramsal bilgi yaratma amacına yöneliktir. Faktör analizinin bilimselliği, kurama yapılan katkıyla saptanır. Bu katkı ortak faktör analizinde yüksek iken temel bileşenler analizinde daha düşüktür. Keşfedici faktör analizinde bilim adamı üç temel bilgiyi kullanarak değişkenlerin yapısal özelliğini ortaya çıkarmaya çalışır. Bunlar; (a) faktör veya bileşen sayısı (total variance explained tablosu), (b) değişkenlerin faktör yükleri veya faktörle olan korelasyon katsayıları (component matrix veya factor matrix tablosu) ve (c) değişkenlerin paydaşlık oranı veya değişkenlerin çıkarılan faktörleri temsil etme oranı (communalities tablosu) değerleridir. Keşfedici faktör analizi sonucunda, öncü veya işaretleyici değişkenlerden (marker variables) bir bölümü elenerek güvenilirliği yüksek olan belirtke değişkenler (salient variables) elde edilir. Alfa güvenilirlik hesaplaması faktöriyel bazda, söz konusu belirtke değişkenlere dayalı olarak yapılır.

Uygulama Koşulları

Testlerin[b] veya test maddelerinin faktöriyel yapısı, kaç faktör içerdiği keşfedici faktör analizi yöntemiyle test edilir. Keşfedici faktör analizi bir taraftan testin içerdiği temel boyutlar hakkında bilgi verirken diğer taraftan maddelerin ve testin güvenilirliğine ilişkin bazı ipuçlarını da araştırmacıya sağlar. O nedenle faktör analizi hem ölçeğin/testin faktör yapısını ortaya çıkarmak ve hem de maddelerin güvenilirliği hakkında fikir sahibi olmak için kullanılır.

---

[a] Batarya sözcüğü genel bir terimdir. Değişik sayıda maddeden oluşan bir ölçek, "yüzey değişkenleri bataryasını" oluştururken, her biri farklı sayıda maddelere sahip bir grup bilişsel testten oluşan ölçüm aracına da batarya adı verilir. Türkçede daha çok ikinci kullanım biçimi yaygındır. Bu kitapta da *batarya* kelimesi, bir grup teste verilen ortak ad olarak kullanılmıştır.

[b] Faktör analizi yapılacak veriler Likert tipi bir ölçeğin maddeleri, sürekli veriler veya bir psikoteknik test bataryasında yer alan değişik bilişsel testlerin toplam veya ortalama puanları olabilir. Veriler eğer, Likert tipi bir ölçeğin maddelerine ait ise faktör analizi polikorik korelasyon katsayılarına dayandırılmalıdır.

Bilim adamı faktör analizi yöntemini uygulayabilmek için verilerin belirli koşulları sağlamış olduğunu kontrol etmelidir. Aşağıdaki paragraflarda bu koşullara ilişkin bilgiler verilmiştir.

Değişken sayısı. Keşfedici faktör analizi yönteminin uygulanabilmesi için Thurstone en az üç değişkenin veya bataryada en az üç test sonucunun bulunması gerektiğini bildirmiştir. Teyit edici faktör analizinde ise değişken sayısı için her hangi bir sınırlama getirilmemiştir.[13] Faktör yükü (component matrix) yüksek olan en az üç değişkene sahip olmayan faktörler yorumlanamazlar. Bu nedenle her bir faktör altında en az üç değişken bulunmalı ve bu değişkenlerin faktör yükleri de genel bir kriter olarak ,40'tan düşük olmamalıdır (Stevens, 1992, aktaran Gliem ve Gliem).[14] Test geliştirme çalışmalarına, bazı maddelerin ortak varyansının düşük olabileceği ihtimali göz önünde bulundurularak üç olması düşünülen madde sayısının en az iki veya üç katı kadar madde ile başlamakta yarar vardır. Analiz sonucunda eğer değişkenlerin faktör yükleri düşük çıkmışsa nihai ölçekte yer alması düşünülen madde sayısının altı veya 10 katı kadar yeni madde geliştirilir. Ölçek veya testteki madde sayısı arttıkça özdeğeri 1'in üzerinde olan faktör sayısı da artar. Bir ölçekteki madde sayısı 50'nin üzerine çıktığında özdeğeri 1'in üzerinde olan faktör sayısı büyük bir olasılıkla 10'u aşacaktır.

Öte yandan ölçekteki/testteki değişkenlerden bazılarını, çoklu doğrusallık özelliği nedeniyle ölçekten çıkarmak gerekebilir. Bunun için her bir göstergeye Kaiser-Meyer-Olkin örneklem uygunluğu[a] testi uygulanır. Maddelere ait KMO test sonuçlarını görebilmek için SPSS'teki mönüde anti-image şıkkı seçili hale getirilir. Teste madde ilave ederek veya madde çıkararak test KMO sonucunun ,60'ın üzerinde kalması sağlanır.

Örneklem hacminin belirli bir büyüklüğe sahip olması. Örneklem hacmi, değişken başına en az beş vak'a düşecek kadar büyük olmalıdır. Comrey ve Lee (1992, aktaran George) örneklem büyüklüğü olarak $n = 50$ rakamını çok zayıf; $n = 100$ rakamını zayıf; $n = 200$ rakamını vasat; $n = 300$ rakamını iyi; $n = 500$ rakamını çok iyi ve $n = 1000$ rakamını ise mükemmel olarak tanımlamıştır.[15] Bilim adamlarının büyük çoğunluğu 100'ün altındaki rakamları faktör analizi için yetersiz ve güvenilmez bulmuşlardır.[16] Literatürü incelediğimizde yöntem bilimcilerin örneklem büyüklüğü için değişik kurallar önerdiklerini görüyoruz. Bunlardan biri 10 kuralıdır. Buna göre de-

[a] Buradaki örneklem uygunluğu kişilere ilişkin değil, maddelere ilişkin alan örneklemesiyle ve bir bütün olan bilgisayara girilen veri matrisi verileriyle ilgilidir. Bazı kaynaklarda ise, "örneklem verilerinin uygunluğu"ndan söz edilmiştir. "Örneklem verileri" kavramı, araştırmaya katılan kişi ve değişken sayısını birlikte içerir.

ğişken başına en az 10 katılımcı bulunmalıdır. Bir diğeri 100 kuralıdır. Değişken başına ya 5 katılımcı olmalı veya en az 100 kişiye ulaşılmalıdır. Bilim adamı kendi küçük örneklem hacmini makul göstermek için literatürden örnek araştırmalar bulmaya çalışmamalı okuyucuların ve diğer araştırmacıların tereddütsüz kabul edebilecekleri örneklem büyüklükleriyle çalışmayı yeğlemelidir.

Farklı örnek kütle verilerinin tek bir havuzda toplanmaması. Bilim adamı faktör analizi yaparken birbirinden önemli ölçüde farklı özellikler gösteren örneklem verilerini tek bir havuzda toplamamalıdır. Çünkü her bir örneklem grubunda faktöriyel yapı farklı bir niteliğe sahip olabilir. Sadece ayrı örneklemlerde yapılan analizlerde faktöriyel yapıların benzer çıkması halinde söz konusu örneklem verileri tek bir havuzda toplanabilir.

Eksik veri analizi. Faktör analizi işlemine girişmeden önce bilim adamı veri matrisini gözden geçirerek eksik veri bulunup bulunmadığını incelemelidir. Faktör analizi değişkenler arasındaki korelasyonlara dayandığından eksik verilerin korelasyon katsayılarını ne şekilde etkilediğine bakmak gerekir. SPSS korelasyon matrisini, eksik verileri liste bazında veya vak'a çifti bazında silerek hesaplar. Korelasyon matrisi her iki yönteme göre iki defa hesaplatılarak sonuçlar arasında önemli bir değişikliğin ortaya çıkıp çıkmadığına bakılır. Bunun için her iki yönteme göre yapılan korelasyon katsayıları hücre bazında birebir karşılaştırmalar yapılarak incelenir. Eğer sonuçlarda önemli bir değişiklik gözükmüyorsa (,05 gibi) liste bazında silme seçeneği tercih edilerek veriler analize hazır hale getirilir. Korelasyon analizi için SPPS'te correlate – bivariate – exclude cases pairwise ve exclude cases listwise komutları çalıştırılır. Eksik verileri analizden çıkarmanın bir diğer yöntemi faktör analizi mönüsündeki Options düğmesini kullanarak Exclude cases pairwise seçeneğini işaretli hale getirmektir. SPSS'te istatistik yazılımın eksik verileri tanıyabilmesi için bu verilerin "99 – veri yok" şeklinde tanımlanması gerekir. Bunun için ilgili sütunda gri alana çift tıklanarak açılan değişken tanımlama matrisinde Missing Values başlığına gidilir. Burada üç nokta halinde görülen alana tıklanarak açılan diyalog kutusundaki herhangi bir alana 99 değeri girilir ve OK tuşuna basılır.

Vak'alarda ayrık veya uç değerlerin bulunmaması. Güvenilir bir faktör analizi için test puanları arasında uç veya ayrık değerlerin bulunmaması gerekir. Ayrık değerlerin bulunma durumu "faktör puanlarıyla" oraya çıkarılan standart *z* değerleriyle saptanır. Bir diğer yöntem Mahalanobis uzaklık ölçüsünün kullanılmasıdır. Yüksek Mahalanobis değeri bulunan vak'alara İ

kukla değişken değeri atanarak bu değişken diğer değişkenlerle regresyon analizine tâbi tutulur. Regresyon analizi sonucu anlamlı çıkmamışsa veya büyük örneklemlerde düşük R^2 değeri elde edilmişse uç değerlerin tesadüfî olduğuna karar verilir.[17] Sonuçlar tam tersine anlamlı çıkmışsa, bu kez söz konusu değerlerin faktör analizi sonuçlarını etkilememesi için uç değer içeren vak'alar analizden çıkarılır.

Kovaryans veya korelasyon matrisinden faktör çıkarılabilmesi. Faktör çıkarılması için uygun olan korelasyon matrisi, değişkenler arasında orta büyüklükte korelasyon katsayılarına sahip olan tablodur. Değişkenlerin mutlak değerleri arasındaki korelasyon katsayıları en az ,30 olmalıdır. Korelasyon katsayıları daha düşük olursa değişken sayısı kadar faktör çıkma olasılığı doğacağından faktör analizinin pratik yararı ortan kalkar. Öte yandan değişkenler arasındaki yüksek korelasyonlar *koşutluk* ve *tekillik* sorununa yol açar. Koşutluk sorunu KMO istatistiği ile çözülebilir. Eğer hesaplanan *Bartlett boyutsallık test* değeri büyükse ve Kaiser-Myer-Olkin örneklem uygunluk ölçüsü de ,60'tan büyük çıkmışsa o zaman, korelasyon/kovaryans matrisinden faktör çıkarılabileceği sonucuna varılır.[18]

Uygun olmayan maddelerin elenmesi. Faktör analizi sonuçlarının güvenilir bir şekilde kullanılabilmesi için araştırmacının uygun olmayan maddeleri eleyerek faktör analizini birkaç kez yeniden hesaplaması gerekir. Uygun olmayan maddelerle birlikte yapılan faktör analizi sonuçları gerçeği tam yansıtmaz. Uygun olmayan maddeler aşağıdaki gibi sıralanabilir.[19]

1. Çıkarılan faktörlerle herhangi bir şekilde ilişkili olmayan değişkenler/maddeler.
2. Faktörlerin dışında diğer maddelerle ilişkili olan değişkenler.
3. Birden fazla faktörle ilişkili olan değişkenler/maddeler.

Bu maddeler, faktör analizi sonuçlarını etkileyerek belirginleşmesi gereken faktörlerin yeterince güçlü bir şekilde ortaya çıkmasına engel olabilir. Bunun için korelasyon matrisi tablosu incelenerek bu tabloda ideal olarak bütün değerlerin veya en azından maddelerin %20'sinden fazlasına ait korelasyon katsayılarının ,30'u aşması gerekir. Az sayıda madde ,30'u aşmamışsa bu maddeler analiz dışı bırakılır. Uygun olmayan maddeleri belirlemenin bir diğer yöntemi anti-imaj korelasyon tablosunun köşegeninde yer alan değerleri incelemektir. Bu değerler "bireysel maddelerin örneklem uygunluğu"

olarak isimlendirilir. Köşegende yer alan maddeler eğer ,50'den küçük ise ölçek veya testten çıkarılarak analize alınmaz. Anti-imaj korelasyon tablosu aynı zamanda bize, faktörlerin ötesinde diğer değişkenlerle ilişkili olan maddeleri de gösterir. Köşegenin dışında kalan değerler eğer ,30'un üzerinde ise bu değişkenler faktörle ilişkili değildir ve bu nedenle seçilen herhangi bir madde analiz dışında bırakılır.[20]

Verilerin parametrik olması. Faktör analizi parametrik veriler üzerinde yapılır. Belirli kategoriler halinde ele alınan sınıflandırma verileri ise parametrik değildir. Sınıflandırma verileri iki, üç, dört, beş veya daha fazla dereceli olabilir. Sınıflandırma verilerinde dereceler arasındaki mesafenin eşit olması ve kategorik ölçüm verileriyle arka palanda ölçülen özellik arasındaki ilişkilerin doğrusal bir niteliğe sahip olması önemlidir.[21] Eğer dereceler arasındaki mesafe eşit değilse ölçüm değerleriyle arka plandaki ölçülen özellik arasındaki ilişkiler de doğrusal değildir. Tam tersine mesafe eşitse ölçüm verileriyle gizli değişken arasındaki ilişkiler doğrusal bir niteliğe sahip olacağından ölçüm verileri yüksek derecede basık veya çarpık çıkmaz.[22]

Sıralı ölçeklerde, sadece 7 dereceli Likert ölçeklerinin eşit aralıklı olma özelliği yüksektir. Bu nedenle yedi dereceli ölçekler genellikle sürekli ölçek verisi olarak kabul edilir. Ancak buna rağmen veriler eğer çok değişkenli normal dağılım özelliği göstermiyorsa "normal dağılım özelliği göstermeyen sürekli değişkenler için" uygulanan istatistik analiz yöntemlerine başvurulur. Araştırmacı verilerin çok değişkenli normal dağılım özelliğini incelemek için ilk aşamada her bir değişkeni tek tek ele alır. Her bir değişkenin normal dağılım özelliğini, basıklık ve çarpıklık değerlerini değerlendirir. Çarpıklık değeri, çarpıklığın standart hatası (Std. Error of Skewness) 2 ile çarpılarak elde edilecek artı, eksi değer içinde kalıyorsa dağılımın yaklaşık olarak normal olduğuna karar verilir. Aynı şekilde basıklık değeri, eğer basıklığın standart hatası (Std. Error of Kurtosis) 2 ile çarpılarak elde edilecek artı, eksi değer içinde kalıyorsa dağılımın yine yaklaşık olarak normal olduğuna karar verilir. (Çarpıklık ve basıklık değerlerinin örneklem büyüklüğünden etkilendiğine dikkat etmek gerekir. Örneğin, 50'den küçük örneklem verilerinde çarpıklık ve basıklık katsayılarına güvenilmez.) Bu incelemede çarpıklıktan çok basıklık derecesi üzerinde durulur.

Maddelerin tek tek dağılımları normal değilse büyük bir ihtimalle çoklu dağılımları da normal çıkmayacaktır. Bununla birlikte, maddeler tek tek bakıldığında normal dağılım özelliğine sahip olsa bile çoklu dağlımda normal olmayan dağılım özelliği gösterebilir. Bu nedenle araştırmacı sadece tekli normal dağılım özelliğini değil, çoklu normal dağılım özelliğini de araştırmalıdır.[23] Maddelerin çoklu dağılım özelliği EQS ve Lisrel gibi yazılımlarla

incelenebilir. SPSS ve SAS yazılımlarında ise çoklu normal dağılım özelliğini belirlemek için özel makrolar geliştirilmiştir.

Araştırmacı beşli veya yedili Likert ölçeği kullanıyorsa faktör analizi için Lisrel'deki polikorik korelasyonla birlikte Ağırlıklandırılmış En Küçük Kareler (weighted least squares – WLS) yöntemini uygulamalıdır. Ancak tatmin edici bir sonuç alabilmek için örneklem büyüklüğünün 2000 gibi büyük rakamlar olması gerekir. Bir diğer yaklaşım, Muthen (1993) tarafından önerilen Mplus yazılımındaki Kategorik Değişken Modeli'ni (categorical variable model – CVM) uygulamaktır. Bu yaklaşımda verilerin ikili veya çok dereceli olması fark etmemektedir. Bununla birlikte literatürde Likert ölçeklerine faktör analizi yöntemini uygulayan ve bunu normal karşılayan bilim adamları da vardır. Bu konuda "Verilerin Metrik Olması" başlığında gerekli açıklamalar yapılmıştır.

Türleri

Keşfedici faktör analizinin hesaplama biçimine dayanan değişik türleri vardır. Ülkemizde sık kullanılan istatistiksel analiz programı SPSS'te faktör analiziyle faktör çıkarma işlemi yedi farklı şekilde yapılır:[a]

1. Temel bileşenler analizi.
2. Ortak faktör analizi.
3. Maksimum olasılık yöntemi.
4. Ağırlıklandırılmamış en küçük kareler yöntemi.
5. Alfa yöntemi.
6. İmaj faktör yöntemi.
7. Genelleştirilmiş en küçük kareler yöntemi.

Keşfedici faktör analizinde hangi yöntemin kullanılacağı araştırmacının varsayımlarına ve amacına göre değişir. Bilim adamı öncelikle şu soruya yanıt vermelidir: Benim amacım nedir? Bu soruya iki şekilde cevap verilebilir. Birinci cevap, "kurama katkı yapmak için değişkenlerin arka planındaki gizli yapıları ortaya çıkarmak veya gizli yapıları doğrulamak" şeklindedir. İkinci cevap ise, "gözlem değişkenlerini farklılaştıran temel boyutları belirlemek" olabilir. Temel boyutlar/bileşenler bir yapının parçalarıdır. Yapı o

[a] İstatistik yazılımı SAS'ta faktör çıkarma işlemi biraz daha farklı yöntemlerle açıklanmıştır: temel bileşenler analizi, temel faktör analizi, tekrarlanmış temel faktör analizi, ağırlıklandırılmamış en küçük kareler yöntemi faktör analizi, maksimum olasılık faktör analizi, alfa faktör analizi, imaj bileşenler analizi ve Harris bileşenler analizi.

parçalarla birlikte anlamlılık kazanır. Yukarıda açıklanan yedi yöntemden ilk ikisi bu cevapların karşılığını temsil eder. Diğer yöntemler ise değişik varsayımların dikkate alındığı hesaplama teknikleriyle alakalıdır. Aşağıdaki paragraflarda bu hesaplama yöntemlerine ilişkin bilgiler verilmiştir.

Temel bileşenler analizi. Bu analiz, indeks türü ölçüm araçlarında gözlem değişkenlerini farklılaştıran "temel boyutları" ortaya çıkarır. Bileşenler, boyutlar veya kavramsal yapının parçalarıdır. Bazı bilim adamları "bileşenler"in faktör olarak nitelendirilmesine karşı çıkmışlar ve hatta *temel bileşenler analizinin* faktör analizi olmadığını belirtmişlerdir. Temel bileşenler analizinin kökleri 1890'lı yıllara kadar uzanır. O yıllarda Galton "gizli özellik" kavramını ortaya atmış ve daha sonra kavram 1901'de Pearson tarafından bilim dünyasına tanıtılmıştır.

Hotelling (1933) ise temel bileşenler analizinin geliştirilmesini sağlamıştır. Bilim adamı, ölçtüğü konunun temel boyutlarını, ögelerini ortaya koymak istiyorsa, çalıştığı veriler an azından eşit aralıklı ölçek niteliğinde ise, verilerde hata varyansı düşükse ve esas amacı bir ölçek, test geliştirmek veya ölçüm maddelerinin hangi başlıklar/boyutlar altında gruplanabileceğini saptamaksa *temel bileşenler analizi* istatistik yöntemini kullanır. Keşfedici faktör analizinde, her bir *temel bileşen* veri grubundaki gözlem değişkenlerinin *doğrusal*[a] *bileşik değişkenler* halinde gruplaşmasını temsil eder ve orijinal değişkenlerin doğrusal kombinasyonlarını temsil eden birbiriyle ilişkisiz[b] yeni değişkenler ortaya çıkarır (Newcomer, 1984, aktaran Cleland).[24] Orijinal değişkenlerin birbiriyle ilişkili olmalarının tam tersine temel bileşenler birbirlerinden bütünüyle bağımsızdır. Temel bileşenler analizi, değişkenleri "temsil etme", "özetleme" ve "toplamla" olgusuyla açıklanır. Analiz, bir dizi değişkeni gruplandırarak her grubu açıklayan farklı faktörleri ortaya çıkarır. Bu nedenle bilim adamı eğer temel bileşenler analizinden yararlanmışsa yorum yaparken değişkenleri *farklılaştıran* veya *gruplayan* faktörlerden söz etmelidir. Çok sayıda faktörün/bileşenin ortaya çıkması araştırmacıya esaslı bir fayda sağlamaz. Maddelerde hasislik veya maddelerin basitleştirilmesi ancak, önemsiz temel bileşenlerin ayıklanması ve toplam varyansı büyük ölçüde açıklayan önemli faktörlerin alıkonmasıyla sağlanabilir. Bununla birlikte elde edilen faktörler, söz konusu değişkenlerdeki "ortak özü" temsil etmez. Çünkü, temel bileşenler analizinde *ortak faktör anali-*

[a] Doğrusallık her bir faktörün tek bir boyutu ölçmesi anlamındadır. Doğrusallık aynı zamanda matris tablosunda vak'ların ve değişkenlerin birbirinden bağımsız olması anlamına gelir.

[b] Birbiriyle ilişkisizdir, fakat kavramsal yapıyla ilişkilidir.

zinde göreceğimiz "ortak varyans", "spesifik varyans" ve "hata varyansı" hep birlikte hesaplamaya katılır. Bu yaklaşımda spesifik ve hata varyansının faktörler arasında dağıldığı varsayılır. Temel bileşenler analizinde, gözlem maddeleri "oluşturucu" göstergeler olarak ele alınmıştır. Göstergeler, gizli bir faktörden etkilenen birimler değil, tam tersine faktörü / boyutu / bileşeni oluşturan ögelerdir. Bu nedenle ölçüm modeli diyagramında, göstergelerle faktörler (bileşenler, oluşmuş yapılar) arasındaki ilişki, göstergelerden faktörlere doğru yönelen oklu çizgilerle gösterilir. Temel bileşenler analizinde ortaya çıkan faktörler/bileşenler arasında yüksek derecede ilişki olması gerekmez. Genelde, ortaya çıkan faktörler (oluşmuş yapılar) birbirinden bağımsızdır.[25]

Temel bileşenler analizini kovaryans veya korelasyon matrisini temel alarak hesaplamak mümkündür. Eğer değişkenler benzer ölçek değerlerine sahipse bilim adamlarının çoğu kovaryans matrisini temel alırlar.[26] Analize alınan değişkenlerin ölçek değerleri birbirinden farklı ise korelasyon matrisiyle çalışılır. Temel bileşenler analizinde, faktör çıkarma yöntemi olarak *varyansı en yüksek değere getirecek döndürme* (varimax) yaklaşımı temel alınmıştır. Bu yaklaşım kısaca *varyans maksimizasyonu* olarak adlandırılır. Döndürme sonunda değişkenleri temsil eden uzaydaki noktalar faktörü temsil eden regresyon doğrusuna daha yakın bir konuma gelmiş olur. Öte yandan varyansın en yüksek değerine getirilmesi gizli değişkenin veya faktörün varyansını (değişkenliğini) artırma anlamına gelir.[27] Faktörün varyansı arttığı ölçüde değişkenleri temsil etme özelliği de artar. Temel bileşenler analizinde birinci faktör; x_1, x_2, x_3 ... x_k gibi kendisini temsil eden belirli sayıda değişkenin doğrusal kombinasyonunu ifade eder.

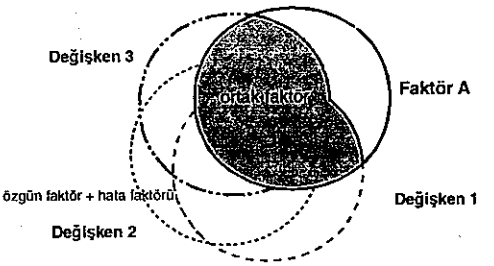
Temel bileşenler analizi, değişik amaçlarla kullanılabilir. Bir araştırmacı bu analiz yönetimiyle test ve ölçek geliştirmeyi hedeflerken, başka bir araştırmacı "bağımsız tahmin değişkeni" olarak faktörleri (bileşenleri) ortaya çıkarmayı hedefleyebilir. Temel bileşenler analizi, göstergelerin değişim aralıkları ve varyansları farklı *R* tipi korelasyon matrisleriyle çalıştığı zaman kullanır.[28] Uygulamada daha çok pazarlama araştırmalarında, yönetim ve organizasyon araştırmalarında örgütsel ve sosyal yapıları ortaya çıkarmaya yönelik olarak kullanıldığı görülür.

Ölçek verileri eşit aralıklı değilse veya Likert ölçeğinde olduğu gibi veriler eşit aralıklı olmak yerine sıralı ölçek verisi olarak kabul edilmişse bu şekildeki sıralı ölçek verilerine dayanan *oluşturucu* ölçüm modellerinde "temel bileşenler analizi" yerine *kısmî en küçük kareler analiz* tekniği kullanılır.[29] Bununla birlikte Likert ölçekleri eğer eşit aralıklı ölçek verisi olarak değerlendirilmişse yine temel bileşenler analizi yöntemi kullanılabilir. An-

cak bu yöntemin sıralı ölçek verilerinde sınırlı bilgi sağlama özelliğine sahip olduğu bildirilmiştir.

Ortak faktör analizi. C. Spearman (1904) tarafından bulunan ve yine parametrik verilerle çalışılan bu analize literatürde değişik adlar verilmiştir. Temel eksenler analizi, ortak faktörler analizi, ortak faktör analizi veya sadece faktör analizi terimlerinin hepsi aynı analiz türünü tanımlar.[a] Terimi Türkçeye daha anlaşılır bir şekilde çevirirsek "değişkenleri etkileyen ortak öz veya ortak özler analizi" olarak isimlendirebiliriz. Ortak faktör, üç veya daha fazla değişkenin birlikte içerdiği ortak özü, paylaşılan ortak kavramsal yapıyı temsil eder. Ortak faktör analizi değişkenlerin arka planındaki gizli yapıyı/yapıları ortaya çıkarmak veya apriori belirlenen faktöriyel yapıyı teyit etmek amacıyla kullanılır.

Bir değişken iki temel ögeden oluşur: *ortak parça* ve ortak olmayan *özge parça.* Ortak olmayan parça da kendi içinde iki alt bölümde değerlendirilir: *spesifik faktör ve hata faktörü.* Spesifik faktör bir tür gizli değişkendir ancak sadece tek bir değişkenin, yani kendisinin varyansını açıklar. Bir ölçekte madde sayısı kadar spesifik faktör vardır, ancak faktör analizi sonucunda spesifik faktörlerin etkileri birleştirilerek sanki tek bir spesifik faktör varmış gibi değerlendirilir (*bk.*, Şekil 8-1).[30]



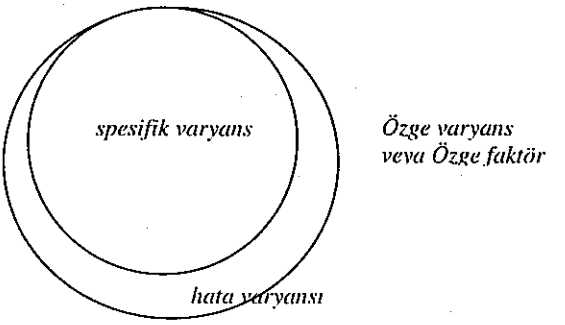
Şekil 8-1. Ortak faktör.

- Ortak faktör analizinde toplam varyans formülü.

Toplam varyans = ortak varyans + özge varyans.
Özge varyans = spesifik varyans + hata varyansı (*bk.*, Şekil 8-2).

---

[a] Faktör analizi konusunda kafa karıştıran terim, *faktör* sözcüğüdür. Bu terim gizli yapı, gizli değişken ve ortak öz anlamlarında kullanılmıştır. Bu anlam *temel bileşenler* ifadesinden farklıdır. Temel bileşenler sözcüğünde, gözlemlenebilir doğrusal bir kombinasyon vardır. Faktör sözcüğü aynı zamanda *faktör matrisi* anlamına gelir. Bu anlamda faktör analizi faktör matrisini kullanarak veri analizi yapan tüm yöntemlere verilen genel bir addır.



Şekil 8-2. Ortak parçası bulunmayan bir değişkenin spesifik varyans ve hata varyansından oluşması.

Ortak faktör analizinde bir değişkenin (veya bir bataryada yer alan bir testin) diğer değişkenlerde/testlerde de var olduğu düşünülen *ortak özü* ne ölçüde içerdiği teması önemlidir. Değişkenler tek bir öz veya birden fazla öz tarafından temsil ediliyor olabilir. Eğer birden fazla öz varsa ortak özlerden her biri farklı bir "gizli yapıyı" temsil eder. Ortak faktör analizinde sadece göstergelerdeki kovaryans (ortak değişkenlik) analiz edilir, spesifik ve hata varyansı hesaplamalara dahil edilmez.

Bir değişken, diğerleriyle paylaştığı *ortak özün* yanında ortak yönü olmayan ögelere de sahiptir. Değişkenin ortak olmayan yönü "kendine özgü bir özelliği" ölçer. Ortak olmayan yön, *spesifik öz* ve *hata* biriminden oluşur. "Spesifik öz" ve "hata" birimleri farklı faktörlerdir. Bazen değişkenin ortak olmayan yönünde hata birimi daha yüksek olabilir. Bilim adamları *spesifik öz* ile *hata birimini* birleştirerek bu ikisine birden *özge faktör* (unique) adını vermişlerdir. Özge faktör, ortak faktörün dışındaki faktör demektir ve diğer değişkenlerden bağımsız olarak sadece kendisiyle ilgilidir. Ancak araştırmalarda ölçüm yaptığımız her bir değişkeni bir çeşit operasyona tabi tutup "ortak öz", "spesifik öz" ve "hata birimi" şeklinde bir ayırıma tabi tutmak imkansızdır. Bu nedenle değişkenlerin her birinin içindeki *ortak öz / özler* faktör analizi yöntemiyle analiz edilerek tahmin edilir. Ortak faktör analizinin iki temel varsayımı vardır:

1. Özge faktörler birbirleriyle ilişkili değildir.
2. Özge faktörler aynı zamanda ortak faktörle de ilişkili değildir.

Faktör analizi sonucunda değişkenler ortalaması 0 ve varyansı ±1,0 olacak şekilde standardize edilerek yeni değerlere dönüştürülür. Değişkenlerin varyansı 1,0 olduğuna göre, yorum yapmaya elverişli olacak hesaplama sonucunda çıkarılmış herhangi bir faktörün varyansı 1,0 birim varyanstan daha büyük olmalıdır ki birden fazla değişkeni temsil etsin. Diğer bir deyişle *özdeğer* $\lambda > 1,0$ olmalıdır. Aksi takdirde çıkarılan faktör sadece tek bir göstergenin varyansını açıklamış olur. Örneğin, bir hesaplamada bir numaralı faktöre ait özdeğerin 2,745 olması bu faktörün birden fazla değişkenin/göstergenin varyansını açıkladığı anlamına gelir. Buna göre bir faktörün özdeğeri, kapsadığı değişkenlerin varyanslarının toplamına eşittir. Bir faktöre ait *toplam varyans değeri* (özdeğer) kadar önemli olan bir diğer konu, bu değerin diğer faktörler arasındaki ağırlığının yüzde olarak ne olduğudur. Yukarıdaki örnekte bir numaralı faktöre ait özdeğer 2,745 olarak tespit edilmişti. Diyelim ki bu değer, toplam sekiz faktörün çıkarıldığı bir ölçekten elde edilmiş olsun. Bir faktörün *açıkladığı varyans yüzdesi*, özdeğer rakamının testten çıkarılan toplam faktör sayısına bölünüp 100 ile çarpılması sonucunda bulunur.

■ Faktörün toplam varyans içindeki yüzde ağırlığı = (2,745/8)100 = %34,31.

Söz konusu 1 numaralı faktör (bileşen) toplam varyansı %34 oranında açıklıyor demektir. Faktörün *varyans yüzdesi*, karar vermede kullanılan önemli bir değerdir. Örneğin bir ölçümde *baskın faktör* tanısı koyabilmek için faktörün varyans yüzdesi değerinin en az ,30 düzeyinde olması ve diğer faktörlerin yüzde ağırlıklarının ise daha düşük seviyelerde kalması gerekir. İkinci ve üçüncü faktörlerin varyans yüzdeleri de aynı şekilde hesaplanır. *Özdeğer* 1'den düşük olduğunda ise bu faktörler tek bir değişkenin varyansından daha azını açıkladığı için faktör olarak kabul edilmez.[31] Bu faktörlerin varyans yüzdeleri çoğunlukla ,20'nin altına düşmüştür. Araştırmacı faktör sayısını belirlerken kümülatif toplam varyansın %70 ilâ %80'ini açıklayan faktörleri "temel faktörler" olarak görebilir.[32] Bazı bilim adamları ise araştırmanın amacına göre değişmekle birlikte bu oranın ,60'lar düzeyinde kalması gerektiğini belirtmişlerdir.[33]

Temel bileşenler analizinde *artık değerlerin* birbirleriyle ilişkili oldukları düşünülürken, ortak faktör analizinde ise, "spesifik faktörler" artık değerlerin işlevine sahiptir, fakat spesifik faktörler bir taraftan birbirleriyle ilişkili olmadıkları gibi diğer taraftan ortak faktörle de ilişkili değildirler. Ortak faktör analizinde her bir ortak faktör en az iki değişkene katkı yapıyor olma-

lıdır.[a] Eğer iki değişkenden daha az değişkene katkı yapıyorsa bu faktör *özge faktör* olarak nitelendirilir.[34] Faktör analizinde gözlemlenen korelasyon katsayılarıyla (observed correlation matrix), faktör analizi sonucunda elde edilen yeniden üretilmiş korelasyon katsayıları[b] (reproduced correlation matrix) arasındaki farklılık *artık değer korelasyon katsayılarını* (residual correlation matrix) verir. Faktör çıkarırken artık değer korelasyon katsayılarına da bakılabilir. Artık değer korelasyon katsayıları ,10'dan küçük ise veya sonuçlar anlamlı değilse korelasyonun faktöriyel yapıyı iyi açıkladığına karar verilir. Eğer ,10'dan büyük ise modele daha fazla faktör alınması düşünülür.[35] Çünkü iyi bir faktör analizinde bu değerlerin düşük olması gerekir. Bununla birlikte bu yaklaşım bir modelin reddedilip reddedilmeyeceğine karar vermek için kullanılan bir test değildir. Çünkü anlamlılık veya anlamsızlık artık değer katsayılarının büyüklüğü kadar örneklem hacmine de bağlıdır.[36]

Temel bileşenler analizi ile ortak faktör analizi arasındaki benzerlik ve farklılıklar. İki yöntem arasındaki farklılık esas olarak kullanılan matematik modellerden kaynaklanır, fakat hesaplamada birbirine benzer sonuçlar elde edilir. Bununla birlikte bilim adamı iki yöntemden birini seçerken veri yapısını ve kullanım amacını birlikte göz önünde bulundurmalı doğru bir tercih yapmalıdır. TBA ile OFA arasındaki benzerlik ve farklılıkları aşağıdaki gibi sıralayabiliriz.

Yapı-bileşen ayrımı. Temel bileşenler analizinin amacı gizli yapıları ortaya çıkarmak değildir. Temel bileşenler analizi, ortak ve özgün bir etki yaratan değişkenleri bir araya getirerek bileşik bir ölçüm ortaya çıkarır. Bileşenler, teknik anlamda faktörler değildir. Ortak faktör analizinde ise gizli yapıların ortaya çıkarılması düşüncesi esastır.

Kavramlaştırma. Ortak faktör analizi sonucunda elde edilen faktörler "ayağı yere sağlam basan", gerçek dünyada görülen, rastlanılan ve konuşulan kavramlardır. Temel bileşenler analizi sonucunda elde edilen "faktörler"

---

[a] Daha önce Thurstone'ın faktör analizi yapabilmek için en az üç değişken olması gerektiği düşüncesini aktarmıştık. Burada ise faktör analizi sonucunda en az iki gözlem değişkeninin varyansına katkıdan söz edilmektedir. Ancak iki değişkenli bu faktör daha sonra bir ölçüm aracı olarak kullanılacaksa Thurstone'ın sözünü ettiği ölçüm yetersizliğiyle karşılaşmak kaçınılmaz olacaktır. Bilim adamlarının çoğunluğu üçten az değişkenin varyansını açıklayan faktörleri ayrı bir faktör olarak kabul etmeme eğilimindedirler.

[b] Faktörlerden üretilen korelasyon matrisi. Faktörler arasındaki korelasyon katsayılarını gösteren (factor correlation matrix) ile aynı anlamda değildir.

ise basit geometrik soyutlamalardır. Bu nedenle gerçek dünyadaki olgulara kolaylıkla uyarlanamaz.

Korelasyon-kovaryans matrisi. Ortak faktör analizinde değişkenler arasındaki kovaryans değerleri dikkate alınırken temel bileşenler analizinde özellikle farklı ölçek değerleri kullanılmışsa korelasyon katsayılarına bağlı olarak hesaplama yapılır.

Temel bileşenler analizinde özdeğeri 1'den büyük olan bileşenleri faktör olarak belirlemek çok güçlü bir yaklaşım değildir. Bu yöntemin faktör sayısını gereğinden fazla veya yetersiz tahmin ettiği bildirilmiştir.[37] Temel bileşenler analizi ve ortak faktör analizi farklı amaçlara hizmet eder. Ortak faktör analizinde değişkenlerin içindeki özlerin hangi faktör veya faktörler altında toplandığı izlenirken, temel bileşenler analizinde orijinal değişkenlerdeki varyansın maksimum bölümü minimum sayıda bileşik değişken üzerinde toplanması amaçlanır. Temel bileşenler analizinde testlerdeki / göstergelerdeki / değişkenlerdeki varyansın veya değişkenliğin tamamı hesaplanırken, ortak faktör analizinde değişkenlerdeki varyansın sadece bir bölümü, yani diğer değişkenlerle "ortak" olan, paylaşılan bölümü hesaplanır. Spesifik faktör varyansı ile hata varyansı hesaplamaya katılmaz.

Temel bileşenler analizinden "test veya ölçek geliştirmek" amacıyla yararlanılırken ortak faktör analizi çoğunlukla "kuram" geliştirme amacıyla kullanılır. Bu yöntemde bilim adamının amacı gizli yapıyı veya gizli yapıları ortaya çıkarmak ve davranışın temel öğelerini ortaya koymaktır.[38]

Temel bileşenler analizini mi yoksa ortak faktör analizini mi seçmemiz gerektiğine karar verirken göz önünde bulundurmamız gereken bir diğer nokta orijinal değişkenlerdeki varyansla ilgili varsayımlarımızdır. Gözlem değişkenleri nispeten az hata içeriyorsa (örneğin, yaş, eğitim yılı, aile üyelerinin sayısı gibi) böyle bir durumda hata ve spesifik varyansın toplam varyansın küçük bir bölümünü oluşturduğu varsayılır ve temel bileşenler analizi yöntemi uygulanır. Ancak gözlem değişkenleri gizli yapıların göstergeleri niteliğinde ise (test puanları veya tutum ölçeklerinde olduğu gibi) veya hata varyansı (spesifik varyans) toplam varyansın önemli bir bölümünü oluşturuyorsa ve araştırmacı gizli yapıları ortaya çıkarmayı amaçlıyorsa böyle bir durumda ortak faktör analizi yöntemi seçilir.[39] Garsona göre (2003), temel bileşenler analizi çok sayıda madde içeren test ve ölçeklerin belirli sayıda faktörü içerecek şekilde kısaltılması amacıyla kullanılırken, ortak faktör analizi kuram geliştirme ve gizli yapıları ortaya koyma amacına yöneliktir.[40] Bilim adamı çalışmasını *yapısal eşitlik modeli* çerçevesinde test ediyorsa ortak faktör / temel eksenler analizi modelini kullanmalıdır. Gizli

yapı veya gizli faktörlerin parametrelerini elde etmek isteyen araştırmacılar temel bileşenler analizini kullanmamalıdırlar (Widaman, 1993, aktaran Garson).[41]

Maksimum olasılık analizi. Maksimum olasılık, Fisher (1921) tarafından geliştirilmiş olan bir tekniktir. Örneklem verileri eğer çok değişkenli normal dağılım özelliği gösteriyorsa *maksimum olasılık* yöntemi, gözlem verilerine dayalı korelasyon katsayılarının olası en yüksek değerlere sahip olmasını sağlar. Maksimum olasılık analizi normal dağılım özelliği gösteren bir ana kütle/örneklem verileri için en iyi uyuşumu veren hesaplama yöntemidir. Analiz; *temel bileşenler faktör analizi* veya *ortak faktör analizi* yapıldıktan ve belirli sayıda faktör ortaya çıkarıldıktan sonra gerçekleştirilir. Diyelim ki böyle bir analiz sonucunda üç faktör ortaya çıkmış olsun. Maksimum olasılık faktör analizi, bu aşamada ek bilgi edinmek için yapılan bir analiz niteliğindedir. Analizle araştırmacı ortaya çıkarılan üç faktörlü modele göre maksimum olasılık analizi sonucunda değişkenleri daha iyi açıklayan başka bir model geliştirilip geliştirilemeyeceğini görmeye çalışır.[42] Bu açıdan teyit edici faktör analizine benzer. Bu analiz ile faktörler arasındaki korelasyon katsayılarını görmek ve faktör yüklerinin anlamlı olup olmadığını test etmek mümkündür. Ancak yöntemin uygulanabilmesi için test/ölçek maddelerinde çok değişkenli normallik özelliğinin sağlamış olması gerekir. Değişkenlerin çarpıklık değeri >2 ve basıklık değeri >7 olmadığı sürece bu analiz uygulanabilir.[43]

Alfa faktör analizi. Bu yöntemde ölçekteki maddeler, muhtemel maddeler evreninden seçilmiş bir örneklem olarak ele alınır. Yöntem, hesaplama sonucunda çıkaracağı faktörlerin alfa güvenilirlik katsayısını maksimum düzeye çıkarma amacına yöneliktir.[44] Bilim adamı ölçekte birden fazla faktör olduğunu düşünüyor ve bu faktörlerin birbirlerine karşılık geldiğini (sözel-sayısal, ast-üst gibi) varsayıyorsa faktör çıkarmak için bu yöntemi kullanır.[45]

İmaj faktör analizi yöntemi. Bu yöntem temel bileşenler analizi ile ortak faktör analizi arasında bir tür uzlaşmayı temsil eder ve maddelerin çoklu regresyon analizi sonuçlarına dayanır. Bazı yazarlar ise imaj faktör analizi yöntemini "ortak faktör analizi" grubunda değerlendirmişlerdir. Araştırmacı, muhtemel faktörlerin birbirlerinin tersi olduğunu (içedönük-dışadönük gibi) düşünüyorsa faktör çıkarmak için imaj faktör yöntemini kullanır.[46]

Ağırlıklandırılmamış en küçük kareler yöntemi. Literatürde daha az kullanılan bu yöntem, gözlem matrisi ile yeniden üretilen korelasyon matrisi arasındaki farklılıkların karelerini minimum düzeye getirmeyi amaçlar.

Genelleştirilmiş en küçük kareler yöntemi. Bu yöntemde de gözlem matrisi ile yeniden üretilen korelasyon matrisi arasındaki farklılıkların kareleri minimum düzeye getirilmeye çalışılır, fakat işlem sırasında her bir madde kendi özgün faktör değeri ile ağırlıklandırılır.[47] Genelleştirilmiş en küçük kareler yönteminin toplanan verilerin dağılımının bilinmediği durumlarda uygulanması önerilmiştir.

Varsayımları

Bilim adamı faktör analiziyle faktör yapısını ortaya çıkarmak ve göstergelerin güvenilirliğini test etmek istiyorsa belirli şartların yerine gelme durumunu kontrol etmelidir. Bunlar faktör analizinin varsayımları veya ön koşulları olarak isimlendirilir ve aşağıdaki gibidir.[48]

Verilerin metrik olması. Gözlem verilerinin eşit aralıklı veya oranlı ölçek verisi niteliğinde olmasıdır. Bu tür rakamlar, "sürekli" veri grubunda değerlendirilir. Likert ölçeğinin 5 veya 7 dereceli maddelerine ait sıralı ölçek verileri ise "kesikli veri" niteliğindedir. Kesikli veriler *gizli faktörlerin sıralı göstergeleri* olarak adlandırılır. LISREL gibi ileri istatistiksel analiz programlarında 15 derecenin altındaki ölçek verileri sıralı ölçek verisi olarak değerlendirilmiştir. Joreskog ve Moustaki'ye göre (2001) sıralı ölçek verilerine faktör analizi yöntemini uygulamak uygun değildir (aktaran Dekkers, 2003).[49] Joreskog ve Moustaki, sıralı ölçek verilerinin arka planında yatan gizli faktörleri ortaya çıkarmak için farklı iki yaklaşım önermişlerdir. Bunlardan birincisi "gizli yanıt değişkeni" (*underlying response variable approach)* yaklaşımıdır. Bu yaklaşımda sıralı ölçek verisine sahip bir değişken, arka planda sürekli dağılım özelliğine sahip ve normal dağılım özelliği gösteren gizli bir değişken tarafından yaratılmıştır. İkincisi ise, "yanıt fonksiyonu" (*response function approach)* yaklaşımıdır (aktaran Cziráky, Tišma ve Pisarović, 2002).[50] Yanıt fonksiyonu yaklaşımı madde-yanıt kuramından geliştirilmiştir ve bu yaklaşıma göre gözlem değişkenlerinin her biri *m* olası yanıt biçimine sahiptir. Model, gizli faktörlerin koşullu bağımsızlığını savunur ve gizli değişkenler için *p* boyutlu yanıt kalıbı oluşturur.[51] Jöreskog ve Sörbom, sıralı ölçek verileri için faktör analizi yerine LISREL programı çerçevesinde *yapısal eşitlik modellerinin* kullanılmasını önermişlerdir.[52] İstatistik yazılımı SPSS'te metrik olmayan ve doğrusal ilişki göstermeyen verilerin faktör analizi için CATEGORIES modülündeki PRINCALS, OVERALS ve

CATREG programları bulunur. Söz konusu yazılımlar sıralı verilerde faktör analizi yapma amacına yönelik olarak hazırlanmıştır, ancak literatürde hem az kullanılmıştır, hem de bu modüller SPSS'in standart paket sürümlerinin içinde bulunmamaktadır. Görüldüğü gibi Likert ölçek verileriyle faktör analizi yapılması konusu tartışmalı ve elde edilen sonuçlar büyük ölçüde kuşkuludur.

Bilim adamlarının bir bölümü Likert ölçek verileriyle faktör analizi yapılması uygulamasına karşı çıkarken, bazıları bu konuda daha serbest bir tutuma sahip olmuşlar ve diğerleri ise ortak faktör analizi yöntemini hariç tutarak, sıralı ölçek verilerinin sadece "temel bileşenler analizi" ile test edilebileceğini belirtmişlerdir.

Merkezî limit teoremi ve Monte Karlo benzetim hesaplamalarına dayalı olarak Baker, Hardyck ve Petrinovich (1966), Borgatta ve Bohrnstedt (1980) kullanılan ölçeğin eşit aralıklı veya sıralı olup olmadığının önemli olmadığını ifade etmişlerdir (aktaran Yu, 2003).[53] Tukey'e (1986) göre de sıralı ölçek verilerinin parametrik testlerde kullanılmasına yasaklama getirilmesinin bir anlamı yoktur.[54] Kim ve Mueller'e (1978) göre eğer arka planda bulunan metrik skalayı ciddi bir şekilde bozmadığı düşünülüyorsa sıralı ölçek verilerinde faktör analizi kullanılabilir (aktaran, Edari, 2002).[55] Ayrıca bu yazarlar, gözlem değişkenlerinin arka planda metrik değişkenlerle olan korelasyon katsayılarının ılımlı (,70) veya düşük olduğunu düşünüyorlarsa ikili veri yapılarında bile faktör analizi yönteminin kullanılabileceğini belirtmişler ve faktör analizinde sıralı ölçek verileri kullanılmışsa böyle bir durumda ortaya çıkan faktörleri yorumlamanın, faktörlere etiket bulmanın zor olacağını söylemişlerdir.[56] Hutcheson ve Sofroniou (1999), keşfedici faktör analizinde *sürekli veri* şartının gevşetilmesinin faktörleri daha iyi yorumlamak amacıyla kullanılıyor olması halinde haklı görülebileceğini ifade etmişlerdir (aktaran Tall, 2002).[57]

Bu bilim adamlarının görüşlerinden sıralı ölçek verilerinde faktör analizi yapılabileceği anlaşılır. Sosyal ve davranışsal bilimlerde Likert tipi ölçek verileriyle faktör analizi varsayımları tam olarak karşılanmasa bile çok sık olarak faktör analizi yapılmıştır. Bu konudaki bir diğer görüş, uygulanacak faktör analizi modeliyle ilgilidir. Araştırmacı, Likert ölçeğini arka planda sürekli dağıldığı varsayılan gizli yapıları ortaya çıkarmak için kullanıyorsa ve bu gizli değişkenlerde "ortak faktör" analizi yöntemini uygulamak istiyorsa böyle bir durumda farklı bir yaklaşımdan hareket etmelidir. Bunun için Likert ölçeğindeki verilerde polikorik korelasyon analizi yöntemi uygulanarak *polikorik korelasyon matrisi* elde edilir ve faktör analizi bu matrise dayalı olarak yapılır. Ancak, gözlem verilerindeki gizli değişkenlerin öngörülen gizli değişken modeliyle ne ölçüde iyi uyuştuğunu görmek için daha

sağlıklı olan yöntem, *teyit edici faktör analizini* veya *yapısal eşitlik modelini* uygulamaktır.[58]

Örneklemin türdeşliği. Faktör analizi yapılacak verilerin toplandığı örneklem verileri ne büyük ölçüde birbirinden uzak ve ayrık bulunmalı ve ne de birbirinin aynı olmalıdır. Büyük ölçüde heterojen birimlerden oluşan örneklemlerde faktör analizi yapmak doğru olmayacağı gibi, bütünüyle homojen birimler de istenen sonucu vermez. Faktör analizi göstergeler arasındaki kovaryans değerlerine dayandığından ölçüm yapılan kişilerin görüşlerinde belirli ölçüde farklılıklar bulunmalıdır. Faktör analizi yapılan veriler, belirli ölçüde yanıt farklılığı içermiyorsa araştırmaya / ölçüme katılan kişilerin heterojenliğinden söz edemeyiz. Faktör analizi, sadece "öbekleşmiş kümesel farklılıkların" bulunduğu koşullarda güvenilir sonuçlar vermez. Ölçümde bu tür kümesel bir gruplaşma yoksa "türdeşlik koşulunun" karşılandığı varsayılır.

Göstergelerin çok değişkenli normal dağılıma sahip olması. Faktör analizi çok değişkenli normallikten sapmalara karşı oldukça güçlü bir teknik olmasına karşılık bu koşul karşılanırsa çözüm daha da genişler ve daha sağlıklı sonuçlar alınır. Faktör analizinde, değişkenlerin normalliği elde edilen sonuçların araştırma yapılan örneklemin dışında ana kütleye genelleme yapılmak istenmesi halinde önem kazanır. Normalde *temel bileşenler analizi* ve *ortak faktör analizinde* doğrudan dağılımla ilgili bir varsayım söz konusu değildir. Ancak örneklem hacmi küçüldükçe normallik değerlendirmesi önem kazanır. Eğer, faktör analizi modüllerinden *maksimum olasılık* yöntemi tercih edilmişse bu modülde verilerin çok değişkenli normal dağılım özelliğine sahip olduğu varsayılır. Faktör analizinde çok değişkenli normallik özelliğinin karşılanıp karşılanmadığı sadece hipotez testi yapılmak istendiği zaman aranır.

Gösterge çiftleri arasındaki ilişkilerin doğrusal olması. Ölçek veya testteki madde çiftleri arasında paralelliğin bulunması anlamına gelir. Paralellik nokta-dağılım grafiği ile test edilir. Değişkenler arasındaki ilişkiler doğrusallıktan uzaklaşmışsa korelasyon katsayıları uygun bir ölçü değildir ve dolayısıyla bu değişkenlere dayalı olarak faktör analizi yöntemi uygulanamaz. Ancak paralellikten / doğrusallıktan küçük ölçüdeki sapmalar faktör analizini etkilemez. Paralellikten büyük ölçüdeki sapmalarda ise, uygun dönüştürme yöntemlerine başvurulur.

Çoklu doğrusallık. Göstergelerin / değişkenlerin çoklu doğrusallık[a] / koşutluk özelliği (r > ,90) veya tekillik özelliği (r = 1,00) gösterip göstermediğinin incelenmesidir. Çoklu doğrusallık veya koşutluk özelliği, faktör analizini olumsuz bir şekilde etkiler. Bu koşul, ters matrisi alınmadığı için *temel bileşenler analizi* için geçerli olmamakla birlikte, *ortak faktör analizi* için mutlaka araştırılmalıdır.[59] Ortak faktör analizinde koşutluk özelliği gösteren tüm değişkenler testten ve analizden çıkarılmalıdır. Çoklu doğrusallık özelliği regresyon analizi ile sınanır. Bunun için her bir değişken geri kalan diğer değişkenlerle regresyon analizine tâbi tutulur. Hesaplama sonucunda bir değişkenin diğer değişkenlerle olan ilişkisini gösteren R^2 değeri düşükse o değişkenin analizden çıkarılması gerekir. Tam tersine R^2 değeri 1'e çok yakınsa bu kez değişkenin analize alınması için yeniden düşünülmelidir. Çoklu doğrusallığı belirlemek için kullanılabilecek bir diğer yaklaşım KMO testini uygulamaktır.

Yu'nun bildirdiğine göre, Micceri (1989) tarafından davranış bilimleri alanında 400 büyük veri yapısında yapılan bir araştırmada veri yapılarının büyük çoğunluğunun çok değişkenli normal dağılım özelliği göstermediği bulunmuştur.[60] Yine Breckler'in (1990) kişilik ve sosyal psikoloji dergilerindeki 72 makale üzerinde yaptığı bir araştırmada, makalelerin sadece %19'unda çok değişkenli normalliğin sağlandığı konusunda bilgi verildiği, %10'unundan azında ise bu varsayımın karşılanamadığının bildirdiği, diğer makalelerde ise hiçbir bilginin verilmediği görülmüştür (aktaran Yu, 2003).[61] Literatür incelemeleri araştırmacıların geçerlilik varsayımlarının karşılanma durumu konusunda çalışmalarında yeterince bilgi vermediklerini göstermektedir. Ancak bu veriler bir gerekçe veya mazeret olarak kullanılmaz. Bilim adamı istatistiksel hesaplamalardan önce faktör analizinin geçerlilik varsayımlarının karşılanma derecesi hakkında okurlarına bilgi vermelidir.

Veri Matrisi

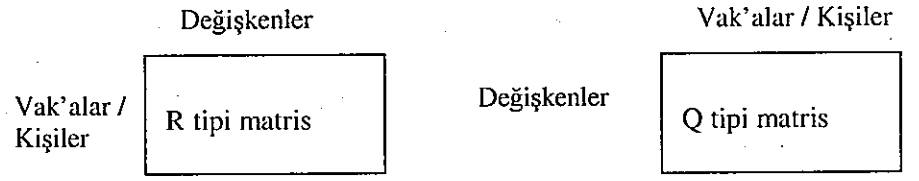
Faktör analizi, veri yapılarının yeniden düzenlenmesi esasına dayanır. Normalde ölçüm verileri istatistiksel analiz programının veri yükleme penceresine "vak'lar x değişkenler" formunda tanımlanır. Ancak veriler her zaman bu şekilde girilmez. Bazen araştırmacılar verileri değişik bir düzenleme içinde bilgisayara tanıtmak isterler. Örneğin test-yeniden test uygulamala-

---

[a] Koşutluk, (veya çoklu doğrusallık) göstergelerin birbirleriyle yüksek derecede ilişkili olması ve sonuçta korelasyon veya regresyon katsayılarının olduğundan daha yüksek çıkmasıdır. Regresyon analizinde de tahmin değişkenleri (bağımsız değişkenler) birbiriyle yüksek derecede ilişkili ise bu özelliğe *koşutluk* adı verilir.

rında veri matrisi ters döndürülmek istenebilir. Her tür veri düzenlemesinde faktör analizi yapılabilir. Verilerin düzenleniş biçimine göre farklı veri matrisleri söz konusudur. Veri matrislerine tanımlama kolaylığı sağlamak için belirli harflerle isimler verilmiştir. Örneğin, R matrisi, Q matrisi gibi. Aşağıdaki paragraflarda bu veri matrislerinin düzenleme biçimine ilişkin bilgiler verilmiştir.

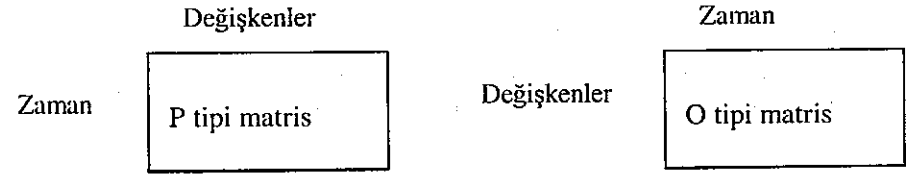
R ve Q matrisleri. R ve Q matrislerinde *vak'a x değişken* düzenlemesi esas alınmıştır. Faktör analizinin en yaygın kullanımı R matrisine dayalı olarak yapılır. Bu uygulamada faktörler değişkenlere yüklenmiştir ve kişiler bazında analiz edilir. Q matrisinde ise faktörler kişilere yüklenmiştir ve değişkenler bazında analiz edilir. Araştırmacı; bireyler, gruplar ve uluslar arasındaki benzerlik yapılarını ortaya çıkarmak istediğinde veya kişileri bağımsız gruplar olarak organize etmek istediğinde Q matrisini kullanır (*bk.*, Şekil 8-3). Ancak günümüzde çeşitli kümeleme tekniklerinin geliştirilmiş olması nedeniyle bu yaklaşım artık daha az tercih edilmektedir.



Şekil 8-3. R ve Q tipi matris veri yapıları.

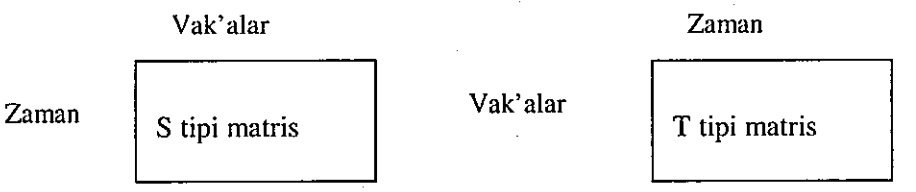
Değişkenler yerine vak'aları kümeleyen Q modeli matrislerde negatif faktör yükleriyle karşılaşma gibi bir sorun söz konusudur. Değişkenlerin geleneksel faktör analizi yöntemiyle incelenmesinde negatif yükler, *değişkenin faktörle negatif yönlü olarak ilişki içinde olduğunu* gösterirken Q modeli faktör analizinde, negatif faktör yüklerinin herhangi bir anlamı yoktur. Negatif yüklü bütün vak'alar ayrı bir grup olarak değerlendirilir.[62]

P ve O matrisleri. P ve O matrislerinde *zaman x değişken* düzenlemesi esas alınmıştır (*bk.*, Şekil 8-4).



Şekil 8-4. P ve O tipi matris veri yapıları.

S ve T matrisleri. S ve T matrislerinde ise *zaman x kişiler* düzenlemesi esas alınmıştır. Zamanla kişiler arasındaki ilişkilerin niteliği araştırılır (*bk.*, Şekil 8-5).



Şekil 8-5. S ve T tipi matris veri yapıları.

İşlem Aşamaları

Ölçek veya testin faktöriyel yapısını ortaya çıkarmayı amaçlayan bilim adamları faktör analizini aşağıdaki adımlar çerçevesinde gerçekleştirirler.

1. *Amacın belirlenmesi.* Gizli yapıları ortaya çıkarma, temel bileşenleri ortaya koyma veya gizli yapılar arasındaki ilişkileri belirleme.
2. *Örneklem sayısının kontrolü.* Değişken başına en az 5 katılımcı verisinin bulunması.
3. *Tanımlayıcı istatistiksel analizler.* Çapraz tablo analiz yöntemi kullanılarak verilerdeki eksik ve uç değerlerin saptanması.
4. *Negatif ifadelerin ters çevrilmesi.* Ölçekte negatif ifadeler varsa bu ifadelerin Transform - Recorde – Into same variables komutuyla pozitif hale dönüştürülmesi.
5. *Korelasyon analizi.* Eksik verilerin anlamlı olup olmadığını belirlemek ve ayrıca korelasyon katsayısı ,30'un altındaki değişikleri ölçekten çıkarmak için uygulanır. Bir diğer amaç verilerin koşutluk ve tekillik özelliğini görmektir.
6. *Varsayımların kontrolü.* Türdeşlik, verilerin metrik olması, çoklu normal dağılım (hipotez testi söz konusu ise), koşutluk (KMO ve anti-image testi), doğrusallık.
7. *Analiz yöntemi.* Ortak faktör analizi veya temel bileşenler analizi.
8. *Faktör çıkarma ilkelerinin belirlenmesi.* Özdeğer 1 eşik değerinin kullanılması, yamaç-birikinti grafiği, açıklanan varyans yüzdesi.

9. *Döndürme yöntemi.* Dik açılı veya eğik açılı[a] döndürme yöntemlerinden hangisinin tercih edildiği ve bu yöntemlerin içinde hangi modelin temel alındığı. Faktörler birbirinden bağımsız değilse eğik açılı döndürme yönteminin tercih edilmesi.
10. *Sonuçların geçerliliği.* Ayrık/uç değerlerin etkisinin analizi, sonuçların doğrulanması için, teyit edici faktör analizi olarak *ortak faktör* veya *maksimum olasılık* analizinin kullanılması, örneklem grubunun iki alt gruba bölünerek faktör analizinin iki ayrı alt grupta tekrar yapılması.
11. *Sonuçların yorumlanması.* Sonuçların kuramsal bilgilerle karşılaştırmalı olarak ele alınması ve değerlendirilmesi.

Analiz ve Çıktılar

Keşfedici faktör analizi değişik istatistiksel analiz yazılımları kullanılarak yapılabilir. Ülkemizde en çok SPSS isimli programın kullanılıyor olması nedeniyle faktör analizi bu program kapsamında ele alınmıştır. Analizin SPSS'te yapılabilmesi için bilim adamının bazı ön bilgilere sahip olması gerekir. Örneğin faktör sayısının kendisi tarafından mı belirleneceği yoksa faktör sayısını belirleme işleminin programa mı bırakılacağı, önemsiz faktör yüklerinin hangi ağırlık değerlerinden itibaren gizleneceği, kovaryans veya korelasyon matrislerinden hangisinin temel alınacağı bu tür gerekli bilgiler arasındadır.

Aşağıdaki paragraflarda önce faktör analizi yapabilmek için mönü tanımlamaları üzerinde durulmuş ve daha sonra hesaplama sonucunda elde edilen çıktılar ve çıktıların yorumları hakkında bilgiler verilmiştir.

Veri ve çözümleme işlemlerinin tanıtılması. İstatistiksel analiz yazılımı SPSS'te keşfedici faktör analizinin yapılması, geliştirilen ana ve alt mönüler sayesinde büyük ölçüde kolaylaştırılmıştır. Bunun için aşağıdaki işlemler yapılır.

1. Ölçek veya testin maddelerine ait değerler programın çalışma sayfasına *veri matrisi* olarak girilir.
2. Analiz mönüsünden faktör analizi şıkkı seçilerek, Factor Analysis diyalog kutusuna değişkenler tanıtılır.

---

[a] Eğik açı, uzayda dik açılı olmayan herhangi bir düzenlemedir. Dar veya geniş açı şeklinde olabilir.

3. Descriptes düğmesine basılarak açılan Factor Analysis: Descriptes diyalog kutusunun Statistics bölümünden Initial solution şıkkı, Correlation matrix bölümünden ise (a) Coefficients, (b) Significance levels, (c) Determinant, (ç) KMO and Bartlett's test of sphericity ve (d) Reproduced şıkları seçili hale getirilir. Bu bölümdeki Coefficients şıkkı *R*-matrisini ve Significance şıkkı korelasyon matrisindeki her bir katsayının anlamlılık değerini verir. Determinant şıkkı, koşutluk ve tekillik özelliğinin belirlenmesi açısından önemlidir. *R*-matrisinin determinantı ,00001'den büyük olmalıdır. Eğer bu değerden daha düşük ise korelasyon matrisi incelenerek $r > ,80$ olan değişkenlerden biri iptal edilir.[63] Seçim işlemlerinden sonra bir önceki Factor Analysis diyalog kutusuna dönülür.

4. Extraction düğmesine basılarak açılan Factor Analysis: Extraction diyalog kutusunun Method bölümündeki şıklardan biri seçilir. Çalışılan veriler parametrik nitelikte ise ve veriler normal dağılım özelliği gösteriyorsa araştırmacının amacına göre (ölçek geliştirmek, kurama katkı yapmak) Principal Components veya Common factor (Principles Axes Factoring) şıklarından biri seçilir. Veriler sıralı ölçek niteliğinde ise, teyit edici faktör analizi yapılmak isteniyorsa ve verilerin çok değişkenli normal dağılım özelliğine sahip olduğu varsayılıyorsa bu kez Maxmimum Likelihood veya Weighted Least Square yöntemi seçilir. Daha sonra Extraction diyalog kutusunun Analyze bölümünden çalışılan verilerin niteliğine göre Correlation matrix veya Covariance matrix şıklarından biri seçili hale getirilir. Eğer değişkenlerin ölçüm birimleri, değişim aralıkları ve varyansları çok farklı ise faktör çıkarmak için *korelasyon matrisi* değerleri baz alınır. Veriler Likert ölçeklerinde olduğu gibi homojen nitelikte ise veya araştırmacı faktör çıkarmak için orijinal değerlerden yararlanmak istiyorsa bu kez *kovaryans matrisi değerleri* temel alınır. Extract bölümünden ise Eigenvalues over şıkkı seçilir. Şıkkın yanındaki kutuya 1 veya Joliffe kriteri temel alınmışsa ,70 değeri girilir. Literatürde daha çok Henry Kaiser tarafından geliştirilen ve kısaca Kaiser kriteri olarak adlandırılın 1 özdeğer kriteri kullanılır. Joliffe kriteri literatürde daha az kullanılmıştır ve liberal bir yaklaşımı yansıtır. Araştırmacı çıkarılacak faktör sayısını önceden kendisi de belirleyebilir. Faktör çıkarma işlemi için "maksimum olasılık" yöntemi tercih edilmişse faktör sayısı araştırmacı tarafından belirlenir. Diyalog kutusunun Display bölümündeki her iki şık da seçili hale getirilir. Böylece ilk aşamada bir taraftan döndürülmemiş faktör çözümü

elde edilirken diğer taraftan yamaç-birikinti grafiği (scree plot)[a] ile faktör sayısını araştırma imkanı elde edilmiş olur. Diyalog kutusunun alt bölümünde yer alan Maximum iterations for convergence bölümünde ön tanımlı olarak belirlenmiş olan 25 sayısı olduğu gibi bırakılır. Döndürme işlemi çoğunlukla bu rakam çerçevesinde yapılır. Bilgisayar programları temel bileşenler analizinde bütün değişkenler için 50'den daha fazla döndürme yapamaz. Tanımlama ve seçme işlemi bittikten sonra önceki diyalog kutusuna dönülür.

5. Bu kez Factor Analysis diyalog kutusundaki Rotation düğmesine basılır ve açılan Factor Analysis: Rotation diyalog kutusunun Method bölümünden Varimax ve Display bölümünden ise Rotated solution kutucukları seçili hale getirilir. Daha sonra önceki diyalog kutusuna geri dönülür.
6. Factor Analysis diyalog kutusundaki Scores düğmesiyle herhangi bir işlem yapılmaksızın Options düğmesine basılır. Factor Analysis: Options diyalog kutusunun Missig Values bölümünde Exclude cases listwise şıkkı seçili hale getirilir. Aynı diyalog kutusunda Coefficient display format bölümünde Supress absolute values less then şıkkı seçilir ve buraya ,40 değeri girilir. Bunun anlamı, ,40'ın altındaki faktör yüklerinin gösterilmemesidir.

Çıktılar ve yorumları. İstatistiksel analiz yazılımı SPSS'ten analiz sonucunda çok sayıda tablo ve iki grafik elde edilir. Bu tablo ve grafikler araştırmacıyı değişik konularda bilgilendirmeye yöneliktir.

Tanımlayıcı istatistiksel analiz tablosu. Bu tabloda değişkenlerin aritmetik ortalama ve standart sapma değerleri vardır.

Korelasyon - kovaryans katsayıları tablosu. İkinci tablo, korelasyon / kovaryans katsayılarını gösterir (Correlation Matrix). *R*-matrisi adı da verilen bu tablo, araştırmacıya değişkenler arasındaki ilişkiler hakkında fikir verir. Bilim adamı bu tabloyu inceleyerek ,30'un altında ve ,90'ın üstündeki korelasyon değerlerini saptamaya çalışır. Tablonun ikinci bölümü, korelasyon katsayılarının anlamlılık değerlerini gösterir.

---

[a] Yamaç-birikinti grafiği. Cattel, görünümünü bir yamacın dibinde biriken taş, toprak ve moloz yığınlarına benzettiği için grafiğe bu adı vermiştir. Grafiğin yamaç bölümü faktörleri açıklarken birikinti bölümü gereksiz, işe yaramayan maddelerle ilgilidir.

KMO ve Barlett küresellik testi tablosu. Üçüncü tablo, KMO ve Barlett küresellik testi ile ilgili sonuçları verir. KMO testi, seçilen "örneklem verilerinin" faktör çıkarmak için uygun olduğunu belirler. Korelasyon ve kısmî korelasyon analizi sonuçlarına dayalı olan test sonuçları 0 ilâ 1,0 arasında değişir. Bu değerin yüksek çıkması, "ölçekteki her bir değişkenin ölçekteki diğer değişkenler tarafından mükemmel bir şekilde tahmin edilebileceği"[64] anlamına gelir. Değerler sıfır veya sıfıra yakın çıkmışsa korelasyon katsayılarının dağılımında bir dağınıklık olduğu için bu değerlere dayalı olarak faktör analizi yapılmaz. Test sonucunun ,50 den büyük olması faktör analizine devam edilebileceği anlamına gelir. Kaiser (1974), KMO testiyle ilgili olarak ,50'den düşük değerin kabul edilmeyeceğini, ,50-,60 değerinin kötü, ,60-0,70 değerinin zayıf, ,70-,80 değerinin orta, ,80-,90 değerinin iyi, ,90'dan büyük değerin ise mükemmel olduğunu bildirmiştir (aktaran, Jackson ve Holland, 2003).[65]

Analizde her bir değişken için KMO test değeri ve tüm değişkenler için toplu (genel) bir KMO değeri elde edilir. SPSS'te, Anti-image correlation matrix başlığı altında elde edilen KMO değerleri değişken düzeyindeki değerlendirmeleri yansıtır. Değişken bazındaki KMO değerleri incelenerek düşük değer gösteren değişkenler düşürülmek suretiyle toplam KMO değerindeki değişkenlik incelenir ve test değerinin ,50 veya ,60'ın üzerinde olmasını sağlayacak değişkenler belirlenir.

Barlett küresellik testi, ki-kare istatistik değerini verir. Bu testte de diğer ki-kare testlerinde olduğu gibi anlamlılık değerine bakılır. Anlamlılık değeri ,05'ten küçük ise *R* korelasyon veya kovaryans matrisindeki verilerin birim matrisinden[a] farklı olduğu sonucuna varılır. Birim matrisinden farklı olması söz konusu korelasyon matrisinden faktör çıkarılabileceği anlamına gelir. Anlamlılık değeri ,05'ten büyük ise matriste paylaşılan varyans olmadığı şeklinde yorumlanır ve söz konusu veri yapısı için faktör analizi yapılmaz.

Anti-imaj matris tablosu. Bu tabloda diğer değişkenlerle paylaşılan varyans hariç tutulmuştur. Anti imaj kovaryans değerleri kısmi kovaryans değerlerinin negatifleridir. Beklenen değerler düşük olmak zorundadır, çünkü bunlar paylaşılan kovaryans değerleri çıktıktan sonra arta kalan değerlerdir. Bu değerler için raporda yorum yapmaya gerek yoktur.

---

[a] Birim matrisi terimi, herhangi bir matriste köşegendeki rakamların dışındaki tüm değerlerin sıfır olması anlamına gelir.

Paydaşlık tablosu. Dördüncü tablo, değişkenlerin faktörle/faktörlerle olan paydaşlık oranını (communalities)[a] verir. Paydaşlık oranı, bir göstergede veya maddede faktörlerin neden olduğu değişkenlik yüzdesidir. Bir başka şekilde ifade etmek gerekirse, bir değişken için birden fazla faktöre ait faktör yüklerinin karelerinin toplamıdır (*bk.*, Tablo 8-1).

■ Paydaşlık oranı = $0{,}323^2 + 0{,}815^2 + (-{,}010)^2 = 0{,}769$.

Tablo 8-1. Paydaşlık Oranı Değerleri

Test maddeleri	Faktör yükleri			
	Faktör I	Faktör II	Faktör III	Paydaşlık
Sözel yetenek	,323	,815	–,010	,769
Sayısal yetenek	,169	,857	–,152	,787
Düzlemsel yetenek	,129	,234	–,123	,702
Yaratıcılık	,086	,012	–,126	,276

İstatistiksel analiz yazılımı SPSS, paydaşlık tablosunu değişik bir düzenleme içinde verir. Bu tabloda iki sütun vardır. Birinci sütun *başlangıç paydaşlık değerlerini* (initials) verir (*bk.*, Tablo 8-2). İkinci sütun ise faktörler çıkarıldıktan sonra elde edilen paydaşlık değerleridir. Bilim adamı, ikinci sütundaki değerleri kullanır.

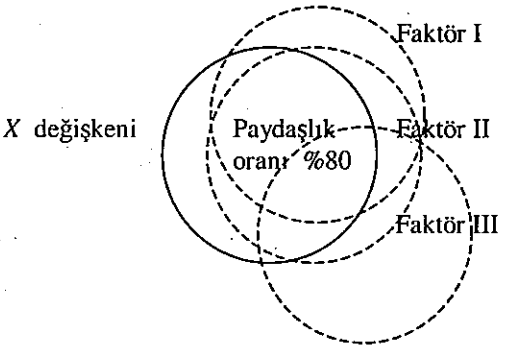
Tablo 8-2. SPSS'te Paydaşlık Oranı Tablosu

Test maddeleri	Initials (başlangıç değerleri)	Communalities (paydaşlık oranı)
1. Sözel yetenek	–,010	0,769
2. Sayısal yetenek	–,152	0,787

Bir değişkene/göstergeye ait paydaşlık değerinin ,799 olması bu değişkendeki değişkenliğin (varyansın) %80 oranında belirlenen faktörlerle açıklanabileceği şeklinde yorumlanır (*bk.* Şekil 8-6). Bir araştırmada örneklem hacmi büyüdükçe paydaşlık oranı da muhtemelen büyük çıkacaktır. Paydaş-

[a] Avrupa Komisyonu'nun *ISI İstatistik Terimler Sözlüğü*'nde kavram, *oransal ortak etken varyansı*; *ortak etken varyansı*; *ortak faktör varyansı* sözleriyle tanımlanmıştır. Ortak kökenlilik, bir göstergede veya maddede faktörün neden olduğu değişkenlik oranı olarak açıklanabilir.

lık büyüklüğü ile örneklem büyüklükleri arasında bir ilişki olduğu belirlenmiştir. Bir maddenin paydaşlık oranı >,60 ise 100'den küçük örneklemlerde dahi o maddenin çıkarılan faktörleri çok iyi temsil ettiği varsayılır. Paydaşlık oranı = ,50 ise 100 ilâ 200 arasında katılımcıya ihtiyaç olduğu anlaşılır. Daha düşük paydaşlık oranlarında ise 500 veya daha büyük örneklem büyüklüklerine ihtiyaç duyulur.[66] Paydaşlık oranı ,20'den düşük ise bu maddeler testten çıkarılarak analiz yeniden yapılır.



Şekil 8-6. Değişken-faktörler paydaşlık oranı.

Açıklanan toplam varyans tablosu (total variance explained). Altıncı tablo analiz sonucunda çıkarılan faktörleri gösterir. Bu tabloda değişkenleri temsil eden tüm faktörler hakkında bilgi vardır. Tablonun ilk sütunundaki *component* sözcüğü değişkenlerden çıkarılan faktörleri tanımlar. Tablonun ikinci bölümü *özdeğerler* hakkında bilgi verir. Bu bölüm de kendi içinde üç ayrı alt bölüme ayrılmıştır. Birinci alt bölümde özdeğerler gözükür. İkinci alt bölümde ise her bir faktöre ait toplam varyansın yüzdesi verilir. Bir faktörün toplam varyanstaki yüzde değeri büyüdükçe faktörün gücü artar. Üçüncü alt bölümde faktörlerin varyans değerleri yığışımlı olarak gösterilmiştir. Tablonun üçüncü bölümü sütun bazında "faktör yüklerinin karelerinin toplamını" verir. Faktör yüklerinin karelerinin toplamıyla özdeğer aynı çıkar ve birbirine eşittir. Ancak ikincisinde sadece çıkarılan faktörler gösterilir. Tablonun dördüncü bölümünde yer alan *Rotation Sums of Squared Loadings* başlığı altında ise, döndürme işlemi sonunda faktörlerin yorumlanabilmesi için iyileştirilmiş olan değerler elde edilir. Döndürme işlemi her bir faktörün özdeğerini değiştirir. Bu bölümün birinci alt sütununda gözüken *total* başlığı altındaki rakamlar iyileştirilmiş olan özdeğerlerdir. Döndürme işlemi sonunda, toplam varyans değişmemekle birlikte faktörlerin açıkladıkları varyans oranlarında değişiklik olur (*bk.*, Tablo 8-3).

Tablo 8-3. Açıklanan Toplam Varyans Tablosu

Total Variance Explained

Components (Faktörler / Bileşenler)	Initial eigenvalues (Başlangıç özdeğerleri)			Extraction sums of squared loadings (Yük kareleri toplamı)			Rotation sums of squared loadings (Döndürülmüş yük kareleri top.)		
	Total (Özdeğer)	Varyansın % si	Birikimli yüzde	Total (Özdeğer)	Varyansın % si	Birikimli yüzde	Total (Özdeğer)	Varyansın % si	Birikimli yüzde
1									
2									
3									
4									

Dik açılı döndürme yöntemi sonucunda elde edilen tablolar. Dik açılı döndürme yöntemi sonucunda yük matrisi tablosu (loading matrix) elde edilir. Ortak faktör analizi seçilmişse çıktılardan Factor Matrix adı ile gözüken faktör matrisi tablosunda faktör yükleri (factor loadings) gösterilmiştir. Faktörler birbirleriyle ilişkili değilse faktör yükleri, gözlem değişkenleriyle faktörler arasındaki ilişkinin gücünü temsil eder. Eğer çıkarılan faktörler birbirleriyle ilişkili ise, faktör yüklerini korelasyon katsayıları olarak yorumlamak çok doğru değildir, ancak yine de bir fikir vermesi açısından bu değerler de korelasyon katsayılarına benzetilebilir.[67] Ağırlığı ,60'ın üzerinde olan faktör yükleri "yüksek"; ,40'ın altında olanlar ise "düşük" kabul edilir. Eğer temel bileşenler analizi tercih edilmişse bu aşamada Component Matrix adıyla *bileşenler matrisi* elde edilir.

Eğik açılı döndürme yöntemi sonucunda elde edilen tablolar. Araştırmacı ortak faktör analizinde eğik rotasyon (oblique rotation) yöntemini uygulamışsa, (a) factor correlation matrix (faktör korelasyon matrisi) (b) pattern matrix (model matrisi) ve (c) structure matrix (yapı matrisi) başlıklarını taşıyan üç farklı tablo elde eder. Faktör korelasyon matrisi faktörler arasındaki korelasyon katsayılarını gösterir. Yapı matrisi, faktör yükü matrisinden başka bir şey değildir. Gözlem değişkenleriyle faktörler arasındaki ilişkinin gücünü gösterir. Model matrisi (pattern matrix), her bir faktöre maddelerden gelen *özge varyans* katkılarını temsil eden katsayılardır. Faktör sayısı arttıkça model matrisi katsayıları düşük çıkar, çünkü açıklanan varyansta daha fazla ortak katkı söz konusudur. Eğik döndürme yönteminde araştırmacı bir faktöre etiket atarken veya isim bulurken hem yapı hem de model matrisi katsayılarına birlikte bakarak karar verir.[68]

Yeniden üretilmiş korelasyon katsayıları tablosu. İstatistiksel analiz yazılımı SPSS'te faktör analizi mönüsü ile ulaşılan Descriptives diyalog kartında Reproduced şıkkı seçildiğinde *çıkarılan faktörleri* açıklayan yeniden üretilmiş korelasyon katsayıları tablosu elde edilir. Yeniden üretilmiş korelasyon katsayıları tablosuyla orijinal korelasyon katsayıları tablosu aynı değildir. Orijinal katsayılardan yeniden üretilmiş korelasyon katsayıları çıkarıldığında ortaya çıkan tabloya ise, "artık değerler korelasyon katsayıları tablosu" adı verilir. Artık değerler, seçilen modelin güvenilir olup olmadığını belirler. Verilerin öngörülen modelle iyi bir uyuşma göstermesi halinde tablodaki değerlerin yeknesak bir biçimde küçük çıkması gerekir.[69]

Temel Kavramlar

Faktör analizinin etkili bir şekilde kullanılması ilgili kavramların ve terimlerin iyi bilinmesine bağlıdır. Araştırmacılar kavramlarla düşünür ve karar verirler. Aşağıdaki paragraflarda bu kavramlar hakkında bilgiler verilmiştir.

Korelasyon matrisi. Korelasyon katsayılarını gösteren tablodur. Korelasyon katsayıları, veri matrisindeki satırlarla sütunlar arasındaki doğrusal ilişkileri gösterir. Korelasyon katsayıları 0'a yaklaştığı ölçüde ilişkinin düşük olduğuna, 1'e yaklaştığı ölçüde ise ilişkinin güçlü olduğuna karar verilir. Negatif işaretli korelasyon katsayıları ilişkinin ters yönlü olduğunu gösterir.

Korelasyon katsayısını yorumlamak için bu değerlerin her biri karesi alınarak 100 ile çarpılır ve yeni bir değer elde edilir. Bu değer, iki değişkendeki ortak olan varyansın yüzdesini verir. Araştırmacı bir değişkendeki değerleri biliyorsa diğer değişkende yüzde kaçlık bir değişiklik ortaya çıkacağını tahmin edebilir.

Korelasyon matrisi, faktörlerin çıkarılması için gerekli olan bir tablodur. Faktörler, korelasyon matrisindeki katsayıların birbirlerine benzer veya farklı olmasına göre belirlenir. Korelasyon matrisindeki değerlerin faktör çıkarmaya uygun olarak yorumlanabilmesi için en az ,30 ağırlığa sahip olması gerekir.

Paydaşlık oranı. Faktör analizinde dikkatle değerlendirilmesi gereken bir diğer kavram, paydaşlık oranıdır (communalities). Paydaşlık oranı bir değişkende, çıkarılan faktörlerin topluca temsil edilme derecesini gösteren yüzde değeridir. Bu orana bakarak değişkenin çıkarılan faktörleri ne ölçüde içerdiğine veya temsil ettiğine karar veririz. Paydaşlık oranı h^2 simgesi ile gösterilir. Analiz sonucunda elde edilen faktöriyel çözüm her bir değişkene ait varyansın en azından yarısını açıklamalıdır. Bu nedenle paydaşlık oranı-

nın ,50 veya daha yukarı olması gerekir.[70] Paydaşlık oranı ,50'den düşük olan değişkenler analizden çıkarılarak faktör analizi yeniden yapılır. Ancak paydaşlık oranı rakamlarını kesin bir şekilde bu rakamla değerlendirmemek gerekir. Paydaşlık oranı düşük olduğu halde faktörle anlamlı bir bağa sahip göstergeler olabileceği gibi paydaşlık oranı yüksek olduğu halde faktörle ilişkisi bulunmayan maddelere de rastlanılabilir.[a] Paydaşlık oranı 1,0'i aşmışsa bu durum, örneklemin çok küçük olduğunu, araştırmacının çok fazla veya çok az faktöre sahip olduğunu gösterir.[71]

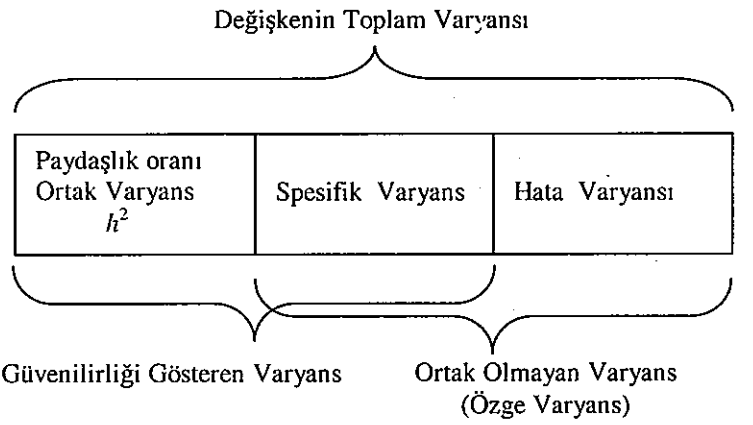
Paydaşlık değerleri aynı zamanda maddelerin/testlerin güvenilirlik göstergeleridir. Bir değişkene ait, faktörlerle ilgili paydaşlık değerinin yüksek veya düşük olması o değişkenin ortak faktörlerle olan ilişkisini ortaya koyar. Paydaşlık değeri düşükse, belirlenen ortak faktörler söz konusu değişkendeki varyansın çok küçük bir bölümünü açıklıyor demektir. Böyle olunca bu değişken söz konusu faktörlerle çok fazla ilgili değildir. Güvenilirliği düşük olan bu maddelerin ölçeğe alınmaması gerekir. Ancak bu maddedeki bilgi veya ifade ölçüm açısından önemli ise o zaman faktör sayısı artırılarak bu değişkenin bir faktör tarafından temsil edilmesi sağlanır. Faktörler çıkarıldıktan sonra bir değişkenin paydaşlık oranının ,50'nin üzerinde olmasına dikkat edilir (Field, 2000).[72] Bir maddenin paydaşlık oranının yüksek olması faktörün yorumlanmasında o maddenin rolünün yüksek olduğunu gösterir.

Herhangi bir değişken, gösterge, test veya madde daha önce belirtildiği gibi kendi bünyesinde üç tür varyans içerir: ortak varyans, spesifik varyans ve hata varyansı.[73] Ortak varyans, faktörün (veya duruma göre faktörlerin) maddede temsil edilme oranıdır. Diğer bir deyişle maddenin paydaşlık oranıdır. Spesifik varyans ortak faktörle/faktörlerle ilgisi olmayan başka bir konuya veya kavramsal alana işaret eder. Hata varyansı ise ölçüm değişkeninin içerdiği düzenleme, ölçme hataları hakkında bilgi verir. Pohlmann bir maddenin değişkenliğini (varyansını) Şekil 8-7'deki gibi açıklamıştır.[74]

Maddenin ortak varyansı ve spesifik varyansı birlikte o maddenin güvenilirlik derecesini gösterir. Maddenin hata varyansı arttığı ölçüde güvenilirliği düşer. Öte yandan, *özge varyans* ortak varyansın dışındaki varyans anlamına gelir. Bu varyans "ayrık varyans" veya "paylaşılmayan varyans"tır. Özge varyans, spesifik varyans ve hata varyansı ögelerinin toplamından oluşur.

---

[a] Bazı bilim adamları sadece paydaşlık oranı ,20'nin altında kalan maddelerin ölçekten/testten çıkarılmasını önermişlerdir.



Şekil 8-7. Maddenin varyans özellikleri.

Kaynak. J.T. Pohlman, "Factor Analsis Glossary [Faktör Analizi Sözlüğü]," <http://www.siu.edu/~epse1/pohlmann/factglos/.> (2.06.2003).

Faktör yükleri. Değişkenlerin her bir faktöre ait özü veya gizli değişkeni içerme oranıdır. Belirli bir orandan daha fazla olarak faktör yüküne (component / factor matrix tablosu) sahip olan maddelerin o faktörü temsil ettiğine karar verilir. Faktör yükü/yükleri, standardize edilmiş regresyon katsayılarıdır ve değişkenle faktör arasındaki korelasyona işaret eder. Lamda simgesiyle (λ_{31}) gösterilen faktör yükleri incelenirken birkaç temel konuya dikkat edilmelidir: (a) faktör yükünün büyüklüğü, (b) çapraz faktör yükü, (c) negatif faktör yükü, (ç) negatif ifadeli maddelerin ayrı bir faktör olarak çıkması.

Faktör yükünün büyüklüğü. Standardize edilmiş regresyon / korelasyon katsayının büyük olması faktörün o madde içinde daha fazla yer aldığı anlamına gelir. Bir maddenin faktör yükü *düşükse*, bunun anlamı o maddenin faktörün ölçtüğü kavramsal yapıyla ilgili olmadığıdır (*bk.*, Tablo 8-4). Ölçümün türüne göre değişmekle birlikte faktör yükü için bilim adamlarının önemli bir bölümü ,40 değerini temel alırlar. Bilim adamı bir maddenin içerik olarak çıkarılan faktörle yakın bir ilişki içinde olduğunu görüyorsa faktör yükünü ,30 gibi bir değere kadar düşürebilir.

Tablo 8-4. Faktör Yüklerinin Süzülerek Gösterilmesi

Testler	Faktör yükleri		
	I	II	III
Sözel yetenek testi	,856		
Sayısal yetenek testi		,689	
Problem çözme yeteneği		,654	
Yaratıcılık testi			,489
Yorumlama testi			,189
Özdeğer			
Toplam varyansın yüzdesi			

Bir maddenin kavramsal yapıyla ilgili olup olmadığına karar vermek için o maddenin faktör yükü en az ±,40 olmalıdır. Titiz araştırmacılar ±,40'ın altındaki maddeleri ölçeğe almazlarken biraz daha serbest hareket eden bilim adamları ±,30'a kadar inmişler ve bu maddelerin de ölçeğe alınabileceğini belirtmişlerdir. Yüzde 30'luk yük, bir değişkende faktör tarafından açıklanan varyansın yaklaşık olarak %10'una tekabül eder.[75] Norman ve Streiner (1994, aktaran Garson) $n \geq 100$ olan örneklemlerde minimum faktör yükünü hesaplamak için bir formül önermişlerdir (*bk.*, Eşitlik 8-1). Ancak bu formül de geçici niteliktedir.[76]

■ Minimum faktör yükü formülü.

$$\text{Minimum FY} = 5{,}152 / \sqrt{(n-2)} \ . \qquad (8\text{-}1)$$

Faktör yükünün anlamı, her bir araştırmada farklı değerlendirilir. Örneğin, ikili maddelerde ,45 faktör yükü, yüksek olarak değerlendirilirken Likert tipi ölçeklerde, bu değerin en az ,60 olması gerektiği belirtilmiştir.[77] Bazı istatistiksel analiz programlarında faktör yükü sınır puanı ön tanımlı olarak ,50 şeklinde belirlenmiştir. Prensip olarak, hesaplama sonucunda faktör yükleri ,40'ın üzerinde çıkmış olan maddelere "belirgin / belirtke değişkenler" (Salient variables) adı verilir. Kim-Yin (2004) ise, faktör yükü değerlerinin örneklem büyüklüğü ile ilişkili oğlunu belirtmiş; faktör yükü ,30 olan maddelerin ölçeğe alınması için örneklem büyüklüğünün en az 350, yüzde 40 faktör yükü için 200, yüzde 50 faktör yükü için 120, yüzde 60 fak-

tör yükü için 85, yüzde 70 faktör yükü için 60 kişilik bir örneklemin yeterli olacağını ifade etmiştir.[78]

Çapraz yükler. Faktör analizinde bazen bir değişkene ait faktör yüklerinin birden fazla faktöre yaklaşık olarak eşit oranda dağılması olgusuyla karşılaşılabilir. Bu durum özellikle eğik döndürme yöntemi için geçerlidir. İngilizcede *cross-loadings* veya *split loadings* terimiyle anlatılan bu olguyu *çapraz-yükler* terimiyle ifade edebiliriz. Bir değişkenin ,40'ın üzerinde[a] olmak üzere birden fazla faktörde eşit veya benzer yük değerlerine sahip olması "karmaşık yapılarla" karşı karşıya olduğumuz anlamına gelir ve bu değişkenler genellikle ölçekten çıkarılır (*bk.*, Tablo 8-5). Ancak değişkenin bir faktördeki yükü oldukça yüksek iken diğer faktördeki yükü düşük ise bu durum çapraz yük olarak değerlendirilmez. Birden fazla faktörde *anlamlı çapraz-yüklere* sahip değerler üç şekilde ele alınabilir.

"Birincisinde, çok sayıda çapraz yüklere sahip bir matris durumuyla karşılaşılmışsa paydaşlık değerlerini daha iyi ortaya çıkması için döndürme yöntemine devam edilir ve faktör sayısı azaltılmaya çalışılır. İkincisinde çapraz yüklere sahip maddelerin ifadelendirme biçimi gözden geçirilir. İfadelerin yüzey geçerliliği dikkate alınarak bu maddeler kendileri için en uygun olan faktörün altında toplanır. Üçüncüsünde ise yüksek çapraz yüklü ifadeler ölçekten/testten çıkarılır ve analiz yeniden yapılır. Bu yöntem az sayıda çapraz yüklü madde olduğu zaman uygulanır."[79]

Hangi yöntemin seçileceği, bilim adamının varsayımlarına ve ölçüm amacına göre değişir. Bilim adamı faktörlerin birbirinden bağımsız olduğunu düşünüyorsa çapraz yüklü değişkenleri iptal etmek daha iyi bir yöntemdir. Faktörlerin bağımlılığı söz konusuysa değişkenlerin anlamlarına bakılarak bu değişken veya değişkenler en uygun faktör altında toplanır.

Negatif işaretli faktör yüküne sahip maddeler. Bazen faktör matrisi tablosundaki faktör yüklerinin (component / factor matrix tablosu) işareti negatif çıkar. Bilim adamı –,40'ın altındaki negatif işaretli küçük faktör yükleriyle ilgilenmediğinden özellikle *büyük ve negatif işaretli faktör yükleri* sorunu üzerinde odaklanır. Negatif faktör yükleri, çıkarılan faktörün söz konusu değişkenlerle ters yönde ilişkili olduğu anlamına gelir.[80] Bilim adamı bazı maddelerde negatif faktör yükleriyle karşılaşmışsa güvenilirlik analizi yapmadan önce bu maddeleri tersine çevirmelidir.[81] Öte yandan negatif faktör

[a] Bazı bilim adamları bu değeri ,30'a kadar çekmişlerdir. Bir değişkenin birinci faktördeki yükü ,40'ın üzerindeyken ikinci faktördeki yükü,30'un altındaysa bu durum çapraz faktör yükü olarak isimlendirilmez.

yükleri, testin iki kutuplu faktöriyel bir yapıya sahip olduğu anlamına da gelebilir.

Tablo 8-5. Çapraz Yüklü Değerlere Sahip Tablo Örneği

	Bileşenler	
	1. Bileşen	2. Bileşen
Yenilik	,009	,959
Detay	,765	,781
Rekabet	,751	,647
Takım	,899	,034
Ahlâk	,972	,023

Analiz sonucunda birden fazla değişkende negatif ve pozitif işaretli yüksek faktör yükleri ortaya çıkmışsa iki kutuplu faktöriyel bir yapıdan şüphelenilir. Faktör yüklerinin hepsi pozitif işaretli ise tek kutuplu faktör yapısı var demektir.[82] Örneğin, araştırmacı içedönük ve dışadönük kişilik yapılarını araştırıyorsa böyle bir durumda iki kutuplu faktör yapısı ortaya çıkabilir. Ancak araştırmacı sayısal ve sözel zekayı ortaya çıkarmak istiyorsa bu iki faktör birbirleriyle düşük derecede ilişkili olamayacaklarından zıt yönlerde yer almayacaktır. Eğer pozitif ve negatif işaretler zıt yönlerde yer almışlarsa böyle bir durumda rotasyon sayısı artırılarak mümkün olduğu kadar daha fazla pozitif yüklü madde sayısı elde edilmeye çalışılır.[83] Negatif işaretli ve küçük değerlikli faktör yükleri aynı zamanda *spesifik faktör* anlamına gelir. Bu faktör yükleri ortak faktörün dışında ve özgün nitelikte başka bir bilgiyle yüklüdür.[84] Dik açılı döndürme yönteminde negatif yükler ters yönde ilişkili olmayı gösterirken, eğik döndürme (oblique rotation) yöntemi uygulandığında yüksek negatif faktör yüklerinin özel bir anlamının olmadığı ve bu nedenle yorumun pozitif işaretli mutlak değerlere göre yapılması gerektiği bildirilmiştir.[85]

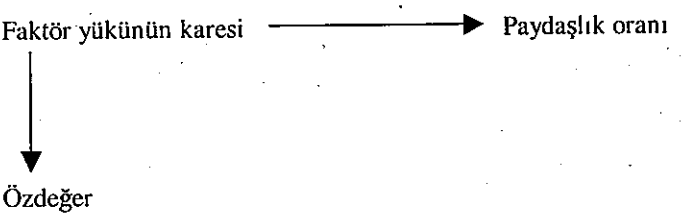
Negatif ifadeli maddelerin ayrı bir faktör olarak çıkması. Bilim adamları işaretleme yanlılığını azaltmak için negatif ve pozitif ifadeli maddelerin birlikte kullanılmasını önermişlerdir. Kişilik ve tutum ölçeklerinde pozitif ve negatif ifadeli maddelerin eşit sayıda olması bir kural haline gelmiş ve pek çok yazar tarafından önerilmiştir (Shiffman ve Jarvik, 1976; Tiffany ve Drobes, 1991; Mueller, 1985; Anastasi, 1988; Mehrens ve Lehmann, 1991,

aktaran Torabi ve Ding).[86] Kişiler eğer negatif ve pozitif ifadeli maddelere aynı şekilde cevap vermişlerse *yanıt yanlılığı* söz konusudur. Bu şekildeki yanlı anketlerin örneklem hacminden çıkarılarak verilerin doğruluğunun artırılması gerekir.[87] Negatif işaretli maddeler bir taraftan yanlı işaretlemeyi azaltırken diğer taraftan hedeflenenlerin dışında ayrı faktörlerin çıkmasına neden olabilir. Bazı bilim adamlarına göre negatif ifadeler pozitif ifadelerin tam tersi olmayabilir. Bu ifadeler bir boyutla ilgili gibi gözüküyor olmalarına karşın başka bir kavramsal yapıya ait olabilir.[88] Araştırmacılar olguyu hem ampirik, hem da kuramsal açıdan ele almışlar ve bu olguya değişik açıklamalar getirmişlerdir. Negatif ifadeli maddelerin birbiriyle ilişkili olması aşağıdaki faktörlerden kaynaklanabilir.

1. Cevaplayıcıların eğitim yetersizliği nedeniyle pozitif ve negatif maddeleri ayırt edemeyerek maddelerin hepsine aynı yanıtları vermiş olmalarından.
2. Cevaplayıcıların negatif işaretli maddeleri işaretleme konusunda isteksizlik duymalarından (negatif madde yanlılığı).
3. Cevaplayıcıların anketi dikkatsiz doldurmuş olmalarından.
4. Bazı vak'alarda negatif ifadelerin, amaçlanan kavramsal yapı yerine ayrı bir faktörle yüklü olma eğilimi taşıyor olmasından.[89]

King'in bildirdiğine göre, Marsh (1996) tarafından 20.000'i aşkın kişi üzerinde yapılan bir araştırmada tek boyut çıkarılmak istenmesine rağmen iki temel faktör ortaya çıkmış ve tersine çevrilmiş negatif ifadeli maddeler ayrı bir boyut altında toplanmıştır.[90] Yine aynı makalede cevaplayıcıların %10 kadar küçük bir kesiminin bile anketi *dikkatsiz doldurmaları* halinde negatif ifadeli maddelerin ayrı bir faktör ortaya çıkarabileceği belirtilmiştir. Bu nedenle bilim adamları araştırmacıları sadece negatif ifadeli maddelere dayalı olarak ortaya çıkan faktörler konusunda dikkatli olmaları için uyarmışlardır.[91] Kişilerin eğitim düzeyleri veya sözel yetenekleri düşük olduğunda, negatif maddeleri işaretleme isteksizliği duymaları halinde negatif işaretli maddelerin ayrı bir faktör altında toplanma olasılığının arttığı anlaşılmaktadır. Bu tür faktörler, gerçek kavramsal yapılar değil, yapay görüntülerdir. Bilim adamı, negatif ifadeli maddelerin gerçek anlamda ayrı bir faktör ortaya çıkması ile katılımcıların eğitim düzeyleri ve işaretleme uygulamalarının sonucunda böyle bir olguyla karşılaşılması arasında ayırım gözetmelidir.

Faktör yükünün karesi. Faktör yükleri, gözlem değişkenleriyle faktörler arasındaki korelasyon katsayılarını tanımlarken faktör yükünün karesi regresyon analizindeki R^2'ye benzetilebilir. Faktör yükünün karesi orijinal değişkende faktörle açıklanan varyansın yüzdesini temsil eder. Bir faktöre ait değişik maddelerin faktör yüklerinin karelerinin toplamı özdeğeri verir.[92] Bir maddedeki değişik faktörlere ait faktör yüklerinin karelerinin toplamıyla ise o değişkenin paydaşlık oranı elde edilir (*bk.*, Şekil 8-8).



Şekil 8-8. Faktör yükü karesi değerlerinden iki yeni farklı değer elde edilmesi.

Etiketleme. Faktör analizi sonucunda ortaya çıkan faktör yüklerine bakılarak, bir faktör altında toplanabilecek gözlem değişkenleri belirlenir. Bundan sonraki sorun söz konusu gizli değişkenlere uygun bir isim bulmaktır. Faktörlere isim bulma işlemine "etiketleme" denir. Faktörlere etiket verme, sonuçları tartışmayı ve yorum yapmayı kolaylaştırır; daha sonraki araştırmacıların faktörlerin isimlerini hatırlamalarını ve yeni modeller geliştirmelerini olanaklı hale getirir. Faktörlere etiket vermeyle bileşenlere etiket verme farklı iki olaydır. Faktör etiketleri gerçek iken bileşen etiketleri bir tür yakıştırmadır. Bileşen etiketleri bilimsel ve kuramsal bir temele sahip değildir. Bir faktörün altındaki değişkenler veya değişkenlerden bir kısmı negatif faktör yüklerine sahipse bu değişkenlerin faktörle ters yünlü olarak ilişkili olduğu yorumu yapılır. Bilim adamı, faktör yükü tablosundaki (component /factor matrix) modelleri üç şekilden birine göre etiketleyebilir; *simgesel olarak, tanımlayıcı bir biçimde* ve *nedensel olarak.*[93]

Simgesel etiketler, herhangi bir anlamı olmayan basit işaretlerdir. Bu işaretlerin amacı farklı faktör yükü modellerini ortaya koymaktır. Örneğin bir araştırmada üç faktör yapısı ortaya çıkmışsa simgesel olarak A, B, ve C harfleriyle veya F1, F2 ve F3 harfleriyle gösterilebilir. İkinci yaklaşım faktörlere tanımlayıcı isimler vermektir. Tanımlayıcı isimler verme bazen oldukça kolay bir şekilde yapılabilirken bazen de oldukça zor bir işlemdir.

Burada araştırmacının sezgileri, görüş ufku, literatür bilgisi ve uygun kavramı bulma konusunda kelime hazinesinin zenginliği gibi etkenler rol oynar. Seçilen kavram anlamlı ifadeleri kapsayan ve bilinen bir sözcük olmalıdır. Bilim adamı orijinal yeni sözcükler türetmeye çalışmamalıdır.[94] Faktör yorumları ve etiketler yüzey geçerliliğine sahip olmalı ve kuramsal bir temele sahip bulunmalıdır. Nedensel etiketlemede ise, bilim adamı gözlem değişkenlerinin nedeni olan ve arka planında yatan gizli değişkeni bulmaya çalışır. Örneğin, yönetici gerilimini araştıran bir bilim adamı yükselme kaygıları, uzmanlaşma ve unvan elde etme konusunda ortaya çıkan yüksek faktör yüklerini "kariyer" kavramıyla etiketleyebilir.

Etiketlemede yararlanılabilecek bir diğer yöntem hakemlerden yararlanmaktır. Ölçüm konusuyla ilgili iki üç hakemle görüşülerek kendilerine faktör yükü modelleri gösterilir. Birinci aşamada bağımsız olarak bu modelleri değerlendirmeleri ve isim bulmaları istenir. İkinci aşamada ise bir araya getirilen hakemler panel çalışması yaparak faktör isimlerini birlikte tartışırlar. Bilim adamı daha sonra bağımsız olarak yapılan değerlendirmelerin ve birlikte yapılan tartışmaların sonucuna göre nihaî kararı kendisi verir.

Döndürülmemiş ve döndürülmüş faktör matrisi. Faktör analizi sonucunda iki tür faktör matrisi elde edilir; birincisi döndürülmemiş faktör matrisi, ikincisi ise döndürülmüş faktör matrisidir. Uygulamada döndürülmemiş faktör analiziyle ilgili bir yorum yapılmaz. Yorumlar daha çok döndürülmüş faktör analizlerine dayandırılır.

Döndürme, faktör eksenlerinin basit ve pratik olarak daha anlamlı çözümler verecek şekilde saat yönünde çevrilmesi anlamına gelir. Döndürme işlemi araştırmanın amacına göre belirlenir. Bilim adamı faktör analizi sonucunda ortaya çıkacak olan faktörlerin birbirinden bağımsız veya birbirine bağımlı olduğunu düşünebilir. Örneğin; genel yeteneği ölçmeye yönelik olarak oluşturulan bir psikoteknik test bataryasında sözel, sayısal, düzlemsel, kelime bilgisi, sözel akıcılık test puanları bulunsun. Bu puanlar üzerinde yapılacak bir faktör analizinde ortaya çıkacak faktörler büyük bir ihtimalle birbirinden bağımsız olacaktır. Ancak araştırmacı kaygıyı ölçen bir ölçek geliştirmişse faktör analizi sonucunda ortaya çıkacak boyutlar muhtemelen birbiriyle ilişkili çıkar. Faktörlerin birbirleriyle ilişkili olup olmamalarına göre döndürme işlemi iki temel grup altında toplanır ve bu gruplar ile içerdikleri döndürme yöntemleri aşağıdaki gibidir.

Dik açılı döndürme yöntemleri (orthogonal). Dik açılı döndürme, bir kavramsal yapıda eğer F_1 ve F_2 şeklinde iki faktör varsa bu faktörlere ilişkin vektör çizgilerinin 90 derecelik açıyla döndürülmesi anlamına gelir. Bu

döndürme sonunda değişkenleri temsil eden noktalar daha anlamlı bir şekilde gruplandırılmış olur. Dik açılı döndürme biçiminde gizli yapıların ilişkisiz olduğu varsayılır. Faktör çıkarma işleminde, eksenler 90 derece doğru açı ile döndürülerek birbirleriyle ilişkisiz olacak şekilde farklı noktalarda konumlandırılır. Bu döndürme biçiminde faktörler arasındaki korelasyon sıfırdır. Bilgisayar çok boyutlu uzayda faktör eksenlerini, değişkenleri temsil eden noktalara olan uzaklıkların karesi en düşük değere gelinceye kadar döndürmeye devam eder ve en uygun noktada durur. Böylece vektörleri, verilerle en iyi uyuşan noktada dondurur. Fakat, bu yöntem faktörler arasındaki "gerçek" ilişkileri tam olarak temsil etmez.[95]

Araştırmacı faktörlerin birbirinden bağımsız olduğunu düşünüyorsa dik açılı döndürme yöntemine başvurur. Yine, faktörlerin her birini bağımsız alt ölçekler olarak kullanacak ve alt ölçek puanları arasında regresyon analizi yapacaksa dik açılı döndürme yöntemini tercih eder. Sık kullanılan dik açılı döndürme yöntemleri aşağıdaki gibidir.

Varimaks (Maksimum değişkenlik – Varimax). Varimaks yöntemi, dik açılı döndürme biçimlerinden biridir. Varimax yönteminde faktör matrisi sütununda bulunan değerlerin karesi alınarak varyans maksimum değerine çıkarılır. Varimaks yöntemi faktör etiketlerine yorumlama kolaylığı getirirken değişkenlerdeki faktör yüklerinin yorumlanmasını zorlaştırır.[96]

Yöntem bir faktör altında toplanabilecek değişkenlerin sayısını minimum düzeye düşürmeye çalışır. Literatür araştırmaları dik açılı döndürmelerde daha çok varimaks yönteminin tercih edildiğini göstermektedir. Birden fazla bağımsız faktör / alt boyut ortaya çıkarmayı amaçlayan araştırmacılar varimaks yöntemini kullanırlar. Araştırmacı eğer genel, baskın bir faktöre ulaşmak istiyorsa veya ölçeğin tek boyutlu olduğunu kanıtlamaya çalışıyorsa bu yöntemi kullanmamalıdır.

Kuartimaks (En büyük çeyrek – Quartimax). Bu yöntem gizli değişkeni açıklamak için ihtiyaç duyulan faktörlerin sayısını minimize etmeye çalışır. Matematiksel olarak dördüncü güç maksimizasyonu anlamına gelen bu yaklaşımda *basit faktör yapısı* ortaya çıkarılmaya çalışılır. Birinci faktörün karmaşıklığı maksimize edilmeye çalışılırken diğer değişkenlerin karmaşıklığı en alt düzeye düşürülür. *Quartimax* yaklaşımının amacı *g* faktörünü ortaya çıkarmak olarak belirlenmiştir.[97] Araştırmacı bu yöntemi alt ölçeklerde tek bir faktörü veya baskın faktörü ortaya çıkarmak istediği zaman kullanır. Bu uygulamada baskın faktörün dışındaki diğer faktörlerin faktör yükleri düşük çıkar.

Ekuamaks (Eşit ölçüde maksimize etme – Equamax). Saunders (1962) tarafından önerilen yaklaşım varimax ve quartimax yöntemlerinin yaptığı işin her ikisini de birlikte gerçekleştirmeye çalışır. Ancak, Tabachnick ve Fidel (aktaran Kinnear) bu yönetimin güvenilmez olduğunu bildirmişlerdir.[98] Literatürde yaygın kullanımı olmayan bir tekniktir.

Orthomax. SAS isimli istatistiksel yazılımda dik açılı döndürme yöntemlerine genel olarak verilen bir addır. SPSS'te olduğu gibi, *Quartimax* ve *Varimax* döndürme yöntemlerini içerir.

Eğik açılı döndürme yöntemleri (oblique rotation). Saunders (1961) tarafından önerilen eğik açılı döndürme yöntemlerinde faktörlerin birbiriyle ilişkili oldukları varsayılır. Teknik, faktörleri beklenenden çok daha fazla ilişkili hale getirir. Eğik rotasyon kullanıldığında en önemli farklılık şudur: Eğik rotasyonda faktör yükleri artık sadece basit bir şekilde faktörlerle değişkenler arasındaki korelasyon katsayıları olarak yorumlanamaz. Eğik döndürme yönteminde iki farklı matrise ihtiyaç vardır: *Faktör yapısı matrisi* (factor structure matrix) ve *faktör modeli matrisi* (factor pattern matrix). Faktör modeli matrisi, her bir değişkenin özge katkılarını belirleyen katsayıları gösterir. Faktör yapısı matrisi ise, faktörlerle değişkenler arasındaki basit korelasyon katsayılarını gösterir. Bu yöntemde, faktörlerin birbirleriyle ilişkili olmaları nedeniyle bu iki matrisin aynı olması gerekmez. Eğik rotasyon yöntemi kullanıldığında yapı ve model matrislerinden hangi matrisin yorumlanması gerektiği konusu tartışmalıdır. Bazı yazarlar, her iki matrise ait verilerin birlikte ele alınıp yorumlanması gerektiğini söylemişlerdir. Rummel ise faktörlerin daha kolay yorumlanmasına imkan sağlaması nedeniyle sadece *model matrisinin* yorumlanmasının yeterli olacağını bildirmiştir (aktaran, Stanek, 2003).[99]

Bilim adamı çıkarılacak faktörlerin birbiriyle ilişkili olup olmadığı konusunda bir fikir sahibi değilse önce eğik döndürme yöntemini uygulamalıdır.[100] Analiz sonucunda faktörler arasındaki ilişkileri saptayan *faktör korelasyon matrisi* verilerini inceleyerek kararının isabetli olup olmadığını belirler. Eğik döndürme yöntemi sonucunda faktörler birbirleriyle ilişkili çıkmışsa veya faktörler arasındaki korelasyon katsayıları yüksekse isabet etmiş demektir. Eğer katsayılar sıfıra yakın değerler olarak elde edilmişse (veya korelasyon katsayısı < ,30 ise) bu kez dik açılı döndürme yöntemi denenir.[101] Karar vermede yararlanılacak ikinci yöntem "teyit edici faktör analizi" sonuçlarından hareket etmektir. Bilim adamı yaptığı teyit edici faktör analizi sonucunda eğer iki faktörün birbiriyle ilişkili olduğunu görmüşse eğik döndürme yöntemlerini, ilişkisiz olduğunu bulmuşsa dik açılı döndürme yöntemlerini uygular.

Eğik döndürme yöntemi, tutum ölçeklerinde faktörler arasındaki ilişkileri daha gerçekçi bir şekilde açıklar, ancak bu yöntemde faktörlerin kavram olarak açıklanmasında veya etiketlenmesinde zorlukla karşılaşılır. SPSS'teki eğik rotasyon yöntemleri aşağıdaki gibidir.

Direkt oblimin (direct oblimin). Eksenleri, 90 derecenin dışında herhangi bir açıyla döndürme yöntemidir. Yöntem faktörlerin birleriyle ilişkili olmasına izin verir. Faktörlerin kendi aralarında ilişkili olma derecesini belirlemeye yönelik olarak belirli bir *delta* değeri belirlenir. Delta 0 veya negatif işaretli herhangi bir değerdir. Sıfır değeri en yüksek derecede birbirleriyle ilişkili olan faktörleri ortaya çıkarırken, büyük negatif değerler dik açılı döndürmeye yakın değerler verir.[102] Örneğin –4 tam bir dik açılı döndürme sonucu verir. SPSS yazılımının Rotation diyalog kartında 0 değeri ön tanımlı olarak verilmiştir. Bu yöntem faktörlerin özdeğerlerini yükseltirken, yorumlanmalarını güçleştirir.[103]

Kuartimin (Quartimin – En küçük çeyrek). SPSS'in direkt oblimin bölümündeki delta değerinin sıfır olarak tanımlanmış biçimiyle olan uygulaması kuartimin yöntemi olarak tanımlanmıştır ($\delta = 0$).

Promaks (Promax). Direkt oblimin yöntemine göre daha hızlı hesaplama yaptığından büyük veri gruplarında kullanılması önerilmiştir. Yazılımda promax döndürme yöntemini kontrol eden kappa değeri ön tanımlı olarak 4 rakamıyla belirlenmiş ve bu rakamın pek çok analiz için uygun olduğu belirtilmiştir. Kappa değeri (κ) 2, 4, ve 6 gibi rakamlar şeklinde tanımlanır. Yüksek değerler faktörler arasında daha fazla korelasyon olmasını sağlar. Promax yöntemi *ilişkili* basit yapıları ortaya çıkarmak için kullanılırken eğik döndürme yöntemini değil, dik açılı döndürme yöntemini uygular. Bu özelliği ile hem dik açılı döndürme yöntemleri ve hem de eğik döndürme yöntemleri içinde sayılabilir. Bu nedenle literatürde melez döndürme yöntemi olarak adlandırılmıştır.

SPSS'te bulunmayan diğer eğik rotasyon yöntemleri ise aşağıdaki gibidir.

Bi-kuartimin (Biquartimin – İkili en küçük çeyrek). Faktör yükleri matrisi eğik döndürme yöntemiyle öyle bir şekilde döndürülür ki, *faktör yükü karesi* en yüksek olan tek bir faktör ortaya çıkar. Diğer değişkenlerin faktör yükü kareleri ise sıfıra yakın değerler olarak elde edilir. *Biquartimin* döndürme yöntemi SAS isimli istatistiksel analiz yazılımında bulunur.

Prokustes (Procrustes). Bu yöntemde gözlem verilerinin varsayılan faktör yapısına ne ölçüde yakın olduğu test edilir. Matematiksel olarak belirli bir *Y* matrisindeki noktaların *X* matrisindeki noktalara tam olarak uyuşum göstermesini sağlamak için dik açılı döndürme imkanı sağlayan bir tekniktir. Yöntemde uyuşma kriteri, hata karelerinin toplamına bakılarak belirlenir.[104] Procrustes döndürme yöntemi SAS isimli yazılımda bulunur.

Harris-Kiser. İstatistiksel analiz yazılımı SAS'ta bulunan bu döndürme yönteminde özel bir formül kullanılır. Bu formül sayesinde özdeğerlerin kareköklerinin gücü belirlenerek bu güçle özdeğer vektörleri yeniden ölçeklendirilir. Hesaplama sonucunda Harris-Kaiser *p* değeri elde edilir. Bu değerin 0 çıkması her bir değişkenin ağırlıklı olarak tek bir faktörle yüklendiği anlamına gelir. Öte yandan 1,0 değeri ise dik açılı döndürme yöntemine eşittir.[105]

Eğik döndürme yönteminin sonucunda SPSS'ten Faktör Korelasyon Matrisi (factor correlation matrix) isimli tablo elde edilir. Bu tablo çıkarılan faktörlerin birbirleriyle ne ölçüde ilişkili olduğunu gösterir.

Dik açılı ve eğik döndürme yöntemlerinin her ikisiyle analiz yapıldıktan sonra hangisinin seçilmesi gerektiği konusunda iki kritere bakılır. Bunlardan birincisi bir faktörün altında kaç değişkenin toplandığıdır. İkincisi ise faktörler arasındaki korelasyonun derecesidir. Bir faktör altında toplanan değişken sayısı azsa, dik açılı ve eğik döndürme yöntemlerinin her ikisi de benzer sonuçlar verir. Ayrıca faktörler arasındaki korelasyon katsayıları sıfıra yakın değerler olarak elde edilmişse, büyük bir ihtimalle her iki yöntem yine benzer sonuçlar verir.[106] Faktörler arasındaki korelasyon katsayısı ,32'den büyükse bu kez eğik döndürme yöntemi tercih edilir.[107]

Basit faktör yapısı. Thurstone tarafından önerilen *basit faktör yapısı* her bir değişkenin sadece veya ağırlıklı olarak tek bir faktör altında konumlanması anlamına gelir. Testin veya ölçeğin basit faktör yapısını ortaya çıkarmak için çoğunlukla döndürme yöntemine başvurulur. Pek çok vak'ada basit faktör yapısı eğik döndürme yöntemiyle elde edilir, fakat bu yöntemin sakıncası faktörlerin birbiriyle ilişkili olması ve bu nedenle de yorumlama güçlüğü yaratmasıdır.[108] Bir ölçeğe ait değişkenler üzerinde faktör analizi yapıldığında ikiden fazla faktör ortaya çıkmakla birlikte bir faktör özellikle *baskın bir biçimde* ortaya çıkmışsa ölçeğin *basit faktör* yapısına sahip olduğunu söyleriz. Maddelerin faktör yükleri bir faktöre yönelik olarak ,40 veya daha fazla faktör ağırlıklarına sahip iken diğer faktörlerdeki ağırlıkları sıfıra yakın değerler olarak çıkmışsa ölçeğin basit faktör yapısına sahip olduğu anlaşılır.[109]

Faktör Çıkarma

Faktör çıkarma işlemi istatistik yazılıma bırakıldığında eğer temel bileşenler analiz yöntemi seçilmişse bilgisayar, değişken sayısı kadar faktör çıkarır. Araştırmacı bir şekilde bu faktörlerden hangilerini alıkoyacağına kendisi karar vermelidir. Faktör analizi yöntemini kullanarak *faktör çıkarma* ve faktör sayısını tespit etme konusunda birkaç önemli yöntem vardır. Aşağıdaki paragraflarda bu yöntemlere ilişkin özet bilgiler verilmiştir.

Toplam varyansın yüzdesi. Kural olarak "total variance explained" tablosundaki toplam varyansın ,90'ının açıklayan *n* sayıda faktör/bileşen testin veya ölçeğin faktöriyel yapı elementi olarak belirlenir. Ancak oran bu kadar yüksek tutulunca faktör sayısının geniş olma ihtimali vardır. Faktör sayısını daha sınırlı tutmak isteyen araştırmacılar toplam varyansın ,80'ini açıklayan faktör sayısını temel alırlar. Toplam varyansın yüzdesini temel almak isteyen araştırmacılar ayrıca her bir bileşenin/faktörün yaptığı katkıyı da göz önünde bulundurmalıdırlar. Literatürde genellikle her bir bileşenin ağırlığı en az %10 olduğu durumda yığışımlı varyans toplamı içine alınır.

Önemsiz varyans değerine kaçıncı faktörde ulaşıldığını araştırma. Önemsiz varyans değeri, %10'un altında kalan rakamlardır. Araştırmacı önemsiz varyans değerini bulmak için verilerde giderek artan biçimde kendi tanımladığı sayıda faktör çıkarma hesaplaması yaptırır.[110] Bunun için verilerde önce 1 faktöre dayalı hesaplama yaptırılır. Daha sonra iki, üç, dört ve beş faktöre dayalı hesaplama yaptırılır. Son faktör sayısında varyans değeri eğeri %10'un altında çıkmışsa ondan bir önceki faktör sayısı veriler için ideal faktör sayısı olarak belirlenir.

Kuramın incelenmesi. Bilim adamı, faktör yapısını belirlemek için öncelikle kavramsal yapıyla ilgili kuramsal bilgileri araştırır. Literatürde örneğin, yönetici gerilimine neden olan faktörler eğer dört grup altında açıklanmışsa bu açıklamaların başka araştırmalarla doğrulanıp doğrulanmadığına bakılır. Eğer dört faktör olgusu genel olarak benimsenmişse araştırmacı bu kuramsal bilgiye dayalı olarak hareket eder ve çıkarılacak faktör sayısını önceden belirleyebilir. Ancak modelde alınan son faktörün varyans değerinin önemsiz olmaması gerekir. Bu yöntem özellikle teyit edici faktör analizi için uygundur. Faktör sayısının kurama dayalı olarak belirlenmesine *apriori kriter* adı verilir.

Yamaç-birikinti grafiği analizi. Yamaç-birikinti grafiğinin adı, şekildeki görüntünün yamaca benzemesi nedeniyle verilmiştir. Düşey kırıklı çizgi-

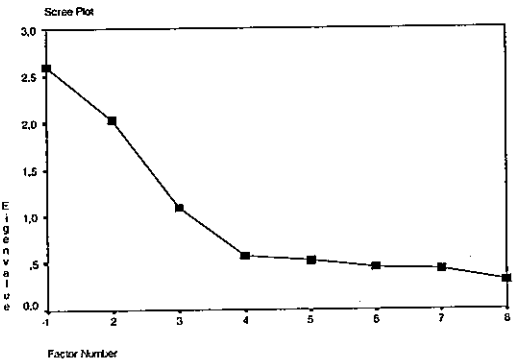
ler yamacı temsil ederken yatık kırıklı çizgiler yamacın dibinde toplanan birikintilere benzetilmiştir. Yamaç-birikinti grafiğinde yatay eksende boyut sayısı, dikey eksende ise özdeğer rakamları gösterilmiştir. Raymond B. Cattell tarafından geliştirilen bu grafikte birinci faktör, en yüksek bileşik varyans değerine sahiptir. Bu yöntemde faktör sayısı, grafiğin düzleşmeye başladığı keskin dirsekten *önceki* "çizgilerin" veya duruma göre "noktaların" sayısına bakılarak belirlenir. Cattell, dirsekte yer alan noktayı faktör sayısına dahil etmemiştir. Fakat daha sonraki araştırmacılar, çoklu dirseğe sahip grafik çizimlerini dikkate alarak düzleşmenin başladığı son noktayı da faktör sayısına dahil etmişlerdir. Grafiğin eleştirilen yönü, aynı görüntüden farklı yazarların çizgi veya noktayı temel alarak değişik sonuçlar çıkarabilmesidir. Bazı liberal yazarlar düzleşmenin başladığı noktayı da faktör sayısına ilave ederler. Şekil 8-9'daki grafik incelendiğinde faktör sayısı 3 veya 4 olarak belirlenebilir. Faktör sayısı çizgiler temel alınırsa 3, noktalar temel alınırsa 4'tür. Yamaç-birikinti grafiği konusundaki bir diğer belirsizlik düzleşme dirseğinin nerede başladığı konusunda ortaya çıkar. Bazı grafiklerde düzleşme dirseği herhangi bir tartışmaya meydan vermeyecek kadar açık iken diğerlerinde araştırmacıyı karar verme güçlüğü içinde bırakabilir. Bu konuda karar verme zorluğunu ortadan kaldırmak için diğer faktör çıkarma kriterleri de dikkate alınmalıdır. Yamaç-birikinti grafiğinin bazen olması gerekenden daha az faktör sayısı verdiği bildirilmiştir. Grafik, tek başına karar vermede yeterli olmadığından diğer karar kriterleriyle birlikte ele alınır. Keskin dirsek noktasındaki faktör sayısı ,70 toplam varyans değeri ve özdeğer > 1 değeriyle ile birlikte değerlendirilebilir. Yamaç-birikinti grafiği genelde Kaiser kriterinden (özdeğer) daha fazla faktör verir.[111] Değişik değerlendirme yöntemlerinde farklı faktör sayılarıyla karşılaşılıyorsa böyle bir durumda yorumlanabilir en iyi çözümü veren ve mümkün olan en düşük faktör sayısı temel alınır. Bir başka yaklaşım, araştırmacının gündemine aldığı veya çıkarmak istediği faktör sayısıdır. Böyle bir durumda araştırmacı yamaç-birikinti grafiği sonuçlarına bağlı olarak hareket etmeyebilir.

Joliffe kriteri. Çok daha az kullanılan liberal bir kuraldır. Joliffe kriterinin yamaç-birikinti grafiğiyle elde edilenin iki katı kadar faktör çıkarabildiği bildirilmiştir. Joliffe kuralında özdeğer ,70'in altındaki bileşenler / faktörler kırpılır.

Yorumlanabilirlik. Bu yöntemde faktörlere isim verme olgusu göz önünde bulundurulur. Birkaç değişik uygulama yapılarak hangi sayıda faktöre daha anlamlı bir şekilde isim verilebildiğine bakılır. Örneğin dört faktör çıkmışsa fakat bunlardan sadece üç tanesi anlamlı olarak isimlendirilebili-

yorsa o zaman sadece üç faktör temel alınır.[112] Bilim adamı yorumlanabilirlik özelliğinin incelenmesinde aşağıdaki faktörlere göre hareket eder.[113]

1. Faktör yükleri ,40'tan büyük olan en az üç değişken var mı?
2. Söz konusu üç değişken bir faktöre birlikte asılırken aynı kavramsal anlama sahip mi?
3. Değişkenler farklı faktörler altında toplanırken gerçekten farklı yapılar ölçülmekte midir?
4. Döndürülmüş faktör modeli basit bir kavramsal yapıyı ortaya çıkarmış mıdır?
5. Tek bir faktör diğerlerine göre daha yüksek bir faktör yüküne sahip midir?
6. Faktörün anlamı kolaylıkla kavranabiliyor mu?



Şekil 8-9. Yamaç-birikinti grafiği.

Özdeğer. Özdeğer (eigenvalue, latent root), bir faktörün toplam varyans içinde sorumlu olduğu varyansın miktarını açıklar. Özdeğer lamda simgesiyle (λ) gösterilir ve istatistiksel olarak değişkenlerin sütun bazında *faktör yükü karelerinin* toplamından oluşur. Farklı sütunlardaki her bir faktör için ayrı ayrı hesaplanır. Analiz sonucunda Kaiser kuralına göre özdeğeri[a] 1'den büyük olan faktörler dikkate alınır, diğer faktörler testin / ölçeğin faktöriyel yapısından çıkarılır. Bazı yazarlar bunun iyi bir kural olmadığını ve kulla-

[a] Kelimenin aslı olan *eigenvalue* teriminin İngilizcede birleşik yazılması nedeniyle Türkçede de birleşik yazım biçimi tercih edilmiştir.

nılmaması gerektiğini bildirmişlerdir. Bu kural test veya maddeler 20 ilâ 50 arasında olduğu zaman güvenilirdir. Madde sayısı 20'den küçük olduğunda Kaiser ölçütü tutucudur, çok az faktör çıkmasına neden olur. Öte yandan 50'den büyük olduğunda ise oldukça liberal sonuçlar elde edilir ve çok sayıda faktör çıkarılır.[114] Bu uygulamadaki yaklaşım, her bir değişkenin girdi olarak en azından 1,0 varyansına sahip olacağıdır (Korelasyon matrisinin köşegeninde yer alan değerler).[115] Temel bileşenler analizinde ilk faktör / bileşen varyansın en önemli bölümünü açıklar. İkinci sıradaki bileşen ikinci en büyük varyansı açıklar ve böylece varyans diğer faktörlerle en alt düzeye kadar açıklanmaya devam edilir.

Faktör analizi yöntemini uygulayan araştırmacı bu tekniğe çok fazla hakim değilse faktör çıkarmak için yukarıda açıklanan tekniklerden ilk etapta *özdeğer* ve *yamaç-birikinti grafiği* değerlerini dikkate almalıdır. Bu değerler eğer birbiriyle uyuşmuyorsa o takdirde *toplam varyansın yüzdesi* değerleri incelenerek değişkenlerin kaç faktör altında toplanacağına karar verilir. Faktör sayısı bu şekilde belirlendikten sonra söz konusu rakam SPSS'te *Factor Analysis: Extraction* diyalog kutusuna tanımlanır ve analiz yeniden yapılır.

Bazen küçük özdeğerler negatif işaretli çıkar. Bu durum özellikle ortak faktör analizi için geçerlidir ki araştırmacılara negatif işaretli bu değerlerle çok fazla ilgilenmemeleri önerilmiştir. Kimi yazarlar ise, negatif işaretli ve sıfıra yakın çıkan özdeğerlerin ham verilerin çoklu doğrusallık özelliğinden kaynaklanabileceği uyarısını yapmışlardır.[116]

Hesaplama yöntemi. R.B. Darlington tarafından önerilen bu yaklaşımda her bir özdeğer *L* için daha sonraki özdeğerlerin toplamı alınarak *S* değeri elde edilir.[117] Her bir ilave faktörün açıklanamayan varyansın yüzde kaçını temsil ettiğini görmek için *L* değeri *S* değerine bölünür (L/S). Öreğin, sekiz değişkene sahip bir ölçekle toplanan veriler üzerinde yapılan faktör analizi sonucunda son dört faktörün özdeğerleri şu şekilde belirlenmiş olsun: ,92, ,25, ,15 ve ,10. Önceki üç faktöre dördüncü bir faktör daha ilave edilirse bu faktörün açıklanamayan varyansın yüzde kaçını temsil ettiğini görmek için ,92/(,92+ ,25+ ,15+,10) işlemi yapılır ve ,647 değeri elde edilir. Bu değerin ,70 veya ,80 gibi oldukça büyük bir değer olması halinde faktör sayısı genişletilebilir.[118] Açıklanan varyansın oranı, bir özdeğerin diğer özdeğerlerin toplamına bölünmesiyle bulunur.

Uyuşma istatistik değerinin dikkate alınması. Bilim adamı faktör sayısına karar verirken *maksimum olasılık* faktör analizi ile elde edilen istatistik analizi sonuçlarından da yararlanabilir. Bunun için önce kaç faktör çıkaracağına karar verir ve daha sonra bu faktör sayısının modeli ne ölçüde iyi açıkladığına bakar.

Her bir faktör altında en az üç değişken bulunması. Faktör çıkarmada dikkat edilecek bir diğer nokta, her bir boyutun yüksek faktör yüküne sahip en az üç değişkene sahip olmasıdır.[119] Thurstone keşfedici faktör analizlerinde bir boyut altında en az üç değişken bulunması gerektiğini belirtmiştir.[120] Guilford da aynı görüşü dile getirmiştir (aktaran, Froman).[121] Araştırmacı bu şartı karşılamak için ölçülmeye çalışılan her bir yapıyla ilgili olarak üç değişkenin en az üç katı kadar madde ile yola çıkmalıdır. Bir yapıya ait değişken sayısı başlangıçta iyi bir şekilde planlanmadığı zaman gerçek bir sorunla karşılaşılır. Bu gibi durumlarda ya iki ve daha az değişken içeren faktörler ölçekten çıkarılır veya anket formu / ölçek yeniden oluşturularak araştırma yeniden yapılır. Velicer ve Fava'ya (1998) göre bilim adamı en az üç değişkeni bulunmayan faktörleri yorumlamaya çalışmamalıdır (aktaran Wuensch, 2003).[122]

Negatif anlamlı maddelerin ayrı bir faktör olarak çıkması. Yapılan araştırmalar tutum ölçeklerindeki negatif işaretli maddelerin faktör analizi sonucunda ayrı bir faktör altında toplanabileceğini göstermiştir (Nunnally, 1970; Goldberg, 1981; Chang, 1994; Maxim, 1999, aktaran Nandakumar, 2003).[123] Bilim adamı negatif işaretli maddelerden yararlanmışsa bu maddelerin ayrı bir faktör altında toplanma ihtimalini göz önünde bulundurmalıdır.

Faktör Puanları

Faktör puanları faktör yüklerinden farklıdır. Faktör puanları her bir kişi için hesaplanan faktörlere ait *bileşik ölçüm değerleridir. Faktör yükleri* değişkenlerdeki faktör ağırlığını temsil ederken, faktör puanları vak'alardaki faktör ağırlığını gösterir. Ortalaması 0 ve standart sapması 1,0 olan standardize edilmiş rakamlardır. Her bir faktör için ayrı ayrı hesaplanır. Faktör puanları koşutluk (çoklu doğrusallık, $r > 0.9$) özelliğine sahip olmadıklarından bu değerler faktör analizinden sonra ikincil düzeyde yapılan analizlerde de kullanılır. Örneğin faktör puanları çoklu regresyon analizinde tahmin değişkeni (bağımsız değişken) olarak kullanılır. Bunun dışında tek yönlü varyans analizinde (TYVA), çok yönlü varyans analizinde (ÇYVA) bağımlı değişken olarak kullanılabilir. Faktörün etkisini görmek veya tahminin değerini belirlemek isteyen bilim adamları bu faktör puanlarını bağımlı değişkenle birlikte regresyon analizine tâbi tutarak hipotezlerini test ederler. Böylece bağımlı değişken üzerinde hangi faktörün daha etkili olduğunu görme imkanı doğar. Bunun için beta değerlerine bakılarak karar verilir.

Faktör puanlarının bir diğer kullanım alanı, TYVA analizinde bağımlı değişken olarak ele alınması ve cinsiyet, yaş, kıdem, eğitim gibi bağımsız değişken gruplarıyla olan ilişkilerinin saptanmasıdır. Bu tür araştırmalarda

bilim adamı iki hususu göz önünde bulundurmalıdır. Birincisi, faktör puanlarının gizli değişkenlere ilişkin değerler olması nedeniyle kesinlik ve belirlilik içermemesidir. Bu puanların ağırlıklandırılmış veya ağırlıklandırılmamış toplamları gerçek anlamda faktör puanları olarak nitelendirilemez. Bu tür puanlar *bileşik ölçüm puanları* veya *ölçek puanları* olarak isimlendirilir. Bu tür puanlar kullanılarak yapılacak analiz sonuçları, bileşik ölçümlerle ilgilidir, gizli değişkenlerin kendileriyle ilgili değil.[124] İkincisi, faktörlerin değişkenlerle ve faktörlerin gruplarla olan ilişkilerinde bir sonuca varmak için olguyu iki aşamalı bir analiz olarak görmek yerine olgunun tek bir yapısal eşitlik modeli (YEM) ile çözülmek istenmesidir. Bu uygulamada faktörler arasındaki ilişkiler ve faktörlerle değişkenler arasındaki ilişkiler klasik yapısal eşitlik modeli ile test edilir. Grupların faktörler açısından nasıl bir farklılık gösterdiğini belirlemek için ise, *çoklu-grup YEM* analizi kullanılır. Oysa araştırmacı önce faktör analizi yapmalı, faktör puanlarını ortaya çıkarmalı ve bu faktör puanlarına dayalı olarak regresyon analizi veya TYVA analizi yapmalıdır. YEM analizini kullanan araştırmacılar birleşik puanları veya faktör puanlarını kullanmadıklarından elde ettikleri sonuçlar sadece gizli değişkenlerle ilgilidir. Bu analiz sonucunda elde edilen bulgular değişkenlerle faktörler arasındaki ve faktörlerle gruplar arasındaki ilişkileri tahmin etmek amacıyla kullanılmaz.[125]

Faktör puanlarının bir diğer kullanım alanı "ayrık değerlerin" saptanmasına yöneliktir. Bunun için veri matrisine yazılan faktör puanları incelenerek –3 ilâ +3 değerlerinin dışında kalan değer olup olmadığına bakılır. Eğer bu değerlerin dışına çıkan bir rakamla karşılaşılmışsa söz konusu vak'ada ayrık değerler bulunduğuna karar verilir. Ayrık değerlerin faktör çözümü üzerinde herhangi bir etkisinin bulunup bulunmadığını görmek için ayrık değer bulunan vak'a veri örnekleminden geçici olarak çıkarılarak analiz yeniden yapılır. Bunun için SPSS'in *select cases* komutu çalışılarak –3 ilâ +3 arasındaki değerleri seçmesi sağlanır. Yapılan faktör analizi sonucunda paydaşlık yüzdeleri değişmiyor, faktör yüklerinin dağılımı her iki analizde de aynı kalıyorsa vak'yı testten çıkarmak için güçlü bir neden yok demektir.

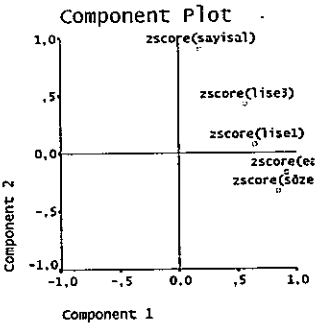
İstatistiksel analiz yazılımı SPSS'te faktör puanlarını yeni değişkenler olarak veri matrisine kaydedilebilmekte ve daha sonra bu yeni değişkenlerle diğer türde başka istatistikî analizler yapılabilmektedir (*bk.*, Tablo 8-6).

Tablo 8-6. Veri Matrisinde Faktör Puanlarının Yer Alış Biçimi

Vak'alar	Değişkenler						Faktör puanları	
	D1	*D2*	*D3*	*D4*	*D5*	*D6*	Faktör I	Faktör II
1	1,00	1,00	1,00	1,00	1,00	1,00	–,751	–,811
2	1,00	3,00	1,00	1,00	1,00	1,00	,113	-,808
3	2,00	4,00	2,00	2,00	2,00	2,00	1,558	,518
4	2,00	3,00	2,00	2,00	2,00	2,00	1,126	,517
5	1,00	3,00	1,00	1,00	1,00	1,00	,113	–,808

Faktör Yükü Grafiği

Faktörlerle değişkenleri yorumlamada ve analiz etmede kullanılan bir grafiktir. Nokta dağılım grafiğine benzer, ancak burada değişkenlere karşı kişiler değil, değişkenlere karşı faktörler vardır. Dikey ve yatay çizgiler faktörleri, noktalar ise değişkenleri gösterir. Bu grafikte değişkenlerin birbirlerine ve faktörlere ne ölçüde yakın olduklarını iki boyutlu "faktör uzayında" görmek mümkün olur (*bk.*, Şekil 8-10).



Şekil 8-10. Faktör yükü grafiği.

Geçerlilikle İlişkisi

Faktör analizi, faktöriyel alt yapıyı belirleme amacının dışında verilerin geçerliliğinin saptanması amacıyla da kullanılır. Keşfedici faktör analizi sonucunda belirlenen değişkenler ve faktörler üzerinde bu kez "teyit edici faktör analizi" yöntemi uygulanır. Bu sınama sonucunda göstergeler belirlenen faktörler üzerine yüksek lamda değerlerine sahip olmuşlarsa maddelerin "uyuşma geçerliliğine" (convergent validity) sahip olduğu söylenir. Geçerli-

lik analizinin bir diğer türü "ayrışma geçerliliği" (discriminant validity) olarak isimlendirilmiştir. Ayrışma geçerliliği için kavramsal yapıyla ilgili, fakat başka bir boyutu ölçen ikinci bir ölçeğe ait veriler de analize katılır. Keşfedici faktör analizi ve eğik döndürme yöntemi ile yapılan analiz sonucunda faktörlerle göstergeler arasındaki korelasyon (faktör yükleri) ,85'in altında kalmışsa ayrışma geçerliliğinin sağlandığı yargısına varılır.[126]

MODELİN GÜVENİLİRLİĞİ İÇİN TEYİT EDİCİ FAKTÖR ANALİZİ

Modelin[a] güvenilirliği, teyit edici (doğrulayıcı) faktör analizine dayanır. Teyit edici faktör analizi, modelin ve faktör yapısının geçerliliği konusunda kuramsal olarak çok daha sağlıklı bilgiler vermesi nedeniyle keşfedici faktör analizinden daha güçlüdür. Teyit edici faktör analizi, bir tür hipotez testidir. Araştırmacı bu yaklaşımda; (a) kuramsal bilgilere dayalı olarak belirlediği gözlem değişkenlerinin gizli faktörlerle, (b) ayrıca gizli faktörlerin de kendi aralarında birbirleriyle ilişkili olduğunu kanıtlamaya çalışır. Duruma göre faktörlerin kendi aralarındaki ilişkiler nedensellik ilişkisine dayanıyor olabilir. Teyit edici faktör analizinde ilişkilerle ilgili tüm varsayımlar önceki araştırma sonuçlarına veya kuramsal bilgilere dayalı olarak belirlenir. Bilim adamı kurama bağlı olarak geliştirdiği modelin gözlem verileri tarafından teyit edilip edilmediğini veya öngörülen modelle gözlem verilerinin ne ölçüde uyuşma gösterdiğini belirlemeye çalışır. Bu açıdan teyit edici faktör analizi kuramsal bilgilerin sınanması ve doğrulanması amacıyla uygulanır. Keşfedici faktör analizinin tersine, ölçüm değişkenleri belirli faktörlere önceden atanmışlar veya sabitlenmişlerdir. Bilim adamı, kurama veya gözleme dayalı olarak saptadığı faktörler arasında ilişki bulunduğunu veya bazı faktörlerin ilişkisiz olduğunu öngörebilir.

Teyit edici faktör analizinde, faktörlerle değişkenler arasındaki ilişkileri analiz etmek için bir faktör altında toplanacak değişken/test sayısı hakkında herhangi bir rakam belirlenmemiştir. Bir faktöre işaret eden tek bir değişken/gösterge/test dahi modelin daha güvenilir olmasını sağlayabilir. Bununla birlikte araştırmacı, diğer şartların eşit olması koşuluyla az değişkenli modellerin daha iyi uyuşma değerleri verdiği, *istatistiksel aldatıcı etkiyi* (statistical artifact) gözden uzak tutmamalıdır.[127] Az değişkenli modeller daha iyi uyuşma değeri vermekle birlikte gerçeği tam olarak temsil etmez. Bunun için çoğunlukla bir faktör altında en az üç madde bulunmasına dikkat edilir.

---

[a] Gözlem değişkenleri ve gizli değişkenler arasındaki ilişkiler ölçüm modelini oluşturur.

Teyit edici faktör analizinde gizli ve görünür değişkenler arasındaki ilişkiler *rota* adı verilen oklu çizgilerle gösterilir. Gizli değişkenler belirli bir rota üzerinde gözlem değişkenlerini etkiler. Her bir rota aynı zamanda gizli değişkenin görünür değişkende temsil edilme ağırlığını veya yükünü gösterir. Bu yük, X değişkenin lamda (lambda) katsayısı olarak ifade edilir ve λx simgesiyle gösterilir. Keşfedici faktör analizinde bu ağırlıkların anlamı "faktör yükü" olarak tanımlanmıştı. Keşfedici ve teyit edici faktör analizinin her ikisinde de lamda, görünür değişkende gizli değişkenin temsil edilme derecesini ifade eder. Gizli faktörde bir birimlik bir değişikliğin görünür değişkende ne kadar bir değişkenliğe yol açacağı hakkında fikir verir. Gözlem değişkeni gizli değişkeni temsil etme derecesinde güvenilir olarak kabul edilir. Lamda değerinin büyük olması ölçüsünde X değişkeni ile gizli faktör arasında güçlü bir ilişki var demektir.

Teyit edici faktör analizini yapmak için iki yaklaşım vardır: geleneksel yaklaşım ve yapısal eşitlik modeli yaklaşımı.[128] Aşağıdaki bölümde bu modeller hakkında bilgiler verilmiştir.

Geleneksel Yaklaşım

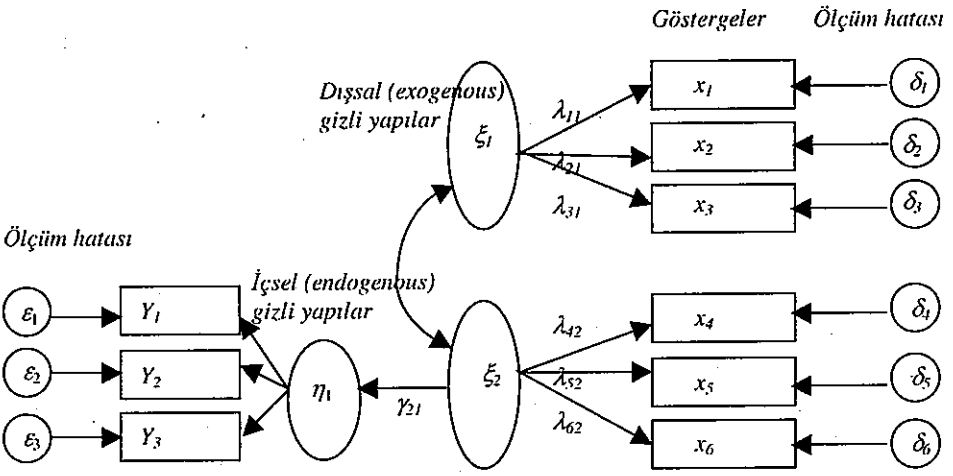
Geleneksel yaklaşımda teyit edici faktör analizini yapmak için, genel amaçlı SPSS, SYSTAT veya SAS isimli istatistiksel analiz programları kullanılabilir. Garson'a göre (2003), *yapısal eşitlik modeli* ve *teyit edici faktör analizinin* her ikisi de faktör çıkarmak için temel bileşenler analizi yerine, temel eksenler analizi (principle axis factoring) yöntemini kullanır.[129] Bu yöntem, araştırmacıya değişkenlere ait faktör yüklerinin öngörülen modelle ilişkili olup olmadığı konusunda fikir verir. Teyit edici faktör analizini kullanmak için tercih edilecek ikinci bir yöntem *maksimum olasılık* (maximum likelihood) yaklaşımıdır. Bu teknik de özel yazılımlara ulaşamayan araştırmacılara basit bir şekilde teyit edici faktör analizi sonuçlarını verir. Ancak sözü edilen geleneksel yaklaşımlar, yapısal eşitlik modeliyle birlikte kullanılırsa araştırmacı incelenen resmi daha gerçekçi bir şekilde değerlendirme imkanı elde eder.[130] Günümüzde teyit edici faktör analizleri daha çok bu amaçla yazılmış özel programlar kullanılarak yapılır. Bunlardan biri SPSS'in, ileri uygulama modüllerinden biri olan AMOS isimli programdır. Bunun dışında teyit edici faktör analizini yapmaya yönelik Lisrel, EQS, Mplus gibi yazılım seçenekleri vardır. Temel modüllerin teyit edici faktör analizi için kullanılması artık terkedilmiştir. Bununla birlikte ileri modülleri kullanma imkanı olmayan araştırmacılar SPSS'te principle axis factoring veya Maximum likelihood hesaplama yöntemini seçerek başlangıç niteliğinde bazı bilgileri edinebilirler. İstatistik yazılımı SAS'ta ise teyit edici faktör analizi ayrı bir hesaplama seçeneği olarak ana mönüler içinde yer almıştır. Teyit edici faktör analizi imkanı sağlayan diğer yazılımları kovaryans temel-

li ve bileşen temelli olarak iki grupta sınıflandırabiliriz. Kovaryans temelli olanlar şunlardır: LISREL, EQS, Mplus, SEPATH, RAMONA, CALIS, MX. Bileşen temelli olanlar ise; PLS-PC, PLS-Graph isimli yazılımlardır.

Yapısal Eşitlik Modeli

Yapısal eşitlik modeli, esas olarak gizli değişkenlerin (faktörlerin) kendi aralarındaki veya gizli değişkenlerle gözlem değişkenleri arasındaki *nedensel ilişkileri* incelemek için geliştirilmiştir, ancak nedensel ilişkilerin yanında kuramda öngörülen ölçüm modelinin geçerliliğini veya doğruluğunu sınamak için de kullanılabilir. Yapısal eşitlik modeli standart istatistik yazılımlarında bulunmaz. Bunun için AMOS, LISREL, EQS gibi özel yazılımları kullanmak gerekir. Yapısal eşitlik modeli, rota diyagramı adı verilen grafiğin çizilmesiyle başlar. Teyit edici faktör analizi için diyagramdaki *gizli faktörler* arasındaki ilişkileri gösteren *oklu dik çizgiler* kaldırılarak yerlerine *oklu eğik çizgiler* kullanılır (*bk.*, Şekil 8-11). İki oklu eğik çizgiler faktör çiftleri arasındaki kovaryansı/korelasyonu temsil eder. Tek oklu dik çizgiler ise, göstergelerle gizli değişkenler arasındaki ve göstergelerle hata terimleri arasındaki regresyon ilişkilerini göstermek için kullanılır.[131] Dış gizli yapılarla göstergeler arasındaki ilişkiler lamda, dış gizli yapılarla iç gizli yapılar arasındaki ilişkiler gamma simgesiyle gösterilir. İki iç gizli yapı arasındaki ilişkileri göstermek için ise beta simgesi (β) kullanılır. Dış gizli yapılar, bağımsız gizli değişkenlerdir. Bu değişkenler diğer gizli yapıları etkileyebilirler, fakat diğer hiçbir gizli yapıdan etkilenmezler. İç gizli yapılar ise, bağımlı gizli değişkenlerdir. Hem diğer gizli yapılardan etkilenirler ve hem de kendileri de diğer gizli iç yapıları etkileyebilirler.

Bilim adamı, ölçüm değişkenlerinin arka planında yatan gizli faktörlerin sayısı ve gizli faktörlerle gözlem değişkenleri arasındaki ilişkiler hakkında gerekli bilgileri literatürden ve önceki araştırma bulgularından elde eder. Bir diğer yaklaşım keşfedici faktör analizi kapsamında "ortak faktör" yöntemiyle belirlediği faktörleri temel almasıdır. Bilim adamı, daha sonra yapısal eşitlik modelini kullanarak bu faktörlerin teyit edilip edilmediğini araştırır.



Şekil 8-11. Yapısal eşitlik modelinde rota grafiği.

İşlem prosedürü. Teyit edici faktör analizine korelasyon veya kovaryans matrisi ile başlanır. Araştırmacı kuramsal verilere bağlı olarak veya elindeki verilere dayanarak ilişkileri açıklayan bir model öngörür. Bu modelde faktörler arasındaki veya faktörlerle değişkenler arasındaki ilişkiler, çiftli karşılaştırma esasına göre belirlenir. Amaç, ilişkinin güçlü olup olmadığını görmektir. Bilim adamı *faktör katsayıları, faktör korelasyon katsayıları* gibi belirli parametreleri, "sabit tutarak" veya "serbest bırakarak" değişik araştırma modelleri oluşturur. Sabit tutma ve serbest bırakma işlemi araştırmacının teorik beklentileri çerçevesinde yapılır.

Gillaspy (1996, aktaran Stapleton) bir parametrenin sabit tutulmasını, araştırmacının kendi beklentisine göre bir parametreye sabit bir değer vermesi şeklinde tanımlamıştır. Böylece araştırmacı bu parametrenin analiz edilmesine izin vermemiş olmaktadır. Parametrenin serbest bırakılması ise analize alınması ve değerinin tahmin edilmesi anlamına gelir. Böylece birden fazla hipotez ve model mevcut veri yapısı çerçevesinde birbirleriyle rekabet eden bir çerçevede analiz edilmiş olur.[132]

Bilgisayar yazılımları kullanılarak birbiriyle rekabet eden modeller tek tek sınanır ve veri yapısını en iyi açıklayan model bulunmaya çalışılır. Analiz sonucunda rekabet eden modellerin veri yapısına ne ölçüde uygun olduğunu gösteren bir dizi istatistik değer elde edilir. Bu istatistik değerlerin tü-

müne *uyuşma istatistikleri* adı verilir. Uyuşma istatistikleri bütün parametreleri eş zamanlı olarak test eder. Araştırmacı uyuşma istatistiklerini inceleyerek önceden belirlenen modellerden hangisinin gözlem değişkenleriyle gizli değişkenler arasındaki ilişkileri en iyi bir şekilde açıkladığına veya hangi modelin gözlem verileriyle *en iyi uyuştuğuna* karar verir. Geliştirilen model eğer verilerle uyuşmuyorsa reddedilir. Model, istatistiksel olarak reddedilmemişse, nedensel yapıları açıklayan ilişkiler geçici bir şekilde temsil edilmiş olur.[133]

Analiz. Yapısal eşitlik modeli analizleri *uyuşma istatistiklerine* dayanır. Uyuşma istatistikleri ise, öngörülen modelin gözlem verileriyle ne ölçüde örtüştüğünü ortaya çıkarmaya yöneliktir. Uyuşmanın başarı derecesini göstermek üzere İngilizce literatürde "iyi uyuşma derecesi" veya "uyuşma mükemmelliği" anlamında *goodness-of-fit* sözcükleri kullanılmıştır. Bu kitapta *goodness-of-fit test* ifadesi kısaca *uyuşma testi* kavramıyla karşılanmıştır. Uyuşma testlerinin en önemlileri şunlardır: ki-kare / serbestlik derecesi oranı, Bentler karşılaştırmalı uyuşma indeksi (KUİ), yakınlık oranı ve uyuşma indeksi (Uİ).

Ki-kare / serbestlik derecesi oranı. Ki-kare mutlak uyuşma indeks değeri olarak sınıflandırılmıştır. Ki-kare ile, "geliştirilen modelin, gözlem değişkenlerine ait kovaryans yapısında ortaya çıkan kalıp/model ile tutarlı olduğu" hipotezi test edilir. Hesaplanan ki-kare istatistik değeri küçük olduğu sürece *uyuşmanın iyi olduğuna* karar verilir. Ki-kare istatistik değeri örneklem büyüklüğüne karşı duyarlı olduğundan, pek çok durumda istatistiğin anlamlı olmasının zayıf uyuşmadan mı yoksa örneklem hacminin yetersiz olmasından mı kaynaklandığı konusunda belirsizlikler ortaya çıkar. Bu tür belirsizlikler nedeniyle modelin uyuşma durumunu belirlemeye yönelik diğer istatistiksel teknikler geliştirilmiştir. Bu teknik, örneklem büyüklüğü 100-200 arasında olduğu zaman bilgi verici olarak değerlendirilmiştir. Ki-kare testi için sıfır hipotezi aşağıdaki gibi belirlenir:

H_0: Teorik ve gözlemlenen kovaryans matris yapıları arasında istatistiksel olarak anlamlı bir farklılık yoktur.

H_1: Teorik ve gözlemlenen kovaryans matris yapıları arasında istatistiksel olarak anlamlı bir farklılık vardır.

Ki-kare istatistiği, "uyuşma indeks değeri" bulunmayan bir yaklaşım olarak değerlendirilmiştir. Hipotez testinin sonucunda *anlamlı bir farklılık bulunduğuna* karar verilmişse modelin uyuşmadığı düşünülür ve reddedilir.

Nispî uyuşma indeksi ve düzeltilmiş uyuşma indeksi. Uyuşma indeksı – Uİ[a], (relative fit indices) model ile açıklanabilen varyans ve kovaryansın nispî miktarıyla ilgili bir ölçüdür. Bu indeks, kaba bir şekilde çoklu regresyon analizindeki R^2'ye benzetilebilir. Hesaplanan Uİ değeri ,90'ın üzerinde olduğu ve 1,00'e yaklaştığı ölçüde modelin verilerle iyi uyuştuğuna karar verilir.

Düzeltilmiş uyuşma indeksi. Düzeltilmiş uyuşma indeksi – DUİ (adjusted fit indices) daha fazla parametreyi serbest bırakarak, daha az kısıtlanmış bir modelde serbestlik derecesini gösteren rakamda yapılan düzeltmeye dayanır. DUİ ve Uİ'nin her ikisi de ki-kare istatistiğine göre örneklem büyüklüğüne karşı daha az duyarlıdır.

Cimrilik oranı (parsimony ratio). Teori oluşturmada veya toplanan verilerin yorumlanmasında karmaşık açıklamalar yerine en basit varsayımların kabul edilmesine bilimde cimrilik (hasislik) adı verilmiştir. Bilimin amacı olgular arasındaki gerçek ilişkileri saptamaktır. Bunu yaparken minimum sayıda varsayıma dayanan basit modellerden[b] hareket etmek çok sayıda varsayıma dayanan karmaşık modellerden daha sağlıklı sonuçlar verir. Literatürde bu olguya "cimrilik ilkesi" adı verilmiştir. Yapısal eşitlik modelinde diğer şartların eşit olması koşuluyla faktörler ne kadar az sayıda göstergeye sahip ise veriler modelle daha iyi uyuşum gösterir. Faktörlerin kapsadığı madde/gösterge sayısı arttıkça veri-model uyuşumu azalır. Az sayıda madde içeren modellerin doğru çıkma olasılığı daha yüksektir ve bu çözümler daha fazla genellenebilir. "Cimrilik oranı" yapısal eşitlik modellerinde kullanılan bir terimdir. Cimrilik oranında öngörülen modele ait tahmin edilen parametrelerin sayısı dikkate alınır. Modeli tanımlamak için kullanılan parametre sayısı azaldıkça model daha basit, anlaşılır ve kavranabilir hale gelir. Parametre sayısının çokluğu ise, modeli karmaşıklaştırır ve kavranabilir olmaktan uzaklaştırır. Cimrilik oranı, öngörülen modelin serbestlik derecesinin, gözlem verilerine dayalı olarak test edilen sıfır hipotezindeki serbestlik derecesine olan oranıdır. Başka bir ifadeyle, gözlemler sonucunda elde edilen gerçeklere ait sayının öngörülen modeldeki gerçeklere ait sayıya olan oranı

[a] Goodness of fit index (GFI).

[b] İstatistikçiler birbiriyle ilgili hipotezler (denenceler) ailesi için "model" terimini kullanırlar.

şeklinde kavramlaştırılabilir. Cimrilik oranı, cimrilik indeks değerine bakılarak yorumlanır ve bu değerin iyi bir uyuşum gösterdiği yargısına varılabilmesi için ,90'dan büyük olması gerekir.[134]

■ Cimrilik oranı formülü.

$$CO = sd\ (model) / sd\ (maksimum, olası\ sd). \qquad (8\text{-}2)$$

Cimrilik oranının uyuşma istatistiği ile çarpılması sonucunda bir taraftan değişkenler arasındaki kovaryansı açıklayan modelin ne ölçüde etkili olduğu saptanırken diğer taraftan önerilen modelin basitlik/cimrilik indeks değeri belirlenmiş olur.

Sonuçlarının yorumlanması. Teyit edici faktör analizi sonuçlarını yorumlarken dikkat edilmesi gereken nokta, birden fazla modelin gözlem sonuçlarıyla uyuşabileceği gerçeğinin göz önünde bulundurulmasıdır (Biddle ve Marlin 1987; Thompson ve Borrello 1989, aktaran Garson, 2003).[135] Bu nedenle öngörülen modelin iyi uyuşma gösterdiğinin saptanması başka iyi uyuşma gösterecek modellerin olmadığı anlamına gelmez. Bilim adamı değişik uyuşma istatistik değerlerini kullanarak ve modelleri birbirleriyle karşılaştırarak yorum yapmalıdır. Öngörülen model uyuşma göstermemişse sabit parametreler serbest bırakılarak veya serbest bırakılanlar sabitlenerek yeni modellere göre uyuşma istatistikleri yeniden hesaplatılır. Bu hesaplamalar sonucunda hangi modelin en iyi uyuşma gösterdiği saptanmaya çalışılır. Açıklamalardan da anlaşılacağı gibi modelin güvenilirliği göreceli bir kavramdır. Bilim adamı modelin hangi koşullarda ve ne ölçüde güvenilir olduğu konusunu karşılaştırmalı analizlere dayalı olarak açıklamalıdır.

ALINTI YAPILAN KAYNAKLAR

[1] A. Weinberg, "Intelligence, [Zeka]," <http://www.chssc.salford.ac.uk/healthSci/psych2000/psych2000/intelligence.htm> (08.02.2004).

[2] T. Bates, "Major Descriptive Theories of Intelligence [Başlıca Tanımlayıcı Zeka Kuramları]," 09 Ağs 2002, <http://www2.psy.mq.edu.au/~tbates/104/104-theories.html> (08.02.2004).

[3] Universitiy of Durham, "Spearman," <http://psynts.dur.ac.uk/notes/Year2/differential/dju0sjm1/Spear.htz> (08.02.2004).

[4] J. Maltby, "Intelligence [Zeka]," <http://www.le.ac.uk/pc/jm148/IntelliLecture2.html> (08.02.2004).

[5] San Francisco State University, "The Nature of Intelligence [Zekanın Niteliği]," <http://online.sfsu.edu/~psych200/unit7/73.htm> (17.09.2003).

[6] Maltby, "Intelligence."

[7] T. Bates, "Major Descriptive Theories of Intelligence [Başlıca Tanımlayıcı Zeka Kuramları]," <http://www.google.com.tr/search?q=cache:ZTAQjrTkPrMJ:www2.psy.mq.edu.au/~tbates/104/104-theories.html+burt+%22group+factors%22+g+spearman&hl=tr&ie=UTF-8&inlang=tr> (17.09.2003).

[8] R.J. Rummel, "Understanding Factor Analysis [Faktör Analizini Anlama]," <http://www.hawaii.edu/powerkills/UFA.HTM> (26.05.2003).

[9] R.B. Darlington, "Factor Analyisis [Faktör Analizi]," <http://comp9.psych.cornell.edu/Darlington/factor.htm> (28.06.2003).

[10] Robin D. Froman, "Elements to Consider in Planning the Use of Factor Analysis [Faktör Analizini Kullanmayı Planlarken Göz Önünde Bulundurulması Gereken Ögeler]," <http://www.snrs.org/members/SOJNR_articles/iss05vol02.pdf> (19.03.2004).

[11] D. Garson, "Factor Analysis [Faktör Analizi]," <http://www2.chass.ncsu.edu/garson/pa765/factor.htm> (11.06.2003).

[12] R. Byman, "Curiosity and Exploration [Merak ve Araştırma]," <http://www.edu.helsinki.fi/oppimateriaalit/byman2003/byman2003.pdf> (28.08.2003).

[13] D. Garson, "Factor Analysis."

[14] R.R. Gliem ve J.A. Gliem, "Job Satisfaction of Civil Service and Administrative and Professional Staff [Sivil Savunma Yöneticilerinin ve Personelinin İş Tatminleri]," <http://aaaeonline.ifas.ufl.edu/NAERC/2001/Papers/gliem.pdf> (28.08.2003).

[15] M.C. George, "Factor Analysis: A Simple Introduction [Factor Analysis: Basit Bir Giriş]," <http://www.psyc.abdn.ac.uk/homedir/amilne/stats/FA.htm> (29.05.2003).

[16] M. Giesen, "Factor Analysis [Faktör Analizi]," <http://www2.msstate.edu/~jmg1/8803/FAOutline.html> (09.09.2003).

[17] Garson, "Factor Analysis."

[18] Cowan University, "Principal Component."

[19] T.Triggs ve S. Moss, "Preparation Before Exploratory Factor Analysis [Keşfedici Faktör Analizini Yapmadan Önce Hazırlık Yapma]," <http://www.med.monash.edu.au/psych/research/rda/Preparation%20before%20EFA.htm> (09.09.2003).

[20] Aynı.

[21] "Practical Approaches to Dealing with Nonnormal and Categorical Variables." <www.ioa.pdx.edu/newsom/semclass/ho_estimate2.doc> (28.06.2003).

[22] Aynı.

[23] Aynı.

[24] D.T. Cleland ve C.W. Ramm, "A Snoypsis

[25] W. Ulaga ve A. Eggert, "Relationship Value in in Business Markets Development of a Measurement Scale [İş Piyasalarında İlişki Değeri: Ölçüm Skalası Geliştirme]," <http://www.mcse.external.xerox.com/isbm/dscgi/ds.py/Get/File-228/2-2003.pdf> (28.06.2003).

[26] P. Ender, "Principal Components Analysis [Temel Bileşenler Analizi]," <http://www.gseis.ucla.edu/courses/ed231a1/notes2/pc1.html> (10.09.2003).

[27] StatSoft Inc., "Principal Components Analysis and Factor Analysis [Temel Bileşenler Analizi ve Faktör Analizi]," <http://www.statsoftinc.com/textbook/stfacan.html> (28.05.2003).

[28] C.P. Garbin, "Factor Analysis [Faktör Analizi]," <http://www-class.unl.edu/psycrs/943/7> (10.06.2003).

[29] Ulaga ve Eggert, "Relationship Value."

[30] L.R. Tucker ve R.C. MacCallum, "Exploratory Factor Analysis [Keşfedici Faktör Analizi]," <http://quantrm2.psy.ohio-state.edu/maccallum/book/ch1.pdf> (28.06.2003).

[31] C.M. Friel, "Factor Analysis [Faktör Analizi]," <http://www.shsu.edu/~icc_cmf/newCJ742/9FACTORANALYSISnew.doc> (29.05.2003).

[32] C.P. Flynn ve S.R. Kunkel, "Principal Components Factor Analysis in the Literature [Literatürde Temel Bileşenler Faktör Analizi]," <http://www.utexas.edu/courses/schwab/sw388r7/ClassMaterials/Principal_Components_Factor_Analysis_in_the_Literature.ppt> (28.08.2003).

[33] Drexel University, "Reliability and Validity of the Progress Questionnaire [Gelişme Anketinin Güvenilirlik ve Geçerliliği]," <http://dspace.library.drexel.edu/retrieve/1327/ch2.pdf> (28.08.2003).

[34] SAS Institute, "Background [Arka Plan]," <http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap26/sect2.htm> (10.06.2003)

[35] R. Shanmugam, "What is Factor Analysis [Faktör Analizi Nedir?]," <http://uweb.txstate.edu/~rs25/5339/slides/factoranalysis.ppt> (29.08.2003).

[36] D. Garson, "Factor Analysis [Faktör Analizi]," <http://www.google.com.tr/search?q=cache:t7tEN7AHojkJ:www2.chass.ncsu.edu/garson/pa765/factor.htm+%22simple+structure%22+communality+factor+analysis&hl=tr&ie=UTF-8&inlang=tr> (10.09.2003).

[37] R. MacCallum, "Commentary on Quantitative Methods in I-O Research [Endüstriyel ve Örgütsel Psikoloji Araştırmalarında Sayısal Yöntemler Üzerine Bir Yorum]," <http://siop.org/tip/backissues/TIPApril98/MacCallum.htm.> (20.06.2003).

[38] Garbin, "Factor Analysis."

[39] University of Texas at Austin, "Factor Analysis Using SAS PROC FACTOR [SAS PROC FACTOR Modülünü Kullanarak Faktör Analizi Yapma]," <http://www.utexas.edu/cc/docs/stat53.html> (10.06.2003).

[40] Garson, "Factor Analysis."

[41] Aynı.

[42] S.M. Boker, "The Factor Models [Faktör Modelleri]," <http://kiptron.psyc.virginia.edu/steve_boker/ColorFactor9507-Science-Reprint/node5.html.> (10.06.2003).

[43] L.R. Fgabrigar ve diğerleri, "Review of Article on Use of Exploratory Factor Analysis [Keşfedici Faktör Analizinin Kullanımı Üzerine Yazılan Makalenin Gözden Geçirilmesi]," <http://core.ecu.edu/psyc/wuenschk/StatHelp/EFA.htm.> (24.06.2003).

[44] University of Illinois, "Principal Component Analysis and Factor Analysis [Temel Bileşenler Analizi ve Ortak Faktör Analizi]," <http://www.uic.edu/classes/epsy/epsy546/Lecture%204%20---%20notes A. 20on%20PRINCIPAL%20COMPONENTS%20ANALYSIS%20AND%20FACTOR%20ANALYSIS1.pdf.> (11.06.2003).

[45] D'Onofrio, "Factor Analysis [Faktör Analizi]," <http://muse.widener.edu/~aad0002/710factoranalysis.html> (10.09.2003).%

[46] D'Onofrio, "Factor Analysis."

[47] University of Illinois, "Principal Component Analysis."

[48] Friel, "Factor Analysis."

[49] Gijs J.M. Dekkers, "Financial and Multidimensional Poverty in European Countries [Avrupa Ülkelerinde Finansal ve Çok Boyutlu Yoksulluk]," 2003, <http://www.ceps.lu/iriss/documents/irisswp41.pdf> (08.02.2004).

[50] D. Cziráky, S. Tišma ve A. Pisarović, "Determinants of the low SME loan... [Düşük SME Borç Uygulamanın Belirleyicileri..]," 2002, <http://www.cerge-ei.cz/pdf/gdn/RRCII_25_paper_01.pdf> (08.02.2004).

[51] K.G. Joreskog ve I. Moustaki, "Factor Analysis of Ordinal Variables: A Comparison of Three Approaches [Sıralı Değişkenlerin Faktör Analizi: Üç Yaklaşımın Karşılaştırılması]," *Multivariate Behavioral Research*, 36, (2001). 347-387.

[52] A. Ferligoj ve A. Mrvar, "Some Methodological Issues with Structural Equation Model [Yapısal Eşitlik Modeli ile İlgili Bazı Metodolojik Sorunlar]," <http://mrvar.fdv.uni-lj.si/pub/mz/mz16/zabkar.pdf> (08.02.2004).

[53] Chong-ho Yu, "Parametric Tests [Parametrik Testler]," 2003, <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/parametric_test.html.> (11.06.2003).

[54] Aynı.

[55] R. S. Edari, "Factor Analysis [Faktör Analizi]," 2002, <http://www.uwm.edu/~edari/methstat/factor.htm> (11.06.2003).

[56] Aynı.

[57] G. Tall, "Factor Analysis [Faktör Analizi]," <http://www.edu.bham.ac.uk/edrt06/factor_analysis.htm> (26.06.2003).

[58] W.S. Sarle, "One Question About Factor Analysis [Faktör Analizi Hakkında Bir Soru]," <http://www.math.yorku.ca/Who/Faculty/Monette/Ed-stat/0158.html> (11.06.2003).

[59] Cowan University, "Principal Component Analsis and Factor Analysis [Temel Bileşenler Analizi ve Faktör Analizi]," <http://www-soem.ecu.edu.au/units/mat5101/BN9.doc> (29.05.2003).

[60] Chong-ho Yu, "Parametric Tests."

[61] Chong-ho Yu, "Parametric Tests."

[62] D. Garson, "Factor Analysis [Faktör Analizi]," <http://www.cfinst.org/suny/readings/factor_analysis_ncstate.pdf> (01.09.2003).

[63] A. Field, "Factor Analysis Using SPSS [SPSS Yazılımını Kullanarak Faktör Analizi Yapmak]," <http://www.cogs.susx.ac.uk/users/andyf/teaching/rm2/factor.pdf> (25.06.2003).

[64] M. Giesen, "What is factor analysis [Faktör Analizi Nedir?]," <http://www2.msstate.edu/~jmg1/8803/FAOutline.html> (18.06.2004).

[65] D.K. Jackson ve T.P. Holland, "Measuring the Effectiveness of Nonprofit Boards [Kâr Amaçsız Örgütlerin Yönetim Kurullarının Etkililiğini Değerlendirme]," 2003, <http://www.arches.uga.edu/~dougjack/Text/Measuring%20the%20Effectiveness%20of%20Nonprofit%20Boards.PDF> (08.02.2004).

[66] K.L. Wuensch, "Factor Analysis [Faktör Analizi]," <http://core.ecu.edu/psyc/wuenschk/MV/FA/FA.doc> (24.06.2003).

[67] V. Gonzales, "Factor Analaysis [Faktör Analizi]," <http://web.uvic.ca/psyc/coursematerial/psyc401.s01/401/Lectures/11_Factor%20Analysis.PDF> (05.09.2003).

[68] Garson, "Factor Analysis."

[69] Napier University Business Schoool, "Advanced Statistics [İleri İstatistik]," <http://www.nubs.napier.ac.uk/nubs/modules/NB71003/Unit10.doc> (17.06.2003).

[70] A.J. Schwab, "Data Analysis and Computers [Veri Analizi ve Bilgisayarlar]," <http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems_Summer2003/PrincipalComponentAnalysis_CompleteProblems_Summer2003.ppt> (28.08.2003).

[71] Gadson, "Factor Analysis."

[72] Field, "Factor Analysis."

[73] J.T. Pohlman, "Factor Analsis Glossary [Faktör Analizi Sözlüğü]," <http://www.siu.edu/~epse1/pohlmann/factglos/.> (2.06.2003).

[74] J.T. Pohlman, "Factor Analsis Glossary [Faktör Analizi Sözlüğü]," <http://www.siu.edu/~epse1/pohlmann/factglos/.> (2.06.2003).

[75] R.B. Hill, "New Look at Selected Employability Skills: A Factor Analysis of the Occupational Work Ethic [İstihdam Edilebilirlik Yeteneklerine Yeni Bir Bakış: Mesleki İş Ahlakı Ölçeğinde Faktör Analizi]," <http://www.coe.uga.edu/~rhill/workethic/jverart.htm> (25.06.2003).

[76] Garson, "Factor Analysis."

[77] Aynı.

[78] C. Kim-Yin, "Exploratory Factor Analysis [Keşfedici Faktör Analizi]," <http://www.ntu.edu.sg/home/akychan/slides/Session5.ppt> (10.07.2004).

[79] Central Queensland University, "Solution [Çözüm]," <http://www.ahs.cqu.edu.au/psysoc/units/53500/pdf/r_sol/week4s.pdf.> (25.06.2003).

[80] The University of Texas at Austin, "Negative Factor Loadings [Negatif Faktör Yükleri]," <http://www.utexas.edu/cc/faqs/stat/general/gen10.html> (31.06.2003).

[81] H.B. Lammers, "Some Comments and Captured Screenshots on Using SPSS to Create an Overall Satisfaction Index [İş Tatmini İndeksi Yaratmak İçin SPSS'in Kullanılmasıyla İlgili Bazı Mülahazalar ve Ekran Örnekleri],"

<http://marketing.csun.edu/lammers/SPSSscreenshots/Proj1/proj1demomain.html.> (31.05.2003).

[82] Faculty of Law and Social Sciences, "More Complex Relationship [Daha Karmaşık İlişkiler]," <http://www.lss.uce.ac.uk/postGrad/rmquant/u3wk12.htm.> (01.09.2003).

[83] A. Yu, "Gabriel Biplot for Principle Component Analysis / Factor Analysis [Temel Bileşenler Analizi ve Faktör Analizinde Gabriel İki Kutupluluk İlkesi]," <http://seamonkey.ed.asu.edu/~alex/computer/sas/biplot.html> (31.05.2003).

[84] H. Kim, "Teamwork and Trait Extremes Based on the Big Five Model of Personality [Büyük Beş Kişilik Modeline Göre Takım Çalışması ve Uç Özellikler]," <http://mars.wnec.edu/eam/AnnualMeetings/Springfield1998/Papers/NancyFrank.html.> (31.05.2003).

[85] Indiana University, "Factor Loadings [Faktör Yükleri]," <http://www.indiana.edu/~nsse/html/mbp/confratable2.pdf.> (01.09.2003).

[86] M.R. Torabi ve K. Ding, "Selected Critical Measurement and Statistical Issues in Health Education Evaluation and Research [Sağlık Eğitiminin Değerlendirilmesi ve Araştırılmasında Kritik Ölçüm ve İstatistiksel Analiz Sorunları]," <http://www.kittle.siu.edu/iejhe/paid/1998/number1/TORABI.HTM.> (31.05.2003).

[87] C.V. King, "Factor Analysis and Negatively Worded Items [Faktör Analizi ve Negatif İfadeli Maddeler]," <http://www.populus.com/tech_papers/fa&_neg_worded.pdf.> (31.05.2003).

[88] J. Yamaguchi, "Positive vs. Negative Wording [Pozitif ve Negatif İfadeli Maddeler]," Rasch Measurement Transactions 11:2 567, 1997, <http://www.rasch.org/rmt/rmt112h.htm> (31.05.2003).

[89] King, "Factor Analysis."

[90] Aynı.

[91] Aynı.

[92] S.M. Smith, "Factor Analysis [Faktör Analizi]," <http://marketing.byu.edu/htmlpages/books/pcmds/FACTOR.html.> (19.06.2003).

[93] R.J. Rummel, "Understanding Factor Analysis [Faktör Analizini Anlama]," <http://www.hawaii.edu/powerkills/UFA.HTM> (31.06.2003).

[94] Aynı.

[95] S.G. Sapp, "Review of Exploratory Factor Analysis [Keşfedici Faktör Analizinin Gözden Geçirilmesi]," <http://www.soc.iastate.edu/soc415a/soc512.efa.html> (26.05.2003).

[96] Aynı.

[97] M. Gardner, "Multivariate Statistics [Çok Değişkenli İstatisitik]," <http://www.ed.utah.edu/Psych/CourseMaterials/7570/MS25/sld010.htm> (19.06.2003).

[98] P. Kinnear, "Factor Analysis : A Simple Introduction [Faktör Analizi: Basit Bir Giriş]," <http://www.psyc.abdn.ac.uk/homedir/amilne/stats/FA.htm> (17.06.2003).

[99] D.M. Stanek, "Factor Analysis [Faktör Analizi]," <http://www.its.ucdavis.edu/telecom/r11/factan.html.> (24.06.2003)

[100] "Factor Analysis [Faktör Analizi]," <http://www.ahs.cqu.edu.au/psysoc/units/53500/pdf/r_lect/week4.pdf.> (25.06.2003).

[101] Kinnear, "Factor Analysis."

[102] University of Illinois, "Principal Components Analysis And Factor Analysis [Temel Bileşenler Analizi ve Faktör Analizi],"
<http://www.google.com.tr/search?q=cache:gUNHK1w5qU4J:www.uic.edu/classes/epsy/epsy546/Lecture%25204%2520---%2520notes%2520on%2520PRINCIPAL%2520COMPONENTS%2520ANALYSIS%2520AND%2520FACTOR%2520ANALYSIS1.pdf+factor+analysis+direct+oblimin+delta&hl=tr&ie=UTF-8> (19.06.2003).

[103] D. Garson, "Factor Analysis [Faktör Analizi],"
<http://www2.chass.ncsu.edu/garson/pa765/factor.htm.> (19.06.2003).

[104] Mathworks, "Procrustes,"
<http://www.mathworks.com/access/helpdesk/help/toolbox/stats/procrustes.shtml> (24.06.2003).

[105] SAS Institute, "PROC FACTOR Statement [PROC FACTOR İfadesi],"
<http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap26/sect6.htm.> (24.06.2003).

[106] K.M. Rennie, "Exploratory and Confirmatory Rotation Strategies in Exploratory Factor Analysis [Keşfedici Faktör Analizinde Keşfedici ve Teyit Edici Döndürme Stratejileri]," <http://ericae.net/ft/tamu/Rota.htm> (20.06.2003).

[107] J.Neill, "Factor Analysis [Faktör Analizi]," 2004,
<http://www.wilderdom.com/301/Lecture7.ppt> (10.07.2004).

[108] Garson, "Factor Analysis."

[109] J. Reynaldo A. Santos ve Max D. Clegg, "Factor Analysis Adds New Dimension to Extension Surveys [Faktör Analizi Geniş Kapsamlı Alan Araştırmalarına Yeni Boyutlar Katıyor]," <http://www.joe.org/joe/1999october/rb6.html> (10.07.2004).

[110] J.D. Brown, "What is an Eigenvalue? [Özdeğer Nedir?],"
<http://www.jalt.org/test/bro_10.htm> (18.06.2004).

[111] D. Garson, "Factor Analysis [Faktör Analizi],"
<http://www2.chass.ncsu.edu/garson/pa765/factor.htm> (18.06.2004).

[112] M.T. Brannic, "Factor Analysis [Faktör Analizi],"
<http://luna.cas.usf.edu/~mbrannic/files/pmet/factor1.htm> (17.06.2003).

[113] D. Suhr, "Reliability, Exploratory and Confirmatory Factor Analysis [Güvenilirlik, Keşfedici ve Teyit Edici Faktör Analizi],"
<http://www.wuss.org/Conference/papers/DA05.pdf> (27.08.2003).

[114] University of Natal, "A Gentle Introduction to Factor Analysis [Faktör Analizine Giriş]," < http://www.psychology.unp.ac.za/2002/> (20.06.2003).

[115] M.T. Brannick, "Factor Analysis [Faktör Analizi],"
<http://luna.cas.usf.edu/~mbrannic/files/pmet/factor1.htm> (26.05.2003).

[116] Wuensch, "Factor Analysis."

[117] R.B. Darlington, "Factor Analysis [Faktör Analizi],"
<http://comp9.psych.cornell.edu/Darlington/factor.htm> (19.06.2003).

[118] Aynı.

[119] Jose Reynaldo ve A. Santos, "PROC FACTOR: A Tool for Extracting Hidden Gems from a Mountain of Variables [PROC FACTOR: Değişkenlerden Gizli Faktörleri Çıkarma]," <http://www2.sas.com/proceedings/sugi23/Stats/p240.pdf> (28.03.2004).

[120] D. Garson, "Factor Analysis [Faktör Analizi]," <http://www2.chass.ncsu.edu/garson/pa765/factor.htm> (28.03.2004).

[121] Robin D. Froman, "Elements to Consider in Planning the Use of Factor Analysis [Faktör Analizi Yöntemini Kullanmayı Planlarken göz önünde Bulundurulması Gereken Ögeler]," <http://www.snrs.org/members/SOJNR_articles/iss05vol02.pdf> (28.03.2004).

[122] K.L. Wuensch, "Factor Analysis [Faktör Analizi]," <http://core.ecu.edu/psyc/wuenschk/MV/FA/FA.doc> (28.03.2004). ; Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. Psychological Methods, 3, 231-251.

[123] R. Nandakumar, "Factor Analysis and Item-Response for Attitudinal Data [Tutumsal Veriler İçin Faktör Analizi ve Madde Yanıtları]," 2003, <http://www.udel.edu/ASA/AttitudeSurveys-HotchkissNotes.pdf> (25.06.2003).

[124] MacCallum, "Commentary on Quantitative."

[125] Aynı.

[126] Garson, "Factor Analysis."

[127] Aynı.

[128] Aynı.

[129] Aynı.

[130] Aynı.

[131] Garson, "Factor Analysis."

[132] C.D., Stapleton, "Basic Concepts and Procedures of Confirmatory Factor Analysis [Teyit Edici Faktör Analizinde Temel Kavramlar ve Prosedürler]," <http://ericae.net/ft/tamu/Cfa.HTM.> (19.06.2003).

[133] Aynı.

[134] D. Garson, "Structural Equation Modeling [Yapısal Eşitlik Modeli]," 2003, <http://www2.uta.edu/sswmindel/S6341/Class%20Lecture%20Sup/SEM/Principles%20of%20SEM.pdf> (27.08.2003).

[135] Aynı.

ÖLÇÜMÜN STANDART HATASI, ORTALAMANIN STANDART HATASI VE FARKLILIK PUANLARININ GÜVENİLİRLİĞİ

Klasik test kuramının temel alındığı çalışmalarda kullanılan *ölçümün standart hatası* (ÖSH) değeri, özellikle bireysel puanlar için önemlidir. Bu nedenle literatürde bu terimin bazen "bireysel puanların standart hatası" olarak isimlendirildiği görülür. Bununla birlikte ÖSH aynı zamanda grup puanlarının ortalaması için veya *özellikleri tanımlanmış* belirli bir örneklemden elde edilen test puanlarının güvenilirliğini saptamak için de kullanılabilir. Ölçümün standart hatası, güvenilirliğe eşit ve hatta bazen güvenilirlik katsayılarından daha önemli bir değer olarak görülmüştür. Herhangi bir ölçüm sonucunda güvenilirlik katsayılarının verilmesi yeterli değildir, onun yanında tercih edilen modele göre test maddelerinin veya kişilere / gruba / örnekleme ait ölçümün standart hata ve güven aralığı (GA) değerlerini de vermek gerekir. Ölçümler sonucunda elde edilen bireysel puanların güvenilirliği elde edilen puanların ne ölçüde hata içerdiğine bağlıdır. Hata oranı yüksekse aynı kişilerde / gruplarda yapılacak daha sonraki ölçümlerde önemli ölçüde puan farklılıkları ortaya çıkar. Ölçümün standart hatası bireyler / gruplar üzerinde yapılacak daha sonraki ölçümlerde ne kadar bir sapma olabileceğini tahmin etmeye yarar. Bireysel puanların istikrarlılığı kadar önemli olan bir diğer konu, grup puanlarının istikrarlılığı ve ayrıca grup puanlarının ana kütleye ne ölçüde genellenebileceğidir. Bu bölümde ölçümün standart hatası, ölçümün güven aralığı, grup puanlarının standart hatası ve farklılık puanlarının güvenilirliği konuları üzerinde durulmuştur.

BİREYSEL PUANLARIN STANDART HATASI

Kişilerin değişik testlerden aldıkları puanlar onların tutumlarını, yeteneklerini, bilgilerini ve becerilerini "kesin" ve "doğru" olarak belirlemez. Bu puanlar yetersiz alan örneklemesi, uygulamanın yeknesak olmaması, kişilerin cevaplandırma yetersizlikleri gibi nedenlerle belirli ölçüde hata ve şans faktörünü içerdiklerinden sadece tahmin değerleridir. Bu nedenle ke-

sinliğe yaklaşmak için bireysel puanların güvenilirliği, standart hata değeriyle birlikte ele alınarak değerlendirilir. Her bir ölçümün değişik nitelikteki tesadüfî hataların etkisinde olması nedeniyle bir kişiye değişik zamanlarda uygulanacak testlerin her birinde kişiler farklı puanlar elde ederler. Araştırmacı çoğunlukla birden fazla ölçüm yapamayacağından tek bir ölçüm sonucuna dayalı olarak bir bireyin puanının ne kadar hata içerdiğini "ölçümün standart hata" (ÖSH) değeriyle tespit eder. Standart hata değeri, bize elde edilen puanın gerçekte hangi puan aralıklarında değişebileceği konusunda fikir verir.

Öğrenciler bazen örnekleme ait ölçümün *standart sapma* değeriyle *ölçümün standart hata* değerini karıştırırlar. Standart sapma, örneklemdeki kişilerin puanlarına ait bir dağılım ölçüsüdür. Buna karşın ÖSH, tekrarlanan ölçümlerde bir kişinin/grubun alacağı puanlardaki sapma miktarını gösteren bir tahmin değeridir. Örneklem verilerinin dağılımıyla ilgili değil, sadece tek bir kişi veya grup üzerinde yapılacak birden fazla ölçüm sonuçlarıyla ilgilidir. Bir kişi / grup üzerinde sonsuz sayıda ölçüm yapılabilmiş olsaydı bu ölçüm sonuçlarının standart sapması veya varyansı ÖSH değerini verirdi. Ölçümün standart hatası, kişiye / gruba ait gerçek puanların varyansını gösterir.

Kullanım Amaçları

Ölçümün standart hatası, bireylerin / grupların puanlarının gerçek durumunu belirlemek ve değişik amaçlı karşılaştırmalar yapmak amacıyla kullanılır. Diğer kullanım amaçlarını aşağıdaki gibi sıralayabiliriz:

1. Bireyin sonraki ölçümlere de temel oluşturacak gerçek puan aralığını görmek için.
2. Farklı iki kişinin puan aralıklarını karşılaştırmak için.
3. Bir kişinin farklı testlerden aldığı puanların standart hatalarının önemli ölçüde değişip değişmediğini görmek için.
4. Kriter referanslı testlerde kesim puanını belirlemek için.
5. Bir kişinin puanını kesim puanıyla karşılaştırmak için.
6. Bir grubun testlerden aldığı puanların ortalamasının güven aralığını tespit etmek için.
7. MYK'de maddenin veya testin bilgi fonksiyonunu belirlemek için.

ÖSH, test/ölçek yerine bireysel puanların ve grup ortalaması puanının güvenilirliği hakkında bilgi verdiğinden danışmanlık yönelimlidir. Örneklem ve ana kütle değerlerinden bağımsız olarak bireylere / gruplara ait ham puanların gerçek değişme aralığını gösterir. ÖSH, *standart sapma* değeri gibi yorumlanır. Değerin küçük çıkması kişinin / grubun puanının kesine yakın olduğunu, büyük çıkması ise belirsizlikler içerdiğini ifade eder.

Güven Aralığı

Ölçümün güven aralığı, herhangi bir kişiye / gruba ait puanın gerçek değerinin hangi rakamlar arasında değişebileceğini görmek için hesaplanır. Gerçek puan, aynı test bir kişiye / gruba sonsuz sayıda uygulandığında elde edilecek puanların ortalaması olduğundan araştırmacı bu puana ulaşmaya çalışır. Fakat gerçek puanın noktasal yeri hiçbir zaman tam olarak belirlenemez, sadece tahmin edilebilir. Bunun için ölçümün standart hatası formülünden yararlanılarak *aralık tahmini* yöntemine başvurulur. Bir kişinin / grubun bir ölçek veya testten elde ettiği puanın güvenilirliği güven aralığı sınırları içinde ele alınarak değerlendirilir. Güven aralığı dar olduğu ölçüde kesinlik artarken geniş olduğu ölçüde kesinlik azalır. Güven aralığının seçimi, olasılık düzeyine göre belirlenir. Davranış bilimleri, psikoloji ve eğitim bilimleri araştırmalarında genelde %95 güven aralığı ile çalışılır.

Örneğin, bir psikometrik test uygulaması sonucunda bir kişinin genel yetenek puanı 52 ve ölçümün standart hatası 4,0 çıkmış olsun. Bu kişinin o testten alabileceği gerçek puanı %95 güven aralığında 44 ila 60 arasında değişir. Bir başka şekilde ifade etmek gerekirse 100 uygulamadan 95'inde kişinin puanları belirlenen puan aralığı arasında kalacaktır. Yüzde ile ifade edilen güven aralığı, deneme sayısı kadar zaman dilimine işaret etmek üzere de kullanılabilir. Böyle bir durumda ölçüm sonuçları zamanın %95'inde belirlenen limitler arasında çıkar.[1] Güven aralığı Eşitlik 9-1'deki formüllere göre belirlenir.

Olasılık düzeyi	*Kritik değerler*	
%68	$GA = X \pm 1{,}00\ (\ddot{O}SH)$	
%90	$GA = X \pm 1{,}64\ (\ddot{O}SH)$	(9-1)
%95	$GA = X \pm 1{,}96\ (\ddot{O}SH)$	
%99	$GA = X \pm 2{,}58\ (\ddot{O}SH)$	

X = Bireysel puan / grup / örneklem puanlarının ortalaması.
$\ddot{O}SH$ = Ölçümün standart hatası.

Bilim adamı, araştırmasında bir test bataryası kullanmışsa bataryanın içerdiği her bir testin ÖSH ve güven aralığı değerlerini ayrı ayrı vermelidir. Bataryadaki testlere ait güven aralıklarıyla ilgili olarak değişik sonuçlara ulaşmak mümkündür:[2]

1. Bataryadaki alt testlerin güven aralığına ait bant uzunlukları çoğunlukla farklı çıkar. Çünkü alt testlerin hata düzeyleri değişkenlik gösterir.
2. Bataryadaki testlerin bant uzunlukları çakışırsa bu durum test puanları arasında önemli bir farklılık olmadığı anlamına gelir.
3. Bataryadaki testlerin bant uzunlukları çakışmaksızın önemli ölçüde birbirinden farklı çıkmış ise test puanları arasında anlamlı ölçüde farklılık olduğu sonucuna varılır.

Güven aralığına ilişkin bant uzunluklarının ve konumlarının çakışması veya farklı çıkması bataryanın özelliğine bağlıdır. Bataryadaki birden fazla test aynı özelliği veya farklı özellikleri ölçmeye yönelik olarak geliştirilmiş olabilir.

Hesaplanması

Ölçümün standart hatası (*ÖSH*), ölçüm puanının *gerçek puandan* ne kadar farklı olabileceğini göstermek için kullanılır. Standart hatanın yüksek çıkması güvenilirliği / kesinliği düşürür. Ölçümün standart hatası, hata puanlarının standart varyansının karekökü dür. Bunun için öncelikle "ölçümün standart varyansı" hesaplanır. Daha sonra hesaplanan bu varyansın kare kökü alınarak ölçümün standart hatası elde edilir (*bk.*, Tablo 9-1).[3] Ölçüm değerlerinin standart hatası bilinirse, elde edilen bireysel veya grup puanlarının güven aralığını tahmin etmek mümkün olur.[4] ÖSH küçük çıktıkça bir test veya ölçüme ait puanlar daha kesin bir değer ifade eder. Gerçek hayatta beklenen değerler tam olarak tespit edilemeyeceğinden Tablo 9-1'deki beklenen değerler fiktif olarak tespit edilmiştir ve eğitim amaçlıdır. Araştırmacılar pratikte *hata puanlarının standart varyansı* yerine alfa değeri, test-yeniden test veya paralel formlar güvenilirlik katsayılarıyla çalışırlar.

Tablo 9-1. Ölçümün Standart Hatası

	X_t	X_∞	X_h
	4	4	0
	3	4	−1
	5	4	1
	1	1	0
	3	3	0
Σ	16	16	0
O	3,2	3,2	0
V	2,2	1,7	,50

Not. X_t = Gözlem puanları, X_∞ = Beklenen değerler, X_h = Hata değerleri.

Hata puanlarının standart varyansını (veya teorik güvenilirlik katsayısını) üç şekilde hesaplarız. Birinci yöntemde, hata puanları varyansını gözlem puanları varyansına bölüp 1'den çıkarırız (*bk.*, Eşitlik 9-2).

$$r_{tt} = 1 - \frac{V_e}{V_t} = 1 - \frac{,50}{2,2} = \frac{2,2 - ,50}{2,2} = \frac{1,70}{2,2} = ,77 \; . \qquad (9\text{-}2)$$

İkinci yöntemde, gerçek puan varyansını gözlem puanları varyansına böleriz (*bk.*, Eşitlik 9-3).

$$r_{tt} = \frac{V\infty}{V_t} = \frac{1,7}{2,2} = ,77 \; . \qquad (9\text{-}3)$$

Üçüncü yöntemde ise gözlem puanlarıyla gerçek puanlar arasındaki ilişkiyi belirleyen korelasyon katsayısının karesini alarak güvenilirlik indeksi değerini buluruz (*bk.*, Eşitlik 9-4).

$$r_{t\infty} = ,879 ,$$

$$r_{tt} = r^2_{t\infty} = (,879)^2 = ,77 . \qquad (9\text{-}4)$$

Her üç yöntemde de "ölçümün standart varyansı" (ÖSV) aynı çıkar. Buradan Eşitlik 9-5'deki uygulamayla ölçümün standart hatası hesaplanır.

$$OSV = V_t(1 - r_{tt}) = 2,2(1 - ,77) = ,50 ,$$

$$OSH = SS_t \sqrt{1 - r_{tt}} = \sqrt{OSV} = \sqrt{,50} = ,70 . \qquad (9\text{-}5)$$

Uygulamada ise, ölçümün standart hatası örneklem verilerine dayandırılır. Örneklem verileri üzerinde güvenilirlik katsayısı alfa veya korelasyon katsayılarıyla hesaplandıktan sonra Eşitlik 9-6'daki formül kullanılır:

$$\ddot{O}SH = SS_x * \sqrt{1 - r} . \qquad (9\text{-}6)$$

SS_x = Ölçümün veya gözlem rakamlarının standart sapması.
r = Ölçümün güvenilirlik katsayısı (alfa katsayısı, paralel formlar, test-yeniden test korelasyon katsayısı).

Ölçümün standart sapma değeri yüksek çıktığı ölçüde standart hata da artar. Diğer yandan, ÖSH ile testin güvenilirlik katsayıları ters yönlü olarak ilişkilidir. Testin güvenilirlik katsayısı yüksek çıktığı ölçüde ÖSH değeri küçük, güvenilirlik katsayısı düşük çıktığı ölçüde ise ÖSH değeri yüksek çıkar. Örneklemdeki kişi sayısı çoğaldıkça ölçümün standart hatası azalıp güven aralığının daraldığı görülür.

Ölçümün standart hatası için kesin bir kriter değer söz konusu değildir. Bu rakam ne kadar küçük olursa o kadar iyidir. Kriter referanslı testlerde ÖSH'yi minimize etmek için kesim puanının aritmetik ortalama veya medyan değeri etrafında belirlenmesi önerilmiştir.[5] Bilim adamı eğer kesim puanını önceden belirlemişse test maddeleri en küçük ÖSH değeri verecek olanlar arasından seçilir.

Klasik ve Modern Test Kuramlarında Ölçümün Standart Hatası

Güvenilirlik analizlerinde nasıl ki *klasik test kuramı* ve *madde yanıt kuramı* gibi farklı paradigmalardan yola çıkılabiliyorsa ölçümün standart hatası olgusunda da aynı kriterlerden hareket edilir. Araştırmacı ölçümün standart hatasını yeknesak bir bütünlük içinde ele almalıdır.

Klasik test kuramında ÖSH. Klasik test kuramında ÖSH, bir testin güvenilirliğinin fonksiyonudur. Diğer bir deyişle bir test güvenilir ise ÖSH değerinin bir anlamı vardır. ÖSH, testin uygulandığı grubun puanların dağılım biçimine göre değişkenlik gösterir ve esas olarak ortalama standart hatayı temsil eder. Klasik test kuramında ÖSH değeri temel alınarak belirlenmiş bir güven aralığı, zımnen tüm yetenek düzeylerinin genel bir ortalamasını yansıtır. Söz konusu güven aralığı, ortalamanın bir standart sapma üzerindeki veya ortalamanın bir standart sapma altındaki gibi belirli yetenek düzeylerine ait değildir. Çünkü testin ortalama puanlarından hareket edildiğinde belirli yetenek düzeylerinde standart hatanın ne olduğuna veya ne olabileceğine ilişkin bir bilgiye sahip değilizdir.

Klasik test kuramında, değişik bireylere ait olan farklı ham puanlar farklı güven aralığı değerlerine sahiptir. Orta derecede puanlara sahip bireylerin ÖSH değerleri düşük çıkarken uç puan alan kişilerin ÖSH değerleri yüksek çıkar. Bunun nedeni testlerde, orta yetenek düzeyindeki kişilere yönelik olarak yeteri sayıda madde varken üstün yetenekli veya düşük yetenekli kişiler için yeterince madde bulunmamasıdır. Uç yetenek düzeylerini ortaya çıkaracak maddelerin testte yeterinden fazla olması halinde tavan-taban etkisinin ortaya çıkması nedeniyle bu maddelerin sayısı genelde sınırlı tutulur. Sonuçta üstün ve zayıf kişilerin yetenek düzeyleri daha az kesinliğe sahip bir biçimde ölçülmüş olur.[6] Bununla birlikte "eğer puanların dağılımı normale yaklaşırsa, testte farklı gruplar için uygun zorlukta eşit sayıda soru varsa ve elde edilen puanlar muhtemel puan ranjını aşmazsa ölçümün standart hatası bütün puan düzeylerinde muhtemelen yeknesak çıkar."[7]

Klasik test kuramında ÖSH ile ilgili bir diğer sorun, ÖSH değerlerinin ham puan birimi cinsinden ifade ediliyor olmasıdır. Bu nedenle farklı testlerden elde edilen ÖSH değerleri ham puanlar temel alınarak birbirleriyle karşılaştırılamaz. Örneğin, bir kişinin 30 soruluk bir testteki ÖSH değeri 6 çıkmışsa bu değer yine 30 soruluk bir başka testten elde edilen ÖSH değeri 4'ten daha kötü olduğu anlamına gelmez.[8] Testler arasındaki "kesinlik" veya "doğruluk" karşılaştırmaları ÖSH değerlerine göre değil, bir ölçek üzerinde standardize edilmiş olan güvenilirlik katsayılarına göre yapılır.

Modern test kuramında ÖSH. Modern test kuramı örneklemden bağımsız olarak madde yanıtları ve bireysel yetenekler üzerinde odaklandığından ÖSH test çapında değil madde bazında değerlendirilir. Tek parametreli modellerde her bir madde için kişinin yetkinliği, dar bir ranj aralığında aynı yetkinlik düzeyindeki çok sayıdaki kişinin puan ortalamaları dikkate alınarak belirlendiğinden ÖSH, farklı yanıt modellerine sahip kişiler / gruplar arasında farklılık gösterirken benzer ana kütlelerde genelleme yapmaya imkan sağlayan bir değerdir.[9] ÖSH, kişinin yetkinlik düzeyi hakkında daha fazla bilgiye sahip olursak (madde bilgi fonksiyonunun yüksek çıkması) en küçük değerini verirken; çok az bilgiye sahip olduğumuzda ise en yüksek değerine ulaşır.

Klasik ölçüm modellerinde bireysel puanlara ait ölçümün standart hatasını hesaplamak için ampirik örneklem verileri temel alınırken madde-yanıt kuramında ve Rasch ölçüm yönteminde buna gerek yoktur. Test maddelerinin kalibre edilmesiyle birlikte her bir ham puana karşılık gelen standart hata değerleri de kendiliğinden hesaplanır. Rumm ve Winstep gibi yazılımlarda maddelerin veya maddelere ait ham puanların özellik/güçlük boyutu üzerindeki "yerleri" logit[a] değerlerle gösterilirken her bir logitin yanında sapmaları gösteren standart hata değerleri de verilir (*bk.*, Tablo 9-2). Söz konusu standart hata değerleri muhtemel en küçük "model" hataları olarak yorumlanır ve raporlanır. Uygulamada bu standart hatalar, gerçek verilerdeki modele alınmamış "gürültü"den kaynaklanan %10 kadar şişkinliği ifade eder.[10] Araştırmacı daha sonra standart hata değerlerine bakarak, istatistiksel olarak anlamlı kaç tane farklı performans seviyesi tespit edilebileceğini araştırır. Testteki tüm hata varyansı ise, maddelerin standart hata değerlerinin ortalaması alınarak verilir. Özel olarak geliştirilen yazılımlar bunun yanında, testin veya ölçeğin "tahmin edilen gerçek varyans değerini" de verir.

Tablo 9-2. Lojistik Birim Cinsinden Yer ve Standart Hata Değerleri.

Maddeler	*Yeri*	*SH*
madde 1	–,18	,18
madde 2	+2,1	,12
madde 2	+1,3	,06
madde 3	–,05	,09

[a] Logaritmik yöntemlerle hesaplanan ve genelde –3 ilâ +3 arasında değişen özel birim değerlerine verilen isim.

Madde-yanıt kuramında ölçümün standart hatası, madde özellikleri eğrisi ve madde bilgi fonksiyonuyla birlikte değerlendirilir. MYK'de "bilgi" terimi güvenilirliğe eşittir. Madde özellikleri eğrisi[a] kartezyen görünüme sahip iki boyutlu bir grafiktir. Grafiğin teta olarak adlandırılan yatay boyutunda kişinin yetenek düzeyi ve maddenin güçlük derecesi birlikte gösterilir. Düşük veya negatif logit değerleri o maddenin kolay olduğunu ve bu nedenle düşük yetenek düzeyindeki kişilerin o maddeye kolaylıkla yanıt vereceğini ifade eder. Yüksek pozitif logit değerleri maddenin daha zor olduğunu ve üst yetenek düzeyine sahip kişilerin o maddeyi yanıtlayabileceği anlamındadır. Teta boyutu, sıfır aritmetik ortalama ve 1,0 standart sapma logit değerine sahiptir. Boyut –3,0 logit değerinden başlayıp +3,0 logit değerine kadar uzanır ve ana kütledeki kişilerin %99,5'inin bu aralık içinde yer aldığı varsayılır. Dikey eksende maddeyi doğru yanıtlama olasılığına ilişkin değerler vardır.[11] Madde bilgi fonksiyonunda ise maddeye ait olarak, (a) güçlük, (b) ayırt etme ve (c) şans faktörlerini birleştiren tek bir indeks değeri oluşturulur. Daha sonra bu indeks değerine dayanan *kemer eğrisi* çizilir. MYK'de yüksek seviyede bilgi, yüksek ayırt etme gücüne ve düşük şans faktörüne bağlıdır. Madde bilgi fonksiyonları bir araya gelerek test bilgi fonksiyonunu (TBF) oluşturur. TBF, bir testin hatalardan görece arındırılmış olduğu anlamına gelir.[12] Test bilgi fonksiyonunun dik ve yüksek olduğu noktada ölçümün standart hatası en düşük değerine sahiptir.

Madde yanıt kuramında, her bir yetenek düzeyi için "beklenen standart hata değeri" önemlidir. Wright (2003) göre, MYK kapsamındaki tek parametreli Rasch modellerinde standart hata değerleri bu açıdan üç farklı şekilde raporlanır.[13] Birinci yöntemde "referans maddenin" standart hatası temel alınır. Kendi içinde tutarlı ve yeniden üretilebilir bir parametre tahmini yapabilmek için tahmin prosedürüne belirli kısıtlamalar getirilir. Bu çerçevede ölçek için öncelikle bir "orjin" belirlenir. Belirli bir maddenin güçlüğü için 0 logit değeri orjin olarak saptanır. Daha sonra diğer maddeler bu orijinle karşılaştırılarak bu orijine göre nispî yerleri belirlenir. Sıfır logit değerine sahip madde mükemmel kesinliğe sahiptir ve bu maddenin logit standart hatası 0,00'dır. Çoğunlukla bu maddenin standart hatası hiç raporlanmaz. Wright, bu yönteme göre hesaplanan SH değerlerinin sınırlı bir değere sahip olduğunu belirtmiş ve ölçek için bir orijin oluşturmanın güzel bir fikir olduğunu fakat seçilen herhangi bir maddenin realiteyi mü-

[a] Literatürde "madde özellikleri eğrisi" terimi yerine bazen "madde-yanıt fonksiyonu" ifadesinin kullanıldığı görülür. Her iki terim aynı anlamdadır.

kemmel bir şekilde yansıtmayacağı görüşünü dile getirmiştir. İkinci yöntem, "ideal" standart hata yaklaşımıdır. Herhangi bir ölçümün en yüksek kesinlik derecesi, diğer ölçümlerin her birinin bilinmesi ve bu ölçümlerde en iyi veri-model uyuşumunun sağlanmasıyla elde edilebilir.[14] Rasch modelini analiz eden BIGSTEPS gibi yazılımlarda "model" standart hata yaklaşımı benimsenmiştir. Temiz verilere sahip, çok iyi yapılandırılmış testlerde "model standart hata" değerlerinin gerçek standart hataya çok yaklaştığı, fakat gerçek değeri olduğundan biraz küçük gösterdiği belirtilmiştir. Wright tarafından önerilen üçüncü yöntem, uyuşmazlık şişkinliğine sahip "gerçek" standart hata değerlerinin raporlanmasıdır. Ölçümlerde bir kişi çok sayıda madde ile teste tâbi tutulmakta ve aynı zamanda bir test maddesi çok sayıda kişiye uygulanmaktadır. Bu süreç içinde verilerin modelle tam olarak uyuşmaması yüzünden standart hata, modelin içermesi gereken "ideal" hatanın bariz bir şekilde üstünde çıkar. O halde standart hatayı, modelin uyuşmazlığı olgusuyla birlikte değerlendirmek gerekir. Uygulamada "gerçek" standart hata kesin olmayan ölçümün üst sınırı temel alınarak belirlenir. Fiilî standart hata ise, "model" ile "gerçek" standart hata değerlerinin arasında bir yerdedir.[15]

Ölçümün Standart Hatasını Etkileyen Faktörler

Klasik test kuramındaki ölçümün standart hatası, formül gereği güvenilirlik katsayısı ile ölçümün standart sapma değerinden birinci derecede etkilenir. Bunun dışında ÖSH etkileyen faktörler Nitko (1996) tarafından aşağıdaki gibi belirlenmiştir (aktaran Doolittle, 2001).[16]

1. Uzun değerlendirme prosedürleri kısa olan prosedürlere göre daha güvenilirdir. Testteki madde sayısının artması, test uygulaması sırasında daha fazla performans gösterme güvenilirliği arttırır ve ölçümün standart hatasını düşürür.
2. Homojen grupların değerlendirilmesi heterojen grupların değerlendirmesine göre daha düşük güvenilirlik sonuçları verir.
3. Test-yeniden test veya alternatif form uygulamaları arasında geçen zamanın uzunluğu test alan kişilerin değişme olasılığını artırır. Zaman uzadıkça testin güvenilirliği azalır ve dolayısıyla ölçümün standart hatası yüksek çıkar.
4. Testin objektif olarak doldurulması hata oranını azaltarak güvenilirliği yükseltir.

5. Farklı güvenilirlik yöntemleri ile elde edilecek değişik katsayılardan farklı *standart hata* değerleri elde edilir.
6. Kriter referanslı testlerin güvenilirlik katsayıları norm referanslı testlerin güvenilirlik katsayılarından genelde daha düşüktür. Kriter referanslı testlerde puanlar dar bir ranj aralığından elde edilir.

ÖSH değerini etkileyen faktörler genel olarak testin güvenilirliğiyle yakından ilgilidir. Güvenilirlik katsayısını iyileştirmeye yönelik olarak alınacak önlemler aynı zamanda standart hata değerlerini de düşürecektir.

Koşullu ve Koşulsuz Ölçümün Standart Hatası

Ölçümün standart hatası, ait olduğu puanlara veya uygulama koşullarına bağlı olarak değişir. Belirli puan dilimleri kapsamında elde edilen ölçüm hataları, "koşullu ölçümün standart hatası" (KÖSH) terimiyle adlandırılmıştır.[17] Puan dilimi sınırlamasına gidilmeden yapılan hesaplama ise, *koşulsuz ölçümün standart hatası* olarak isimlendirilir. Klasik ÖSH hesaplamasında ölçüm hatasının kişilerin değişik yetenek düzeyleri boyunca sabit olduğu varsayılır. Ancak bu ön kabul gerçekçi değildir. Testte madde sayısı arttığı ve maddeler daha fazla bilgilendirici olduğu ölçüde ölçüm hatası azalırken güvenilirlik artar. KÖSH, özel olarak seçilmiş belirli puan düzeyindeki ÖSH'dir. Bilim adamları KTK yaklaşımında ölçekten/testten alınan toplam puanlar dizisinin orta bölümünde yer alan kesite ait ÖSH değerlerini KÖSH olarak belirlemeyi tercih ederler. Çünkü, ölçeğin orta bölümünde yer alan puanlar dizinin üst ve alt ucunda yer alan puanlara göre kişilerin başarılarını daha doğru ölçer (*bk.*, Tablo 9-3).

Bir teste ait sonuçların kesinlik/doğruluk derecesi (testin kesinliği), *koşullu ölçümün standart hatası* değeriyle daha iyi saptanır. Yapılan bazı araştırmalarda *koşullu ölçümün standart hatası* değeriyle çalışıldığında testte başarılı sayılanların oranında %5'lik bir artış gözlenmiştir.

"KÖSH değerini tahmin etme işlemleri literatürde birkaç çalışmada ele alınmıştır. Kolen, Hanson ve Brennan (1992) güçlü bir, *gerçek puan* modeli önermişlerdir. Kolen, Zeng ve Hanson ise prosedüre MYK açısından yaklaşmışlardır. Lee, Brennan ve Kolen (2000) ölçek puanları için KÖSH'yi tahmin eden belirli sayıda prosedürü gözden geçirerek kapsamlı bir inceleme sunmuşlardır. Kolen ve Wang (1998) bileşik puanlarda KÖSH'yi madde-yanıt kuramına göre hesaplayan bir işlem süreci tanıtmışlardır. Lee (2001) ise, ikili ve çok dereceli maddeler için KÖSH'yi tahmin imkanı sağlayan çok değişkenli/çok terimli hata modelini önermiştir. Brennan ve Lee (1997) bileşik ölçek puanları için KÖSH'yi belirleyen bir hesaplama yöntemi geliştirmişlerdir (aktaran Ban, Lee ve Hanson)".[18]

Madde-yanıt kuramında, ölçümün *kesinlik derecesi* olarak maddelerin bilgi fonksiyonunun toplamından oluşan *test bilgi fonksiyonu* (TBF) ve *testin standart hatası* (TSH) değerleri kullanılır. MYK'da, belirli test puanları aralığı ölçümün amacına ve bilgi verme fonksiyonuna göre belirlenir. Örneğin, personel seçimi amacıyla kullanılan ve en yüksekten aşağıya doğru sıralanan puan dizisinde ölçümün kesinliği, gizli değişken boyutunun üst bölümünde yer alan puanlardır. Çünkü kullanım amacı açısından bu puanların bilgi fonksiyonu daha yüksektir. Bununla birlikte genel ana kütleye uygulanan standardize edilmiş testlerin çoğunda en yüksek ölçüm kesinliği, orta erimdeki puanlar üzerindedir. Bu testlerde kesinlik yüksek ve düşük puanlara doğru gidildikçe azalır. Orta erimdeki puanların temel alınarak hesaplandığı KÖSH değeri "kısıtlanmış" standart hatayı gösterir. Orta erimde testin uç noktalarındakine göre daha fazla madde bulunması nedeniyle KÖSH değeri daha düşük çıkar.

Belirli bir özellik düzeyinde koşullu ölçümün standart hatası, söz konusu düzeyde gizli özelliği (θ) yansıtan bilgi düzeyinin (I) ters karekökü̈ne eşittir (*bk.*, Eşitlik 9-7).[19] Madde-yanıt kuramında standart hata klasik test kuramında olduğu gibi bir istatistik değer değil, yetenek düzeyinin fonksiyonudur.

■ Madde-yanıt kuramında koşullu ölçümün standart hatası formülü.

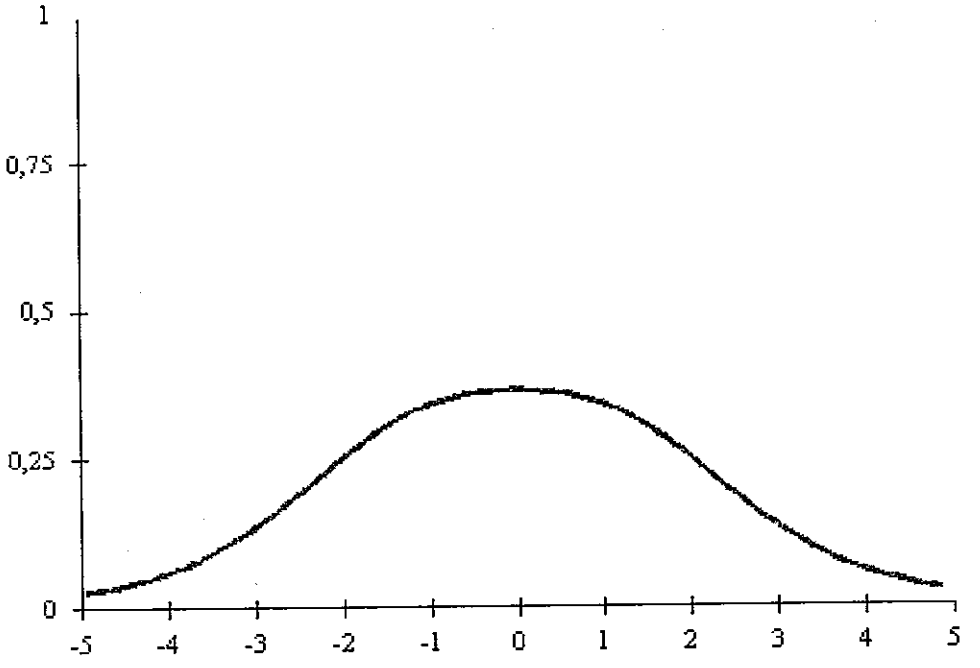
$$KÖSH(\theta) = \frac{1}{\sqrt{I(\theta)}}. \qquad (9\text{-}7)$$

Koşullu ölçümün standart hatasını belirleyebilmek için öncelikle "madde bilgi fonksiyonunu" (MBF) ve daha sonra "test bilgi fonksiyonunu" (TBF) incelemek gerekir. Madde bilgi fonksiyonu, maddenin ayırt etme (a-parametresi), güçlük derecesi (b-parametresi) ve şans faktörünü (c-parametresi) hep birlikte dikkate alarak söz konusu fonksiyonu kemer eğrisi ile açıklar. Rasch modelinde MBF Eşitlik 9-8'deki formülle hesaplanır ve elde edilen değerler daha sonra kemer eğrisiyle gösterilir.

$$I_i(\theta) = P_i(\theta)\,[1 - P_i(\theta)]. \qquad (9\text{-}8)$$

I_i = Madde bilgi fonksiyonu.
θ = Gizli özellik.
P_i = Maddenin olasılığı.

Madde bilgi fonksiyonunu gösteren *kemer eğrisi* basık, dik veya normal dağılım eğrisi şeklinde olabilir. Sivri noktasının konumu ise delta yetenek boyutu üzerinde sıfır noktasına odaklı veya bu noktanın sağında veya solunda odaklanmış bulunabilir. Şekil 9-1'de MBF eğrisi sıfır yetenek noktası üzerinde odaklanmıştır ve sivrilik açısından ise 0,25 noktasındadır. Bu açıdan tek parametreli Rasch modellerinde alabileceği en yüksek değerine sahiptir. İki ve üç parametreli modellerde ise, kemerin sivriliği maksimum değerini alır. Buna göre grafik incelendiğinde maddenin bilgi fonksiyonu orta düzeydeki yetenek düzeyine sahip kişiler için anlamlı bir dağılıma sahiptir ve aynı zamanda söz konusu maddenin bilgi verme fonksiyonu da söz konusu model için kesinlik derecesindedir. Kemer eğer daha yatık bir şekilde çıkmış olsaydı böyle bir durumda maddenin bilgi verme fonksiyonu zayıflayacak ve teste daha iyi bilgi verecek başka maddeler bulunması gerekecekti. MYK'ye göre test oluşturulurken belirlenen özellik boyutu üzerinde en yüksek kesinlik derecesini sağlayan veya en fazla informatif olan maddeler ölçeğe alınır.



Şekil 9-1. Madde bilgi fonksiyonu.

Test parametreleri açısından değerlendirilirse ölçeğe, daha yüksek *a-yırma parametresine* sahip olan maddeler alınmalıdır. Bu maddeler kişiler hakkında daha fazla bilgi verir ve bu nedenle a-parametresi yüksek ve aynı zamanda b-parametresi 0,0 olan maddeler araştırılır. Bir uygulamada b-parametresi 0,0'a yakın bir diğer iken a parametresi yeterince yüksek değilse bu maddeler ölçeğe alınmaz. Maddelerin bilgi fonksiyonu testin oluşturulma amacı çerçevesinde değerlendirilir. Test; tarama (belirli bir puanın altında ve üstünde olanları ayırt etme), geniş bir kitlede uygulama veya yetenek ölçeğinin belirli bir başarı bölgesinde (koşullu bir düzey) ölçüm yapmak amacıyla oluşturulmuşsa madde özellikleri eğrisi ve madde bilgi fonksiyonu ve test bilgi fonksiyonu buna göre yorumlanır. Tarama testlerinde, test bilgi fonksiyonu eğrisi, belirlenen *sınır puanı yetenek düzeyinde* en yüksek zirveye sahip olmalıdır. Bunun yanında arzulanan test özellikleri eğrisi ise, istenen sınır yetenek düzeyinde orta-gerçek puana sahip olmalı eğri bu yetenek düzeyinde mümkün olduğu kadar dik bir görünüme sahip bulunmalıdır. Madde güçlük parametreleri ise sınır yetenek düzeyi etrafında çerçevelenmelidir.

Tablo 9-3. Klasik Test Kuramında Koşullu Standart Hata Değerleri

Sözel yetenek		*Sayısal yetenek*	
Test puanı	*KÖSH*	*Test puanı*	*KÖSH*
250	6	200	8
245	10	180	9
243	10	170	10
240	11	165	12
230	12	140	17

Bireysel Ölçümlerin Noktasal Gerçek Puan Tahmini

Bilim adamları ölçümün standart hatası ile elde ettikleri güven aralığı tahmin değerlerini her zaman yeterince işlevsel bulmadıklarından bir kişiye ait ham puanın noktasal olarak *gerçek puan* tahmin değerini de hesaplamak isterler. Bu puanlar noktasal olarak ifade edilse de, onların birer "tahmin değerleri" olduğu gerçeğini göz ardı edemeyiz. Bireysel ölçümlerin noktasal gerçek puan değerlerini hesaplayabilmek için dizideki puanların aritmetik ortalama değerinin, testin güvenilirlik katsayısının ve kişinin ham puanının bilinmesi gerekir (*bk.*, Eşitlik 9-9).

$$TGP = \bar{X} + [\, r \,.\, (X - \bar{X})] \,. \qquad (9\text{-}9)$$

TGP = Tahminî gerçek puan.
$\bar{X}$ = Serideki değerlerin aritmetik ortalaması.
X = Kişinin ham puanı.
r = Testin güvenilirlik katsayısı.

Noktasal gerçek puan tahmin değerini bir örnekle açıklayabiliriz. Bir öğrenci sözel yetenek testinden 40 almış olsun. Grubun ortalaması 52 ve testin güvenilirlik katsayısı ise ,76'dır. Bu kişinin tahminî gerçek puanı Eşitlik 9-10'daki gibi belirlenir:[20]

$$\begin{aligned} TGP &= 52 + ,76 * (40 - 52) \,, \\ TGP &= 52 + (-9,12) \,, \\ TGP &= 42,88 \,. \end{aligned} \qquad (9\text{-}10)$$

Puan Karşılaştırmalarında ÖSH Değerlerinin Kullanılması

Daha önce ham puanlara dayalı olarak ÖSH değerleri arasında karşılaştırma yapılamayacağı belirtilmişti. Ancak puanlar standardize edildikten sonra bu tür karşılaştırmaları yapmak mümkündür. Ölçümün standart hatası; (a) bir testten iki bireyin aldığı puanların ve (b) bir bireyin farklı iki testten aldığı puanların karşılaştırılması amacıyla kullanılabilir. Bunun için her iki teste ait puanların aynı ortalama ve standart sapmaya sahip olacak şekilde standartlaştırılması gerekir.

Bir testten farklı iki kişinin aldığı puanların karşılaştırılması. Bu yöntemde, ÖSH değerlerine bağlı olarak her iki kişi için güven aralığı (GA) bantları oluşturulur. GA bantları çakışıyorsa iki kişinin puanlarının birbirinden önemli ölçüde farklı olmadığı sonucuna varılır.

Bir bireyin farklı iki testten aldığı puanların karşılaştırılması. Bu yöntemde ise, ÖSH değerlerine bağlı olarak her iki test sonucu için güven aralığı (GA) bantları oluşturulur. Yine GA bantları çakışıyorsa söz konusu kişinin A ve B testinden aldığı sonuçların birbirinden önemli ölçüde farklı olmadığı sonucuna varılır. Çakışma çok az ise veya hiç çakışma yoksa testlerin farklı özellikleri ölçtüğü sonucu çıkarılır.

Bazı araştırmacılar test sonucu karşılaştırmalarına yönelik olarak testlerin aritmetik ortalama değerleriyle birlikte ÖSH değerlerini vermenin bir

anlamı olmadığını belirtmişlerdir. Bu bilim adamlarına göre, iki test puanı arasında anlamlı bir farklılık olup olmadığını görmek istersek anlamlılık testi yapmak, bu teste ait *p* değerlerini ve güven aralığı değerlerini vermek daha doğrudur. Çünkü ÖSH değerlerinin ne ölçüde çakışmazsa anlamlı bir farklılık yarattığını söylemeye imkan sağlayacak bir kriter yoktur. ÖSH, deneysel araştırmalarda istatistiksel anlamlılığı göstermez.[21] Burada ÖSH değerinin istatistiksel anlamlılık amacıyla kullanılmasına karşı çıkılmaktadır. İki teste ait puanların birbirinden önemli ölçüde farklı olup olmadığını belirlemek için ki-kare uygunluk testi ile sonuca gitmek kuşkusuz daha doğru bir yaklaşım olur. Ancak ortalama puanıyla birlikte ÖSH değerinin birlikte verilmesinin amacı aynı örneklemde yapılacak diğer ölçümlerde ne kadar bir sapma görülebileceğini belirlemeye yöneliktir.

GRUP PUANLARININ STANDART HATASI

Standart hata, bireysel puanların yorumlanmasında olduğu kadar grup veya sınıf puanları ortalamasının gerçek puan varyansını tahmin etmek için veya ana kütle varyansını belirlemek için de kullanılır. Burada iki yön vardır. Birincisi *grubun gerçek puan varyansını* belirlemeye yönelik iken ikincisi *grubun ortalama puanının ana kütle ortalamasını ne ölçüde yansıttığıyla* ilgilidir. Birincisi "ölçümün standart hatası" olarak isimlendirilirken ikincisi "grup ortalamalarının standart hatası" olarak nitelendirilir. Ölçümler sınıf öğrencileri, okul öğrencileri, semt veya şehir öğrencileri üzerinde yapılmışsa söz konusu gruplara ait puanların ortalamalarına dayanan güven aralığı değerleri bireysel öğrencilerin güven aralığından daha dar veya küçük çıkar. Çünkü grup değerleri gerçek puan değerlerine daha yakındır.

Bilim adamı, ölçüm yaptığı gruba ait puan ortalamasının standart hatasını iki nedenle görmek isteyebilir. Aynı grupta tekrar tekrar yapılacak ölçümlerde ne kadar bir sapma ortaya çıkabileceğini belirlemek istiyorsa "ölçümün standart hata" (ÖSH) formülünden hareket eder. Tam tersine benzeri gruplardan elde edilecek ortalama puanları arasında ne kadar bir sapma olduğunu veya sonuçların ana kütleye ne ölçüde genellenebileceğini merak ediyorsa "ortalamanın standart hata" (OSH) formülünü kullanır. APA standartlarına göre yapılan herhangi bir ölçümde gruba ait "ölçümün standart hata" değerlerini araştırma raporunda vermek gerekir.

Örneğin, bir sınıfın matematik sınavı ortalaması 3,7 çıkmışsa ÖSH'nin hesaplanmasıyla bu sınıftaki gerçek puanların değişim aralığını görmek mümkün olur. Eğer sınıfın ortalama puanı tüm şubelere genellenmek isteniyorsa OSH formülüyle elde edilen değer diğer sınıflarda ne kadar bir sapma ortaya çıkabileceğini belirler. Bu uygulamada bir taraftan sınıfın puan ortalaması ve gerçek puanın güven aralığı saptanırken diğer taraftan

testin ana kütlede ne ölçüde tutarlı sonuçlar verebileceği hakkında bir fikir edinilmiş olur. ÖSH ve OSH değerleri küçük çıkmışsa daha sonra yapılacak ölçümlerde ve diğer gruplarda yapılacak ölçümlerde muhtemelen benzer ortalama değerleriyle karşılaşılacaktır. Test yayımlayan ticarî kuruluşlar geliştirdikleri ölçüm araçlarının güvenilirlik ve geçerlilik katsayıları yanında değişik norm grupları için ÖSH ve OSH değerlerini de verirler. ÖSH ve OSH değerleri, testin/alt testin puan ortalamalarının güven aralığı konusunda uygulayıcıları bilgilendirerek, onların değişik kriterlere göre gerçek puan değerleri hakkında fikir sahibi olmalarını sağlar. Ortalamanın standart hatası Eşitlik 9-11'deki formülle hesaplanır.

$$OSH = \frac{SS}{\sqrt{n}} . \qquad (9\text{-}11)$$

n = Ölçüm yapılan vak'a/öğrenci sayısı.
SS = Standart sapma.

Ortalamanın standart hatasının yüksek çıkması test sonuçlarının ana kütleden elde edilebilecek sonuçları daha az temsil ettiği anlamına gelir. Ölçümün standart hatasında olduğu gibi ortalamanın standart hatası değeri için de güven aralığı hesaplanabilir. OSH değerinin güven aralığı, belirli bir alfa düzeyinde, $x \pm Z_{\alpha/2}$ (OSH) formülü ile belirlenir. Bu formülde $Z_{\alpha/2}$ simgesel gösterimi, %95 güven aralığında (α = ,05) normal kantil değeri (,025) = 1,96 olarak; %99 güven aralığında (α = ,01) normal kantil değeri (,005) = 2,58 olarak belirlenir. Bir çalışmada ölçüm yapılan örneklem hacmi büyüdükçe OSH değeri küçülür.

GÜVENİLİR DEĞİŞİM İNDEKSİ

Güvenilir değişim indeksi – (GDİ) N.S. Jacobson ve P. Truax (1984) tarafından geliştirilmiştir. Jacobson, farklı zamanlarda yapılan ölçümler sonucunda bireyin puanlarında görülen değişmenin tesadüfî hatalar ve şans faktöründen mi yoksa kliniksel olarak anlamlı bir değişimi mi yansıttığını görmek için bir hesaplama yöntemi geliştirmiştir.[22] Bireydeki değişimler onun rahatsızlanması veya iyileşmesi nedeniyle, testi uygulama yönteminin değişmesi sebebiyle, fizyolojik ritim bozukluğu nedeniyle, kişinin diyet yapması veya ilaç alması nedeniyle, kan vermesi nedeniyle veya testin kendisindeki değişkenlik nedeniyle ortaya çıkabilir. Güvenilir değişim indeksi araştırmacıya ortaya çıkan farklılığın anlamlılığı hakkında yorum yapma imkanı sağlar (*bk.*, Eşitlik 9-12, 9-13).

■ Güvenilir değişim indeksi.

$$\textit{Güvenilir değişim indeksi} = \frac{\textit{son test değeri} - \textit{ön test değeri}}{\textit{ölçümün standart hatası}} . \quad (9\text{-}12)$$

$$\textit{Ölçümün standart hatası} = \textit{Ön test sonuçlarının standart sapması} \times \sqrt{1-r} . \quad (9\text{-}13)$$

Christensen ve Mendoza, Eşitlik 9-12 ve Eşitlik 9-13'te verilen formüllerde değişiklik önermişler ve formülün paydasında yer alan "ölçümün standart hatası" yerine iki test puanları arasındaki "farkların standart hatası değeriyle" çalışılmasını istemişlerdir (aktaran Jacobson).[23] Hesaplama sonucunda indeks değeri eğer, %95 güvenilirlik düzeyinde ±1,96'dan yüksek çıkarsa gerçek bir değişimden söz edebiliriz. Böyle bir durumda farklılık puanları güvenilirdir. Tam tersine ±1,96'dan düşük çıkmışsa ölçüm sonuçlarına güvenilmez.[24] Güvenilmezlik, ölçüm aracının veya ölçüm verilerinin güvenilir olmamasından değil, ölçüm uygulamasından veya değişimin önemli olmamasından kaynaklanmış olabilir.

FARK PUANLARININ GÜVENİLİRLİĞİ

Fark puanlarının güvenilirliği (FPG), değişik iki ölçümden elde edilen puanlar arasındaki fark değerlerinin ne ölçüde güvenilir olduğu hakkında fikir verir. Ölçümler aynı özelliği saptamaya çalışan iki farklı test, iki alt test, ön-test son-test uygulaması veya test-yeniden test uygulaması şeklinde gerçekleştirilmiş olabilir. İki test arasındaki fark puanlarına, *kazanç puanları* adı verilir. Bilim adamları yaptıkları çalışmanın niteliğine göre fark puanlarının yüksek veya düşük çıkmasını isteyebilirler.

Örneğin, araştırmacılar kimi zaman uygulanan test sonuçlarına göre değil; zeka testi-yıl sonu başarı puanı, zeka testi-yabancı dil puanı gibi iki test sonucu arasındaki fark puanlarının düşük olup olmadığına bakarak karar vermeyi düşünürler. Bazen de fark puanları, test-yeniden test uygulamasıyla gelişmeyi veya değişmeyi belirlemeye yönelik olarak saptanır. Bu uygulamada fark puanlarının yüksek çıkması istenir. İlk ölçümde bireyler eğer yüksek puanlar almışsa daha sonraki ölçümlerde gelişmeyi / değişmeyi saptamak zorlaşır. Bu nedenle fark puanları her zaman istenen sonucu vermez. Fark puanlarının kullanıldığı bir diğer alan, kişilerin algıları/düşünceleri ile eylemleri/davranışları arasındaki açığın görülmek isten-

mesidir. Fark puanlarının güvenilirliği Eşitlik 9-14'deki formülle hesaplanır:

■ Fark puanlarının güvenilirliği.

$$r_f = \frac{r_x s_x^2 + r_y s_y^2 - 2R_{xy} s_x s_y}{s_x^2 + s_y^2 - 2R_{xy} s_x s_y} . \qquad (9\text{-}14)$$

r_f = Farklılık puanlarının güvenilirliği.
r_x = Birinci testin güvenilirlik katsayısı.
r_y = İkinci testin güvenilirlik katsayısı.
r_{xy} = Birinci ve ikinci test arasındaki korelasyon.
s_x = Birinci ölçüme ait puanların standart sapması.
s_x^2 s_x = Birinci ölçüme ait puanların varyansı.
s_y = İkinci ölçüme ait puanların standart sapması.
s_y^2 s_x = İkinci ölçüme ait puanların varyansı.

Eğer her iki ölçümün standart sapması aynı ise Eşitlik 9-15'deki formül kullanılır.

$$r_f = \frac{(r_1 + r_2)/2 - r_{12}}{1 - r_{12}} . \qquad (9\text{-}15)$$

r_f = Fark puanlarının güvenilirliği.
r_1 = Birinci testin güvenilirliği.
r_2 = İkinci testin güvenilirliği.
r_{12} = Birinci testle ikinci test arasında gözlenen korelasyon katsayısı.

Fark puanların güvenilirliğini bir örnekle açıklayabiliriz. Bir sınıfta öğrencilere Matematik ve Fen Bilgisi sınavları yapılmıştır. Öğrencilerin sınavdan aldıkları puanlar arasındaki farkların düşük olması istenmektedir. Bilim adamı, farklılık puanlarının Matematik ve Fen Bilgisi sınavlarının güvenilirlik katsayısı kadar yüksek olup olmadığını görmek istemektedir. Fark puanlarının güvenilirliği yüksek çıkarsa, farklılık puanlarının anlamını daha güçlü bir şekilde savunabileceğini düşünmektedir.

Matematik testinin güvenilirliği (r_1) = ,91.
Fen Bilgisi testinin güvenilirliği (r_2) = ,90.
Matematik ve Fen Bilgisi testi puanları arasındaki korelasyon (r_{12}) = ,45.

$$r_f = \frac{(,91 + ,90)/2 - ,45}{1 - ,45} = ,70 . \qquad (9\text{-}16)$$

Eşitlik 9-16'da görüldüğü gibi fark puanlarının güvenilirliği, hesaplama yapılan iki testin güvenilirlik katsayısından daha düşük çıkmıştır. İki test puanları arasındaki korelasyon katsayısı yükseldiği oranda farklılık puanlarının güvenilirliği daha düşük çıkar. Örnekte, iki test arasındaki korelasyon katsayısı ,65 çıkmış olsaydı bu kez farklılık puanlarının güvenilirliği ,57 olarak gerçekleşecekti. Ön test – son test uygulamalarında korelasyon katsayısının düşük çıkması katılımcılardaki farklı gelişme olgusunu gösterir. Kişilerin ön-testteki rank sıralarıyla son-testteki rank sıraları arasında zaman içinde köklü değişiklikler olmuştur.

Farklılık puanlarının düşük güvenilirliğe sahip olması puanlardaki varyansı yapay olarak kısıtladığından araştırmacıyı yanlış yönlendirebilir. Bilim adamı farklılık puanlarını daha az güvenilir olması nedeniyle büyük bir dikkatle kullanmalıdır. Farklılık puanlarının güvenilirliği, ham puanlar standart z puanlarına dönüştürülerek hesaplanır. İki test sonucuna ait puanlar arasındaki korelasyon katsayısı büyüdükçe farklılık puanlarının güvenilirliği düştüğünden bu durum pratik hayatta gerçek bir sorun olarak görülmüştür.[25] Cohen ve Cohen (1983) değişimi ölçmeye yönelik olarak farklılık puanlarını kullanma uygulamasından kaçınmışlardır (aktaran Darvin, 2004).[26] Bu nedenle olsa gerek bazı araştırmacılar farklılık puanlarının güvenilirliği için kovaryans analizi veya regresyon analizi yönteminin kullanılmasını önermişlerdir.

Pilot araştırma ve esas araştırma sonuçları arasındaki anlamlı bir farklılık olup olmadığını veya iki puan dizisi arasındaki ilişkileri merak eden araştırmacılar bunun için "farklılık puanlarının güvenilirliği" yerine t-testi, varyans analizi ve regresyon analizi gibi istatistikî yöntemlerden yararlanabilirler. Pilot araştırma uygulamasıyla esas araştırma (veya ön-test son-test) sonuçları arasındaki farklılık sistematik ve tesadüfi hatalardan kaynaklanır. Tesadüfî hatalar daha çok örneklem hatasına dayanır. Pilot araştırmada nispeten küçük ve esas araştırmada ise daha büyük örnek kütlede çalışıldığından bu farklılık ortaya çıkmış olabilir. Sistematik hata ise pilot araştırmada örneklemin yanlı olarak seçilmesi sebebiyle ortaya çıkar. Araştırmacılar bu

aşamada çoğunlukla "kolayda örnekleme" yöntemine başvurduklarından sonuçlar belirli bir yanlılığa sahiptir. Öte yandan bir test aynı kişilere ikiden fazla kez uygulanmışsa, ölçüm sonucunda elde edilen veriler arasındaki tutarlılık/güvenilirlik fark puanları yerine *tekrarlanmış TEYVA* ile yapılır.[27]

FARKLILIKLARIN STANDART HATASI

Farklılıkların standart hatası bir kişinin iki testten aldığı puanların değerlendirilmesi veya iki kişiye ait puanların karşılaştırılması için kullanılır. Bir kişi aynı zaman diliminde farklı yetenekleri ölçen iki test almışsa veya bir testi aldıktan sonra eğitimle kendisini yetiştirip ikinci bir test daha almışsa bu testlerden alınan puanların birbirinden gerçek anlamda farklı olup olmadığı fark puanlarının standart hatası ile hesaplanır. Fark puanlarının standart hatasını hesaplamak isteyen bir araştırmacı aşağıdaki sorulardan hareket eder.

1. Bir kişinin A testinden aldığı puanlar B testinden aldığı puanlarla uyumlu mudur?
2. Bir kişinin A testinden aldığı puan başka bir kişinin aynı testten aldığı puanla benzerlik göstermekte midir?
3. Bir kişinin A testinden aldığı puan başka bir kişinin B testinden aldığı puanla karşılaştırılabilir mi?

Fark puanlarının standart hatası terimini daha öz bir şekilde, "farklılıkların standart hatası" (FSH) şeklinde de ifade edebiliriz. Farklılıkların standart hatası Eşitlik 9-17'deki formülle hesaplanır.

$$FSH = \sqrt{\left(OSH_a^2 + OSH_b^2\right)}. \qquad (9\text{-}17)$$

OSH_a^2 = Birinci ölçümün standart hatasının karesi.
OSH_b^2 = İkinci ölçümün standart hatasının karesi.

Örneğin, yapılan araştırma sonucunda bir kişi A testinden 300 ve B testinden ise 350 puan almış olsun. Yapılan hesaplamada FSH ise 30 çıkmış bulunsun. Bilim adamı bu değerleri göz önünde bulundurarak bireyin yeteneklerinde gerçek anlamda bir farklılık bulunup bulunmadığına belirli

bir anlamlılık düzeyinde Z değeriyle FSH değerini çarparak ($Z_{\alpha/2} \times$ FSH) karar verir. Bunun için alfa anlamlılık düzeyi araştırmacının liberal veya tutucu bir tutum takınmasına göre 1 standart hata veya 2 standart hata değeri çerçevesinde hesaplanır [(1,00* FSH) / (1,96*FSH)]. Tutucu tutumlara sahip araştırmacılar %95 güvenilirlik düzeyinde 1,96 değerini (veya yaklaşık olarak 2 rakamını) temel alırlar. Eğer %95 güvenilirlik düzeyinde çalışılmışsa FSH değeri 60 olacaktır. Birinci testin değeri temel alınarak puanların 240 ilâ 360 arasında değişebileceği görülür. Anlamlı bir farklılık olabilmesi için ikinci testin puanının 360'den yüksek olması gerekir. Puan 350'de kaldığından iki test puanları arasında anlamlı bir farklılık olmadığına karar verilir. Ancak liberal bir yaklaşımla bir standart sapma temel alınsaydı bu kez %68 güven aralığında iki test sonucu arasında anlamlı bir farklılık olduğu sonucuna varılırdı.

ALINTI YAPILAN KAYNAKLAR

[1] California Department of Education Standards and Assessment Division, "Alignment, Validity, and Reliability of the Spring 2000 Golden Stat Examinations [Bahar 2000 Altın Devlet Sınav Sonuçlarının Denkleştirilmesi, Geçerliliği ve Güvenilirliği]," <http://www.google.com.tr/search?q=cache:YYWUqdXx-rEJ:www.cde.ca.gov/statetests/gse/admin/gsereliabilityrpt.pdf+%22standard+error+of+measurement%22+Confidence+intervals+reliability&hl=tr&ie=UTF-8&inlang=tr> (10.08.2003).

[2] B. Simmerok, "Standard Error of Measurement [Ölçümün Standart Hatası]," <http://home.apu.edu/~bsimmerok/WebTMIPs/Session6/TSExmpl.htm> (09.08.2003).

[3] F.N. Kerlinger, *Foundations Behavioral Research*, New York: Holt, Rinehart and Winston, 1973, 452-453.

[4] Becker, "Reliability and."

[5] G. Cunningham, "Questions for Someone who Knows More About Statistics Than I Do [İstatistik Hakkında Bazı Sorular]," <http://www.google.com.tr/search?q=cache:OzqoAfv-X3cJ:www.interversity.org/lists/arn-l/archives/may2002/msg00537.html+reliability+rasch+%22standard+error+of+measurement%22&hl=tr&ie=UTF-8&inlang=tr> (05.08.2003).

[6] L.S. Wang, "Standard Error of Measurement [Ölçümün Standart Hatası]," <http://psy.ccu.edu.tw/testroom/Reliability_Part_Two.doc> (12.08.2003).

[7] John Michael Linacre, "The New Rules of Measurement [Yeni Ölçüm Kuralları]]," <http://www.google.com.tr/search?q=cache:O1w_0YH4xLYJ:www.rasch.org/rmt/rmt132e.htm+reliability+rasch+%22standard+error+of+measurement%22&hl=tr&ie=UTF-8&inlang=tr> (05.08.2003).

[8] Wang, "Standard Error."

[9] John Michael Linacre, , "The New Rules."

[10] Institute for Objective Measurement, "Separation, Reliability and Skewed Distributions [Ayrılma, Güvenilirlik ve Çarpık Dağılımlar]," <http://www.rasch.org/rmt/rmt144k.htm> (05.08.2003).

[11] J. Bielinski, M. Thurlow, J. Minnema, ve J. Scott, "How Out-of-Level Testing Affects the Psychometric Quality of Test Scores [Seviye Dışındaki Test Soruları Test Puanlarının Psikometrik Kalitesini Nasıl Etkiler]," <http://education.umn.edu/nceo/OnlinePubs/OOLT2.html.> (15.08.2003).

[12] Reed A. Castle, "The Relative Efficiency of Two-Stage Testing Versus Traditional Multiple Choice Testing Using Item Response Theory in Licensure [Klasik Çoktan Seçmeli Test Yerine İki Aşamalı Test Uygulamasının MYK ile Testi]," <http://dwb.unl.edu/Diss/RCastle/ReedCastleDiss.html.> (15.08.2003).

[13] B. Wright, "Which Standard Error? [Hangi Standart Hata]," <http://www.rasch.org/rmt/rmt92n.htm.> (16.08.2003).

[14] Aynı.

[15] Aynı.

[16] P. Doolittle, "The Three Necessities [Üç Gereklilik]," <http://www.google.com.tr/search?q=cache:FmADJLAv1FAJ:edpsychserver.ed.vt.edu/resources/pdf/assessment2.pdf+%22standard+error+of+measurement%22+pdf&hl=tr&ie=UTF-8&inlang=tr> (09.08.2003).

[17] Test of English For International Communication, "TOEIC Technical Manual [TOEIC El Kitabı]," <http://www.google.com.tr/search?q=cache:11EX336twbUJ:www.toeic.com/pdfs/TOEIC_Tech_Man.pdf+csem+conditional&hl=tr&ie=UTF-8&inlang=tr> (13.08.2003).

[18] Jae-Chun Ban, Won-Chan Lee ve Bradley A. Hanson, "The Effect of Component Combining Methods on Conditional Standard Error of Measurement and Classification Consistency for Composite Scores [Koşullu Standart Hata Yönteminin Bileşik Ölçümler Üzerindeki Etkisi ve Bileşik Puanlar İçin Tutarlı Sonuçlar Vermesi]," <http://tigersystem.net/aera2002/viewproposaltext.asp?propID=3595> (13.08.2003).

[19] R. Chris Fraley, Niels G. Waller ve Kelly A. Brennan, "An Item Response Theory Analysis of Self-Report Measures of Adult Attachment [Yetişkin Bağlılığını Ölçmede Öz Değerlendirme Ölçüm Araçlarının Madde Yanıt Kuramı ile Analizi]," <http://www.uic.edu/~fraley/fwb2000.pdf> (14.08.2003).

[20] Ron Gillam, "Reliability [Güvenilirlik]," <http://www.google.com.tr/search?q=cache:StIqKHSltw0J:www.utexas.edu/courses/pena/reliability.html++reliability+standard+error+difference+scores&hl=tr&ie=UTF-8&inlang=tr> (12.08.2003).

[21] W.G. Hopkins, "Mean +/- SD or Mean +/- SEM [Oratalama +/- Standart Sapma mı yoksa Oratalama +/- Ölçümün Standart Hatası mı?]," < http://www.sportsci.org/resource/stats/meansd.html> (16.08.2003).

[22] "The Reliable Change Index [Güvenilir Değişim İndeksi]," <http://www.medal.org/adocs/docs_ch39/doc_ch39.26.html> (05.08.2003).

[23] Neil S. Jacobson ve d. , "Methods for Defining and Determining the Clinical Significance of Treatment Effects [Tedavi Etkisinin Kliniksel Anlamlılığını Belirleme ve Tayin Etme Yöntemi]," <http://homepage.psy.utexas.edu/homepage/faculty/Telch/Research%20Design%20Class/Assigned%20Readings/Clinical%20Trials/ClinicalsignificanceB.pdf> (24.01.2004).

[24] I. Becker,"Statistical and Clinical Significance [İstatistiksel ve Klinik Anlamlılık]," <http://www.google.com.tr/search?q=cache:FKQaB04BzOQJ:web.uccs.edu/lbecker/Psy590/clinsig.htm+%22Reliable+Change+Index%22+&hl=tr&ie=UTF-8&inlang=tr> (09.08.2003).

[25] H.S. Kopeikin, "Psychology:Lecture 4 [Psikoloji: Ders 4]," <http://mentor.lscf.ucsb.edu/course/fall/psych121/lecture/121lec04.htm> (09.10.2002).

[26] darwin@ub.fu-berlin.de, "Test-Retest Reliability [Test-Yeniden Test Güvenilirliği]," <http://darwin.inf.fu-berlin.de/2003/48/AppendixB.pdf> (24.01.2004).

[27] C. Yu, Reliability of Self Reported Data [Kişisel Bildirime Dayanan Verilerin Güvenilirliği]," <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/memory.html> (09.10.2002).

KRİTER REFERANSLI ÖLÇÜMLERDE GÜVENİLİRLİK ANALİZLERİ

Eğitim sektöründe ve iş hayatında yapılan ölçümlerin önemli bir bölümünde önceden belirlenen başarı standartları temel alınır. Daha sonra elde edilen sonuçlar bu başarı standartlarına göre yorumlanır veya değerlendirilir. Önceden belirlenen başarı standartları, literatürde genel bir terim olarak *kriter* sözcüğüyle karşılanmıştır. Kriter temelli testler, norm temelli ölçümlerden farklı güvenilirlik analizlerini gerektirir. Ölçümü yapan araştırmacının odaklandığı nokta test maddelerinin iç tutarlılığı değil, test sonuçlarına göre "standartları karşıladı-karşılamadı" şeklinde verilen sınıflandırma kararının isabet derecesi ve sonraki ölçümlerde bu ayrıştırmanın ne ölçüde benzer çıktığıdır. Bu bölümde bir taraftan kriter puanların güvenilir bir şekilde "belirlenmesi" konusu üzerinde durulmuş diğer taraftan ise kriter temelli puanların "güvenilirlik" analizlerine değinilmiştir.

GENEL

Kriter referanslı test kavramı (KRT), ilk kez Robert Glaser'in (1963) çalışmalarında isimlendirilmiştir.[1] Kriter referanslı testler norm referanslı testlerden farklıdır. Bu testler daha çok sertifika almak için, başarı veya başarısızlığı saptamak için, personel seçmek ve terfi edecek[a] işgörenleri belirlemek için kullanılır. Norm referanslı testlerde araştırma yapılan ana kütlelerin veya örnek kütlelerin her biri için ayrı ayrı varyans hesaplamaları yapılarak ölçümün güvenilirliği saptanmaya çalışılırken kriter temelli testlerde verilerin ana kütleye ilişkin temsil edicilik özelliğine sahip olup olmadığı araştırılmaz. Kriter değerlerin temel alındığı ölçümlerde örneklem hacimleri görece daha dardır. Araştırma / ölçüm yapılan ve belirli bir kriteri tutturan örneklemdeki bireyler büyük ölçüde birbirlerine benzerler.

---

[a] Kriter referanslı testler terfi amacıyla kullanıldığında bir kişi bu testlerde eğer başarısız olmuşsa aradan belirli bir süre geçmeden (bu süre üç ay ilâ iki yıl arasında değişebilir) söz konusu testi tekrar alamaz.

Kriter referanslı testler, belirli düzeydeki (sınıf, okul kademeleri, meslekî bilgi) bir başarıyı ölçtüğünden bazı yazarlar tarafından "başarı testleri"[a] olarak isimlendirilmiştir. Başarı testi kavramı daha çok eğitim bilimlerinde okul giriş veya okul bitirme sınavlarındaki testler için kullanılır. Ancak bilgi, yetenek ve beceri testleri başarıyı belirlemek üzere norm temelli olarak da yorumlanabilir. Bu nedenle literatürdeki yaygın isimlendirme biçiminden hareket ederek bu kitapta "kriter referanslı test" ifadesini kullanmayı tercih ettik. Kriter referanslı test kavramı literatürde *kesim puanlı test, standart referanslı test* kavramları anlamında da kullanılır. Bazı yazarlar *kesim puanlı test, kriter referanslı test* ve *standart referanslı test* kavramları arasında ayrım gözetirler. Kesim puanlı testlerde belirli bir puanının altında kalanlar ret edilirken, kriter referanslı testlerde asgarî kesim puanı yerine belirli ölçütlere göre oluşturulmuş kriter puanlar dikkate alınır. Kriter puanlar, minimum yeteneğe sahip bireyi değil, belirli bir görevde beklenen başarı düzeyini tanımlar. Standart referanslı test ise "içerik standartlarına" dayanır. Özellikle ilköğretim ve orta öğretim okullarında bir öğrencinin her sınıfta bir dersle ilgili olarak hangi konuları bilmesi gerektiği içerik standartlarıyla belirlenmiştir. Bu okullarda ayrıca içerik standartlarının yanında başarı standartları da belirlenerek "temel", "yeterli" ve "ileri" grubunda değerlendirilebilmesi için bir öğrencinin içerik standartlarını ne ölçüde bilmesi gerektiği saptanmıştır. Bu anlamda standart referanslı test, standart içerikle ilgili başarı durumunu belirler.[2]

Saptanan kriter puanları belirli tipteki görevler, bilgi ve yetenekler için oldukça uygun iken (ehliyet sınavı, doktora yeterlilik sınavı, yabancı dil yeterlilik sınavı vb.) başka türdeki yeteneklerin değerlendirilmesinde uygun olmayabilir. Örneğin, bilişsel yetenek testlerinde (dikkat, algılama, düzlemsel ilişkiler, sözel akıcılık vb.) kişileri ayırt edecek bir kriter puanı tam olarak belirlemek veya asgarî bir kesim puanı saptamak çok zordur. Öte yandan, kriter puan veya kesim puanı norm grubunun özelliklerine göre de değişebilir. Bu kitapta "kriter referanslı test" kavramı; kesim puanlı testleri, kriter veya standart temelli testleri de içerecek şekilde geniş anlamlı olarak ele alınmıştır.

---

[a] Kriter referanslı testlere literatürde ayrıca "alan referanslı test", "yetkinlik testi" (mastery test) , "yeterlilik testi", "standart temelli test" gibi adlar da verilmiştir. Yetkinlik testi kriter referanslı test kavramına göre daha dar bir anlama sahiptir ve kişilerin yetkin olanlar ve olmayanlar şeklinde iki gruba ayrılması anlamına gelir.

ÖLÇÜM ALANININ TANIMLANMASI

Kriter referanslı test sonuçlarının güvenilirlik değerlendirmesi konusunu ele almadan önce bu tür testlerin özelliklerini tanımakta yarar vardır. Kriter referanslı testler norm referanslı testlere göre spesifik bir alandaki bilgiyi veya beceriyi ölçmeye yönelik olarak çok sayıda soru içeren ölçüm araçlarıdır. Norm temelli testlerde ölçüm konuları geniş bir yelpazeye dağılıp (geniş kavramsal alan) her bir konudan içeriğinin genişliğine göre 2 ilâ 10 arasında soru belirlenirken kriter referanslı testlerde ölçüm alanı ne kadar geniş olursa olsun kavramsal yapı alt boyutlara / konulara bölünerek her bir konuda o konunun/becerinin iyi öğrenildiğini saptamaya yönelik olarak en az 6-12 veya daha güvenilir bir sonuç elde etmek için 12-24 soru hazırlanır. Kriter referanslı testlerde ölçüm alanı, *konu içeriği uzmanları*[a] (KİU) tarafından daraltılarak içeriğin sınırları net bir şekilde belirlenir. Kavramsal yapının veya ölçülmek istenen bilgi, yetenek ve beceri (BYB) alanının sınırları çizilerek duruma göre konular, üniteler veya beceriler sınıflandırılır. Böylece ölçümün hangi konulardan, ünitelerden oluştuğu ve her bir ünitenin kaç soruyla ölçüleceği, bu soruların yeterli olup olmadığı ön incelemeye tâbi tutulur. Kriter referanslı testlerde ölçüm alanı yeterli miktarda soru çıkarmaya uygun olarak oldukça dar ve spesifiktir. Bilim adamı geniş bir alanı kapsayan kriter referanslı test oluşturmak istediğinde, bunun için her bir alt konuyu yeterince temsil edecek çok sayıda sorudan oluşan bir test bataryası hazırlamak durumundadır. Böyle bir durumda ise test gereğinden fazla maddeye sahip olacağından bazı yazarlar günümüzde gerçek anlamda kriter referanslı bir test bulunmadığını ifade etmişlerdir.

---

[a] Konu içeriği uzmanlarının belirlenmesi ve seçimi önemli bir konudur. Bu kişilerin sadece kendi konularında uzman olmaları yeterli değildir. Araştırmacı konu içeriği uzmanlarının belirlenmesinde çeşitlendirme yapmalı ve tesadüfî seçim yöntemiyle değişik alanlardan, farklı cinsiyetlerden ve farklı bilim dallarından uzmanları hakem olarak belirlemelidir. Üniversitelerdeki tek bir ana bilim dalı öğretim üyelerini konu içeriği uzmanı olarak belirlemek yeterli olmayabilir. Konu içeriği uzmanları bilfiil konuyla ilgili bir işte çalışıyor olmalı ve ölçüm yapılan konuda en az bir yılı aşmış bir iş deneyimine sahip bulunmalıdırlar. Üniversitelerde profesörler ve doçentler varken, özel olarak araştırılan konuyla ilgili olması durumu dışında, araştırma görevlilerini veya doktorasını yeni vermiş kişileri konu içeriği uzmanı olarak seçmek doğru değildir. Araştırmacı konu içeriği uzmanlarını tanıtırken bu kişilerin uzmanlık alanlarını, unvanlarını, kaç yıldır söz konusu mesleği icra ettiklerini, yaşlarını, cinsiyetlerini, adres ve telefon numaralarını da raporuna almayı ihmal etmemelidir.

ÖLÇÜM DERECELERİNİN BELİRLENMESİ

Kriter referanslı testlerde başarıya ilişkin ölçüm kriteri genelde *geçti-kaldı* veya *başarılı-başarısız* şeklinde ikili olarak belirlenir. Ancak bu yönteme itiraz edilmiştir. İnsanlar bilgi-yetenek-beceri (BYB) açısından iki dereceli bir boyut üzerinde değil çok dereceli bir boyut üzerinde sıralanırlar. Kişileri, belirli kriter puanları temel alarak daha ileri düzeyde bir sınıflandırmayla *yeterli*, *vasat* ve *vasatın altında* olmak üzere üçlü veya *zayıf*, *orta*, *iyi*, *çok iyi* gibi dörtlü gruplar içinde de sınıflandırabiliriz. Sınıflandırma derecelerinin optimum sayısı için genel geçerli bir kural yoktur. Araştırmacı derece sayısını kendi değerlendirmesine veya kurumun gereksinimine göre kendisi belirler. Ehliyet alma, lisans sertifikasına sahip olma, personel seçimi gibi belirli konularda ikili sınıflandırma biçimi uygunken öğrenci başarılarının gruplandırılmasında daha çok üçlü veya dörtlü sınıf aralıkları kullanılır. Buna göre kriter referanslı testlerin başarı değerlendirme yaklaşımlarını iki grupta değerlendirmek gerekir:

1. İkili sınıflandırmaya dayalı değerlendirmeler.
2. Çok dereceli sınıflandırmaya dayalı değerlendirmeler.

Bilim adamları, bireylerin ölçüm sürecinde iki dereceli kriter referanslı testlerin kullanılmasından mümkün olduğu kadar kaçınılmasını tavsiye etmişlerdir (Shepard 1979 ve Green 1981, aktaran Linn).[3] Çünkü değerlendirmenin kesim puanına göre yapılması ve kişilerin başarılı veya başarısız olarak iki gruba ayrılması çok doğru değildir. Kişilerin başarıları *geçti-kaldı* gibi bir kırılma noktası yerine, derecelendirilmiş bir boyut üzerinde sıralanırsa daha gerçekçi bir değerlendirme yapılmış olur (*bk.*, Tablo 10-1).

Tablo 10-1. Elli Maddelik Bir Testte Derecelendirilmiş Başarı Düzeyleri

Başarı düzeyleri	*Doğru madde sayısı*	*Doğru madde yüzdesi*
İleri	40-50	80-100
Yeterli	30-39	60-79
Temel	20-29	40-59
Sınırda	10-19	20-39
Yetersiz	0-9	0-19

Literatürde değerlendirme uygulamaları, üç grup içinde sınıflandırılmıştır. Bunlardan birincisinde *mutlak başarı* puanları, ikincisinde, *alan puanları* ve üçüncüsünde ise *derecelendirilmiş yeterlilik düzeyi* puanları temel alınır. Mutlak başarı puanında bir kişinin yeterliliği süreklilik gösteren bir boyut üzerinde, örneğin 10 dereceli bir ölçek üzerinde belirlenir. On puan, bir kişinin en üst düzeyde yeteneğe sahip olduğunu gösterirken 1 puan kişinin zayıf olduğunu veya öğrenmenin henüz başlangıç aşamasında olduğunu ortaya koyar. Alan puanı da mutlak başarı puanına benzer, farklı olarak ölçüm konusuyla ilgili maddeler daha dikkatli ve özenli olarak sadece söz konusu alanın tamamını kapsayacak şekilde seçilmiştir. Derecelendirilmiş yeterlilik düzeyi puanları ise, test alan kişinin önceden belirlenmiş bir sınır değeri veya çoklu derecelendirme söz konusu ise sınır değerlerini aşıp aşmadığıyla ilgilidir. Kişiler test sonucuna göre *yeterli* veya *yeterli değil* şeklinde ikili olarak veya *vasat, iyi, çok iyi* şeklinde derecelendirilmiş olarak belirli yeterlilik gruplarına ayrılırlar. Yeterlilik sınır değerini veya değerlerini belirlemeye yönelik olarak literatürde 30'dan fazla yöntem önerilmiştir. Bu yöntemlerin büyük çoğunluğu ikili ayrıştırmayı temel alan kesim puanı hesaplamalarına dayanır.

DEĞİŞKENLİĞİN DÜŞÜK OLMASI

Kriter referanslı testlerde ölçüme katılan bireyler arasındaki değişkenlik düşüktür. Örneğin, ilköğretim okulundaki yedinci sınıfın öğrencileri bilgi, yetenek ve beceri açısından büyük ölçüde birbirlerine benzerler. Aynı şekilde sekreterlik işi için başvuran adaylar da gazete ilanında belirlenen şartları taşıdıkları ölçüde birbirine benzer niteliklere sahiptirler. Bu tür testleri alan kişi sayısı görece az olduğundan bilgi, yetenek ve beceri açısından kişilerin kendi aralarında büyük farklılıklar görülmez. Varyansın düşük olması sonuçta korelasyon katsayılarını da etkiler. Bu nedenle az sayıda kişi üzerinde ölçüm yapılan kriter referanslı testlerde test-yeniden test ve paralel formlar puanlarına dayalı olarak yapılan korelasyon analizi sonuçları anlamlı çıkmaz, diğer bir deyişle bu sonuçlar güvenilir değildir. Korelasyon katsayısını sadece örneklem hacminin küçük olması faktörü değil, aynı zamanda kriter puanının kendisi de etkiler. Bu tür testlerde puanlar düşük veya yüksek uçta yoğunlaştığından dağılım eğrisi sağa veya sola çarpıktır.

KULLANIM ALANLARI

Günümüzde kriter referanslı testler daha çok eğitim sektöründe ve iş hayatında kullanılır. Türkiye'de temel öğretim, orta öğretim ve yüksek öğretimde geçme-kalma durumu asgarî başarı puanına (kriter) göre belirlenmiştir. Bu puan örneğin temel öğretimde 2 iken, yüksek öğretimde 50, 60 ve bazen de uygulanan sisteme göre değişiklik göstererek 30 gibi bir değer olabilmektedir. İş hayatında ise kriter puanlar daha çok personel seçimi ve terfi uygulamalarında kullanılan "bilgiye dayalı testler" veya "akademik yeteneği ölçen testler" için söz konusudur. İş hayatındaki uygulamalarda standart bir kesim puanının varlığından söz edilemez. Teamüller bilimsel gerçekliği değil, performans tercihlerini yansıtır. ABD gibi gelişmiş ülkelerde personel seçiminde uygulanan kesim puanları sık aralıklarla mahkemelerde dava konusu olmuştur. Bu puanın gereğinden fazla yüksek veya düşük belirlenmesi yeterlilik değerlendirmesini geçersiz hale getirir.

Okullarda Kriter Referanslı Testler

Okullarda kullanılan kriter referanslı testler öğrenci başarılarının değerlendirilmesine yöneliktir. Değerlendirme sonucunda, öğrencileri 100 üzerinden 70 puan (5 üzerinden 3,5 veya 4) almaları halinde başarılı sayan ölçüm araçlarıdır. Yetmiş puan kriteri eğitim sektöründe yeterliliği belirlemek için yaygın bir şekilde kullanılan keyfî bir değerdir ve bu nedenle de çok sağlıklı değildir. Kesim puanını rasyonel bir şekilde belirlemek için yeterli zamanı olmayan okullar geçici bir süre için öğrencilerini bu değere göre sınıflandırırlar. Kriter referanslı test uygulamasında öğrenciler sınıf veya birden fazla şube varsa şubeler ortalamasına göre değil, önceden belirlenmiş başarı eşik değerine göre iki grup halinde sınıflandırılırlar: eşiği aşanlar ve aşamayanlar. Kriter referanslı testlerde öğrenciler arkadaşları yerine yetenek, bilgi ve beceri açısından önceden belirlenmiş kriter puana göre yarışırlar. Norm temelli testlerde belirlenen "standart", grup ortalaması iken; kriter referanslı testlerde "standart" baraj olarak belirlenen ve belirli formüllerle ortaya çıkarılan puandır.

Öğrencilere yönelik olarak hazırlanacak kriter referanslı testlerde bir konunun veya ünitenin iyi öğrenilip öğrenilmediğini belirlemek için o konuyla ilgili olarak en az altı soru sorulmalıdır. Sadece üç dört soru ile bir öğrencinin kesim puanının aşıp aşmadığına, konuyu öğrenip öğrenmediğine karar vermek sağlıklı bir değerlendirme olmaz. Kriter referanslı test oluştururken testin uygulanacağı hedef kitle de göz önünde bulundurulur. Öğrenciler il bazında, bölge bazında veya ülke bazında test alıyor olabilirler. Bu nedenle kriter referanslı test geliştirme çalışmalarında hedef kitlede

pilot uygulama yapılması ve maddelerin temsil edici örneklem üzerinde sınanması gerekir. Öğrencilere uygulanan kriter referanslı testlerde eşik değerin farklı test uygulamalarında farklı sonuçlar vermemesi için keyfî olarak belirlenen ,70 değeri yerine çoğunlukla *değiştirilmiş Angoff yöntemi* kullanılır. Bu yöntem sayesinde başarı eşiği her bir sınav için farklı olarak belirleneceğinden öğrenciler bir sınavdan geçmeyi diğer sınava göre daha kolay veya daha zor olarak değerlendirmezler.

İş Hayatında Kriter Referanslı Testler

Kriter referanslı testler, eğitim sektörünün dışında işletmelerde personelin seçilmesi ve terfi ettirilmesi amacıyla da kullanılabilir. İş hayatında kullanılan bilişsel testler genel karakteri itibariyle norm temelli olmasına karşılık, spesifik bir alana ilişkin yeterliliğin, bilgi ve beceri sahibi olunduğunun belirlenmek istenmesi durumunda kriter referanslı testlerden yararlanmak daha doğrudur. Norm referanslı testler okullarından yeni mezun olan kişilerin yetenek ve becerilerini ölçmek için uygun iken, iş ortamında deneyimli kişilerin belirli yetenek ve becerilere sahip olup olmadıklarını belirlemek için kriter referanslı testler kullanılır.

İş hayatında bilişsel yeteneklerin dışında bazen fiziksel yeterliliği belirlemek üzere de kriter referanslı testlere başvurulabilir. Örneğin; polis memurlarının, itfaiyecilerin, askerî öğrencilerin dayanıklı ve güçlü olmaları kriter referanslı testlerin kesim puanlarına göre belirlenir. İnsan kaynakları biriminin daktilo/sekreter seçiminde dakikada en az 40 kelime yazma ve her 100 kelimede sadece beş hata veya en fazla %5 oranında hata yapma gibi bir kriteri seçim standardı olarak belirlemesi her zaman mümkündür. Norm referanslı testlerde testi alan kişiler için "vasat", "zayıf" değerlendirmesi yapılırken; kriter referanslı testlerde işi "yapabilir" veya "yapamaz" kesin yargısına varılır.

Kesim puanının belirlenmesi. Yeterlilikler, iş hayatında duruma göre birden fazla KRT ile ölçülebilir. Bu amaçla bir test bataryası uygulanmışsa kesim puanı / kriter puanı ya, tek tek test sonuçlarına göre veya batarya toplam puanına göre belirlenir.[a] Araştırmacı testlerin tahmin puanlarını (ham puanları) farklı yaklaşımlara göre birleştirme imkanına sahiptir. Bu konuda sık başvurulan birleştirme yöntemleri aşağıdaki gibidir:[4]

[a] Aşamalı eleme yöntemi uygulanıyorsa kesim puanı her bir test için ayrı belirlenir.

1. Çoklu regresyon analizi.
2. Çoklu kesim puanı uygulaması.
3. Aşamalı eleme yöntemi.
4. Kombinasyon yöntemi.
5. Profil denkleştirme yöntemi.

Kitabın amacı, seçim bataryasındaki testlere ait tahmin puanlarını birleştirme tekniklerini tanıtmak olmadığından bu bölümde sadece kesim puanlarının belirlenmesi konusu üzerinde durulmuştur. Kesim puanları tek bir tahmin göstergesine veya çoklu tahmin göstergelerine dayalı olarak belirlenebilir.

Tek bir tahmin göstergesine dayalı olarak kesim puanı belirleme yöntemleri. Bu grupta iki temel teknik vardır: Thorndike tarafından önerilen tahminî hasat yöntemi ve beklenti grafiği metodu.[5]

Thorndike'ın "tahminî hasat" yönteminde *hasat* sözcüğü, çürüklerin elenmesi sonucunda işletmeye alınan toplam personel sayısını tanımlar. Bu yöntemde araştırmacı gazetelerde ilân edilen boş pozisyon sayısı ile bu pozisyonlara müracaat eden kişi sayısını göz önünde bulundurur. Örneğin, 20 boş pozisyon için 160 kişinin iş müracaatında bulunabileceğini tahmin etsin. Böyle bir durumda, başvuran adayların işe girme oranı 20/160 formülüne göre belirlenir. Seçim oranı %12,5 olduğundan kesim puanı 88. yüzdelik dilim olarak saptanır. Ölçüm hatalarının etkisini gidermek için ,88 yüzdelik dilimden eksi bir standart sapma ÖSH değeri çıkarılarak fiilî kesim puanı saptanır. "Eksi bir standart sapma ÖSH" değerini kullanmamız ölçüm hatalarının etkisini azaltmak istememiz nedeniyledir.

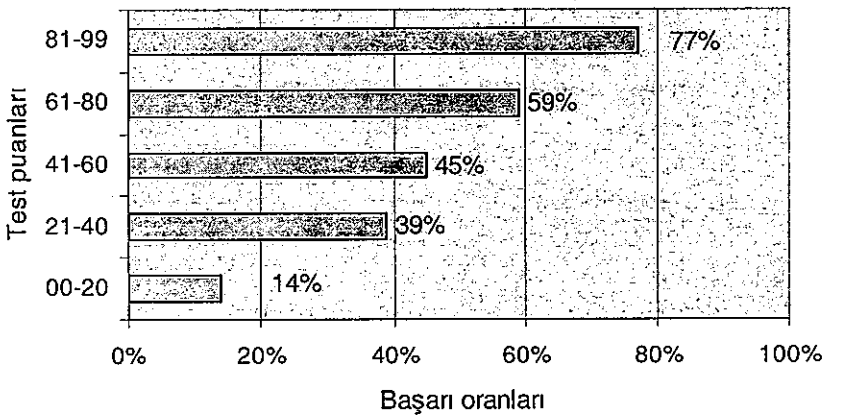
Kesim puanını belirlemede kullanılan tek göstergeli bir diğer yaklaşım, "beklenti grafiği" metodudur. Beklenti grafiği, beklenti tablosuna dayalı olarak çizilir (*bk.*, Tablo 10-2). Bu yaklaşımda X puanına sahip bir adayın Y başarı düzeyini tutturma olasılığının ne olduğu araştırılır. Örneğin, bir işletmede 100 makine operatörümüz bulunsun. Bu makine operatörlerine "mekanik bilgi testi" uygulamış olalım. Aynı zamanda söz konusu makine operatörlerinin dönem sonunda elde edilen başarı puanlarına sahip olduğumuzu varsayalım. Performans ölçüsü 5 puan temel alınarak 4 ve 4'ün üzerinde olanlar yeterli, düşük olanlar ise yeterli olmayan kişiler olarak değerlendirilmiş bulunsun. Bu verilere dayalı olarak beklenti grafiği beş adımda çizilir.[6]

Birinci adımda dönem sonu performans puanları dikkate alınarak kişiler *yeterli* ve *yeterli olmayanlar* şeklinde iki gruba bölünürler. İkinci aşamada yeterli ve yeterli olmayanların test puanları belirlenir. Üçüncü aşamada test puanlarının dağılımı kartil veya kantil değerlerine (dörtte birlik veya beşte birlik puan dilimlerine) dönüştürülür. Dördüncü aşamada ise yeterli gruptaki bireylerin yüzdesi saptanır.

Tablo 10-2. Beklenti Tablosu

Test puanları	Yeterli değil	Yeterli	Toplam	Yeterli bireylerin oranı
00-20	12	2	14	%14
21-40	11	7	18	%39
41-60	11	9	20	%45
61-80	9	13	22	%59
81-99	6	20	26	%77
Toplam	49	51	100	

Beklenti tablosunun oluşturulmasından sonra "test puanları" ve "yeterli bireylerin oranı" değerleri dikkate alınarak beklenti grafiği çizilir (*bk.*, Şekil 10-1).



Şekil 10-1. Beklenti grafiği.

Beklenti grafiği bize aynı zamanda kesim puanının hangi dilimde tutulması gerektiği konusunda bilgi verir. Kesim puanının üst %20'lik dilimde tutulması halinde personelin başarılı olma şansının %77 olacağı tahmin edilir. Öte yandan ölçüm hataları faktörünü göz önünde bulundurmak ve kişilerin işe giriş şanslarını artırmak için kesim puanı ölçümün bir standart hatası kadar düşük tutulur.

Çoklu tahmin göstergelerine dayalı olarak kesim puanını belirleme yöntemleri. Çoktan seçmeli soruları içeren *bilgi testleri* ve bu testlerden oluşan bataryalar için uygundur. Çoklu kesim puanı yaklaşımı, sınır puanının belirlenmesinde soruların zorluk derecelerini temel alır. Bu grupta Angoff, değiştirilmiş Angoff ve Ebel yöntemi gibi yaklaşımlar bulunur. Sınır puanı belirlenirken ya belirli bir teknik seçilir veya bataryadaki her bir test için değişik kesim puanı yöntemleri uygulanarak bunların ortalaması alınır. Bazen ortalama yerine performansı en iyi temsil etme özelliğine sahip olan kesim puanı tekniğiyle elde edilen değerler geçme / başarı oranı olarak belirlenir. Tek bir kesim puanı yöntemi bütün durumlar için en iyi çözümü vermediğinden kesim puanı mümkün olduğunca maksimum iş başarısı standardını sağlayacak şekilde farklı yöntemler sınanarak belirlenmelidir. Hesaplama yapmaya zamanı olmayanlar için genel uygulama, bu puanın 100 üzerinden 70 olarak belirlenmesi şeklindedir, fakat daha önce de belirtildiği gibi bu yöntem herhangi bir hesaplamaya dayanmadığından sağlıklı bir yaklaşım değildir.

Dikkat edilmesi gereken hususlar. Belirlenen kesim puanlarının çalışanlar/gözlemciler tarafından gerçekten "yeterlilik" sınır düzeyi olarak algılanması önemlidir. Bu puanların düşük tutulması veya gereğinden çok daha yüksek belirlenmesi personel/gözlemciler tarafından normal karşılanmaz. Belirlenecek sınır/eşik puanı başarılı personelin beklentilerine uygun olmalı ve *yeterlilik* düzeyi açısından makul karşılanmalıdır.

Son yıllarda Türk bilimsel yazınında "yeterlilik" sözcüğü ile "yetkinlik" sözcükleri birbirine karıştırılmaktadır. Bunun nedeni dünya literatüründe de kavramın aynı belirsizliğe sahip olmasıdır. İngiltere'de "Meslekî Nitelikler Çalışma Grubu" (1986) yeterliliği (competence), "kişilerin ne bildikleriyle ilgili değil, ne yapabildikleriyle ilgili bir kavram" olduğunu belirtmiştir.[7] Bu tanımda belirli bir zaman süresi içinde kişinin ne yapabileceği veya yaptığı konusu üzerinde odaklanılır. Söz konusu çalışma grubuna göre, yeterlilikte önemli olan sonuçlardır. Yetkinlik kavramı (competency) ise daha dar bir çerçeveye sahiptir. Çalışma grubuna göre,

yetkinlik atomistik bir kavramdır, bu nedenle sadece belirli yetenekleri, davranışları veya göstergeleri isimlendirmek için kullanılır. İnsanlara ait *yeterlilikler* daha sonra istenirse birbirinden ayrı bir dizi faaliyet, yetenek, bilgi, beceri ve davranış ögelerine ayrılabilir. Sınırları daraltılmış, küçültülmüş veya atomize edilmiş bu yetenekler *yetkinlikleri* ifade eder.[8] Yetkinlikte *davranışsal göstergeler* önemlidir.[9] İngiliz kaynaklarının tersine ABD kaynaklarında yeterlilikler, "objektif olarak gözlenebilen davranışlar" olarak tanımlanırken, yetkinlikler "başarının temelinde veya arka planında yatan faktörler" olarak ifade edilmiştir. Örneğin; karar verme, analiz gücüne sahip olma, planlama, kişiler arası iletişim yeteneği, ikna etme özelliği, takım çalışmasından yana olma, stres ve engellenmelerin etkisini çabuk atlatabilme, enerji dolu olma, işi sahiplenme duygusu yetkinliklerle ilgilidir.[10]

Görüldüğü gibi yetkinlik ile yeterlilik kavramları arasındaki sınır çok net değildir. Uluslar Arası Çalışma Örgütü (ILO) kaynaklarında iki kavram geçişimli olarak kullanılmıştır. Türkçede kavramın kullanıldığı yer önemlidir. Genel kullanım ile spesifik kullanım arasında ayrım gözetmek gerekir. Genel kullanımda yetkinlik ile yeterlilik kavramları birbirinin yerine kullanılabilir, fakat spesifik kullanımda ayırım gözetmek doğru olur. Örneğin, insan kaynakları uzmanları yetkinlik ile yeterlilik kavramları arasında ayrım gözetirler. İnsan kaynakları uzmanlarına göre yeterlilik, bir kişinin mesleğini icra edebilmesi için sahip olması gereken minimum bilgi, yetenek ve beceri standardını tanımlar. Yetkinlik ise, asgarî iş gereklerinin üzerinde, çalışanların sahip olmaları veya kazanmaları gereken tutum ve davranışları belirtir.[11] Yeterlilik kavramında, bir kişinin işiyle ilgili olarak ortaya koyduğu *birincil sonuçlar* üzerinde odaklanılırken, yetkinlik kavramında aynı kişinin tutum ve davranışlarıyla ortaya koyduğu *genel sonuçlara* önem verilir. Bu anlamda yetkinlik; meslek ahlakını benimseme, gerekli olgunluğa erme ve meslekte mükemmelleşmedir. Yetkinlikte; meslekle ilgili diğer alanlar arasında bağlantı kurma, görgü, duyarlılık, özen gösterme, titizlik, ilişkiler, sorumluluk, adanmışlık ve bağlılık vardır. Yetkinlik, yaşam boyu devam eden sürekli eğitim ve deneyimle gelişir. Yetkinlik, belirli bir mesleği 3-7 yıl gibi belirli bir süre icra ettikten sonra kişinin kendisini çevresindekilere bir şekilde kanıtlamasıyla başlar. Bu kanıtlama, yöneticilerin ve diğer çalışanların toplu onayları şeklinde olabilir veya girilen sertifikasyon sınavlarında, terfi sınavlarında, unvan-derece sınavlarında başarılı olunarak elde edilir. Ancak yetkinlik, sadece sınavda başarılı olunarak elde edilebilecek mekanik bir süreç değildir. Uygun tu-

tum ve davranışlarla desteklenmediği sürece sınav başarısı kişiye yetkinlik saygısını kazandırmaz. Yetkinlikte yeterlilikten farklı olarak uzmanlaşma vardır. Uzmanlaşma konuya hakim olmayı, kişinin kendisine ve başkalarının da o kişiye güvenmelerini ifade eder.

Kriter referanslı testlerde "yeterliliğin" ölçümü söz konusudur. En genel anlamıyla bireysel yeterlilik, sınırları belirlenmiş spesifik bir alanda başarısız olmamayı tanımlar. Bir kişinin asgarî düzeyde de olsa işinde başarılı olacağını ve iş yapma yeterliliğine sahip olduğunu belirtir. ABD'de *Uniform Guidelines on Employee Selection Procedures* isimli etik kurallar kitabında "kabul edilebilir yeterlilik puan sınırı"nın işgücü çevrelerindeki normal başarı standardı "beklentilerine" uygun olması gerektiği vurgulanmıştır.

Kabul edilebilir yeterlilik puan sınırı genelde *konu içeriği uzmanlarının* test maddeleri üzerinde yapacağı incelemeler sonucunda belirlenir. Konu içeriği uzmanları bir taraftan test maddelerinin işle ilgililiğini saptarlarken diğer taraftan her bir test maddesinin doğru yanıtlanma oranı hakkında kendi tahminlerini ortaya koyarlar. Bu süreçte konu içeriği uzmanlarının maddelerin işle ilgililiği konusunda tam bir mutabakat içinde olmaları gerekmez. Uzmanların %50'sinin bir maddenin işle ilgili olduğunu belirtmeleri pek çok vak'ada yeterli bir kanıt olarak değerlendirilir. Fakat daha güvenli sınırlarda çalışmak isteyen araştırmacılara %70 oranı önerilir. Diğer bir deyişle, eğer yedi uzmanla çalışılıyorsa bu uzmanların en azından beşinin aynı görüşte bulunmalarının sağlanması gerekir.[12] İlgililik çalışmasından sonra konu içeriği uzmanları her bir maddenin asgarî yetenek düzeyine sahip kişiler tarafından hangi oranda yanıtlanabileceğine karar verirler. Bu aşamada değişik kesim puanı hesaplama yöntemlerinden yararlanılır. Kesim puanı *başarılı-başarısız* şeklinde ikili veya *başarısız, minimum yeterliliğe sahip, yeterli* ve *üstün* şeklinde derecelendirilmiş olabilir. Personel seçimi ve terfi işlemlerinde öncelikle "yeterlilik" kavramının tanımının yapılması gerekir. İşetme eğer asgarî yeterliliğin üzerinde bir başarıyı hedefliyor ve başarı standardı için gerçekten bu başarı düzeyi gerekli ise bunu açık bir şekilde tanımlamalıdır. Bu, minimum yeterlilik düzeyinin üstünde bir üst kademeyi oluşturur. Personel seçim testlerinde kesim puanları işin gerektirdiği performans gereğine göre bazen minimum yeterlilik düzeyi temel alınarak bazen de bir üst düzeye göre belirlenir. Yurt dışındaki uygulamalarında işin gerektirdiği performans minimum yeterliliği gerektiriyorsa ve işletme ise minimum yeterlilikten daha yüksek bir başarı standardı belirlemişse sınava katılıp kazanamayan adayların mahkemelere

başvurup test sonuçlarını dava etme ve sınavı iptal ettirtme hakları vardır. Karşıt kültürel çevrelerde doğru olan yaklaşım, performans kriterinin normal koşullardaki başarı beklenti düzeyine uygun olarak belirlenmesidir.

Güvenilirlik değerlendirmesi. İş hayatında uygulanan kriter referanslı testlerde en çok test-yeniden test, gözlemci içi değerlendirme ve gözlemciler arası değerlendirme güvenilirliği yöntemleri uygulanır. Çünkü değerlendirmenin odak noktası ölçüm aracının iç tutarlılığı değil, verilen *başarılı-başarısız* kararına dayalı olarak yapılan sınıflandırma uygulamasının güvenilirliğidir.

KESİM PUANININ BELİRLENMESİ

Kriter referanslı testlerde başarı standardını veya kesim noktasını belirlemek için çok sayıda yöntem önerilmiştir. Bunlardan sık kullanılanları; (1) normal dağılım eğrisi, (2) Angoff yöntemi (1971), (3) Ebel yöntemi (1972) (4) Nedelsky yöntemi (1954), (5) madde haritası yöntemi (item mapping), (6) zıt gruplar yöntemi (contrasting groups) ve (7) ayraç (bookmark) yöntemleridir. Literatürde hangi kesim puanının kullanılması gerektiği konusunda bilim adamları belirli bir mutabakattan çok anlaşmazlık içindedirler. Her bir yöntemin avantajları ve yetersiz kaldığı durumlar söz konusudur. Araştırmacılar, bilim adamları ve test merkezlerinin yetkilileri kesim puanını belirleme yönteminin seçilmesine kendi özel amaçlarını, test uygulama çalışmasının büyüklüğünü, maliyet ve zaman ögesini dikkate alarak karar verirler. Çünkü başarı standartlarını belirleme işi bir "bilim" olmaktan çok bir "sanat" olarak değerlendirilmiştir.[13] Glass (2003) kriter puanlarının belirlenmesine yönelik yaklaşımları altı başlık altında toplamıştır.[14]

1. Diğerlerinin başarı puanlarının temel alınması.
2. Yüz rakamından geriye doğru sayma.
3. Diğer kriter puanlarının üzerine ekleme yapma.
4. Minimum yeterlilik düzeyini belirleme.
5. Teorik yaklaşımlardan hareket ederek karar verme.
6. Yöneylem araştırmaları yöntemi.
7. Madde merkezli değerlendirme.

Aşağıdaki paragraflarda kesim puanını belirlemede Glass'ın sınıflamasına bağlı olmadan sık kullanılan belli başlı yöntemler değerlendirmeye alınmıştır. Kriter referanslı test puanlarının güvenilirliği kesim puanıyla birlikte değerlendirilmelidir. Kesim puanı eğer kişileri istikrarlı bir biçimde ayrıştırmıyorsa böyle bir durumda diğer yöntemler de denenerek maksimum düzeyde güvenilirlik sağlayacak kesim puanı yöntemi bulunmaya çalışılır.

Diğerlerinin Başarısı

Bu yöntemde kesim puanı belirlenirken kendilerine test uygulanan kişilerin ham puanları temel alınır. Bir kişinin başarılı veya başarısız sayılmasında temel alınacak eşik değer;

1. ham puanların medyan değerine,
2. aritmetik ortalama değerine,
3. %50 yüzdelik dilimine göre,
4. puanların ,90 *sınıf eşdeğerlik değerine* göre veya sadece
5. keyfî olarak belirlenen ,70 başarı puanına

göre saptanabilir. Bazı uygulamalarda medyan veya aritmetik ortalama değeri yerine aritmetik ortalamanın 1,5 standart sapma altındaki değer minimum yeterlilik düzeyi olarak belirlenmiştir.[a] Okullarda ve üniversitelerde genellikle ,50 başarı puanı kesim puanı olarak kabul edilir. Eğri (körv) sisteminin uygulandığı okullarda ise başarı standardının ,35'e hatta ,30'a kadar düşürüldüğü görülür. Yüzde 70 eşik değeri ise, personel seçim testlerinde, meslekî eğitim testlerinde, terfi testlerinde, sertifika programlarında ve lisans testlerinde kullanılır.[15] "Diğerlerinin başarısına göre" kriter puanı belirleme, aslında *norm referanslı* test uygulamasına benzer. Farklı olan yönü ilk kez uygulandığında norm grubunun temel alınması daha sonraki uygulamalarda ise elde edilen değerin bir kriter olarak kabul edilmesidir. Diğerlerinin başarısı temel alınırken *minimum yeterlilik* ile, *yeterlilik* ve *üst düzey yeterlilik* puanının ne olacağına konu içeriği uzmanlarının birlikte karar vermeleri gerekir. Aritmetik ortalamadan 1,5 standart sapma-

[a] Bu değer üniversitelerdeki not verme sistemlerinde D puanına karşılık gelir. Bu sınır düzeyindeki kişiler tüm öğrencilerin yaklaşık %24'ünü temsil eder. Ortalamadan negatif yönde iki standart sapma ise F başarısız derecesini temsil eder.

nın çıkarılmasıyla elde edilen değer minimum yeterlilik düzeyini belirlerken ortalamaya ,05 standart sapmanın eklenmesiyle elde edilen değer *yeterliliği* ve ortalamadan 1,5 standart sapma değerindeki pozitif fark ise *üst düzey yeterliliği* belirleyebilir. Sınır puanını belirleme işlemi, genelde bir icra uygulaması veya politika uygulaması olarak ortaya çıkar. Eşik puanı, okullarda "zorunlu bir icra" uygulaması iken, işletmelerde "politika" olarak ortaya çıkar. Zorunlu bir icra uygulaması olması nedeniyle okullarda minimum yeterlilik düzeyi ön plandadır. İşletmelerde ise politika gereği olarak veya mevcut başarı kanıtları dikkate alınarak başarı standardı bir çok işte minimum yeterlilik düzeyinin bir kademe üstünde belirlenir.

Yüz Puandan Geriye Doğru Sayma

Bu yöntemde önce testin amacı belirlenir ve test maddeleri bu amaca yönelik olarak hazırlanır. Testin amacı katılımcıların belirli bir sürede eğer maddelerin %95'inine doğru yanıt vermelerini sağlamaksa test maddeleri bu amaca göre oluşturulur. Bilim adamı cevaplayıcıların aynı süre içinde maddelerin %80'ine doğru yanıt vermelerini istiyorsa daha kolay olan bazı maddeleri testten çıkarır. Böylece test maddelerinin belirli bir sürede cevaplandırılma oranı dikkate alınarak katılımcıların %70 veya %60'ı başarılı olacak şekilde oluşturulur. Bir testteki maddelerin yüzde kaçının "asgarî yetenek düzeyine sahip" kişiler tarafından doğru cevaplandırılması gerektiği *konu içeriği uzmanları* tarafından belirlenir. Ancak burada bazı güçlükler söz konusudur ve bu güçlüklerin yenilmesi gerekir. Örneğin, incelemeye alınan bir test maddesi için bir uzman %90 kriterini getirmişken diğeri %70 kriteri üzerinde ısrar ediyorsa bilim adamı uzmanların görüşlerini birleştirmek için değişik formüllerden yararlanma yoluna başvurmalıdır. Fakat her halde belirlenen kesim puanının "akla yatkın" ve makul bir performans için "beklentilere uygun" olması gerekir. Uzmanların verdikleri puanları birleştirmede daha çok Angoff, Ebel ve Nedelsky yöntemleri uygulanır.

Diğer Kriter Puanlarının Üzerine Ekleme Yapma

Bu yöntemde kriter puan, haricî bir başarı faktörü, dış başarı puanı temel alınarak veya o puanın üzerine ekleme yapılarak belirlenir. Örneğin, ölçüm yapılan alana göre hangisi temel alınıyorsa piyasadaki belirli odalara bağlı sertifikalı kuaförler, berberler, kozmetikçiler veya diş teknisyenleri belirlenerek bu kişiler üzerinde test uygulanır. Söz konusu kişiler kendi mesleklerinde odanın belirlediği kritere göre "yeterli" olan bireylerdir. Bu kişilere KRT uygulanmasıyla elde edilecek ortalama değer veya ortalama değerden ,05 veya 1 standart sapma kadar daha yüksek bir değer kriter puan olarak

belirlenir.[16] Ancak bu yöntem odanın belirlediği kriter puanı ile kriter referanslı test puanlarının kesim noktasının farklı çıkabilme sakıncasına sahiptir. Örneğin, KRT ile elde edilecek ortalama değer odanın belirlediği kriter değerden daha düşük çıkabilir.

Madde Merkezli Değerlendirme

Literatürde Angoff (1971), Ebel ve Nedelsky (1954) modelleri madde merkezli değerlendirme yöntemi olarak adlandırılmıştır. Bu yöntemlerde *konu içeriği uzmanlarının* maddelere ilişkin yargıları toplanarak maddeler söz konusu yargılar çerçevesinde değerlendirilir. Madde merkezli değerlendirmede önemli olan, konu içeriği uzmanlarının kendi deneyimlerine ve bilgilerine dayanarak yaptıkları zorluk değerlendirmesidir. Aşağıdaki paragraflarda madde merkezli kriter puan belirleme yöntemlerinden Angoff, değiştirilmiş Angoff, genişletilmiş Angoff, Nedelsky, Ebel ve ayraç modelleri üzerinde durulmuştur.

Angoff modeli. Model, ilk kez William Angoff (1971) tarafından kendisinin hazırlamış olduğu *Educational Measurement* adlı kitabında tanıtılmış olması nedeniyle bu isimle anılır. Angoff yönteminde konu içeriği uzmanlarına *minimum yeterlilik düzeyine* sahip kişileri düşünmeleri ve daha sonra bu kişilerin söz konusu maddeyi hangi oranda cevaplandırabileceklerini soru formunun üzerine yazmaları söylenir.[a] Zihinsel ve fiziksel yetenek testlerinde ise KİU'lar testi önce kendileri alırlar ve bu süreç içinde asgarî yetenek düzeyine sahip kişilerin her bir maddeye yüzde kaçının doğru yanıt verebileceğini tahmin etmeye çalışırlar. Araştırmacılara *geçti-kaldı* değerlendirmesine dayanan Angoff yöntemi için aşağıdaki prosedürü uygulamalarını öneririz:[17]

---

[a] Panele katılan içerik uzmanlarından bazıları bu süreç içinde kendilerini rahatsız eden *minimum yetkinlik* düzeyi kavramı yerine "vasat başarıya sahip" kişilerin veya öğrencilerin göz önünde bulundurulması teklifini getirebilirler. Ancak bu yaklaşım doğru değildir. Angoff yönteminde minimum yetkinlik, başarısızlığın üstündeki en düşük yeterliliği tanımlar. Bu grubun içinde vasat öğrenciler veya kişiler de bulunur. Minimum yetkinlik, geçmeyi veya başarılı sayılmayı belirleyen öğrenme/yetkinlik düzeyidir. Bu düzeye sahip olmayan kişiler başarısız olarak değerlendirilirler. Genelde panelistler arasında orta/vasat yetenek düzeyindeki kişileri düşünerek yüksek geçme oranı belirleme gibi bir eğilim vardır. Bunun için başarılı sayılabilecek en düşük yeterliliğe sahip öğrencilerin göz önünde bulundurulması ve geçme oranının buna göre daha düşük değerlerde saptanması kendilerine hatırlatılmalıdır. Sadece derecelendirilmiş yetkinlik değerlendirmesinde *vasat, iyi, çok iyi* gibi gruplar için farklı yüzde değerleri saptanır.

1. Testin kapsadığı konuya hakim olan ve öğrencileri (işi ve çalışanları) tanıyan en az beş hakem (uzman) bulunuz. Hakemler asgarî yetenek düzeyine sahip olan öğrencileri / kişileri iyi tanıyan ve onları doğru bir şekilde değerlendirebilecek bireyler olmalıdırlar.
2. Hakemleri Angoff yönteminin uygulanma biçimi konusunda bilgilendirip eğitiniz. Değerlendirme sürecini kendilerine tanıtınız. Bu işlemin kaç gün süreceği, değerlendirmenin kaç oturum halinde yapılacağı hakkında kendilerine bilgi veriniz.
3. Hakemlere asgarî yetenek düzeyine sahip kişilerin/öğrencilerin kim olduğu konusunda bilgi veriniz ve bu konuda ortak bir düşünceye sahip olmalarını sağlayınız. Gerekiyorsa bunun için bir iki sayfalık bir bilgilendirme broşürü hazırlayınız.
4. Hakemlere her bir soru için asgari kabiliyete / yetenek düzeyine sahip öğrencilerin en az yüzde kaçının doğru yanıt vermesi gerektiğini sorunuz ve gerçekçi bir tahminde bulunmalarını isteyiniz. Bunun için her bir sorunun yanına birincisi ilk değerlendirme ve ikincisi ise düzeltilmiş değerlendirme için kullanılmak üzere iki cevaplama kutucuğu koyunuz.
5. Hakemlere soruları beşerli veya duruma göre onarlı gruplar halinde değerletiniz. (Değişiklik yapılmış Angoff modelinde ikinci bir raunt daha düzenlenir. İkinci rauntta hakemlerin bir test maddesine verdiği puanlar ile panel grubunun puanları arasında ,20'den fazla farklılık varsa puanlamayı kendi aralarında tartışmaları ve ortak bir karara varmaları söylenir. Tartışma sonucunda ortaya çıkan düzeltilmiş puanlar yeniden yazılır. *bk.*, Tablo 10-3).
6. Düzeltilmiş puanları temel alarak hakem puanlarının ortalamasını alınız ve standart sapma değerini belirleyiniz. Böylece hakem grubuna ait her bir maddenin asgarî yetenek düzeyine sahip bir kişi tarafından doğru yanıtlama oranını bulunuz.
7. Maddelerin doğru yanıtlama oranlarını toplayınız ve madde sayısına bölerek test sorularına yüzde kaç oranında doğru yanıt verilmesi gerektiğini belirleyiniz.

Bu yöntemle bir öğrencinin veya testi alan kişinin testten başarılı olabilmek için 100 üzerinden en az kaç puan alması gerektiği, soruların yüzde kaçını doğru yanıtladığına bakılarak belirlenir. Modelde, *minimum yeterlilik düzeyine* sahip öğrencilerin başarı oranlarından hareket edilerek asgarî

test başarı oranına ulaşılır. Angoff modelinin hesaplanmasını Eşitlik 10-1 ve Eşitlik 10-2'deki formüllerle gösterebiliriz.

$$\text{Maddenin yanıtlanma oranı tahmini} = \frac{\sum P_t}{N_h} \quad (10\text{-}1)$$

$$\text{Kesim puanı} = \frac{M_o}{N_m} \quad (10\text{-}2)$$

P_t = Hakemlerin tahmin ettikleri oranlar.
N_h = Hakem sayısı.
M_o = Madde oranlarının toplamı.
N_m = Testteki madde sayısı.

Angoff yönteminin avantajı hesaplanmasının kolay olması ve test sorularının zorluk ve kolaylık derecesine göre kesim puanının değişebilmesidir. Angoff modelinin daha sonraki yıllarda değişik versiyonları ortaya çıkmış, bilim adamları modele uygulamayı zenginleştiren değişik ögeler eklemişlerdir. Öte yandan, model her bir yöntemde farklı sonuçlar vermesi nedeniyle eleştirilmiş ve hakemlerin asgarî yetenek düzeyine sahip aday[a] kavramını farklı değerlendirebildikleri olgusuna dikkat çekilmiştir. Modelin bir diğer olumsuz yönü zaman alması, maliyetli olması ve uzmanlarda değerlendirme yaparken yorgunluğa neden olmasıdır. Bu eleştirilere karşın Mehrens (1995) Angoff yöntemini standart belirlemede makul bir yaklaşıma sahip olması, kullanımının kolaylığı ve standardın psikometrik özelliklere sahip olması nedeniyle tavsiye etmeye devam etmiştir (aktaran Kinesmetric böl.).[18] Öte yandan AERA, APA ve NCME tarafından geliştirilen *Test Standartları* kitabında Angoff yönteminin kişisel yargılar nedeniyle esaslı bir şekilde yanlı sonuçlar verdiği düşüncesi ve iddiasına yer verilmemiştir. Belirli bazı yetersizliklerine karşın yöntem literatürde yoğun uygulama alanı bulmuştur.[19]

Değişiklik yapılmış Angoff modeli. Değişiklik yapılmış Angoff modeli Amerikan Yüksek Mahkemesinin *ABD ve South Carolina* davası sonu-

[a] Minimum yetkinlik düzeyi okullarda çoğunlukla "C" ve "B" puanı alan ve ara sıra "A" puanı alan öğrenciler olarak düşünülür. İlköğretim okullarında 2 puan alan öğrencilerdir. Bu kişiler çoğunlukla 2 bazen 3 ve çok az dersten ise 4 puan almışlardır.

cunda vermiş olduğu karara bağlı olarak geliştirilmiştir. Büyük ölçekli değerlendirme ve test çalışmalarında kullanılan bu yaklaşımda orijinal Angoff modelinden bazı farklılıklar söz konusudur. Orijinal Angoff modelinde tek raunt değerlendirme yapılırken değiştirilmiş Angoff yönteminde değerlendirme iki veya daha fazla raunt içinde gerçekleştirilir. Değişiklik yapılmış modelde; panelistler bir arada bulundukları eğitim uygulamasından sonra gerçekleşen ilk rauntta yargılarını grup tartışması yapmadan bağımsız olarak verirler ve daha sonra ikinci rauntta bir araya gelerek panelin değerlendirmesiyle kendi değerlendirmelerinin karşılaştırmasını yaparlar.[20] Bu Karşılaştırma ve değerlendirme sonucunda başlangıçta belirledikleri oranları revize ederek düzeltirler. Değişiklik yapılmaya izin verilmesi nedeniyle bu yönteme *değişiklik yapılmış Angoff modeli* adı verilmiştir.

Tablo 10-3. Birinci Raunt Sonunda Elde Edilen Sonuçların Hakemler Tarafından Gözden Geçirilmesi

Maddeler	İlk değerlendirmenizdeki *p* değerleri	Panel değerlendirme sonuçları			Gözden geçirilmiş *p* değerleri
		ortalama	Standart sapma	medyan	
1	,65	,65	,06	,65	,65
2	,68	,70	,07	,65	,70
3	,70	,65	,10	,63	,65
4	,65	,65	,12	,64	,65
5	,75	,70	,05	,70	,70
6	,70	,65	,06	,62	,65
7	,70	,65	,04	,65	,65
8	,65	,65	,08	,62	,65

Bu yöntemin bir diğer farklılığı test bankasından alınan sorularının değerlendirilmesinde soruların güçlük düzeyinin göz önünde bulundurulmasıdır. Güçlük düzeyinin dikkate alınması test sonuçlarının güvenilirliğini büyük ölçüde arttırır. Her testin geçme sınır değeri güçlük derecesi dikkate alınarak belirlenir. Testin farklı versiyonları öğrencilere uygulanmış olsa bile güçlük derecesi dikkate alındığından test sonuçları daha adilâne değerlendirilmiş olur.[21] Orijinal modelde uzmanlar bir maddenin doğru yanıtlanma oranı hakkında 0 ilâ 1,00 arasında bir değer verebilirlerken değişik-

lik yapılmış modelde uzmanlar doğru yanıtlama oranı hakkında tam olarak serbest bırakılmamışlar ,05 ilâ ,95 arasındaki (,5; ,20; ,40; ,60; ,75; ,90; ,95) yedi kategoriden birini seçmek zorunda bırakılmışlardır. Ayrıca maddenin yanıtlanma oranı hakkında tahmin yürütemedikleri durumda hakemlerin "Bilmiyorum" şıkkını işaretlemelerine izin verilmiştir.[22] Böylece hakemler toplam sekiz kategori üzerinden işaretleme yapabilmektedirler.

Bu modelde kesim puanı daha sonra orijinal Angoff yönteminde olduğu gibi hesaplanır. Mahkeme kararına dayanan değişiklik yapılmış Angoff modelinde standart veya orijinal uygulamaya göre puanlar ölçümün iki üç standart hatası kadar daha düşüktür. Buradaki amaç puanların bir ölçüde düşük tutulması ve kazanan kişi sayısının artırılmasıdır. Değişiklik yapılmış Angoff yönteminde bir dizi istatistiksel değerler ve insan faktörü göz önünde bulundurulur ve söz konusu faktörler aşağıdaki gibidir:[23]

1. Ölçümün standart hatasının büyüklüğü.
2. Gerçekten kaliteli bir adayı elemenin kalifiye olmayan bir adayı almaya göre taşıdığı risk veya hata yapmanın ne ölçüde önemli zararlara yol açacağı.
3. Angoff paneline katılan hakemlerin değerlendirmelerinde ne ölçüde tutarlı oldukları.
4. Boş pozisyonlara yönelik başvuru taleplerinin yoğunluğu.
5. Söz konusu işte çalışanların cinsiyet ve ırk dağılımları.

Bu yöntemde Amerikan mahkemesi saptanan Angoff değerinden ölçümün bir, iki veya üç standart hatası kadar azaltılmaya gidilmesini minimum geçme puanı olarak belirlemiştir. Ancak uygulamada, yüksek kalite standardını tutturmak için hakemler ortalaması, azaltmak yerine ölçümün 2 standart hatası kadar artırılarak geçme puanı belirlenmektedir. Bu konuda standart bir uygulama bulunmadığından araştırmacı kendi yaklaşımını içinde bulunduğu şartları dikkate alarak kendisi belirleyecektir.

Değişiklik yapılmış Angoff yönteminde panelistler çalışmalara *minimum yeterlilik düzeyi* kavramını tanımlayarak başlarlar. Okullarda başarı düzeyi genellikle iki veya üç düzey olarak belirlenir. İki düzey geçme-kalma durumunu belirler. Üç düzey ise derecelendirilmiş bir değerlendirmedir.[a] Bu değerlendirmede dereceler; (1) standartları karşılamıyor, (2)

[a] Derecelendirme yaklaşımları farklı olabilir. Örneğin Amerika'da Eğitimsel Gelişmenin Ulusal Değerlendirmesi örgütü (National Assessment of Educational Progress – NAEP)

yeterli ve (3) üstün başlıkları altında toplanır. Uzmanlar daha sonra testi incelerler ve hatta gerekiyorsa testi kendilerine uygularlar. Uzmanlar birinci rauntta her bir test maddesini inceleyerek maddenin güçlük düzeyini belirlerler. Böylece test maddesini minimum yeteneğe sahip kişilerin yüzde kaçının cevaplandırabileceği saptanmış olur. Bu aşamada eğer derecelendirilmiş kriter yaklaşımı temel alınmışsa uzmanlar test maddesini yeterli öğrencilerin yüzde kaçının doğru yanıtlayacağını ve üstün öğrencilerin yüzde kaçının doğru yanıtlayacağını belirler. İkinci rauntta panelistler birinci raunttaki yargılarını kendi aralarında tartışırlar ve gerektiğinde yaptıkları değerlendirmelerde düzeltme yoluna başvururlar. Bu aşamada panelistler öğrencilerin neyi bilip bilmediklerini değil neyi bilmeleri gerektiği üzerinde dururlar. Bu tartışma sonunda panelistlerden ikinci bir değerlendirme daha yapmaları istenir. Birinci raunttan sonra araştırmacılar farklı yöntemleri deneyebilirler. Bazı uygulamalarda birinci raunttan sonra belirli sayıda öğrenciye pilot araştırması yapılması ve bu araştırmanın sonunda elde edilen *p* değerlerinin hakemlere verilmesi yöntemine başvurulur. Hakemler bu değerlere bakarak kendi *p* değerlerini değiştirebilirler veya aynı kalmasında ısrar edebilirler. Üçüncü rauntta panelistler fikir birliğine ulaşmaya veya görüşlerini birleştirmeye çalışırlar. Bu aşamada kişilerin fiilî başarı oranları panelistlerin değerlendirmesine sunulur. Değişiklik yapılmış Angoff modelinin son adımı her bir test maddesinin yargılara dayalı *p-değerlerinin* toplamını almak ve panelist sayısına bölmektir. Panelistler arasında uyuşmazlık varsa bu durum ortalama alma veya çoğunluğun görüşünü kabul etme yöntemiyle çözülür.

Normal dağılım eğrisi dışındaki diğer yöntemlerde kesim puanını belirlemek için 7 ilâ 10 arasında uzman, hakem veya sertifikalı kişi belirlenir. Bu kişiler duruma göre üç ilâ beş gün arasında bir araya gelerek test maddelerini tek tek inceleyerek değerlendirmeye tâbi tutarlar ve söz konusu maddeleri alt düzeyde yeterliliğe sahip bireylerin cevaplandırma oranı hakkında tahminde bulunurlar. Angoff yönteminde maddelerin cevaplandırma oranlarını belirleyecek konu içeriği uzmanlarının kaç kişi olması gerektiği konusunda ABD'de genellikle mahkeme kararları temel alınmıştır. Bu uzmanlar sadece yanıtlama oranını değil aynı zamanda eğer test işle ilgiliyse maddelerin işle ilgili olma durumunu da belirlerler.

Angoff yönteminde konu içeriği uzmanlarının (KİU) sayısının ne olması gerektiği konusunda değişik görüşler ileri sürülmüştür. Mehrens ve Popham, KİU sayısının 20-25 olması gerektiğini söylerken Livingston ve

---

öğrencileri yetkinlik açısından dört grupta sınıflandırmıştır: *temelin altında*, *temel*, *yetkin*, *ileri*. Bu sınıflandırma biçimini kendi değerlendirme yaklaşımımıza benzeterek *alt-orta*, *orta*, *iyi* ve *çok iyi* öğrenciler başlıkları altında toplayabiliriz.

Zieky (1982) beş kişilik bir grubun dahi yeterli olabileceğini belirtmişlerdir (aktaran Hurtz ve Hertz).[24] Norcini ve arkadaşları ise KİU sayısının 5'ten 10'a kadar artmasının verilerin güvenilirliğini büyük ölçüde arttırmadığını bulmuşlardır (aktaran Hurtz ve Hertz).[25] Bu konuda yapılan benzeri diğer araştırmaların sonucundan beş uzmandan yararlanma fikrinin bazı ölçümlerde aşırı iyimser karşılanabilecek iken, pek çok durumda 20-25 konu içeriği uzmanından yararlanmanın daha sağlıklı sonuçlar vereceği anlaşılmaktadır. Okullarda uygulanan bilgi testleri için genellikle 20 civarında KİU'dan yararlanılır.

Değişiklik yapılmış Angoff modeli bir çok açıdan eleştirilmiştir. Uzman veya panelist sayısının 10 veya 20 gibi belirli bir sayıyla sınırlı tutulmasının bütün uzmanları temsil etmeyeceği ve bu uzmanların yapmış oldukları değerlendirmelerin istatistiksel açıdan anlamlı olmayabileceği iddia edilmiştir. Yönteme getirilen bir diğer eleştiri, yaklaşımın değerlendirme için 2-3 gün gibi bir zaman ayrılmasına neden olması ve yoğun emek gerektirmesidir. Ayrıca yönteme doğru tahmin değerleri elde etme konusunda da değişik eleştiriler getirilmiştir.

Genişletilmiş Angoff modeli. R.K. Hambleton ve B.S. Plake (1995) tarafından önerilen bu modelde, konu içeriği uzmanlarına en düşük yeterlilikteki bireylerin yanıtlama oranını tahmin etmek yerine en düşük yeterlilikteki bir bireyin söz konusu maddeyi yanıtlayıp yanıtlayamayacağı sorulur.[26] Hakemlerin yanıtlayabilir şeklindeki cevapları 1 yanıtlayamaz şeklindeki cevapları 0 olarak kodlanır. Daha sonra hakemlerin ortalaması alınarak söz konusu maddenin yanıtlama oranı saptanır (*bk*, Tablo 10-4).

Tablo 10-4. Değiştirilmiş ve Genişletilmiş Angoff Yönteminde Başarı Standartlarının Saptanması.

	Hakemler					Ort.
	1	2	3	4	5	
Angoff yöntemi (yüzde)	65	70	65	75	75	,70
Genişletilmiş Angoff (beklenen puan)	1	1	0	1	1	,80

Angoff modelinin güvenilirliği. Angoff yöntemiyle saptanan kesim puanının güvenilirliğini test etmek için paneldeki uzmanların vermiş oldukları puanların ortalaması fiilî test sonuçlarıyla karşılaştırılır. Maddele-

rin güçlük puanlarıyla panelin vermiş oldukları puanlar arasındaki korelasyonun yüksek olması Angoff kesim puanının güvenilir olduğunu gösterir. Buna göre panelde yer alan konu içeriği uzmanlarının ölçüm yapılan grubu çok iyi tanıdıkları ve iyi bir tahminde bulundukları yorumu yapılır.

Nedelsky yöntemi. Leo Nedelsky (1954) tarafından geliştirilmesi nedeniyle yönteme kısaca *Nedelsky modeli* adı verilmiştir. Bu yöntemin temelinde yatan felsefe, madde seçeneklerinin güçlük derecesidir. Bu nedenle Nedelsky, test alan kişinin performansının soruya ait yanlış şıkların (çeldiricilerin) doğru şıkka yakın veya uzak olmasına bağlı olduğunu söylemiştir. Yanlış şıklar eğer doğru şıkka çok yakın ise test maddesi zor, yanlış şıklar doğru şıktan oldukça uzak ise test maddesi o denli kolay olarak algılanır. Yöntemde çeldiricilerin doğru şıkka ne ölçüde yakın olup olmadığı değerlendirilir.

Nedelsky yönteminde konu içeriği uzmanının görevi, çoktan seçmeli sorulardan oluşan her bir test maddesini çeldiriciler açısından incelemektir. Minimum yeterlilik düzeyine sahip bir öğrencinin söz konusu çeldiricileri ayırt etmesi ve doğru yanıtı bulması konusu üzerinde odaklanılır. Bu yöntem, soru maddesi derecelendirilmiş çeldirici özelliğine sahip olduğunda daha iyi çalışır. Nedelsky prosedürünün değiştirilmiş farklı bir versiyonu konu içeriği uzmanlarına hangi maddelerin kesim puanının üzerinde kalan öğrenciler tarafından cevaplandırılabileceğini sormaktır.[27]

Nedelsky tekniğinde çeldiricilerin fonksiyonu dikkate alınarak test maddeleri bir dereceleme ölçeği üzerinde puanlandırılır. Ölçeğin dereceleri şu şekilde belirlenmiştir:

1. İki puan = En iyi yanıt veya doğru yanıt.
2. Bir puan = Akla yatkın inandırıcı yanıt. Kişi doğru yanıtla bu yanıt arasında tercih yapma konusunda zorluk çekebilir.
3. Sıfır puan = İnandırıcı olmayan yanıt. Kişi kolaylıkla bu yanıtın yanlış olduğunu görebilir.

Bu ölçek sadece standart düzlemlerde uygulanır. Puanlama her zamanki gibi ikili sisteme bağlı olarak yapılır. Çeldiricinin seçilmesi halinde sıfır, doğru yanıtın seçilmesi halinde 1 puanı verilir. Daha sonra dereceleme ölçeğine bağlı olarak Eşitlik 10-3'teki formüle göre minimum geçme indeksi (MGİ) adı verilen değer hesaplanır.[28]

- Minimum geçme indeksi formülü.

$$MGİ = \frac{2}{\text{Ağırlıkların toplamı}} - \frac{1}{5\ (\text{Ağırlıkların toplamı})} . \qquad (10\text{-}3)$$

Bir soruda, birinci şık inandırıcı yanıt, ikinci şık doğru yanıt ve üçüncü, dördüncü şıklar inandırıcı olmayan yanıtlar olsa minimum geçme indeks değeri Eşitlik 10-4 ve 10-5'teki gibi hesaplanır.

$$MGİ = \frac{2}{1+2+0+0} - \frac{1}{5\ (1+2+0+0)} . \qquad (10\text{-}4)$$

$$MGİ = ,67 - ,07 = ,60 . \qquad (10\text{-}5)$$

İki şıkkın elenmesiyle geriye kalan diğer iki şıkkın doğru olma olasılığı %50'dir. Bu yöntemde test soruları zor olduğu ölçüde geçme-kalma sınır puanı düşük tutulurken kolay olduğu ölçüde ise yüksek belirlenir. Nedelsky yönteminde keyfî bir sınır puanı veya cevaplama yüzdesi belirlenmemiştir. Sınır puanı yıldan yıla ve soruların güçlük derecesine göre değişir. Bu açıdan sınır puanının uygulamada %60 ilâ %70 arasında değiştiği görülmüştür.

Yöntemin uygulanmasında uzmanlara şu yönde bir talimat verilir: Her bir test maddesini okuyarak cevap şıklarından öğrencinin/adayın yanlışı kolaylıkla ayırt edebilecekleri şıkların üstüne bir çizgi çizin ve geriye kalan maddeleri şık sayısına göre oranlayınız. Örneğin eğer sorunun 5 şıkkı varsa ve sorulardan iki tanesinin üzeri çizilmişse kolay yanıtlama oranı 2/5'tir (*bk.*, Tablo 10-5). Madde içerik uzmanları bu şekilde tüm maddeleri işaretledikten sonra kendi aralarında küçük bir müzakere yaparak kullandıkları değerlendirme kriterlerini tartışırlar. Bu süre içinde aynı zamanda geçici olarak kullanılan sabit *k* değeri konusuna (yukarıdaki formülde 5 olarak belirlenmişti) karar verirler.

Bu yöntemin uygulanmasında iki konuya dikkat edilir. Bunlardan birincisi, uzmanlara doğru yanıtın söylenip söylenmeyeceğidir. Bazıları uzmanların doğru yanıtı kendilerinin bilmeleri gerektiğini ifade ederken, diğerleri bu uygulamadaki amacın uzmanları test etmek olmadığını sadece soruyu doğru bir şekilde cevaplandıracak öğrenci oranını tespit etmek olduğunu söylerler. Dikkat edilmesi gereken ikinci konu uzmanların bu değerlendirmeyi bir grup olarak mı yoksa bireysel olarak mı yapacaklarıdır. Değerlendirme eğer grup olarak yapılacaksa uzmanların oybirliği ile karar verip vermeyecekleri konusu ayrıca düşünülmeli ve bu konuda da belirli bir karar verilmelidir.[29]

Tablo 10-5. Nedelsky Yöntemi

Test maddeleri	İptal edilen seçenekler	Çıkarılmayan seçenek sayısı	Karşılılık	Beklenen puan
1	1xxxx	1	1/1	1,00
2	x2x4x	2	1/2	,50
3	123xx	3	1/3	,33
4	x234x	3	1/3	,33
5	xx2x5	2	1/2	,50
6	x2345	4	1/4	,25
7	1xxx5	2	1/2	,50
8	12xx5	3	1/4	,33
9	12x4x	3	1/4	,33
10	xx3xx	1	1/1	1,00
			Toplam	5,07

x = İçerik uzmanlarının oylarının çoğunluğuyla inandırıcı bulunmayan şıklar.

■ Nedelsky yöntemine göre kesim puanının hesaplanması.

$$\text{Kesim puanı} = \frac{5{,}07}{10} * 100\,, \qquad (10\text{-}6)$$

$$\text{Kesim puanı} = 50{,}7\,. \qquad (10\text{-}7)$$

Adayın 10 sorudan en az beşini doğru yanıtlaması halinde başarılı sayılması söz konusudur. Bu yöntemde sınav veya testin uygulanmasından sonra konu içeriği uzmanları geçerlilik çalışması yaparlar. Bu çalışmada adayların şıkları kendi belirledikleri gibi işaretleyip işaretlemedikleri incelenir ve varsa testten çıkarılması gereken sorular belirlenir.[30]

Ebel yöntemi. Ebel (1979) kriter referanslı testlerin de bir istisnasıyla diğer testler için yapılan güvenilirlik analizlerine tâbi tutulacağını bildirmiştir (aktaran, Paccioretti, 2003).[31] Ebel yönteminde hakemler her bir ifadeyle ilgili olarak iki konuda karar verirler. Bu kararlardan birincisi söz konusu maddenin ölçüm alanıyla veya işle ilgisinin ne olduğudur. Hakemler bu konudaki kararlarını dört seçenekten birine göre verirler: *gerekli, önemli, kabul edilebilir* ve *şüpheli.* İkinci karar konusu ise, asgarî yetenek düzeyine sahip bir kişi için test maddesinin güçlük seviyesinin ne olduğudur.

Hakemler değerlendirmelerini üç seçenek içinde yaparlar: *kolay, orta, zor.* Uzmanlar soruların yanıtlarını 3 x 4 büyüklüğündeki bir tablo üzerine yerleştirirler. Daha sonra 12 hücredeki her bir sorunun asgarî yetenek düzeyine sahip kişilerin cevaplama yüzdeleri belirlenir. Sınır puanı, her bir hücredeki soru sayısı doğru yanıt yüzdesiyle çarpılarak ve çarpım sonuçları toplanarak bulunur.[32]

Ayraç modeli. Değiştirilmiş Angoff modeline gelen eleştiriler nedeniyle 1996 yılında Dan Lewis ve CTT/McGraw-Hill test şirketi yetkilileri ayraç (bookmark) adı verilen yöntemi geliştirmişlerdir.[33] Ayraç modelinde, testteki her bir maddeyle ilgili olarak bilişsel yargılara dayanan apriori bir *p-değeri* belirlenmez. Tam tersine model öğrencilerin/kişilerin fiilî test sonuçlarına dayanır. Model ikili ve çok dereceli olarak puanlanan verileri başarılı bir şekilde ele alıp aynı madde-yanıt ölçeği üzerinde değerlendirme kapasitesine sahiptir.[34] Ayraç modeli deney temellidir ve bu nedenle yargılama işini basitleştirir. Çünkü bu yöntemde hakemlerden *en düşük yetenek düzeyine sahip kişiler* için nokta temelli bir tahminde bulunmaları istenmez. Tam tersine onlardan öğrencilerin / kişilerin fiilî başarı durumlarını göz önünde bulundurarak test maddeleri içinde onların yetenek düzeylerini aşan maddeleri işaretlemeleri istenir.

Yöntemde maddelerin güçlük derecesine uzmanlar değil öğrenciler / kişiler karar verirler. Ayraç modelinde öncelikle bir pilot araştırma yapılır. Bu araştırmanın sonucunda test maddeleri madde-yanıt kuramına göre test

edilir.[a] Madde-yanıt kuramı test sonuçları temel alınarak maddeler en kolaydan en zora doğru sıralanır ve sorular bir "kitapçık" haline getirilir. Konu içeriği uzmanları sınır çizgisinde yer alabilecek en düşük yeterliliğe sahip bir kişi veya öğrencinin hangi özelliklere sahip olduğunu tanımlarlar. Daha sonra hakemler veya konu içeriği uzmanları test kitapçığındaki maddeleri inceleyerek minimum yetenek düzeyine sahip öğrencilerden oluşan grubun hangi soruları cevaplandıramayacaklarını belirleyerek o maddelerin bulunduğu yere bir işaret / ayraç koyarlar. İlk işaretleme yapıldıktan sonra gerçek fiilî *p* değerleri hakemlere verilir ve bu değerlendirmeyi söz konusu bilgiyi göz önünde bulundurarak yeniden yapmaları istenir. Bunun üzerine hakemin yapacağı ikinci işaretleme aynı maddede veya başka bir madde üzerinde olabilir. Kesim puanı ikinci değerlendirmenin sonucuna göre belirlenir. Daha sonra her bir hakemin yapmış olduğu değerlendirmeler toplanarak puanların ortalaması alınır ve testin kesim puanı belirlenir.

Yöntemin avantajı; uzmanların tartışmalarına daha fazla imkan sağlaması, farklı soru formatlarını tek bir format altında birleştirmesidir. Olumsuz yönü ise gereğinden fazla zaman kaybettirmesi ve ampirik verilere ihtiyaç duyulmasıdır.

Karma uygulama veya sentezleme yöntemi. Bu uygulamada konu içeriği uzmanları 20 kişilik bir grup halinde oluşturularak daha sonra iki gruba bölünür ve bir grup değiştirilmiş Angoff yöntemini uygularken diğer grup ise Ebel, zıt gruplar veya ayraç modelini uygular ve sonuçlar karşılaştırılarak aralarında önemli bir farklılık bulunup bulunmadığına bakılır. Sonuçta değişik standart belirleme yöntemlerinin sonuçları birleştirilerek tek bir standart belirlenir.

İnceleyici Merkezli Değerlendirme

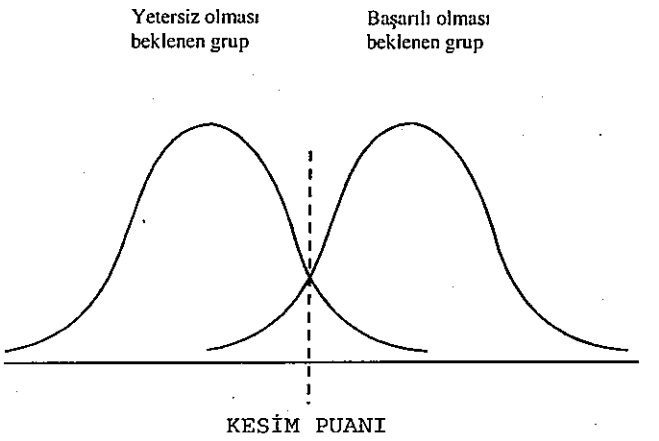
İnceleyici merkezli değerlendirmelerde odak noktası maddeler veya test değil, değerlendirmeyi yapan hakemler veya konu içeriği uzmanlarıdır. Bu yaklaşımın başlıca yöntemleri zıt gruplar, sınır çizgisi (borderline) yöntemi, genelleştirilmiş inceleyici merkezli (generalized examinee centered) değerlendirme yaklaşımlarıdır. Bu bölümde bunlardan sadece zıt gruplar yöntemi üzerinde durulmuştur.

---

[a] Maddelerin zorluk değerlendirmesi için klâsik test kuramının mı yoksa madde-yanıt kuramının mı temel alınacağı eğer MYK temel alınacaksa 1p, 2p ve 3p modellerinden hangisinin kullanılacağı, hangi yazılımdan yararlanılacağı araştırmacının çözmesi gereken teknik sorunlardır.

Zıt gruplar yöntemi. Zıt gruplar (borderline/contrasting groups) yönteminde test maddeleri pilot araştırma çerçevesinde ölçüm yapılacak kitleyi temsil eden bir örnekleme uygulanır. Örneklemdeki üyeler rasgele seçilirler. Konu içeriği uzmanları bu aşamada test puanları dışında diğer dış verileri (karne not ortalamasını, öğrencilerin devam ettikleri şubenin başarılı bir şube olması, farklı testlerden aldıkları başarı notları gibi faktörleri) göz önünde bulundurarak adayları yetenekleri açısından *yeterli* ve *yetersiz* şeklinde iki zıt veya farklı gruba bölerler. Söz konusu bölme "beklenen" yeterliliği ve/veya yetersizliği tanımlar (*bk.*, Şekil 10-2).

Daha sonra adayların pilot araştırmadaki test puanları belirlenir. Test puanları ise "fiilî" yeterliliği gösterir. Fiilî test puanları beş puan aralıklarla en büyükten en küçüğe doğru sıralamaya sokulur. Her bir puan diliminde yeterli ve yetersiz gruptaki kişi sayısı belirlenir. Yetersiz olması beklenen gruptaki kişilerin sayısıyla yeterli olması beklenen gruptaki kişilerin sayısının aşağı-yukarı birbirine denk geldiği nokta kesim puanı olarak saptanır (*bk.*, Tablo 10-6; ayrıca *bk.*, Şekil 10-3). Bu yöntemin olumsuz yönü, kişilerin yeterli ve yetersiz şeklindeki zıt gruplar halinde sınıflandırılmasında öznel yargıların temel alınıyor olmasıdır. Sınıflandırmanın iş veya başarıyla olan bağlantısı dolaylıdır.



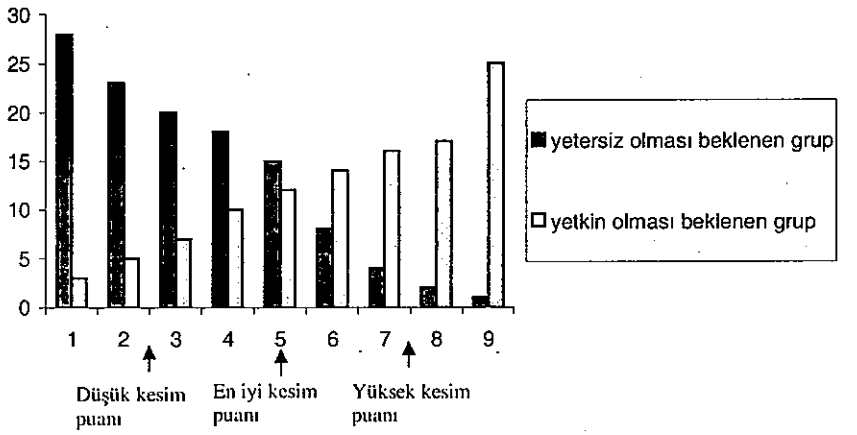
Şekil 10-2. Yeterli olması beklenen grupla yetersiz kalması beklenen grubun anlamlı bir biçimde kesişme noktasının kesim puanı olarak belirlenmesi.

Kaynak. C. van der Vleuten, "Borderline Approaches to Performance Assessment [Başarı Değerlendirme Uygulamalarında Sınır Çizgisi Yaklaşımı]," AMEE Konferansı, 29 Ağustos - 1 Eylül 2002, Lizbon.

Tablo 10-6. Zıt Gruplar Yöntemi

Puan aralığı	Adayların sayısı		Yeterli adayların oranı
	Yeterli	Yetersiz	
46-50	5	0	1,00
41-45	14	1	,93
36-40	25	7	,78
31-35	22	10	,69
26-30	17	12	,59
21-25	4	12	,25
16-20	3	14	,18
11-15	2	15	,12
6-10	1	18	,05
0-05	0	3	0

Tablo 10-6 incelendiğinde zıt grupların buluşma noktasının ,59 oranı olduğu görülür. Bu oran yanlış sınıflandırmanın en düşük olduğu noktayı temsil eder.



Şekil 10-3. Zıt gruplar yönteminde kesim puanının belirlenmesi.

Analitik Yöntemler

Analitik yöntemler başlıca iki grupta değerlendirilir: Birincisi kümeleme analizi yöntemidir. Bu yaklaşımda test alıcılarının ne yapmaları gerektiği üzerinde değil, ne yapabilecekleri konusu üzerinde odaklanılır.[35] İkincisi

ise madde-yanıt kuramına dayanan madde haritası modelidir. Bu kitapta sadece madde haritası modeli üzerinde durulmuştur.

Madde haritası modeli. Madde haritası modeli, yeterlilik düzeyinin belirlenmesinde grafik araçlarını kullanan bir yaklaşımdır. Bunun için test maddeleri pilot araştırma yapılarak belirli bir gruba uygulanır. Her bir maddenin madde-yanıt kuramına göre güçlük *p* değerleri belirlenir ve daha sonra maddeler güçlük derecelerine göre "madde haritası" adı verilen bir grafik üzerinde sergilenir. Madde içeriği uzmanları grafiği inceleyerek en düşük yeterliliğe sahip kişilerin bu maddelere doğru cevap verirken en azından %50 şansa sahip olup olmadıklarını değerlendirirler. Eğer adayın yeteneği maddenin güçlük derecesine eşitse o zaman maddeye doğru yanıt verme ihtimali veya şansı %50 demektir. Yöntemin güçlü yönleri, istatistiksel bir temele sahip olması, sınır çizgisindeki adayın portresini net bir şekilde ortaya koyması ve araştırmacıya çok fazla zaman kaybettirmemesidir. Zayıf yönü ise bir ölçüde karmaşık olan Rasch modelinin temel alınması ve hesaplamada ampirik verilere ihtiyaç bulunmasıdır.[36]

Normal Karşılanan Kabul Edilebilir Yeterlilik Düzeyi Beklentisi

ABD'de *Personel Seçim Prosedürü Kılavuzu*'nda (Uniform Guidelines on Employee Selection Procedures) test sonucunda belirlenecek kesim puanının *normal karşılanacak kabul edilebilir yeterlilik düzeyine* göre belirlenmesi önerilmiştir. Buradaki sorun normal, kabul edilebilir yeterlilik düzeyinin nasıl belirleneceğidir. Bunun için farklılıkların standart hatası değerinden (FSH) yararlanılabilir (*bk.*, Eşitlik 10-8; 10-9).

- Farklılıkların standart hatası = $ÖSH * \sqrt{2}$ (10-8)

- Ölçümün Standart hatası = $SS\sqrt{1-r}$ (10-9)

SS = Hakem puanlarının standart sapması.
r = Hakem puanlarının güvenilirlik katsayısı.

Farklılıkların standart hatası bize hakemler tarafından belirlenen ortalama kesim puanının güven aralığını verir. Belirli bir puan düzeyinden ne kadar uzaklaşabileceğimizi belirlemek ve farklı bir düzeyi ölçüp ölçmediğimizi saptamak için puanları 1, 2 veya 3 standart hata değeri çerçevesinde analiz ederiz.[37] Bir standart hata farklılıkların %68'ini, 2 standart hata %95'ini ve 3 standart hata ise farklılıkların %99'unu açıklar. İki puanın

birbirinden önemli ölçüde farklı olup olmadığı FSH değerine bakılarak belirlenir.

Kesim Puanlarının Kullanılmasından Kaçınılması Gereken Yerler

Kesim puanları sadece yeterliliği açık bir şekilde ortaya çıkardığı durumlarda kullanılmalıdır. Deneysel kanıtlar eğer kesim puanının altında kalan kişilerin büyük çoğunluğunun başarısız olduğunu ortaya koymuyorsa kesim puanı uygulamasına başvurulmamalıdır. Bunun yanında kesim puanları cinsiyet ve yaş gruplarında ayrıca etnik gruplarda sistematik bir şekilde bazı kesimleri dışarıda bırakıyor olmamalıdır.[38] Böyle bir durumda kriter referanslı testlerin "olumsuz etki" sonucuna yol açtığından söz edilir. İşletmelerde personel seçim kararlarında sadece kriter referanslı test sonuçları değil; mülakat, referans, okul başarısı ve iş müracaat formlarındaki veriler de dikkate alınır.

Kesim Puanı Modelleri ve Güvenilirlik

Seçilen kesim puanı modelleriyle güvenilirlik katsayıları arasında belirli ilişkiler olduğu saptanmıştır. Testin orta güçlükte olması veya daha zor olması gibi durumlarda kriter puan modelleri farklı sonuçlar verebilir. Orta güçlükteki maddelere sahip kriter referanslı testlerde ,05 anlamlılık düzeyinde Nedelsky prosedürünün en yüksek güvenilirlik katsayısını verdiği bulunmuştur. Oldukça zor test maddelerinden oluşan kriter referanslı ölçümlerde ise değişik eşik değer modelleri arasında ve ,05 anlamlılık düzeyinde Sheehan ve Davis prosedüründen[a] en yüksek güvenilirlik katsayısı elde edilmiştir.[39] Bu sonuçlar test maddelerinin zorluk derecelerinin ve kriter olarak seçilen eşik değer modelinin güvenilirlik katsayıları üzerinde etkili olduğunu göstermektedir.

KRİTER REFERANSLI TEST PUANLARIN GÜVENİLİRLİĞİ

Kriter referanslı test puanlarının güvenilirliğinde araştırmacı, kişilerin belirlenen kesim/kriter puanına göre doğru bir şekilde sınıflandırılıp sınıflandırılmadığıyla ilgilenir. Kriter referanslı testlerde norm referanslı testlerde gördüğümüz her tür güvenilirlik analizini yapmak doğru değildir. Örneğin, bu tür testlerde Cronbach alfa güvenilirlik analizi yapılmaz.[40] Test soruları yeterliliği belirlemeye yönelik çoktan seçmeli bilgi testi niteliğinde ise ve araştırmacı maddelerin güçlük derecelerinin eşit olduğunu düşünüyorsa bu

[a] Bu prosedür kitapta tanıtılmamıştır.

tür testlerde iç tutarlılık açısından KR-21, güçlük derecelerinin eşit olmadığını düşünüyorsa KR-20 güvenilirlik analizlerini kullanabilir. Fakat asıl ilgi odağı test sorularının iç tutarlılığı değil, kişilerin doğru bir şekilde ayrıştırılma durumudur.

Kriter referanslı testlerde çoğunlukla; (a) test-yeniden test, (b) paralel formlar güvenilirlik analizi, (c) gözlemci içi ve (ç) gözlemciler arası değerlendirme güvenilirliği yöntemleri uygulanır. Kriter referanslı testlerde yüksek iç tutarlılık katsayılarına ulaşılmaya çalışılmaz. Önemli olan, mümkün olduğu kadar daha fazla kişinin testte başarılı olması ve aynı zamanda insanların doğru bir şekilde sınıflandırılmasıdır. Bu açıdan güvenilirlik, "sınıflandırma kararı" sonucunda ortaya çıkan portrenin doğuluk derecesidir.[41] Kriter referanslı testlerden elde edilen puanların güvenilirliği için değişik istatistiksel analizlerden yararlanılır ve bunlar aşağıdaki gibidir:

1. Uyuşma oranı.
2. Ki-kare istatistiği.
3. Phi istatistiği.
4. Kappa istatistiği.

Bu istatistik analizler test-yeniden test verilerini, paralel formlar verilerini veya gözlemciler arası değerlendirme verilerini kullanarak belirli bir katsayı ortaya koyar veya ilişkinin anlamlı olup olmadığı konusunda bizi bilgilendirir.

Test-yeniden Test Tasarımı

Daha önce yeterlilik puanları güvenilirliğinin bu puanlara dayalı olarak yapılan sınıflandırmanın zaman içinde istikrarlı sonuçlar verip vermediğiyle test edildiğini belirtmiştik. Birden fazla ölçümde (test veya gözlemci değerlendirmesi şeklinde olabilir) kişilerin aynı oranda yeterli ve yeterli değil şeklinde sınıflandırılması sonuçların güvenilir olduğunu gösterir. Bunun için araştırmacının KRT'leri aynı örnek kütleye farklı iki zaman diliminde uygulaması gerekir, fakat bu prosedürü uygulamak oldukça zordur. Aradan bir hafta 10 gün gibi belirli bir süre geçtikten sonra aynı kişilere testi ikinci kez uygulamak her zaman mümkün olmayabilir. Test-yeniden test prosedürü daha çok okul ortamları için uygundur. Shifflett'e (2003) göre, KRT'de kesim puanı kendilerine test uygulanan grubun büyük bir kesiminin yetenek düzeyine yakın bir yerde tespit edilmişse güve-

nilirlik düşük çıkar. Öte yandan kesim puanı büyük ölçüde düşük veya yüksek belirlenirse bu kez de puanların *geçerliliği* soruşturulabilir bir nitelik kazanır.[42] Bu nedenle kesim noktasının gerçekçi bir şekilde belirlenmesi KRT için hayati derecede önemlidir.

Paralel Formlar Tasarımı

Kriter referanslı testlerin güvenilirliğini sınamak için yararlanılabilecek ikinci yöntem paralel formlar modelini denemektir. Bu uygulamada kişilere aynı kesim puanına sahip iki paralel test verilir. Birincisi ilk kez geliştirilen test, ikincisi ise daha önce güvenilirliği saptanmış olan testtir. Birinci test sonucunda ayrışan kişilerin sayısıyla ikinci test sonucuna göre ayrışan kişilerin sayısı eşit ise testin güvenilir olduğuna karar verilir. Gerçek hayatta paralel iki test oluşturmak oldukça güç olduğundan Subkoviak (1984) tarafından geliştirilen ve tek test sonucunun temel alındığı güvenilirlik analizi yöntemine başvurulabilir.

Gözlemci İçi Değerlendirme Tasarımı

Bu uygulamada kriter referanslı değerlendirme aynı gözlemci tarafından t_1 ve t_2 zamanında olmak üzere (bir hafta 10 günlük bir zaman boşluğu bırakılabilir) iki kez tekrarlanır. Değerlendirmelerin sonucunda kişiler tutarlı bir şekilde ayrışıyorsa gözlemci içi değerlendirmelerin güvenilir olduğuna karar verilir.

Gözlemciler Arası Değerlendirme Tasarımı

Gözlemciler arası değerlendirme kişilerin veya maddelerin puanlandırılmasına dayanır. Maddelerin değerlendirilmesinde sorular zorluk / kolaylık açısından ikili veya çok dereceli olarak sınıflandırılır. Aynı uygulama kişilerin sınıflandırılması için de geçerlidir. Kişilerin değerlendirilmesinde bireyler *geçti-kaldı* şeklinde ikili veya çok dereceli olarak sınıflandırılırlar. Örneğin iki gözlemci maddeleri asgarî yetenek düzeyine sahip bir kişi açısından zor-kolay şeklinde veya dereceli olarak "A grubu, B grubu ve C grubu öğrenciler/kişiler yanıtlayabilir" şeklinde sınıflandırmışlarsa ve bu sınıflandırmada hakemler arasında tutarlılık varsa testin güvenilir olduğuna karar verilir.

Tek Test Sonucuna Dayalı Sınıflama Tutarlılığı

Sınıflama tutarlılığı tek bir test uygulamasının sonuçlarına dayalı olarak da yapılabilir. Bu konuda Subkoviak ve Peng (1988) tarafından geliştirilen prosedür uygulanır. Aynı kişilere farklı zamanlarda iki veya daha fazla test

uygulamanın oldukça güç olması nedeniyle güvenilirlik analizlerinin bazen tek bir test sonucuna bağlı olarak yapılması gerekir. Literatürde tek test sonucuna dayalı olarak yapılan güvenilirlik analizine "kararın tutarlılığı"[a], "kararın doğruluğu" veya "kararın uygunluğu" adları verilmiştir. Kararın tutarlılığını tahmin eden yöntemlerde test maddelerinin eşit ağırlığa sahip oldukları, ikili derecelendirildikleri (doğru-yanlış) ve doğru bir şekilde cevaplandırılan soruların toplam puanının temel alındığı varsayılır.[43] *Başarılı - başarısız* veya *yeterli - yeterli değil* kararı tek bir test sonucuna bakılarak verildiğinde muhtemelen bazı adaylar başarılı gözükmelerine rağmen gerçekte başarısız (yanlış pozitif) ve bazı adaylar da başarısız gözükmelerine rağmen gerçekte başarılı (yanlış negatif) olabilirler. Yanlış pozitif ve yanlış negatif kararının her ikisi de problemlidir. Yanlış pozitif ve yanlış negatif kararlar literatürde "eşik kaybı" (threshold-loss) terimiyle adlandırılmış ve eşik kaybı indeks değerleriyle isabetsizliğin derecesi belirlenmeye çalışılmıştır. Bir başka anlamda eşik kaybı, yeterlilik değerlendirmesinde ortaya çıkan sınıflandırma hatasıdır. Eşik kaybı değişik analiz yöntemleriyle hesaplanabilir. Eğer test-yeniden test şeklinde iki farklı test uygulaması yapılmışsa "uyuşma oranı" formülü ile "kappa" formülüne başvurulur. Ancak bu yaklaşım kararlılık analizi değil, istikrarlılık analizidir. Ölçüm sonuçları eğer tek bir teste bağlı ise bu kez Subkoviak tarafından geliştirilen yöntem uygulanır ve bu yaklaşım karar tutarlılığını verir.

Subkoviak ve Peng kendilerinden önce Huynh tarafından geliştirilen yeterlilik sınıflandırması formülünden hareket etmişler ve oldukça karmaşık olan Huynh hesaplama prosedürünü bir ölçüde basitleştirerek kendi özel hesaplama yaklaşımlarını geliştirmişlerdir. Literatürde bu prosedür sonucunda elde edilen oranlara "Subkoviak ve Peng yaklaşık değerleri" adı verilmiştir. Subkoviak ve Peng yaklaşık değeri ($\hat{p}_0$), test alan kişilerin sonuçta *yeterli – yeterli değil* şeklindeki sınıflandırmasının ne ölçüde doğru olduğunu belirleyen bir tahmin değeridir. Söz konusu tahmin değeri $\hat{p}_0$, "doğru yerleştirme oranı" (DYO) veya "karar tutarlılığı indeks değeri" (KTİD) olarak da isimlendirilebilir. Karar tutarlılığı indeks değeri kesim puanına, testin uzunluğuna ve puanların değişkenlik göstermesine karşı duyarlıdır. Kesim puanı ortalamaya yakın olarak belirlendiğinde KTİD daha düşük çıkar.[44] Örneğin, $\hat{p}_0$ değerinin ,79 çıkması kişilerin %79 ora-

[a] *Tutarlılığı* sözcüğü, "uygunluk" ve "doğruluk" anlamında kullanılmıştır. Tutarlılık, doğru değerlendirme ile doğru sınıflandırma yapma anlamına gelir. Tutarlılık ile istikrarlılık kelimelerinin anlamları aynı değildir. İstikrarlılık "oturma", "durulma", "yerleşme" ve "hep aynı sonuçları verme" anlamlarına gelir.

nında *yeterli – yeterli değil* şeklinde doğru olarak sınıflandırıldıkları ve bu sınıflandırmada %20'den fazla sınıflandırma hatası veya eşik kaybı olduğu anlamına gelir ($1-\hat{p}_0$). Bu uygulamada kesim puanına göre sınıflandırılan her beş kişiden biri yanlış bir şekilde *yeterli* veya *yeterli değil* şeklinde nitelendirilmiştir.

Subkoviak bir testte veya alt testte kişilerin yeterli veya yeterli değil şeklinde iki gruba ayırma işleminin güvenilir bir şekilde yapılabilmesi için en az altı maddenin / sorunun bulunması gerektiğini bildirmiştir. Altı madde kişinin bir konu içeriğine hakim olup olmadığını belirlemek için yeterlidir.[45] Bu görüş daha sonra Webb (1999, aktaran Impara) tarafından da desteklenmiştir. Webb bir kişinin spesifik bir ölçüm alanında yeterli sayılabilmesi için söz konusu altı sorunun en az dördünü doğru yanıtlaması gerektiğini belirtmiştir.[46] Ancak Webb bu rakamı bütün test durumları için önermemiş duruma göre daha fazla test maddesinden yararlanılabileceğini söylemiştir. Bilgi testleri genelde çoklu içerik boyutlarına sahip olduklarından her bir boyutun karşılaştırılabilirlik ve amaca uygunluk açısından makul bir içeriğe sahip olması ilkesinden hareket edilir. Bu yöntemde altı maddenin ortalaması üç ve standart sapması bir madde ise kesim puanı 4 madde olarak belirlenir ve 4 maddenin uyuşma oranının (veya doğru yerleştirme oranının) ,77 olacağı tahmin edilmektedir. Subkoviak tek bir maddenin güvenilirlik katsayısını ,10 olarak belirlemiştir. Prosedürde, altı maddeden oluşan en küçük bir testte *doğru yerleştirme oranının* veya *uyuşma oranının* en az ,63 olacağı tahmin edilmektedir. Kesim noktası ortalamadan 1 standart sapma kadar artırıldığında uyuşma oranın ,77'ye 1,5 *ss* kadar artırıldığında ise ,88'e çıkacağı tahmin edilmektedir.[a]

---

[a] Kriter referanslı testlerin güvenilirliği literatürde değişik makalelerde inceleme konusu yapılmıştır. Konuya ilgi duyan okurlara bu kaynakları tanıtmayı uygun bulduk. Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion referenced test. Journal of Educational Measurement, 15, 265-276.; Subkoviak, M.J. (1988), "A Practitioner's Guide to Computation and Interpretation of Reliability Indices for Mastery Tests," Journal of Educational Measurement, 25, 47-55.; Subkoviak, M.J. (1984). Estimating the reliability of mastery-nonmastery classifications. R.A. Berk (Ed.), A Guide to criterion-reference test construction (ss. 267-291). Baltimore: John Hopkins University Press.; Livingtson, S. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measruement, 9, 13.26.; Hambleton, R., Mills, C. ve Simon, R. (1983). Determining the lengths for criterion referenced tests. Journal of Educational Measurement, 20(1), 27-38.; Brennan ve Kane (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14, 277-289.; Swaminathan H., Hambleton R.K. ve Algina J. (1974) "Reliability of criterion-referenced tests: a decision-theoretic formulation", Journal of Educational Measurement, C. 11, N. 4, ss. 263-267; Huynh, H. (1976). On the reliability of domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Subkoviak'ın geliştirdiği karar tutarlılığı yaklaşımı $\hat{p}_0$, şans faktörünün etkisini ortadan kaldırmadığından bu değerin ayrıca Kappa değeriyle birlikte değerlendirilmesinde yarar vardır. Feldt ve Brennan'a göre $\hat{p}_0$ kararın uygunluğunu gösterirken, kappa değeri kararın tutarlılığına testin katkısını yansıtır.[47]

Kriter referanslı test puanları ve örneklem büyüklüğü. Kriter referanslı test puanlarının güvenilirliğini belirlemek için kişilerin test puanlarıyla fiilî performans puanları karşılaştırılır. Bu iki puan dizisi arasında en azından ,30 düzeyinde bir korelasyon olması gerekir. Araştırmacının ,05 anlamlılık düzeyinde ve tek kuyruklu bir hipotez testinde geçerli bir sonuç elde edebilmesi için örneklem büyüklüğünün en az 20 olması gerekir. Kuşkusuz örneklem büyüklüğünün daha büyük olması sonuçların daha sağlıklı olmasını doğuracaktır.[48]

Testin içindeki alt bölümlerin güvenilirliği. Kriter referanslı bir test bazen kendi içinde küçük alt testlerden (testçik) oluşur. Böyle bir durumda her bir alt testin kendi özel amacı vardır ve güvenilirlik analizi bu alt testlerin her biri için ayrı ayrı yapılmalı ve sonuçlar buna göre raporlanmalıdır. KRT eğer alt testlerden oluşuyorsa kişilerin ölçüm yapılan konu hakkında yeterli olup olmadığına güvenilir bir şekilde karar verebilmek için alt testte en az altı madde bulunmalıdır.[49]

GÜVENİLİRLİK ANALİZLERİ

Kriter referanslı testlerde uygulanabilecek istatistiksel güvenilirlik analizleri belirli başlıklar altında toplanmıştır. Bunlar; uyuşma oranı, ki-kare testi, phi korelasyon katsayısı ve Kappa testidir.

Uyuşma Oranı

Uyuşma oranı, (proportion of agreement / percentage of agreement) kişilerin her iki ölçümde doğru bir şekilde sınıflandırılma oranı toplamlarının toplam test edilen kişi sayısına bölünmesiyle elde edilen bir değerdir. Uyuşma oranı kesim puanına veya dereceli ölçek puanlarına göre belirlenebilir.

Uyuşma oranının kesim puanına göre belirlenmesi. Bu yöntemde kişiler, yapılan farklı iki ölçümde sınır değerinin altında kalıp kalmama durumuna göre iki gruba ayrılırlar. Geçenler 1, kalanlar ise 0 olarak kodlanır.

Her iki test sonucunda kişilerin aynı oranda kesim puanının üzerinde veya altında kalması testin güvenilir olduğunu gösterir. Traub (1994) bu olguyu bir matris çizelge üzerinde göstermiştir (aktaran, Fairchild, 2003).[50] Bunun için araştırmacı gerçek pozitif ve gerçek negatif değerleri gösterecek bir çapraz tablo veya matris çizelge hazırlar. Bu çizelgede birinci ve ikinci ölçümde geçenler A hücresinde, birinci ve ikinci ölçümde kalanlar ise D hücresinde gösterilirler (*bk.*, Tablo 10-7). Bu iki hücre her iki ölçümdeki uyuşma sayısını gösterir. B ve C hücreleri ise uyuşmamayı temsil eder.

Tablo 10-7. Uyuşma Oranı Çapraz Tablosu

		1. gün / 1. test uygulaması	
		Geçenler	Kalanlar
2. gün / 2. test uygulaması	Geçenler	*Birinci ve ikinci gün geçenlerin hepsi bu hücrede yer alırlar* (A)	(B)
	Kalanlar	(C)	*Birinci ve ikinci gün kalanların hepsi bu hücrede yer alırlar* (D)

Çapraz tabloyu düzenlemenin amacı, güvenilirliği sağlamak için A ve D hücrelerindeki değerleri maksimize etmek, B ve C hücrelerindeki değerleri ise minimize etmektir (*bk.*, Eşitlik 10-10; 10-11).

■ Uyuşma oranı (UO/P[ii]) formülü.

$$UO = P = P_A + P_D / N . \qquad (10\text{-}10)$$

veya;

$$UO = \frac{P_A + P_D}{A + B + C + D} = \frac{\text{Uyuşma sayısı veya oranı}}{\text{Uyuşma + uyuşmama sayısı}} . \qquad (10\text{-}11)$$

Hesaplanan *P* değeri, kişiler gerçek anlamda "yeterli" ve "yeterli olmayanlar" şeklinde iki eşit gruba ayrılırlarsa yüksek çıkar, fakat kişilerin çoğunluğu kesim puanına yakın olursa bu kez *P* oranı düşük çıkar. Uyuşma

[ii] İngilizce *proportion* [oran] kelimesini temsil eden simgedir. Bazı kitaplarda bu simgenin kullanılmış olması nedeniyle ayrıca gösterilmiştir.

oranının şans faktörünü dikkate almaması nedeniyle kullanılması her zaman doğru olmayabilir. Bu nedenle araştırmacının uyuşma oranı değeriyle Kappa değerini birlikte vermesinde yarar vardır.

Uyuşma oranının dereceli ölçek puanlarına göre belirlenmesi. Dereceli ölçek puanlarında birden fazla hakem kişileri veya maddeleri belirli boyutlarda; 1 = *yetersiz*, 2 = *vasat*, 3 = *iyi* ve 4 = *mükemmel* şeklinde puanlandırır. Bu şekilde birden fazla değişkende ve çok sayıda kişi tarafından yapılan değerlendirmede öncelikle hakemlerin uyuşma sayısı tespit edilir. Bu rakam toplam hakem/uzman sayısına bölünerek uyuşma oranı bulunur. Daha sonra değişik değişkenlere ait uyuşma oranlarının ortalaması alınarak genel uyuşma ortalaması bulunur.

■ Örnek bir mülâkat değerlendirme formu.

Değerlendiricinin ismi:
Değerlendirme tarihi:

Adaylar	Giyimi	Konuşması	Özgüveni	Amaç odaklılığı
1. kişi	3	3	4	4
2. kişi	4	5	4	3
3. kişi	5	5	5	4
4. kişi	4	4	3	3
5. kişi	4	4	4	4

Değerlendirme ölçütleri: 1 = yetersiz, 2 = vasat, 3 = iyi, 4 = mükemmel.

Diyelim ki dört hakem değerlendirme yapmış olsun. Uyuşma oranının hesaplanabilmesi için sonuçların özetleneceği ikinci bir tablo daha hazırlanır. Bu tabloda değerlendiricilerin verdikleri puanlarda uyuşma sayıları gösterilir. Uyuşma, belirli bir değişkende hakemlerin kişiye aynı puanı vermeleri anlamına gelir.

- Birinci kişi için uyuşma değeri ortalaması

	Değerlendiricilerin uyuşma sayısı	Değerlendirici sayısı	Uyuşma oranı
Giyimi	2	4	,50
Konuşması	3	4	,75
Özgüveni	4	4	1,00
Amaç odaklılığı	3	4	,75
Genel ortalama			,75

Uyuşma oranının %70'in üzerinde olması nedeniyle kişiyle ilgili olarak yapılan değerlendirmenin güvenilir olduğuna karar verilir. Anlamlı ve yorumlanabilir katsayılar ,50 ilâ 1,00 arasında değişir. Uyuşma oranı, şans eseri denk gelme olgusunu dikkate almadığından sınıflandırma kriteri ayrıca Kappa katsayısı ile de test edilir.

Kappa Katsayısı

Kappa, hipotez test eden istatistiksel bir test değil, bir tür korelasyon katsayısıdır. Diğer korelasyon katsayılarından farklı olan yönü çıkabilecek negatif katsayıların yorumlanamaz oluşudur. Yorumlanabilir katsayılar 0 ile 1,00 arasında değişir. Araştırmacılar Kappa analizinde güvenilirlik için en az ,60 oranını tutturmaya çalışırlar. Kappa katsayıları aşağıdaki gibi yorumlanır:

1. Zayıf uyuşma = < ,20 .
2. Kabul edilebilir = ,20 - ,40 .
3. Orta derecede uyuşma = ,40 - ,60 .
4. İyi uyuşma = ,60 - ,80 .
5. Çok iyi uyuşma = ,80 - 1,00 .

Kriter referanslı testlerde güvenilirlik katsayılarını tam olarak vermek için uyuşma oranı ve şans faktörü açısından düzeltilmiş Kappa değerlerinin her ikisi birlikte raporlanır. Cohen Kappa katsayısında *uyuşma oranından* farklı olarak şans faktörünün etkisi azaltılmaya çalışılmıştır. Bu nedenle Kappa katsayıları uyuşma oranlarına göre daha düşük değerlerdir. Fakat literatürde Kappa katsayısının şans faktörünü ortadan kaldırdığı iddiasının çok gerçekçi olmadığı ifade edilmiştir. Bazı bilim adamları karar veren hakemlerin şans faktöründen etkilenmediklerini ortaya koyacak daha kesin

deliller sunulması gerektiği görüşündedirler.[51] Kappa katsayısının hesaplanabilmesi için uyuşma oranında olduğu gibi değerlendirmenin ikili veya çok dereceli olma durumu dikkate alınır.

İkili değerlendirmelerde Kappa katsayısının hesaplanması. Bunun için iki gözlemcinin / değerlendiricinin adaylara verdikleri *geçer / kalır* veya *başarılı / başarısız* puanları dikkate alınır. İkili değerlendirmelerde Kappa katsayısının hesaplanabilmesi için araştırmacılara aşağıdaki adımları öneririz.[52]

1. aşama. Kesim puanını belirleyiniz. Kesim puanı, grubun büyük bir kesiminin sahip olduğu yetenek düzeyine yakın bir yerde belirlenmişse güvenilirlik düşük çıkar. Bu nedenle kesim puanının bir ölçüde yüksek tutulmasında yarar vardır. Fakat kesim puanı aşırı derecede düşük veya yüksek tutulursa bu kez testin geçerliliğinin azalması tehlikesiyle karşılaşılır.

2. aşama. Testi, gruba farklı iki zamanda uygulayınız ve kişileri başarılı olup olmama durumuna göre bir çizelge üzerinde sınıflandırınız (*bk.*, Tablo 10-8).

Tablo 10-8. Farklı İki Gözlemci Değerlendirmesinin veya Farklı İki Ölçüm Sonucunun Çizelge Haline Getirilmesi

Adaylar	Birinci gözlemci	İkinci gözlemci
1. kişi	Başarılı	Başarılı
2. kişi	Başarılı	Başarılı
3. kişi	Başarısız	Başarılı
4. kişi	Başarılı	Başarısız
5. kişi	Başarısız	Başarısız
6. kişi	Başarısız	Başarısız
7. kişi	Başarılı	Başarılı
8. kişi	Başarılı	Başarısız
9. kişi	Başarısız	Başarısız
10. kişi	Başarılı	Başarılı

3. aşama. Birinci ve ikinci ölçümde geçme-kalma oranlarını göstermek üzere 2x2 şeklindeki çapraz tabloyu oluşturunuz.

4. aşama. Çapraz tablonun ana köşegenindeki gözlem oranlarını (veya sayılarını) toplayınız. Bu rakamlar daha sonra Kappa formülünde kullanılacaktır.

5. aşama. Kappa formülünde yer alan şans eseri uyuşma değerlerini (oran veya frekans olarak) hesaplayınız (*bk.*, Tablo 10-9). Bunun için satır ve sütun toplam değerlerini çarparak genel toplam değerine bölünüz.

Tablo 10-9. Kesim Puanına Göre Kappa Testi İçin Frekans Verileri

		1. gün / 1. test uygulaması		Toplam
		Geçenler	Kalanlar	
2. gün / 2. test uygulaması	Geçenler	12 (8,6)*	2	14
	Kalanlar	4	8 (4,6)*	12
	Toplam	16	10	26

* Şans eseri uyuşma beklentisi değerleri.

Kappa testi için çalışılan verilerin niteliğine göre iki farklı formülle çalışılabilir: frekans formülü ve oranlar formülü (*bk*, Eşitlik 10-12; 10-13).

- Frekans formülü.

$$\kappa = \frac{f_o - f_c}{N - f_c}. \qquad \kappa = \frac{20 - 13{,}2}{26 - 13{,}2}, \qquad \kappa = 0{,}531. \qquad (10\text{-}12)$$

Formülde f_o ana köşegende yer alan gözlem puanları uyuşma sayılarının toplamını, f_c ise, yine ana köşegende yer alan *şans eseri uyuşma sayılarının toplamını* temsil eder. Gözlem puanları uyuşma oranı ana köşegendeki frekans sayıları toplanarak bulunur (20 = 12 + 8). Şans eseri uyuşma sayıları satır ve sütun toplamları çarpımlarının genel toplama bölünmesiyle elde edilir [(8,6 = (16 * 14) / 26; 4,6 = (10 * 12) / 26].

■ Kappa oranlar formülü.

$$\kappa = \frac{p_o - p_c}{1 - p_c}. \qquad (10\text{-}13)$$

Formülde P_o ana köşegende yer alan gözlem puanlarına ait uyuşma oranları toplamını, P_c ise şans eseri uyuşma oranları toplamını temsil eder (*bk.*, Tablo 10-10).

Tablo 10-10. Kesim Puanına Göre Kappa Testi İçin Oransal Veriler

		1. gün / 1. test uygulaması		Toplam
		Geçenler	Kalanlar	
2. gün / 2. test uygulaması	Geçenler	,46 (,33)*	,08	,54
	Kalanlar	,15	,31 (,18)*	,46
	Toplam	,61	,39	1,00

* Şans eseri uyuşma beklentisi değerleri.

$$\kappa = \frac{p_o - p_c}{1 - p_c}. \quad \kappa = \frac{,77 - ,51}{1 - ,51}, \quad \kappa = 0{,}531. \qquad (10\text{-}14)$$

Çok dereceli değerlendirmelerde Kappa katsayısının hesaplanması. Bunun için iki gözlemcinin / değerlendiricinin (veya iki farklı ölçüm sonucunun) adaylara verdikleri *derecelendirilmiş* puanlar dikkate alınır. Bu yöntemde uygulanacak Kappa formülleri ayrıdır. Sadece veri matrisi 2x2 şeklinde değildir, matris 3x3 veya 4x4 şeklinde derece sayısına göre belirlenir. Çok dereceli değerlendirmeleri istatistik analiz programları çerçevesinde yapmak isteyen araştırmacılar bu amaçla geliştirilmiş olan "ağırlıklı Kappa katsayısı"ndan yararlanabilirler.

SPSS ve Kappa. Kappa hesaplamasını SPSS'te yapmayı düşünen araştırmacılar her iki değişkenin de eşit sayıda kategoriye sahip olmasına dikkat etmelidirler. SPSS'te Kappa istatistiği Statistics/Crosstabs bölümünde bulunur. SPSS ayrıca kapa değerinin standart hatasını da hesaplar.

Ağırlıklı Kappa Katsayısı

Kappa istatistiğinin bir diğer türü "ağırlıklandırılmış Kappa" yöntemidir. Araştırmacının Kappa yerine ağırlıkla Kappa yöntemini kullanmaya karar vermesi verilerin kategorik veya sıralı olmasına bağlıdır. Cohen Kappa değerinde uyuşmazlığın şiddeti dikkate alınmadığından, kısmi bir uyuşmanın sağlandığı dereceli ölçümlerde ağırlıklandırılmış Kappa kullanılır.[53] Ağırlıklı Kappa, kategoriler arasında nispî farklılıkları gösteren ağırlıkların kullanılması suretiyle hesaplanan basit Kappa'nın genelleştirilmiş bir şeklidir. Bilim adamı kesim puanını geçti-kaldı gibi iki derece yerine ikiden fazla kategoriyi sıralı ölçek niteliğine dönüştürüp üç veya daha fazla dereceye göre belirlemişse ağırlıklı Kappa yönteminden yararlanır.[54] Bu yaklaşımda hakemlerin kriter puanları değerlendirmesi geçme-kalma durumuna göre değil, derecelendirilmiş olarak saptanır. Örneğin, iki hakem yapmış oldukları gözlemler sonucunda adayları "sınır çizgisinde = 1", "yeterli = 2", ve "üstün = 3" şeklinde derecelendirerek değerlendirmiş olsun. Ölçümde 50 kişilik bir aday grubu değerlendirildiğinde birinci hakemle ikinci hakem arasındaki uyuşma durumu sayılarak bu uyuşma sayısı "ağırlık" olarak atanır (*bk.*, Tablo 10-11).

Tablo 10-11. İki Hakemin Vermiş Oldukları Puanlar Arasındaki Uyuşma Durumuna Bakılarak Ağırlıkların Saptanması

Birinci hakem	İkinci hakem	Ağırlık
1	1	5
1	2	7
1	3	7
2	1	8
2	2	8
2	3	5
3	1	5
3	2	2
3	3	3

Bu yöntemde istatistik analiz ham veri matrisine göre değil, yeniden düzenlenmiş üç sütunlu veri matrisine göre yapılır. Yeniden düzenlenmiş veri matrisinde sadece dokuz vak'a ve üç değişken vardır. Kappa istatistiğini SPSS'te yapmak için Data ve Weight mönüleri kullanılarak öncelikle "ağırlık" değişkeni frekans değişkeni olarak atanır. Daha sonra Crosstabs

mönüsü kullanılarak diğer iki değişken arasında Kappa istatistiği hesaplatılır. Sonuç birinci ve ikinci hakemin benzer puanları verdikleri ölçüde Kappa değeri yüksek çıkar ve anlamlılık değeri de sıfıra yaklaşır.

Hakemlerin 1, 2 ve 3 puan verme açısından uyuşma oranları ise çapraz tablonun köşegeninde yer alan hücreler incelenerek belirlenir. Kappa istatistiğinde hakemlerin "tesadüfen" veya "şans eseri" uyuşma gösterip göstermediklerini görmek için Expected count (EC) satırlarına bakarız. Tek başına Count satırları uyuşmanın gerçekliği konusunda bize tam bir fikir vermez. Expected count hesaplaması Eşitlik 10-15'teki formülle hesaplanır:

- Beklenen değer formülü.

$$EC = (\text{sıra toplamı} * \text{sütun toplamı}) / N. \qquad (10\text{-}15)$$

Buna göre çapraz tablo çıktısının köşegeninde yer alan expected count sayılarının toplamı beklenen frekansı verir. Sadece count değerlerinin toplamı ise gözlem değerleri frekansını verir. Bu değerlerden hareket ederek Kappa katsayısı Eşitlik 10-16'daki formülle hesaplanır.

- Kappa katsayısı formülü.

$$k = (Oa - Ea) / (N - Ea). \qquad (10\text{-}16)$$

O_a gözlenen uyuşma sayılarının toplamıdır. Formüldeki Ea ise beklenen uyuşma değerlerinin toplamını gösterir. Tablodaki değerler toplanarak formüldeki yerine konduğunda ,562 Kappa değeri elde edilir. Bu rakam şans faktöründen arındırılmış olan bir değerdir (*bk.*, Tablo 10-12).

Tablo 10-12. Kappa Değerinin Formül Aracılığıyla Hesaplanması

		2. gün		Toplam
		Yeterli	Yetersiz	
		A	**B**	
	Yeterli	5	2	(W)
		(,23)	(,09)	(,32)
1. gün		**C**	**D**	
	Yetersiz	7	8	(Y)
		(,32)	(,36)	(,68)
		(X)	(Z)	22
		(,55)	(,45)	(1,00)

Gözlem değerleri oranı, *Pg*: ,23 + ,36 = ,59 .
Beklenen değerler oranı, *Pb*: $(x)W + (Z)Y$.
Beklenen değerler oranı, *Pb*: ,55*,32 + ,45* ,68 ,
Beklenen değerler oranı, *Pb*: ,17 + ,30 ,
Beklenen değerler oranı, *Pb*: ,47 .

$$\text{Kappa} = Pg - Pb \,/\, 1 - Pb \,. \qquad (10\text{-}17)$$

$$\text{Kappa} = ,59 - ,47 \,/\, 1 - ,47 = ,12 \,/\, ,53 = ,23 \,. \qquad (10\text{-}18)$$

Kesim puanının doğruluğunu iki hakem yerine ikiden fazla hakemin verdiği değerleri karşılaştırarak yapmak istediğimizde gizli özellik modellerinden (latent trait models) yararlanırız.

Ki-kare Analizi

Ki-kare bağımsızlık testi[a], nominal verilere sahip iki değişken arasındaki ilişkilerin önemli olup olmadığını belirleyen anlamlılık testidir. Ki-kare analizi güvenilirlik ve geçerliliğin her ikisini de belirlemek için kullanılabilir. Eğer aynı değişkene ait farklı zamanlarda yapılmış iki ölçüm sonucu karşılaştırılıyorsa güvenilirlik; bir ölçüm sonucu kriter değişkene karşı değerlendirmeye alınıyorsa geçerlilik analizi yapılıyor demektir.

[a] Ki-kare testi, uygulama amacına göre literatürde değişik isimlerle adlandırılmıştır. İki bağımsız değişken arasındaki ilişkileri belirlemek için daha çok ki-kare bağımsızlık testi ifadesi kullanılır. Bunun yanında bazı yazarlar Pearson ki-kare testi ifadesini, kimi yazarlar da çapraz tablo ki-kare testi deyimini tercih ederler.

Birinci ölçüm sonuçlarıyla (geçenler / kalanlar) ikinci ölçüm sonuçları (geçenler / kalanlar) arasında anlamlı bir ilişki / benzerlik varsa ($p < ,05$) kriter referanslı test sonuçları güvenilirdir. Ki-kare analizinde sıfır hipotezi "birinci ölçüm sonuçlarıyla ikinci ölçüm sonuçları arasında anlamlı bir ilişki yoktur" veya "birinci ölçüm sonuçlarıyla ikinci ölçüm sonuçları birbirinden bağımsızdır" şeklinde belirlenir. Hesaplama sonucunda ki-kare olasılık değeri ,05 veya daha küçük çıkmışsa H_0 hipotezi reddedilir ve iki değişken arasında (birinci ölçüm sonuçlarıyla ikinci ölçüm sonuçları arasında) anlamlı bir ilişki / benzerlik olduğuna karar verilir.[a] İki test sonucu birbirinden bağımsız ise veya benzerlik yoksa kriter referanslı test sonuçları güvenilir değildir. Sonuçların anlamlılığına, p olasılık değerine bakılarak karar verilir.

Araştırmacı, 2x2 şeklindeki tablodan (geçenler / kalanlar x geçenler / kalanlar) yararlanmışsa serbestlik derecesi 1 olacağından Yates düzeltme formülünden yararlanır. Bunun için SPSS'te "süreklilik düzeltmesi" değeri (continuity correction) temel alınır. Süreklilik düzeltmesi, ki-kare değerine göre daha tutucu olan bir değerdir. İki değişken arasında anlamlı bir ilişki olduğuna daha zor karar verilir. Ki-kare testi hakemlerin değerlendirmesinde veya ölçüm sonuçlarında şans faktörünü dikkate almaz.

Ki-kare istatistiğini 10'dan küçük[b] gruplarda uygulamak doğru olmayabilir. Bu tür gruplarda güvenilirliği belirlemek için Fisher kesin testi uygulanır. İstatistiksel analiz yazılımı SPSS, 2x2 tablosunda herhangi bir hücredeki beklenen frekans sayısı <5 veya örneklem büyüklüğü <20 olduğunda Fisher kesin değerini otomatik olarak hesaplar.

Ki-kare testinin uygulanabilmesi için belirli varsayımların karşılanması gerekir. Bunlar aşağıdaki gibidir:

1. Sınıflandırma puanlarının her biri birbirinden bağımsız olmalıdır.
2. Her bir gözlem sadece tek bir hücrede yer almalıdır. Diğer bir deyişle bir kişi ya geçmiş veya kalmış olmalıdır.
3. Örneklem hacmi analiz yapmaya imkan verecek büyüklükte olmalıdır. Her bir hücredeki frekans sayısı bazı bilim adamlarına göre en az 10 olmalıdır.[55] Fakat bilim adamlarının çoğunluğu 2x2 tablolarında en az beş veya daha fazla gözlem değeri olma kriterini te-

[a] İlişki veya benzerlik, iki değişken arasındaki korelasyona işaret eder. Korelasyonun yüksek olması ilişki olduğu anlamına gelir.

[b] Bazı istatistikçiler bu sayıyı 20 olarak belirlemişlerdir.

mel almışlar ve hiçbir hücrenin sıfır frekansa sahip olmaması gerektiğini bildirmişlerdir. Öte yandan 2x2'den büyük tablolarda ise hücrelerin %80'inin 5 veya daha fazla gözlem değeri içermesi gerekir.[56] Hücre büyüklüğü koşulu sağlanmadığında Yates düzeltme formülü uygulanır veya bilgisayar sonuçlarından Yates düzeltme değerine göre yorum yapılır.

4. Çift yönlü hipotez kurulur..
5. Ki-kare testinde nominal, sıralı veya eşit aralıklı ölçek verileri kullanılır.[57]
6. Ölçüm verileri temelde benzer dağılıma sahip olmalıdır.

Phi Katsayısı

Phi katsayısı, 0 ve 1 şeklinde kodlanan *X* ve *Y* değişkenleri arasındaki ilişkileri araştıran Pearson korelasyon katsayısıdır (*bk.*, Tablo 10-13). Ki-kare, sıfır hipotezini yanlışlamaya çalışan bir anlamlılık testi iken, phi bir tür korelasyon katsayısıdır ve korelasyon katsayısı gibi yorumlanır. Phi korelasyon katsayısını hesaplamak için ki-kare değeri *N*'e bölünür ve çıkan değerin kare kökü alınır (*bk.*, Eşitlik 10-19; 10-20).[58]

■ Ki-kare değeri ile phi korelasyon katsayısının hesaplanması.

Örnek: $\chi^2 = 83{,}424$, $N = 160$.

$83{,}424/160 = {,}5214$, $\sqrt{{,}5214}$, (10-19)

Phi korelasyon katsayısı = ,72 . (10-20)

SPSS ve phi katsayısı. Phi katsayısı, SPSS'te Crosstabs mönüsü altında Phi and Cramer's V seçeneği işaretlenerek hesaplanır. Cramer V değeri iki değişken arasındaki ilişkinin gücünü gösterir. Korelasyon katsayısı 0 ilâ 1 arasında değişir ve 1 gerçek bir ilişki olduğu anlamına gelir. Nominal değişkenlerle çalışıldığında negatif bir ilişki çıkma olasılığı yoktur. Araştırmacı 2x2 şeklindeki tabloyla çalışıyorsa phi katsayısını veya Cramer V korelasyon katsayısını verebilir, çünkü her iki değer de aynıdır. Ancak

tablo 2x2'den daha büyük ise bu kez yalnızca Cramer V değeri rapor edilir, çünkü phi katsayısına ilişkin sonuçlar doğru olmayabilir.

Tablo 10-13. Phi (φ) Korelasyonu

	0	1	Toplam
0	a	b	a + b
1	c	d	c + d
Toplam	a + c	b + d	a + b + c + d

■ Phi korelasyon katsayısının hesaplanması.

$$\text{Phi } (\varphi) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (10\text{-}21)$$

Phi korelasyon katsayısı hesaplanırken herhangi bir hücredeki beklenen değer 5'ten az ise Fisher kesin testi sonuçları dikkate alınır. (Lehner 1996, aktaran Packard).[59] Phi katsayısının negatif çıkması halinde bu işaret dikkate alınmaz.[60]

Küme İçi Korelasyon Katsayısı

Kriter referanslı testlerde kişiler üzerinde iki ölçüm yerine üç veya daha fazla ölçüm yapılmışsa; kişileri iki hakem yerine üç veya daha fazla hakem değerlendirmişse (bu değerlendirmenin ikili veya çok dereceli olarak yapılması durumu değiştirmez) sınıflandırmanın güvenilirliği için bu kez küme içi korelasyon analizi yöntemine başvurulur. Bu konuda daha fazla bilgi "Güvenilirlik ve Korelasyon Analizleri" bölümünden elde edilebilir.

ALINTI YAPILAN KAYNAKLAR

[1] National Council on Measurement in Education , "The Rise and Fall of Criterion-Referenced Measurement [Kriter Referanslı Ölçümlerin Yükselişi ve Düşüşü]," <http://www.enc.org/resources/records/0,1240,005605,00.shtm> (08.02.2004).

[2] FairTest@aol.com, "Criterion- and Standards- Referenced Tests [Kriter ve Standart Referanslı Testler]," <http://www.fairtest.org/facts/csrtests.html> (08.02.2004).

[3] R.L. Linn, "Performance Standards [Başarı Standartları]," <http://www.google.com.tr/search?q=cache:RlipBstioVAJ:epaa.asu.edu/epaa/v11n31/v11n31.pdf+Angoff+Nedelsky+Ebel&hl=tr&ie=UTF-8&inlang=tr> (16.09.2003).

[4] Gatewood ve Field, "Human Resource Selection [İnsan Kaynakları Seçimi]," <http://www.southalabama.edu/mcob/sboyar/staffing/sfaffing_pp/CH06.PPT> (24.09.2003).

[5] Gatewood ve Feild, "Human Resource Selection [İnsan Kaynakları Seçimi]," <http://www.southalabama.edu/mcob/sboyar/staffing/sfaffing_pp/12 > (27.12.2003).

[6] J.K. Palmer, "Making the Selection Decision [Seçim Kararı Verme]," <http://www.psychology.eku.edu/Palmer/590790/SectionM.DOC> (27.12.2003).

[7] Infed, "competence and competency [Yeterlilik ve Yetkinlik]," <http://www.infed.org/biblio/b-comp.htm> (02.07.2004).

[8] Aynı.

[9] PTUK, "Extending the Play Continuum [Oyun Boyutunun Genişletilmesi]," <http://www.playtherapy.org.uk/PSM1.htm> (02.07.2004).

[10] D. Scholarios, "Job Analysis and Design [İş Analizi ve Tasarımı]," <http://www.hrm.strath.ac.uk/teaching/classes/41203-core/documents/Semester2jobanalysis.pdf> (02.07.2004).

[11] CIPD, "Competencies or competences? [Yeterllilik mi yoksa Yetkinlik mi?]," 2004, <http://www.cipd.co.uk/subjects/perfmangmt/competnces/comptfrmwk.htm> (02.07.2004).

[12] R.E. Biddle, "How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing [Terfi, Eğitim, Sertifika ve Lisanslama Uygulamalarında Kullanılan Bilgi Testlerinde Eşik Değer Nasıl Belirlenir?], <http://www.biddle.com/resources/articles/cutoff.pdf> (24.09.2003).

[13] Kinesmetrics Laboratory, "New Perspective and Practice in Setting Performance Standards [Başarı Standartlarını Belirlemede Yeni Bakış Açıları]," <http://www.kines.uiuc.edu/labwebpages/kinesmetrics/Presentations/cut-off_03/web_pdf/cutoff_AAHPERD_03_final1.pdf> (07.12.2003).

[14] G.V. Glass, "Standards and Criteria [Standartlar ve Kriter]," <http://www.google.com.tr/search?q=cache:rdSEMbEVT9YJ:www.wmich.edu/evalctr/pubs/ops/ops10.html+criterion+test+Nedelsky++Angoff&hl=tr&ie=UTF-8&inlang=tr> (16.09.2003).

[15] R.E. Biddle, "How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing [Terfi, Eğitim, Sertifika ve Lisans Testlerinde Kullanılan Bilgi Testlerinde Kesim Puanları Nasıl Belirlenir]," <http://www.google.com.tr/search?q=cache:cFKSX2j9DCUJ:www.biddle.com/resources/articles/cutoff.pdf+criterion+test+Nedelsky++Angoff&hl=tr&ie=UTF-8&inlang=tr> (16.09.2003).

[16] G.V. Glass, "Standards and Criteria Redux [Standartlar ve Kriter]," <http://glass.ed.asu.edu/gene/papers/standards/> (06.12.2003).

[17] Gail. C. Delicio, "Standard Setting: Deciding on the Mastery/Non-Mastery Standard [Standart Belirleme: Yetkin ve Yetkin Olmama Standartlarına Karar Verme]," t.y., <http://www.gcd.clemson.edu/Main808/Notes808/StdSetng.htm> (18.12.2002).

[18] University of Illinois, Kinesmetric bölümü, "The Angoff Method and Its The Angoff Method and Its Extensions [Angoff Yöntemi ve Angoff Yönteminin Uzantıları]," <www.kines.uiuc.edu/labwebpages/kinesmetrics/Presentations/ cut-off_03/web_pdf/cutoff_AAHPERD_03_final2.pdf> (27.12.2003).

[19] Aynı.

[20] A. Kramer1, A. Muijtjens, K. Jansen1, H. Düsman1, L. Tan1 ve Cees van der Vleuten, "Comparison of a Rational and an Empirical Standard Setting Procedure for an OSCE [OSCE İçin Rasyonel ve Ampirik Standart Belirleme Prosedürlerinin Karşılaştırılması]," <http://www.blackwell-synergy.com/links/doi/10.1046/j.1365-2923.2003.01429.x/abs/> (25.09.2003).

[21] California Water Environment Association, "How Test Pass Points Are Determined [Testlerde Geçme Noktası Nasıl Belirlenir?]," <http://www.cwea.org/tcp/passpoints.htm> (25.09.2003).

[22] G. W. Bracey, "Literacy in the Information Age [Bilişim Çağında Okur-yazarlık]," <http://www.pdkintl.org/kappan/kbra0009.htm> (25.09.2003).

[23] S.L. Bell, "Protective Service Physical Ability Tests [Koruyucu Hizmetlerde Fiziksel Yetenek Testleri]," <http://www.ipmaac.org/conf00/bell.pdf> (25.09.2003).

[24] G. M. Hurtz ve Norman R. Hertz, "How Many Raters Should Be Used For Establishing Cutoff Scores With The Angoff Method? [Angoff Yönteminde Kesim Puanlarını belirlemek İçin Kaç Değerlendiriciden Yararlanmak Gerekir?]," <http://147.46.94.112/e_journals/pdf_full/journal_e/e21_199959601.pdf> (24.09.2003).

[25] Hurtz ve Hertz, "How Many Raters."

[26] C. W. Buckendahl ve d., "Setting Minimum Passing Scores on High-Stakes Assessments... [Geniyş Katılımlı Test Uygulamalarında Minimum Geçme Puanının Belirlenmesi]," <http://www.unl.edu/BIACO/AERA/buckendahl99.pdf> Ayrıca *bk.*, University of Illinois, Kinesmetric bölümü, "The Angoff Method and Its The Angoff Method and Its Extensions [Angoff Yöntemi ve Angoff Yönteminin Uzantıları]," <www.kines.uiuc.edu/labwebpages/kinesmetrics/Presentations/ cut-off_03/web_pdf/cutoff_AAHPERD_03_final2.pdf> (27.12.2003).

[27] B. F. Green, "Setting Performance Standard [Başarı Standartlarını Belirleme]," <http://www.ipmaac.org/mapac/meetings/2000/berrtgre.pdf> (16.09.2003).

[28] L.J. Gross, "How The National Board Establishes Its Pass-Fail Standards on Multiple-

choice Tests [Ulusal Test Kurulu Çoktan Seçmeli Testlerde Geçme-Kalma Standartlarını Nasıl Oluşturuyor?], <http://www.optometry.org/articles/passfail.pdf.> (24.09.2003).

[29] L. Dunn, "MAPAC News," <http://www.ipmaac.org/mapac/newsltrs/fallnl1999.pdf> (24.09.2003).

[30] D. Paccioretti, "Setting the Passing Score [Kesim Puanının Belirlenmesi]," 2003, <http://www.caslpa.ca/english/certification/dan.asp> (24.09.2003).

[31] Aynı.

[32] Dunn, "MAPAC News."

[33] C. Horn, M. Ramos, I. Blumer ve G. Madaus, "Cut Scores: Results May Vary [Kesim Puanları : Sonuçlar Çok Değişik Olabilir]," <http://www.bc.edu/research/nbetpp/statements/M1N1.pdf> (05.12.2003).

[34] V. L.Kiplinger, "Advantages/Concerns About the Bookmark Procedure [Belirteç Prosedürünün Avantajları ve Bu Yöntemle İlgili Düşünceler]," <ttp://www.google.com.tr/search?q=cache:Oq8bL9O2kM4J:www.ctlt.org/documents/datamining/training_session_support_materials/reference_8.doc+%22modified+angoff%22&hl=tr&ie=UTF-8&inlang=tr> (25.09.2003).

[35] Kinesmetrics Laboratory, "New Perspective and Practice in Setting Performance Standards [Başarı Standartlarını Belirlemede Yeni Bakış Açıları]," <http://www.kines.uiuc.edu/labwebpages/kinesmetrics/Presentations/cut-off_03/web_pdf/cutoff_AAHPERD_03_final1.pdf> (07.12.2003).

[36] Assessment Systems Inc, "Establishing Passing Standards Without Gambling [Oyun Oynamadan Geçme Standartlarının Belirlenmesi]," <http://www.clearhq.org/Witt_02.ppt> (28.09.2003).

[37] IPMA Assessment Council, "Protective Service Phiysical Ability Tests [Koruyucu Hizmetler Fiziksel Yetenek Testleri]," <http://www.ipmaac.org/conf00/bell.pdf> (27.12.2003).

[38] Graduate Management Admission Council, "Using Scores to Assess Individuals [Bireylerin Değerlendirilmesinde Puanların Kullanılması]," <http://www.gmac.com/gmac/TheGMAT/Scores/UsingScorestoAssessIndividuals.htm> (27.12.2003).

[39] S. Meechan, "A Study of Reliability and Validity of Classifying Subjects as Mastery and Non-mastery of Criterion-Referenced Tests [Kriter Referanslı Testlerde Kişilerin Sınıflandırılmasının Güvenilirlik ve Geçerlilik Çalışması]," <www.clib.psu.ac.th/acad_41/msur1.htm> (28.12.2003).

[40] J.D. Brown, "The Cronbach Alpha Reliability Estimate [Cronbach Alfa Güvenilirlik Tahmini]," <http://www.jalt.org/test/bro_13.htm.> (17.09.2003).

[41] A. J. Fairchild, "Instrument Reliability and Validity [Araç Güvenilirliği ve Geçerliliği]," <http://www.jmu.edu/assessment/wm_library/Reliability_validity.pdf> (27.09.2003).

[42] B. Shifflett, "Reliability [Güvenilirlik]," 2003, <http://www.geolog.com/msmnt/mrelobj.htm> (22.10.2003).

[43] M. J. Young, "Estimating the Consistency and Accuracy of Classifications [Sınıflandırmaların Doğruluk ve Tutarlılıklarının Tahmin Edilmesi]," <http://www.cse.ucla.edu/CRESST/Reports/TECH475.PDF> (23.10.2003).

[44] Nuclear Medicine Technology, "Technical Appendix To Arrt's Annual Report Of Examinations ARRT'ın Yıllık Sınav Raporlarına Teknik Ek]," [<http://www.arrt.org/website/newsite/Psychometrics/AnnualReportofExamsAppendix2002.pdf> (23.10.2003).

[45] A. T. Roac, Stephen N. Eliot, Norman L. Webb, "Alignment Analaysis [Denkleştirme Analizi]," <http://www.wcer.wisc.edu/publications/workingpaper/paper/Working_Paper_No_2003_2.pdf> (27.09.2003).

[46] J. C. Impara, "Alignment: One Element of an Assessment's Instructional Utility [Denkleştirme: Eğitsel Feydayı Değerlendirmede Bir Öge]," <http://www.unl.edu/BIACO/NCME/Alignment%20revised.pdf> (27.09.2003).

[47] Illinois State Board of Education, "1999 Technical Manual [1999 Teknik El Kitabı]," <http://www.isbe.net/assessment/PDF/1999ISATTech.pdf> (23.10.2003).

[48] S.L. Bell, "Protective Service Physical Ability Tests [Koruyucu Hizmetlerde Fiziksel Yetenek Testleri]," <http://www.ipmaac.org/conf00/bell.pdf> (06.12.2003).

[49] N. L. Webb ve John Smithson, "Alignment Between Standards and Assessments in Mathematics for Grade 8 in one State [Sekizinci Sınıfların Matematik Derslerinde Standartla Değerlendirmeleri Denkleştirme]," <http://www.google.com.tr/search?q=cache:xZEMiuBp3TcJ:facstaff.wcer.wisc.edu/normw/99AERAAlignment.pdf+Subkoviak&hl=tr&ie=UTF-8&inlang=tr> (26.09.2003).

[50] Fairchild, "Instrument Reliability."

[51] J. Uebersax "Kappa Coefficients [Kappa Katsayıları]," <http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm> (30.11.2003).

[52] J. Shifflett "Reliability [Güvenilirlik]," <www.geolog.com/msmnt/mrelobj.htm> (30.11.2003).

[53] M. Friendly, "Plots for two-way Frequency Tables [İki Yönlü Frekans Tabloları İçin Grafikler]," <http://www.math.yorku.ca/SCS/Courses/grcat/grc3.html> (30.11.2003).

[54] G. Hripcsak, "Evaluation of function [Fonksiyonların Değerlendirilmesi]," <http://www.cpmc.columbia.edu/edu/G4060/hw-f03/G4060%20HW2-key.doc> (10.10.2003).

[55] Evaluation Software Publishing, "Reliability [Güvenilirlik]," <www.educationadvisor.com/ocio2001/Reliability_n.doc> (29.11.2003).

[56] D. Garson, "Chi-Square Significance Tests [Ki-Kare Anlamlılık Testi]," <http://www2.chass.ncsu.edu/garson/pa765/chisq.htm> (28.12.2003).

[57] Garson, "Chi-Square Significance."

[58] J. R. Brown, "Inter-rater Agreement and Inter-rater Reliability [Gözlemciler Arası Uyuşma ve Gözlemciler Arası Güvenilirlik]," <http://www.stevens.clarion.edu/ncate/standard_2/Inter-rater_research.htm> (30.11.2003).

[59] J. Packard, "Observer Reliability [Gözlemci Güvenilirliği]," <http://canis.tamu.edu/wfscCourses/WFSC620/exercises/ExerciG.htm> (28.11.2003).

[60] D.J. Krus "The Phi Correlation and the Chi Square Ratio [Phi Korelasyon Katsayısı ve Ki-kare]," <http://www.visualstatisticsstudio.com/Workbook/phi_correlation.htm> (28.12.2003).

NİTEL ARAŞTIRMALARDA GÜVENİLİRLİK

Bilimsel araştırmaların ikinci önemli grubunu niteliksel (veya kısaca isimlendirmek gerekirse "nitel") araştırmalar oluşturur. Nicel (niceliksel) araştırmalarla toplanamayan bazı önemli veri ve bilgiler derinlemesine mülâkat yönteminin uygulandığı, ayrıntılı gözlem çalışmalarının yapıldığı veya içerik analiziyle anlam derinliklerinin sorgulandığı nitel araştırma yöntemleriyle toplanır. Büyük ölçüde rakamlara dayanmıyor olsa bile güvenilirlik konusu nitel araştırmalar için de önemlidir. Literatürde bilim adamları nitel araştırmalarda güvenilirlik konusundan çok "geçerlilik" konusunu ön plana çıkarmışlardır. Fakat, yine de nitel araştırmaların bir bölümünde güvenilirlik bilgilerini vermemek araştırmanın değerini soruşturulabilir hale getirir. Bu bölümde nitel araştırma türleri hakkında verilen kısa açıklayıcı bilgilerden sonra her biri için yapılabilecek güvenilirlik analizleri konusuna değinilmiştir.

GENEL

Nitel veriler; mülâkatlar, gözlemler, uygulamalı proje araştırmaları, odak grubu araştırmaları, arşiv çalışmaları ve içerik analizi sonucunda elde edilen bilgi ve değerlendirmelerdir. Nitel veriler sözel içerikli araştırmalardan elde edilir. Bilim adamı, araştırdığı alanla ilgili olarak daha önceden geliştirilmiş kuramsal bir bilgi birikimine sahip değilse, bir konu, olgu veya bir insan grubu ilk kez inceleniyorsa veya mevcut incelemeler ve kuramsal bilgi birikimi yetersizse niceliksel araştırma yerine niteliksel içerikli bir araştırma yapmayı tercih edebilir. Tüme varım (endüksiyon) yönteminin[a] uygulandığı bu araştırmalarda sonuçlar çoğunlukla rakamlara dayandırılmaz; tersine gözlemler, mülâkatlar, kişisel değerlendirmeler ve

[a] Endüksiyon yönteminin uygulandığı araştırmalardan elde edilen bulgular bir olguya ait başlangıç, hazırlık veya deneme niteliğindeki bilgilerdir. Bu nedenle de bu bilgilere ihtiyatla yaklaşılır. Bu bilgiler bir bütünün parçası olabilir veya bir kuramın üretilmesinde işe yarayacak taslak niteliğindeki ilk bulgulardır.

maz; tersine gözlemler, mülâkatlar, kişisel değerlendirmeler ve sözel açıklamalar belirli iddialar için kanıt olarak gösterilir. Araştırmacı çevresini, dış dünyayı ayrıntılı bir şekilde gözlemleyerek daha soyut fikirlere, kavramlara veya genellemelere ulaşmaya çalışır. Çoğunlukla birinci tekil şahıs kipiyle ve aktif cümle yapısı kullanılarak yapılan yorumlar hipotezler şeklinde ortaya konmaz. Sonuçlar, incelenmesi gereken yeni araştırma önerileri veya önermeleri şeklindedir. Nitel araştırma yapmaya karar veren bir araştırmacı üç farklı amaçtan hareket eder:[1]

1. Gözlemlerden hareket ederek yeni bir kuram geliştirmeye çalışmak, tüme ulaşmak.
2. Mevcut kuramın bilinmeyen yönlerini veya ayrıntılarını ortaya çıkarmak için daha dikkatli bir eleme yapmak, kuramda derinleşme veya genişletme faaliyetine girişmek.
3. Mevcut kuramın geçerliliğini test etmek.

Bilim adamları nitel araştırmaları, nicel araştırmaların zayıf bir alternatifi olarak değil, yerine göre nicel araştırmalar kadar değerli bir veri toplama aracı olarak değerlendirmişlerdir. Bazı araştırmacılara göre tanımlayıcı nitelikteki frekans verileri de niteliksel araştırma grubuna girer. Nitel verilerin kodlanarak daha sonra nicel veriler haline dönüştürülmesi, hipotez test edilmediği sürece o verileri nitel veri olmaktan uzaklaştırmaz.[a]

Nitel Araştırmaların Güvenilirliği

Nitel araştırmaların geçerlilik ve güvenilirliği, araştırmacının elde ettiği kayıtlarla veya yaptığı yorumlarla gerçek hayattaki grubun, kişinin veya kurumun gerçeklerinin örtüşme derecesine bağlıdır. Kayıtlar ve yorumlar gerçeğine uygun olduğu ölçüde geçerli sayılır ve sınamalarda aynı çıktığı ölçüde ise güvenilirdir. Araştırmacı gözlemlerini, elde ettiği bilgileri yorumlarıyla çarpıttığı ölçüde araştırma verileri güvenilmez olarak değerlendirilir. Nitel araştırmalarda bilim adamı yanlı bir tutum içinde olmamalı,

---

[a] Bilimsel araştırmaların nitel-nicel şeklinde ikili bir ayrıma tâbi tutulması konusunda bilim adamları farklı görüşler ileri sürmüşlerdir. Miles ve Huberman'a (1984) göre sosyal realitenin bir bölümünde nicel araştırma yöntemini uygulamak mantıksal olarak uygun değildir. Bazı konularda sadece nitel içerikli veriler toplanabilir. Onların bu görüşlerine karşılık Kaplan, 1964; Richardt ve Cook, 1979, 1980; Walker ve Evers, 1988 gibi bir çok bilim adamı bu tezin mantıksal bir temelinin olmadığını ileri sürmüşlerdir (aktaran, Bezruczko, (2004). N. Bezruczko, "Faulty Thinking by Educational Researchers [Eğitimsel Araştırmacıların Hatalı Düşünceleri]," <http://www.rasch.org/rmt/rmt43e.htm> (15.02.2004).

incelediği olguyu kendi gerçekliği içinde ele almalıdır. Araştırmacı, dış gözlemcidir. Dış bir gözlemci olarak veri toplamadaki başarısı konuya hâkim olmasına, terminolojiyi bilmesine, olaya ve insanlara sempatik bir şekilde yaklaşmasına bağlıdır. Deneyimli olma doğru bilgilere ulaşmayı sağlar. Böylece daha sonraki dönemlerde yapılacak incelemelerde benzer sonuçların elde edilme olasılığı artar. Güvenilirlik; aynı kişi, grup veya kurum (kurumlar) üzerinde yapılacak daha sonraki incelemelerde benzer veya aynı sonuçların elde edilmesidir. Bu açıdan araştırmacı; ölçüm hatalarını, yanlış anlamaları veya yanlış değerlendirmeleri mümkün olduğu kadar azaltmaya çalışmalıdır.

Nitel araştırmalarda geçerlilik ve güvenilirlik konusu nicel araştırmalardan farklı bir şekilde ele alınmıştır. Nitel araştırmalarda aynı olgunun iki kez ölçülemeyeceği belirtilmiş, iki kez ölçülse bile sonuçların kesinlikle aynı çıkmayacağı ifade edilmiştir. Altheide ve Johnson (1984) ve Leininger'e (1994) göre, geçerlilik ve güvenilirlik analizleri nicel araştırmalar için uygundur, nitel araştırmalarda bu tür analizler yapılamaz (aktaran Morsa ve d. (2003).[2] Bu nedenle nitel araştırmalarda "güvenilirlik" (reliability) terimi yerine "dayanıklılık" (dependability)[a] teriminin kullanılması önerilmiştir. *Dayanıklılık*, araştırmacının araştırma süreci içinde değişen olguların veya olayların araştırmayı ne şekilde etkilediğini veya etkileyebileceğini açık, net bir şekilde ortaya koyması ve böylece aynı araştırmayı tekrarlamayı düşünen diğer bilim adamlarına sağlam bir araştırma zemini sunmasıdır. Sonuca etki eden faktörleri ve olgudaki değişim sürecini başarılı bir şekilde ortaya koyduğu ölçüde bilim adamının yaptığı araştırmanın "dayanıklı" veya "sağlam" olduğu söylenir.[3] Lincoln ve Guba (1985) nitel araştırmaların geçerlilik ve güvenilirliği için *doğruluk* (trustworthiness) terimini kullanmışlar ve *doğruluğun* da dört kriter çerçevesinde sağlanabileceğini söylemişlerdir: *inanılırlık, aktarılabilirlik, dayanıklılık ve teyit edilebilirlik* (aktaran, deWet ve Erasmus, 2003).[4] Aşağıdaki paragraflarda bu ölçütlere ilişkin özet bilgiler verilmiştir.

İnanılırlık. İnanılırlık, araştırma sonuçlarının kendilerinden bilgi toplanan kişilerin bakış açısıyla doğru ve güvenilir olmasıdır.[5] İnanılırlık özelliği altı temel öge ile belirlenir: (a) veri doygunluğu elde edilinceye kadar sitede uzun süreli bir araştırma yapma, (b) çok yönlü etkileri de dikkate alarak gözlemleri ısrarlı bir şekilde sürdürme, (c) farklı sorular sorarak,

[a] Terimi Türkçeye değişik şekillerde çevirebiliriz. Kanıtlanabilirlik, yinelenebilirlik, gerçeklik, sağlamlık, doğruluk, istikrarlılık, düzenlilik ve gerçeğe uygunluk terimleri belirli koşullarda hep aynı anlama gelir.

farklı yöntemlere başvurarak ve farklı kaynaklardan yararlanarak üçleme yöntemini uygulama, (ç) teyp bandı, video çekimi, el yazısı mektuplar gibi uygun kanıtları referans gösterme, (d) ast veya üst konumunda olmayan benzer statüdeki meslektaş görüşlerini yansıtma, (e) araştırma yapılan gruptaki kişilerin görüşlerini alarak yazılanları onlara kontrol ettirme.[6] Trochime (2004) göre, *inanılırlık* veya *itibarlılık* sadece kendilerinden bilgi toplanan kişilerin bakış açısıyla değerlendirilebilir. İlgili kişiler eğer araştırmayı onaylıyorlarsa, değer veriyorlarsa, saygı duyuyorlarsa çalışma itibarlıdır.[7] Guevara ve Mendias'a (2002) göre ise, nitel araştırmanın itibarı "meslektaş incelemesi", "sunumu" veya "savunmasıyla" gerçekleştirilebilir. Bu kişiler ya araştırmacının proje danışmanlarıdır veya projeyle ilgisi olmayan bağımsız araştırmacılardır.[8] Meslektaşlar araştırmayı incelemeli, kullanılan malzemeyi analiz etmeli, hipotezleri test etmeli, araştırmacının düşünce ve mülâhazalarını dinleyerek onun yönelim ve yöntemini haklı bulmalıdırlar. Nitel bir araştırmanın inanılırlığı düşükse bu araştırma aktarılabilir olma özelliğini de yitirmiş demektir.

Aktarılabilirlik. Aktarılabilirlik, bulgular ve sonuçların benzer düzlemlere, durumlara, ana kütlelere veya olaylara genellenebilmesidir. Araştırmanın yöntem bilim başlığında verilen ayrıntılı açıklamalara dayalı olarak çalışmanın başka vak'alarda sınanabilirliği veya uygulanabilirliğidir. Niteliksel istatistik araştırma sonuçlarını başka düzlemlere genellemek için çeşitli teşebbüslerde bulunulmuştur, ancak sadece sınırlı ölçüde sonuç alınabilmiştir. Twining'e göre (1999) aktarılabilirlik genellenebilirlik değildir. Genellenebilirlikte tüm çevrelerde uygulanabilirlik söz konusu iken aktarılabilirlikte sadece benzer çevreler düşünülür.[9] Yin ise, "istatistiksel genellenebilirlik" kavramıyla "analitik genellenebilirlik" kavramları arasında ayırım yapmış, niteliksel araştırmalarda sadece analitik genellenebilirliğin söz konusu olabileceğini belirtmiştir (aktaran, Dereshiwsky, 2004).[10] Araştırmacı aktarılabilirlik koşulunu sağlamak için araştırma uygulamasını okuyucularına kapsamlı bir şekilde tanıtmalıdır. Aktarılabilirlik özelliğinin yükü orijinal araştırmacı üzerinde değil, aynı yöntemi kullanarak araştırmayı tekrar etmek isteyen diğer araştırmacıların üzerindedir. Bu nedenle nitel bir araştırmanın aktarılabilirlik özelliğine sahip olup olmadığını en iyi aynı araştırmayı tekrar etmek isteyen diğer bilim adamları saptayabilir.

Dayanıklılık. Dayanıklılık, sunulan bilgi ve bulguların aradan geçen zaman içinde geçerliliğini korumasıdır. Verilerin istikrarlılığı ve tutarlılığı *dayanıklılığı* gösterir. Zaman içinde araştırma yöntemlerinde yeni gelişme-

ler olur ve farklı metodolojiler geliştirilir. Nitel araştırmayla elde edilen bulguların bu yöntemlerden ve değişikliklerden etkilenmemesi sonuçların dayanıklı olduğu anlamına gelir. Nitel araştırmanın sağlam ve dayanıklı olması için araştırmacı kayıt tutmalı, konuşmaları banda almalı ve izin alabilirse video çekimleri yapmalıdır. Bant çözümleri daha sonra kendisiyle mülâkat yapan kişiye okutularak onayı alınmalıdır. *Dayanıklılık*, araştırmayı kanıtlarla destekleme ve istenildiğinde bunları gösterebilme ölçütüdür. Nitel araştırma, aynı kişiler üzerinde iki kez tekrarlanamayacağına göre bilim adamı, araştırmanın dayanıklılığını belirli kanıtlarla ortaya koymalıdır. Nicel araştırmalardaki "güvenilirlik" kavramı, nitel araştırmalarda "dayanıklılık" kavramıyla karşılanmaya çalışılmıştır. Nitel araştırmalarda, nicel araştırmalarda olduğu gibi daha sonraki zamanlarda bire bir yinelenme söz konusu olmayacağından araştırmacı dayanıklılığı sağlamak için sadece sonuçları vermemeli aynı zamanda sonuçların hangi faktörlerden etkilenebileceğini, değişim süreci içindeki koşulları ve araştırmanın kısıtlarını da net bir şekilde ortaya koymalıdır.

Teyit edilebilirlik. Teyit edilebilirlik, araştırmada yapılan yorumları ve ulaşılan sonuçları "araştırmacı yanlılığı" açısından değerlendirmektir. Her bir araştırmacının kendine özgü bir yaklaşımı vardır. Aynı olgu farklı araştırmacılar tarafından okuyuculara bir ölçüde farklı bir şekilde tanıtılır. Teyit etmede; *tarafsızlık, nötr olma* ve *objektivite* araştırılır. Bunun için ham veriler bağımsız bir meslektaşa veya bir panel grubu üyelerine verilerek inceletilir. İnceleme sonucunda söz konusu kişilerin benzer yorumlara veya sonuçlara ulaşıp ulaşmadıklarına bakılır. Araştırmacının yaptığı yorumlar ve ileri sürdüğü tezler eğer başkaları tarafından da teyit ediliyorsa veya destekleniyorsa nesnellik sağlanmıştır.[11] Teyit edilebilirliği sağlamanın bir diğer yöntemi bağımsız bir kişiye "şeytanın avukatı" rolünü oynatmaktır. Bu kişinin yapacağı bütün eleştiri ve sorulara kanıtlarla cevap verilebiliyorsa yorum ve sonuçların doğrulaması yapılmış demektir. Teyit edilebilirliği test etmenin bir başka yöntemi "doğrulama denetimi" yapmaktır. Doğrulama denetiminde dış gözlemciler araştırmanın başlamasından itibaren bilim adamının yorumlarını, bulgularını ve sonuçlarını bir proje kütüphanesi oluşturarak izlemeye devam ederler ve her bir aşamayı ayrı ayrı denetlerler. Proje kütüphanesi oluştururken araştırmacı ilişkisel veri tabanlarından, Excel türü çalışma çizelgelerinden, kelime işlem yazılımlarından veya duruma göre İnternet teknolojilerinden yararlanabilir.[12] Pek çok araştırmacı bunun için çalışmanın başlangıcından itibaren tüm çalışmalarını not ettikleri kronolojik bir indeks oluşturma yoluna başvurmuştur. Dış gözlemci veya gözlemciler bu indeksle birlikte araştırmacının kullandığı teyp

ve video bantlarını, araştırma notlarını, kullandığı mektupları, çalışma hipotezlerini, mülâkat yapılan kişilerin listesini inceleyerek veri ve kanıtların kendilerini aynı sonuçlara ulaştırıp ulaştırmayacağına karar verirler. Bir denetim çalışması için, dış gözlemcilere bir hafta ile on gün arasında bir sürenin yeterli olacağı bildirilmiştir.[13] Yapılan çalışma eğer içerik analizinde olduğu gibi kavramların kodlanmasına dayanıyorsa birbirinden bağımsız farklı iki kodlayıcının benzer "kod ağaçları" üretmesi halinde de teyitleşme sağlanmış olur. Halpern (1983) doğrulama denetimi için altı grup ham verinin incelenmesini önermiştir (aktaran Siegle, 2004).[14]

1. Ham veriler (video ve teyp bantları, yazılı alan notları, belgeler, araştırma sonuçları).
2. Veri azaltma ve analiz yapmakta kullanılan ürünler (çalışma hipotezleri, özetler, kuramsal notlar, kavramlar, düşünceler ve tasarım taslak notları).
3. Verileri yeniden oluşturma ve birleştirme notları (geliştirilen temalar, bulgu ve sonuçlar, nihaî rapor).
4. Süreç notları (yöntem bilim notları, güvenilirlik notları, doğrulama denetimi notları).
5. Niyet ve eğilimlere ilişkin notlar (soruşturma anket taslakları, araştırma önerileri, beklenti veya niyeti tanımlayan planlar).
6. Araç geliştirme bilgileri (pilot uygulama formları, uygulama çizelgeleri, gözlem formları, nihaî anket formları).

Trochim'in (2004) Lincoln ve Guba'dan aktardığına göre, *inanılırlık* kavramı nicel araştırmalarda iç geçerliliğe, *aktarılabilirlik* kavramı dış geçerliliğe, *dayanıklılık* kavramı güvenilirliğe ve *teyit edilebilirlik* kavramı ise nesnelliğe karşılık gelir (*bk.*, Tablo 11-1).[15] "Geleneksel değerlendirme kriterleri çerçevesinde dayanıklılık ve teyit edilebilirlik farklı nosyonlar olmasına karşılık nitel araştırmalarda her iki özellik de benzer teknikler kullanılarak gerçekleştirilir."[16] Araştırmacılara dayanıklılık ve teyit edilebilirlik özelliklerinin her ikisini de sağlamak için "denetim ve izleme" tekniğini kullanmaları önerilmiştir. Lincoln ve Guba'nın geliştirdiği geçerlilik ve güvenilirlik kavramlarının arka planındaki felsefî temel "yapısalcılık"[a]

[a] Yapısalcılık (constructivism). Eğitim bilimlerinde bilginin öğrencilerin kafalarına aktarılmadığını, tam tersine bilginin öğrenci tarafından kendi zihninde inşa edildiğini savunan felsefî akım. Öğrenci, mevcut ve geçmiş bilgilerine dayalı olarak kafasında yeni fikirleri ve

olduğundan bu kriterlerin her tür nitel araştırmaya uygulanamayacağı belirtilmiştir.[17]

Tablo 11-1. Geçerlilik ve Güvenilirliğe Karşı Doğruluk

Kriter	*Nicel araştırmalar (Geçerlilik/Güvenilirlik)*	*Nitel Araştırmalar (Doğruluk)*
Gerçek değer	İç geçerlilik [a]	İnanılırlık / İtibarlılık
Uygulanabilirlik	Dış geçerlilik [b]	Aktarılabilirlik
Tutarlılık	Güvenilirlik	Dayanıklılık
Tarafsızlık	Nesnellik	Doğrulanabilirlik

[a] İç geçerlilik: Araştırmacının dış bağımsız değişkenlerin hepsini kontrol altında tutması ve sadece incelediği veya ele aldığı değişkenin belli bir sonuca yol açtığını kanıtlaması. Dış bağımsız faktörler kontrol değişkenleridir.

[b] Dış geçerlilik: Araştırmanın sonuçlarının diğer yerlere, kişilere ve zamanlara genellenebilmesidir.

İngiltere ve Avrupa'daki bilim adamları nitel araştırmalar için "geçerlilik" ve "güvenilirlik" kavramlarını kullanmaya devam etmelerine karşılık ABD'de bu kavramları nitel araştırmalar için kullanan yazar ve bilim adamı sayısı daha azdır.[18]

Nitel Araştırmaların Türleri

Nitel araştırmalar bilim adamının belirli bir olgu, söylem, metin, grup, toplum veya kişi hakkında ayrıntılı bilgi edinmek istemesi veya olguyu derinlemesine keşfetmek istemesi üzerine yapılır. Keşfedici araştırma bulgularına dayalı olarak daha sonra gerekiyorsa nicel araştırmalar yapılır. Literatürde nitel araştırma türleri belirli bir mutabakattan uzak olarak değişik başlıklar altında sınıflandırılmıştır. Nitel araştırma türleri antropoloji, sosyal antropoloji, sosyoloji, psikoloji, dil bilimi (lenguistik), işletme yönetimi, tıp, hemşirelik bilimleri gibi disiplinlerde daha fazla uygulama alanı bulmuştur. Burada belirli bir sınıflamaya bağlı olmadan literatürde sık karşılaşılan nitel araştırma, analiz ve inceleme türleri ele alınmıştır.

---

kavramları oluşturur. Terim, Türkçede aynı zamanda *oluşturmacılık, türetimcilik* kavramlarıyla karşılanmıştır.

1. Biyografi araştırmaları.
2. Fenomenoloji araştırmaları.[a]
3. Temelli kuram araştırmaları.[b]
4. Etnografya araştırmaları.[c]
5. Odak grubu araştırmaları.
6. Vak'a etüdü araştırmaları.
7. Kritik olay araştırmaları.
8. Eylem araştırmaları.
9. İçerik analizi araştırmaları.
10. Tarihsel araştırmalar.
11. Söylem analizleri.
12. Mülâkat araştırmaları.
13. Gözlem araştırmaları.
14. Herüstik değerlendirmeler.
15. Tefsir, yorumlama araştırmaları (Hermeneutics [he-mü' nüü:reks]).
16. Estetik soruşturmalar.

Nitel araştırmaların sayısını daha da çoğaltmak mümkündür. Sayılan türlerden bazılarını araştırma türü olarak değil, bilgi toplama aracı olarak gören yazarlar vardır. Literatürde yukarıda sayılanların içinde en çok biyografiler, fenomenoloji[d] araştırmaları, temelli kuram araştırmaları, etnografya araştırmaları, eylem araştırmaları ve vak'a etüdü çeşitli incelemelere

---

[a] Olgu bilim/fenomenoloji. Görünen olgu ve olayların gizemini aklı kullanarak açığa çıkarmayı ve tasvir etmeyi hedefleyen bilim dalı. Yirminci yüzyılda Edmund Husserl tarafından geliştirilen felsefe ekolü. Belirli bir grubun kendi yaşamlarının bazı yönlerini nasıl değerlendirdiklerini onların bakış açılarıyla sergilemeye çalışma.

[b] Temelli kuram. Alanda yapılan belirli bir proje temeli üzerine bina edilmiş kuram. Burada temel; sosyal durumlar, gözlemler veya olgulardır. Seçilen belirli bir teori, niceliksel araştırmalarda olduğu gibi sosyal durumlara uygulanmaz, tam tersine teori insanların da karıştığı veya içinde bulunduğu sosyal durumlardan, düzlemlerden çıkarılır. Teori, sosyal sistemden çıkarılan "gerçeklere" dayandırılır.

[c] Budun bilim / kavim bilim / Etnografya. Belirli bir kültür veya alt kültüre bağlı bir insan topluluğunun tüm etkinliklerini (maddî yaşama bağlı teknikler, toplum düzeni, dinî inançlar, iş ve toprağı işleme aletlerinin bir kültürden öbürüne devri, akrabalık ilişkileri) tasvir edici bir yaklaşımla inceleyen beşerî bilimler dalı (*bk.*, Büyük Larousse, 3875).

[d] MS-Word kelime işlem yazılımında kavram *fenomonoloji* şeklinde yazılırken TDK *İmlâ Kılavuzu*'nda *fenomenoloji* şeklinde ifade edilmiştir.

konu olmuştur. Ancak, bilim adamı incelediği konunun niteliğine, koşulların gerektirdiği duruma göre hangi nitel araştırma türlerinden yararlanacağını kendisi belirler.

BİYOGRAFİ ARAŞTIRMALARI

Biyografi araştırmaları tek bir kişinin, özelliği olan bir grubun, bir kurumun veya bir uçak, bir gemi gibi özel anlamı olan nesnelerin yaşamını konu edinir. Bu yaklaşımda bir peygamberin, meşhur bir devlet adamının, ünlü bir din adamının, sanayicinin, öğretim üyesinin veya bir sanatçının yaşamı değişik yönleriyle tarihsel süreç içinde bir bütün olarak ele alınıp incelenir. İlgili kişinin başından geçen olaylar, diğer insanlardan farklı olan özellikleri, inançları, değerleri, davranışları, düşünceleri varsa eserleri (kitapları, makaleleri, çevirileri) ve sanatı karşılaştırmalı olarak ele alınır. Biyografik araştırmalar yaşamı anlatılan kişinin (veya kişilerin) doğrudan kendi ağzından dinlenip anlatılmışsa birincil kaynak, kütüphane ve literatür çalışması yapılarak değişik kaynaklardan bilgi toplanmak suretiyle anlatılmışsa ikincil kaynak niteliğindedir.

Biyografilerin güvenilirliği, türüne, birincil veya ikincil kaynak olma özelliğine ve araştırmanın genişliğine göre farklı düzeylerde ele alınır. Eleştirel biyografilerde araştırmacı sadece kişinin yaşamını anlatmaz, aynı zamanda söz konusu kişinin sanatını, düşüncelerini veya yönetim biçimini kritik eder. Biyografik bilgiler oldukça nesnel olarak hazırlanabilmesine karşılık kişinin sanatı, yönetim biçimi ve düşüncelerine ilişkin fikirlerin güvenilirliği her zaman tartışma konusudur. Araştırmacı ileri sürdüğü fikirlerin yansız olduğunu kanıtlamak için bu fikirleri başka görüşlerle ve kaynaklarla desteklemek zorundadır. Bunun için kitaplardan, İnternet'ten, ansiklopedilerden, dergilerden, belgelerden ve diğer kaynaklardan yararlanır.

Otobiyografiler ise bireylerin kendilerini tanıttıkları yazılardır. Birincil kaynak niteliğinde olmasına karşılık yazarın önyargılarından, yetersiz bilgiye sahip olmasından, değerlerinden ve beklentilerinden etkilenir. Yazarın samimiyeti, dürüstlüğü tek başına yeterli değildir. Kişiyle yapılan mülâkat ve röportajların güvenilirliği verilen bilgilerin başka kaynaklarla teyit edilmesine bağlıdır. Banda alınan konuşmalar araştırmacı tarafından dikkatli bir şekilde çözümlenip düşünce ve fikirlerin birbiriyle tutarlı ve anlamlı olup olmadığı incelenir.

Yaşamı incelenen kişiye ait mektuplar, koleksiyonlar ve diğer kişisel belgeler biyografik çalışmaların bir diğer türünü oluşturur. Araştırmacılar bunları derleyerek önemli bir kanıt olarak okuyucuların bilgisine sunarlar.

Birincil nitelikteki bu kaynakların güvenilirliği kendi orijinal bağlamı içinde değerlendirilir. Bilim adamları biyografik çalışmalarda güvenilirlik olgusundan çok geçerliliğin önemli olduğunu belirtmişlerdir.[19] Güvenilirliği daha az önemseyen bu bilim adamları biyografik çalışmalarda "gerçekçilik" ve "yapısalcılık" kavramlarını ön plana çıkarırlar. Gerçekçilik nosyonunda, yaşamı ele alınan kişiyle ilgili olarak objektif bilgiye ulaşmak önemlidir. Yapısalcılıkta ise hikayeyi anlatan kişinin "rivayet etme" biçimi, olayı "anlatma" ve "şekillendirme" tarzı önemlidir. Biyografik araştırmalarda hayatı anlatılan kişi belirli bir çerçevede "anlamı yaratan" kişidir. Güvenilirlikten çok bu anlam önemlidir. Araştırmacı metodolojik farklılıklar ve kuramsal varsayımlar yerine, kişinin yaşamına ışık tutma amacı üzerinde odaklanmalı ve kültürel çerçeveyi doğru bir şekilde aktarmalıdır.[20]

FENOMENOLOJİ ARAŞTIRMALARI

Herhangi bir kişinin başından geçen bir olgunun veya bir olayın (terör, deprem, araba kazası, şampiyonluk, buluş, hastalanma, transfer olma, piyango çıkma vb.) iç yüzünün, iç dinamiklerinin incelenmesi ve yaşanan olgunun bütün yönleriyle tam olarak açığa çıkarılmasıdır. Olgusal araştırmalarda kişi veya grup davranışları ilgili üyelerin kendi bakış açılarından değerlendirilerek kendilerine ve dünyaya nasıl baktıkları ele alınır. Bu yaklaşımda, "en iyi yaşayan bilir, o halde ondan öğrenelim" felsefesinden hareket edilir. Araştırmacı dış bir gözlemci olarak "realite" hakkında herhangi bir yorumda bulunmaz. Fenomenoloji araştırması yapabilmek için örneklem büyüklüğünün en az 6 olması gerekir.[21]

Fenomenoloji (olgu bilim) 20. yüzyılın felsefî akımlarından biridir. Bu yaklaşımda bilinç düzeyinde yaşanan deneyimlerin tanımlanması üzerinde durulur. Söz konusu tanımlama yapılırken kurama başvurulmaz, genelden özele çıkarımda bulunma yöntemi uygulanmaz ve diğer disiplinlerin varsayımlarından yararlanma gibi bir davranış içinde de olunmaz. Araştırmacı sadece ilgili kişinin bilinç düzeyinde yaşadığı deneyimleri aktarır.[22] Campbell'e göre (1998) fenomenolojinin kökeni Alman filozofu Edmund Husserl (1913) ve Fransız fenomenoloji bilimcisi Merleau-Ponty'nin görüşlerine dayanır. Bu kişilerin yaklaşımları "klasik tarz" olarak isimlendirilmiştir. Yine Campbell'e göre Van Manen fenomenolojiyi "canlı olarak yaşanan deneyimlerin esasının araştırılması" biçiminde açıklamıştır.[23] Husserl fenomenolojiyi *olgunun yapısal özelliklerinin bilinç düzeyinde* araştırılması olarak görmüştür. Bilinçlilik, nesnelerin kendilerini dışarıdan görmeleri anlamına gelir. Bu yaklaşımda, bütün her şey dışarıda bırakılarak sadece olgunun içeriği hakkında düşünülür. Husserl bu tefekkür biçimini

"fenomenolojik redüksiyon" (indirgemecilik)[a] olarak isimlendirmiştir. Olgusal indirgemecilik literatürde aynı zamanda "parantez içine alma" veya "parantezleme" ifadesiyle karşılanmıştır. Pratik anlamda ifade etmek gerekirse parantezleme, olguyu araştıran kişinin ön yargılarını, eğilimlerini, felsefesini, ideolojisini, dinini ve hatta sağ duyusunu bir kenara bırakıp olguyu ne ise olduğu gibi aktarmasıdır.[24] Olgusal indirgemecilik sonuçta kişiyi zihnin var olmadığı düşüncesine götürür. Husserl'e göre fenomenolojik düşüncede herhangi bir şeyin var olduğu düşüncesinden hareket edilmez. Düşünülen nesnelerin gerçekten var olup olmadığı düşüncesi bir kenarda bırakılır.[25]

Fenomenolojik (olgu bilim) çalışmalarının başlıca yedi özelliği olduğu bildirilmiştir.[26]

1. Olgu bilimcileri, gözlenemeyen maddelerin veya olguların kabul edilmesine yanaşmazlar. Bunları spekülatif düşünce biçimleri olarak görürler.
2. Olgu bilimcileri "Doğacılığa"[b] karşı olma eğilimi içindedirler. Doğacılık, Kuzey Avrupa'da Rönesans'tan sonra modern doğa bilimleri ve teknolojinin gelişmesine paralel olarak gelişme göstermiş bir düşünce akımıdır.
3. Olgu bilimcileri sadece akıl yürütme, algılama ve sezme yoluyla bilgiye ulaşma sürecini temel alırlar.

---

[a] İndirgemecilik bir tür mantık yürütme biçimidir. Karmaşık gerçekleri, olguları basit bir biçimde bütünün parçalarıyla veya en küçük birimiyle açıklama girişimi. Materyalist felsefede ruhsal iç dünyanın, zihnin veya bilincin basit bir açıklamayla nöronların aktivitesine bağlanması. Ontolojik redüksiyonizm iddiasında ise, varlığın sadece madde ve enerjiyle açıklanması. Son yıllarda, irrasyonel bu düşünce biçimine karşı *sistem düşüncesi* ve *bütüncüllük* yaklaşımları geliştirilmiştir. Bütüncüllük yaklaşımında gerçek bir bütün olarak vardır, sadece parçalarına bakılarak açıklanamaz. Hayat, zihin ve bilinç sistemler halinde doğar. Bu olguyu sadece sinirler, hücreler ve atomlarla açıklamaya çalışmak irrasyonel bir düşünce tarzıdır. Bu konuda daha fazla bilgi için *bk.*, Wikipedia, "Reductionism [İndirgemecilik]," <http://en. wikipedia.org/wiki/Scientific_reductionism> (22.02.2004).

[b] Doğacılık, bütün varlıkların kaynağı olarak doğayı temel alan maddeci felsefî akım. Dünyadaki her şeyi doğa ile açıklamaya çalışır. Bu felsefî paradigmada tüm varlıklar madde ve fiziksel olgulara indirgenir. Doğacılık, Yüce Varlık kavramını ret etmesi nedeniyle aynı zamanda ateizmdir. Doğacılar, Tanrı kavramını ret etmeleri nedeniyle "ahlak" konusunda göreceli bir bakış açısına sahiptirler. Onlara göre Tanrı olmadığından evrensel ahlak ve davranış standartları da yoktur. *Bk.*, Naturalism-1, "Naturalism [Doğacılık]," <http://www.naturalism-1.com/> (22.02.2004).

4. Olgu bilimcileri doğal ve kültürel dünyadaki nesnelerin açık ve bilinebilir olma özelliğinin yanında sayılar gibi fikirsel nesnelerin dahi açığa çıkarılabileceğini ve bu nesnelerin bilgisine ulaşılabileceğini varsayarlar.
5. Olgu bilimcileri soruşturma ve araştırmaların "karşılaşma" teması üzerinde odaklaşması gerektiği düşüncesine sahiptirler.
6. Olgu bilimcileri "nedeni", "amacı" veya "temeli" gibi açıklamalarda evreni tanımlama rolünün neredeyse fotoğraf doğruluğunda görsel bir imajla önceden belirlenmesi gerektiği görüşündedirler.
7. Olgu bilimcileri transandantal fenomenolojik zaman diliminin veya indirgemenin yararlı veya mümkün olup olmadığını tartışma eğilimi içindedirler.

Sosyal bilimlerdeki olgu bilim araştırmalarında felsefî bir temelden hareket edilir. Araştırmaya alınacak katılımcılar söz konusu olguyu yaşamış kişiler arasından çok dikkatli bir şekilde seçilirler. Kişisel deneyimlerin araştırmacı tarafından ortaya çıkarılması oldukça güç bir iştir. Araştırmacı ilgili bireyin kişisel deneyimlerini nasıl ve ne şekilde doğru bir şekilde ortaya çıkaracağına ve okuyucularına sunacağına bir şekilde karar vermek zorundadır.[27] Bunun için Spiegelberg, Vankaam, Giorgi, Colaizzi, Husserl, Heidegger, Gadamer ve Ricoeur gibi bilim adamlarının inceleme modellerinden hareket edilir. Söz konusu modellere ilgi duyan araştırmacılara yazarların orijinal eserlerine başvurmalarını öneririz. Bu yöntemde kişinin duyguları, düşünceleri günü gününe izlenerek dünyayı nasıl algıladığı, hangi iç deneyimlere sahip olduğu raporlanır.

Olgu bilim araştırmalarının doğruluğu, yapılan çalışmada araştırmacının hassas ve titiz bir tutum takınmasına bağlıdır. Toplanan veri ve bulguların "gerçeği" yansıtması yararlanılan yöntem bilimin açık bir şekilde ortaya konmasını gerektirir. Bu tür araştırmalarda tesadüfî örnekleme yöntemiyle değil, maksatlı örnekleme yöntemiyle çalışılır. Araştırmaya sadece ilgili olguyu yaşamış olan kişiler katılır. Araştırmacı ilgili kişiye veya kişilere mülâkat çözümlerinin doğrulmasını yapmak için belirli mekanizmalar geliştirmeli, gerekiyorsa kendisiyle seri mülâkatlar yapmalıdır. Olgu bilim çalışmalarında belirli bir hipotezden yola çıkılmaması nedeniyle araştırmacı incelediği olgunun "esasını" ortaya çıkarmak için ilgili kişide "açıklık" ve "isteklilik" duygusu yaratmalıdır. Titiz ve dikkatli bir çalışma yapılabilmesi için *parantezleme* yaklaşımıyla belirli bir zaman dilimi üzerinde odaklanılır.[28] Niteliksel araştırmaların geçerlilik ve güvenilirliğinde

temel alınan kriterler olgu bilim araştırmaları için de geçerlidir. Bu tür çalışmalarda doğruluğu ve dayanıklılığı sağlamak için aşağıdaki konulara dikkat edilir:[29]

1. Yanıtlayıcının verdiği cevaplar dikkatli bir şekilde incelenerek bunların *deneyimler* mi yoksa *teorik bilgiler* mi olduğuna bakılır.
2. Bant çözümleri ilgili kişiye okutularak "gerçek deneyimlerini" yansıtıp yansıtmadığı sorulur ve doğrulaması yapılır.
3. İncelenen olgu hakkında *negatif* bir tanımlama yapılarak cevaplayıcının samimiyeti ve içtenliği soruşturulur.

Olgu bilim araştırmaları, "nedir, neler oldu, neler yaşadın?" sorularının cevabını almaya yönelik olarak yapılan derin bir iç görü ve iç deneyimin aktarıldığı çalışmalardır. Kişinin yaşantılarını doğru ve yansız bir biçimde aktarabilmesi için her hangi bir dış etkiden, tehditten veya korkudan uzak olması gerekir. Araştırmacının rolü sadece sözel ifadeleri değil, beden diline ilişkin ip uçlarını da okuyucularına aktarmaktır. Olgu bilim incelemelerinde araştırmacı aktarılan bilgileri tekrar tekrar okuyarak gerçeği yalın bir şekilde ortaya çıkarmaya çalışır.

Fenomenolojik araştırmalar büyük ölçüde etnografya araştırmalarıyla çakışma göstermesine karşılık bazı olgu bilim araştırmacıları etnografya araştırmacılarından farklı olarak kendilerinin "insan bilincini oluşturan simgesel anlamları araştırdıklarını" belirtmişlerdir.[30]

TEMELLİ KURAM ARAŞTIRMALARI

Sahadan toplanan verilere ve yaşanan deneyimlere dayalı olarak kuram geliştirmeyi hedefleyen çalışmalardır. 1960'lı yıllarda Glaser ve Strauss tarafından geliştirilen bu yaklaşımda araştırmacı başından geçen bir dizi deneyime bağlı araştırma sorularından hareket ederek bir sonuca ulaşmaya çalışır. "Temelli kuram" şeklindeki ifadelendirme biçimi, sahada yapılan gözlem ve mülâkatlara dayalı olması nedeniyle Glaser ve Strauss tarafından önerilmiştir. 1998 yılında Strauss ve Corbin güvenilirlik ve geçerlilik kaygılarıyla önceden belirlenmiş kategorilere dayanan "öngörücü temelli

kuram" yaklaşımını geliştirmişler ve bunu 2000 yılında Charmaz'ın tanıttığı "yapısalcı temelli kuram"[a] yaklaşımı izlemiştir.

Glaser ve Strauss *temelli kuram* yaklaşımını, simgesel ve anlamsal etkileşim[b] perspektifiyle sosyal fenomeni inceleme yöntemi olarak değerlendirmişlerdir.[31] Temelli kuram, önceden inceleme yapılmamış bâkir araştırma alanlarında uygulanır. Örneğin, dünyanın diğer uluslarıyla karşılaştırıldığında Türk insanının iş yerlerinde çalışırken birbirlerine veya büyüklerine "abla" ve "ağabey" şeklinde hitap ettikleri görülür. Bu hitap şekli Türk işletmelerinde istisnaî bir uygulama değildir. Özel ve kamu tüm Türk işletmelerinde gördüğümüz bu davranış biçiminden yola çıkarak bir kuram geliştirebilir ve bu kuramın adını da "Türk tipi etkileşim" olarak belirleyebiliriz. Temelli kuram araştırmalarında toplanan verilerin başlangıçta belirlenen bir kuramı doğrulayıp doğrulamadığı araştırılmaz. Bilim adamı gözlem, mülâkat, anket verilerini gözden geçirerek ve süreç içinde araştırma sorularını değiştirerek kafasındaki bir sorunu çözmeye çalışır.

Araştırma Süreci

Temelli kuram araştırması yapmak isteyen bilim adamı kafasındaki belirli bir sorundan hareket eder veya gözlemlerine dayalı olarak belirli bir inceleme süreci içine girer. Başlangıçta konu hem geniş hem de büyük ölçüde belirsizdir. Süreç içinde konu daraltılarak belirli boyutlarda ele alınmaya başlanır. Temelli kuram soyut fikirler yerine pratik faaliyetler, davranışlar veya eylemler üzerinde odaklanır. Bilim adamının amacı soruna, olguya bir açıklama getirmek ve çalışmanın sonucunda bir teori ortaya çıkarmaktır. Araştırmacı bu süreç içinde olguyu açıklarken neyin önemli olduğunu ve neyin ise önemsiz olduğunu dikkatli bir düşünceyle

[a] Yapısalcı Temelli Kuram (Constructivist Grounded Theory). Charmaz, temelli kuram yönteminde objektivist ve yapısalcı yaklaşımlar arasında ayrım gözetir. Objektivist yaklaşımda dış dünyanın somut gerçekleri göz önünde bulundurularak tarafsız bir gözlemcinin bu gerçekleri göreceği varsayımından hareket edilir. Bağımsız gözlemci, incelediği verileri niteliğine bağlı olarak gruplandırır ve belirli boyutlar altında toplar. Yapısalcı yaklaşımda ise veriler kendi kendilerine konuşmaz. Tam tersine, araştırmacı veya gözlemci gördükleriyle etkileşim içine girerek verileri kendisi yaratır. Veriler zihinsel faaliyetin/inşanın sonucudur.

[b] Simgesel etkileşim, sosyolojide önemli bir kuramsal bakış açısıdır. Bu bakış açısında simgelerin ve kelimelerin anlamları üzerinde durulur. Simgesel etkileşimde kişinin benlik tasarımı diğer kişilerin tepkisiyle belli olur. Bu yaklaşıma göre toplumlar gerçekte bireylerin etkileşimleri sonucunda oluşur. Yaklaşım "yapı" ve "süreklilik" üzerinde değil, bireylerin sosyal konumları ve diğerleriyle yaptığı davranışsal sözleşmeler üzerinde odaklanır. Simgesel etkileşimde araştırmacının ilgi odağı yüz yüze yapılan bireysel etkileşimlerdir. Etkileşim sosyologları, veri toplamak için anket ve mülakat yerine katılımlı gözlem yöntemini tercih ederler. Bu konuda daha fazla bilgi için *bk.*, <http://web.grinnell.edu/cours es/ soc/ s00/ soc 111 -01/IntroTheories/Symbolic.html> (23.02.2004).

ve neyin ise önemsiz olduğunu dikkatli bir düşünceyle sezgilerine dayalı olarak belirler. Bunu yaparken ön yargılarından, eğilimlerinden ve kişisel varsayımlarından sıyrılarak olguyu objektif bir biçimde ele almaya çalışır.[32] Temelli kuram geliştirme çalışmalarında kendileriyle görüşme yapılan örneklem büyüklüğünün en az 30-50 kişiden oluşması gerektiği belirtilmiştir.[33] Chang'a göre (2004) temelli kuram geliştirme çalışmaları belirli aşamalarda gerçekleştirilir ve bu aşamalar aşağıdaki gibidir.[34]

- Birinci aşama.

1. Konuyla ilgili bir araştırma örneklemi belirlenir.
2. Örneklemde doygunluk[a] sağlanıncaya kadar gözlem, yapılandırılmamış mülâkat çalışması ve kapsamlı bir literatür taraması yapılır.
3. Gözlem, mülâkat ve literatür taramasına dayalı olarak bir hikaye yazılır ve bu hikayede yazılan her sayfa dikkatli bir şekilde okunur.
4. Daha sonra hikayedeki fikir ve kavramlar üzerinde düşünülür.
5. Orijinal düşünce ve kavramların bir listesi çıkarılır.

- İkinci aşama.

1. Düşünce ve kavramların arkasındaki temel fikirler nedir, sorusu sorulur.
2. Temel fikirler literatürdeki bulgular da dikkate alınarak anlamlı bir şekilde gruplandırılır, kategorize edilir.

- Üçüncü aşama.

1. Konuyla ilgili başka bir uzman bulunarak bu kişiye aynı hikaye anlatılır. Birinci ve ikinci adımı bağımsız bir şekilde kendisinin tekrarlaması istenir.
2. Temelli kuramın belirli bir güvenilirliğe sahip olması için diğer uzmanların da kodlama işlemi sonucunda benzer ana fikirlere, temalara, kategorilere ulaşmaları gerekir.

---

[a] Veri doygunluğu. Toplanan verilerin yeni bir tema veya konuyu ortaya koymaması.

■ Dördüncü aşama

1. Diğer uzmanın hikayede geçen *başlıklar*, *alt başlıklar* (temalar) ve kategoriler konusundaki görüşleri alınır ve bu başlıkları onaylayıp onaylamadığı sorulur.
2. Diğer uzmanın araştırmada kullanılan *kavramları* onaylayıp onaylamadığı araştırılır.

■ Beşinci aşama.

1. Metin baştan itibaren tekrar gözden geçirilir ve ana fikrin sunuluş ve savunma biçiminin ne ölçüde güçlü olduğu araştırılır.
2. Bağımsız kişiden veya uzmandan da aynı şeyi yapması istenir.
3. Bu aşamada gerekiyorsa ilgililerle tekrar mülâkat yapılır.

■ Altıncı aşama.

1. Araştırmacının düzenlemesi ile diğer uzmanın düzenlemesi (kategoriler, kavramlar ve konu başlıkları) büyük ölçüde çakışmıyorsa (gözlemciler arası değerlendirme güvenilirliği ,70'ten küçükse) temanın veya belirlenen temel boyutların taraflar açısından tam net anlaşılmadığına karar verilir.
2. Açık ve net olmayan temalar, alt başlıklar, boyutlar yeniden gözden geçirilir, ilgililerle tartışılır; kavramlar ve boyutlar yeniden belirlenir.

Görüldüğü gibi temelli (gerekçeli) kuram geliştirme çalışmalarında güvenilirliği sağlamak için araştırmacıdan bağımsız olarak başka kişilerin de aynı kavşak noktasında buluşmaları önem kazanmaktadır. Gözlemciler arası değerlendirme güvenilirliği esas olarak ana boyutlarda yapılır. Alt boyutlar için de gözlemciler arası değerlendirme güvenilirliği yapmak yöntemi oldukça karmaşık ve uzun bir süreç haline getirebilir.

Analiz

Temelli kuram çalışmalarında verileri değerlendirmek ve yorumlamak için (a) boyut analizi, (b) açık kodlama, (c) eksenli kodlama ve (ç) seçici kodlama yöntemleri uygulanır. Açık kodlamada verilerden çıkan başlıklara ve

ve kavramlara göre kodlama yapılır. Araştırmacı daha önceden belirlediği kavramlara veya ön kabullere göre hareket etmez. Eksen kodlamasında, belirlenen kategoriler ve alt kategoriler arasında bağlantılar kurulur. Böylece teorik çatı daha derinlemesine incelenmiş olur. Seçici kodlamada ise odak kategori ile ilgili kategoriler arasında yapısal ilişki kurulur. Açık kategorilerden biri odak kategori olarak kabul edilip bu kategorinin ortaya çıkmasına neden olan koşullar, ara değişkenler, çevresel faktörler ve ortaya çıkan sonuçlar yapısal bir bütünlük içinde ele alınır.[35] Bu yöntemlerin uygulama biçimleri için okuyuculara ilgili literatüre başvurmalarını öneririz.

Temelli kuram, tümevarımsal bir yaklaşım olmakla birlikte bilgi vericilerin deneyimlerine dayalı olarak bu süreçte endüksiyon ve dedüksiyon yöntemlerinin her ikisi de kullanılabilir. Boyut analizi veri toplama işlemiyle birlikte eş zamanlı olarak yapılır. Bu süreçte araştırmacı bant çözümlerini alır, doğrulamasını yapar ve bilgileri belirli boyutlar altında toplar veya kategorileştirir. Her bir boyut olguyu açıklayan soyut kavramlar şeklindedir. Verileri belirli boyutlar altında toplamanın amacı kavramsal yapının genişliğini görmektir.[36]

Glaser ve Strauss tarafından 1960'lı yıllarda geliştirilen yaklaşım 1980'lı yıllarda yaygınlaşmış ve teknikle ilgili olarak iki sorun gündeme gelmiştir. Bunlardan birincisi kodlama sorunu ve ikincisi ise, katılımcıların sesi veya görüşü sorunudur. Glaser geliştirdiği kod-gösterge modelinde daha çok ikinci sorunla ilgilenmiştir. Strauss ise, kodlama paradigması modelinde araştırmacıların kodlama sorunlarına ayrıntılı bir şekilde açıklama getirmiştir.

Temelli Kuram ve Güvenilirlik

Tüme varım yöntemiyle üretilen kuramın güvenilirliği tekrarlanabilirlik özelliğiyle ölçülür. Eğer başkaları tarafından da aynı veri malzemesi kullanıldığında benzer kategoriler elde ediliyorsa temelli kuram güvenilirdir. Bunu sağlamak için gözlemciler arası güvenilirlik değerlendirmesi yöntemine başvurulur. Ancak verileri sınıflandırma veya kategorize etme tutarlılığı tek başına yeterli olmayabilir. Temelli kuram yaklaşımı üzerinde bazı değişiklikler yaparak genişleten Haig, güvenilirliğin veriler üzerine değil fenomen üzerine bina edilmesi gerektiğini savunmuştur. Ona göre araştırmacı sosyal olgu üzerinde odaklanmalı ve bu olguyu doğru bir şekilde kategorize etmeye çalışmalıdır. Haig'e göre verilerin güvenilirliği söz konusu fenomenin var olduğunu gösterir. Ancak, fenomenle ilgili iddiaların haklı görülebilmesi için yapılan açıklamalar arasında belli bir mutabakatın olması gerekir. Tek başına güvenilirlik yeterli değildir. Yapılan açıklama-

lar tutarlı, bağlantılı ve fenomeni net bir şekilde ortaya koyuyor olmalıdır (aktaran, Kinach).[37]

Güvenilirliği artırmak için başvurulacak bir diğer yaklaşım, gözlem yapılan her bir vak'ada üçleme yöntemine başvurmak ve elde edilen verileri ek belgeler ve kanıtlarla desteklemektir. Üçleme yönteminde bazı tutarsızlıklarla karşılaşılmışsa araştırmacı bu konuda "yansıtıcı" bir tutum içinde olmalı, tutarsızlığın nedenlerini araştırmalı ve düşüncelerini okuyucularla paylaşmalıdır.

ETNOGRAFYA ARAŞTIRMALARI

Kültürel antropolojinin temel araştırma yöntemi olan *folklor ve etnografya* araştırmalarında belirli gruplar veya kültürler bir süre incelenerek, gözlemde bulunarak grup üyeleriyle mülâkatlar yapılarak söz konusu grupların veya kültürlerin temel özellikleri saptanır. Etnografik çalışmalarda araştırma soruları genellikle kültür ve davranış arasında bir bağ kurularak başlar ve daha sonra zaman içinde kültür-davranış ilişkilerinin nasıl bir seyir gösterdiği incelenir.[38] Etnografya araştırmaları başlangıçta antropolojik bir çalışma türü sayılırken son yıllarda yönetim araştırmalarında da kullanılmaya başlanmıştır. Etnografya araştırmalarını yapan kişilere "alan araştırmacıları" adı verilir. Alan araştırmacıları inceledikleri grupla bir süre birlikte yaşayarak onların davranışlarını, düşüncelerini, değer yapılarını anlamaya çalışırlar. Bu süreçte katılmalı gözlem, mülâkat ve belge toplama yöntemlerini uygulayarak onlardan biri gibi hareket ederler. Kendi kavramsal çatılarını bir kenarda tutarak olguları inceledikleri grubun bakış açısından yansıtmaya çalışırlar. Bunun için siteye ait belgeleri toplarlar, site yöneticilerinin ve site halkının görüşlerini derlerler. Etnografya araştırmalarında anket uygulama ve sonuçları yüzdesel olarak açıklama gibi yöntemlere başvurulmaz. Etnografya araştırması yapmayı düşünen bir araştırmacı bu amaçla en az 30-50 kişiden oluşan bir örneklem grubuyla çalışmalıdır.[39] Etnografik araştırma türleri literatürde değişik başlıklar altında sınıflandırılmıştır. Burada bir fikir vermesi açısından Garson (2004) tarafından yapılan sınıflandırma temel alınmıştır:[40]

1. *Makro etnografya araştırmaları.* Geniş kapsamlı kültürel grupları inceleyen araştırmalar.
2. *Mikro etnografya araştırmaları.* Dar kapsamlı kültürel grupları inceleyen araştırmalar.

3. *Emik perspektifli araştırmalar.* İncelenen grup üyelerinin kendi bakış açılarından kendilerini tanıttığı çalışmalar.
4. *Etik perspektifli araştırmalar.* Araştırmacının kendi bakış açısından kültürel bir grubu tanıttığı çalışmalar.

Etnografya araştırmalarında güvenilirlik, objektif olma ve sadece gerçeği yansıtmayla sağlanabilir. Bunun için bu tür çalışmalarda bilgiler araştırmacının saptadığı kişilerden değil grubun belirlediği kişilerden alınır. Bununla birlikte antropologlar uzun zamandan beri etnografik araştırmaların objektif nitelikli araştırmalar olmadığını kabul etmişlerdir.[41] Objektif nitelikte çalışma olamayacağına dayanak olarak da şu tezleri ileri sürmüşlerdir.[42]

1. Etnografya araştırmaları yoruma dayanır.
2. Bütün site alanları etnografyacılar için "yabancı" değildir. Bu nedenle de etnografyacılar bazen inceledikleri grubun "yerli" halkından biri olabilirler.
3. Etnografya araştırmaları tekrarlanamaz ve yeniden üretilmez.
4. Etnografya araştırmaları niceliksel araştırmalarda olduğu gibi çok sayıda vak'aya dayanmaz.

Etnografyacılar, etnografik araştırmalarda bir tür kültürel rölativizm düşüncesiyle hareket ederler. Kültürel rölativizm, dünyadaki kültürlerin tek bir bakış açısından değerlendirilemeyeceği anlamına gelir. Araştırmacı istemese de belirli bir şekilde "pozisyon" almaya koşullanmıştır ve bu nedenle incelediği gerçeği sadece "kısmen" görebilir. Etnografyacıların araştırmalarında "yansıtıcı" olmaları istenir. Yansıtıcılık, araştırmanın hangi koşullarda yapıldığının, verilerin nasıl toplandığının, hangi faktörlerden etkilenildiğinin okuyuculara tam olarak aktarılmasıdır. Yansıtıcılıkta araştırmacı sanki bazı itirazlara cevap veriyormuşçasına niyetini, düşüncelerini ve tepkilerini açıklar. Yansıtıcılık daha itinalı, özenli, dikkatli ve titiz bir çalışma ortaya konmasını sağlar. Araştırmacı; veri toplama, analiz etme ve yorumlamada takip ettiği yöntemi tam olarak açıklarsa, karşılaştığı güçlükleri ve kişisel değerlendirmelerini yansıtırsa okuyucular kültürel fenomeni kafalarında daha gerçekçi bir şekilde değerlendireceklerdir.

Etnografik araştırmalarda yansızlığı sağlamak ve gerçeği daha doğru bir şekilde ortaya koymak için başvurulacak bir diğer yöntem üçleme yakla-

şımıdır. Etnografyacı site belgelerinden, gözlemlerinden, literatür bilgilerinden ve mülâkat tekniklerinden birlikte yararlanmalıdır. Üçlemede başvurulacak bir diğer yaklaşım konun birden fazla sitede araştırılmasıdır. Tek bir site, tek bir grup, tek veri toplama aracı ve tek bir yöntem olguyu tanımlamakta genellikle yetersiz kalır.

Etnografik araştırmaların güvenilirliğini artıran bir diğer öge, araştırmaya ayrılan süredir. Derinlemesine mülâkat tekniğinin uygulanması ve bazı gerçeklerin zaman içinde kavranabilmesi nedeniyle araştırmacı araştırma grubuyla oldukça uzun bir süre bir arada bulunmalı, onların düşündüklerini, ne hissettiklerini anlamaya çalışmalıdır. Kısa süre içinde yapılan araştırmalar bilgi toplama ve anlama açısından yetersiz kalır.

Etnografik araştırmaların nesnelliği, iç güvenilirlik ve dış güvenilirlik terimleriyle bağıntılandırılmıştır. Nunan (1992), iç güvenilirliği araştırmacının "veri toplama, veri analizi ve yorumlamada tutarlılığı" ve dış güvenilirliği ise, "bağımsız araştırmacıların orijinal araştırmacının elde ettiği benzer sonuçlara ulaşma derecesi" olarak açıklamıştır (aktaran, Hadley, 1996).[43] LeCompte ve Goetz'e (1982) göre etnografik araştırmanın güvenilirliği projede birden fazla araştırmacı kullanılarak artırılabilir. Birden fazla araştırmacı başka sitelerde benzer yöntemler kullanarak yaptıkları araştırmalarda benzer bulgulara ulaşmışlarsa araştırma iç güvenilirliğe sahiptir. Araştırmanın dış güvenilirliğinin güçlendirilmesi ise, araştırmacının projeye katılan bireyler, araştırma süreci, araştırma koşulları, bilgi toplama amacıyla kullanılan formlar, analiz yöntemleri, sınıflandırma tipolojileri hakkında açık ve ayrıntılı bir şekilde bilgi vermesiyle sağlanabilir (aktaran Hadley, 1996).[44] Araştırmacının bu konularda dikkatli ve titiz olması, özenli bir dil kullanması, araştırmanın tekrarlanmasına imkan verecek ayrıntıları ortaya koyması halinde incelemenin dış güvenilirliğe sahip olduğu söylenir.

ODAK GRUBU ARAŞTIRMALARI

Odak grubu araştırmaları, belirli bir konuyla ilgili olarak dikkatlice seçilmiş az sayıda kişinin bir araya getirilmesi ve bu kişilerin konuyu kendi aralarında tartışmaları suretiyle konunun değişik yönlerine ışık tutmalarıdır. Derinlemesine mülâkat ve tartışma yönteminin uygulandığı bu tür çalışmalarda katılımcılar araştırılan konuyla ilgili olarak ne düşündüklerini ortaya koyarlar ve tepkilerini gösterirler. Odak grubu araştırmalarında sadece görüşler değil; tutumlar, hisler, inançlar, deneyimler ve değerler de ortaya konur.[45] Nicel araştırmalardan farklı olarak odak gurubu araştırmalarında katılımcıların bilinç altındaki sezgilerini ve duygularını ortaya çı-

karmak önemlidir.[46] Odak grubu araştırmaları aşağıdaki amaçlarla kullanılır:[47]

1. Hipotez üretme.
2. Mülâkat çizelgeleri geliştirme.
3. Temel sorunları belirleme.
4. Ölçek ve testler için madde geliştirme
5. Yeni temalar oluşturma.
6. Sayısal değerlere ışık tutma.
7. Geri besleme alma.

Odak grubu üyeleri tartışılan konuyu iyi bilen, birbirine benzer özelliklere sahip kişilerden oluşur. Uzak yerlerden gelip zaman ayırdıkları için kendilerine belirli bir ücret ödenen bu kişilerin birbirlerini tanımamaları ve aralarında ast-üst ilişkisinin bulunmaması gerekir. Araştırma konusuna göre odak grubuna katılacak kişilerin demografik özellikleri ayrıca önem kazanır. Odak grubuna bazen aynı bölgeden ve bazen de farklı bölgelerden kişilerin alınması gerekli olabilir. Bilim adamı, odak grubu çalışmalarında genellikle birden fazla odak grupla ve birden fazla seans yaparak çalışır. Örneğin, işletmesinde yeniden yapılanma çalışması yapmak isteyen bir firma sahibi üç farklı odak grubu oluşturabilir: sendikalı işçiler, sendikasız işçiler ve yöneticiler. Odak gruplarında belirli bir konu her birinde 6-12 kişi bulunan farklı gruplarda tartışılır. Konunun en az 3-5 grupta tartışılması ve elde edilen sonuçların yorumlanması olguya ait resmin daha gerçekçi bir şekilde görülmesi imkanını sağlar.[48] Bununla birlikte farklı odak gruplarının sonuçlarını rakamsal anlamda karşılaştırmak mümkün değildir.

Odak grup çalışması, grup dinamikleri konusunda eğitim almış, grubu idare etmesini bilen ve grup psikolojisinden anlayan kişiler tarafından yönetilir. Literatürde bu kişilere "moderatör" veya "facilitator" (kolaylaştırıcı) adı verilmiştir. Grubu yöneten *kolaylaştırıcı* araştırılan konu hakkında uzmanlık bilgisine sahip olmalıdır. Odak grubu seans tartışmaları, *kolaylaştırıcının* kendisinin hazırladığı soru listesine göre veya müşteri tarafından hazırlanmış olan plana göre yürütülür. Odak gruplarında ortaya çıkan veri ve bilgilerin kalitesi *kolaylaştırıcının* sorduğu soruların kalitesine ve cevaplayıcıların konuya odaklanma derecesine bağlıdır. Odak grubu araştırmalarında açık uçlu ve birden fazla yanıtı olan sorular sorulur, *evet-hayır* yanıtının verildiği kısa sorulardan ise kaçınılır. Odak grubu *seri mülâkat yöntemi* de değildir. Kolaylaştırıcı soruyu ortaya attıktan sonra katılımcıla-

rın konuyu kendi aralarında serbestçe tartışmalarına izin vermelidir.[49] Odak grubu çalışmaları iki saat sürecek şekilde planlanır ve bunun yaklaşık 90 dakikası sadece grup tartışması için ayrılır. Kolaylaştırıcı seans içinde üyelerden birine tartışmaları not alma görevi verebileceği gibi; tartışmaların teybe, videoya kayıt edilmesini veya ayna arkasından izlenmesini de sağlayabilir. Ancak bilim adamları seans içinde not alma yerine seans sonrasında rapor tutma uygulamasına gidilmesini önermişlerdir. Odak grubu tartışmalarının raporu, seanslar kapandıktan sonra kolaylaştırıcı tarafından hazırlanır. Raporda açığa çıkan en önemli temalara vurgu yapılır. Rapor tartışmacıların gündeme getirdikleri önemli konulara ilişkin öneri maddeleriyle sona erer.

Odak grubu araştırmalarının sonuçları diğer niteliksel araştırma sonuçlarında olduğu gibi daha geniş ana kütlelere genellenemez. Sonuçlar katılımcılar ve grup yöneticisi arasındaki etkileşimi yansıtır. Kolaylaştırıcının profesyonel olmaması tartışmacıları yanlış sonuçlara götürebilir.[50] Odak grubu tartışmalarında ortaya çıkan görüşleri analiz etmek, değerlendirmek ve yorumlamak oldukça zor ve karmaşık bir süreçtir. Bu nedenle diğer niteliksel araştırmaların içinde güvenilirlik ve doğruluk açısından daha zayıf bir tekniktir. Bilim adamları odak grubu araştırmalarını nicel araştırmalarının yanında bilgi toplamaya veya mevcut bilgileri derinleştirmeye yönelik ek bir araç olarak görme eğilimindedirler. Odak grubu verilerinin *çoklu veri toplama yöntemi* çerçevesinde analiz edilmesi güvenilirliği artırır. Birden fazla odak grubundan yararlanma ve birden fazla seans yapma bazı bilgilerin teyit edilmesine olanak sağlar.

VAK'A ETÜDÜ

Vak'a etüdü (vak'a incelemesi); belirli bir kişi, grup veya belirli bir kurum hakkında kendi doğal ortamlarında ayrıntılı olarak yapılan inceleme, değerlendirme ve elde edilen sonuçları yorumlama çalışmalarıdır. Vak'a etütlerinde sık sık katılımcılar kendi kendilerini anlatırlar veya kendi uygulamalarını tanıtırlar. Vak'a incelemesinde bilim adamı, evrensel gerçekleri bulmaya çalışmadığı gibi elde ettiği sonuçları belirli ana kütlelere genelleme yoluna da başvurmaz. Bu nedenle sosyal bilimlerdeki diğer araştırmalara göre daha az nesnel ve bilimsellik derecesi daha düşüktür. Vak'a araştırmaları değişik şekillerde yapılır:

1. Olay sayısı açısından
   a) Tek vak'alı araştırmalar

b) Çok vak'alı araştırmalar

2. Olayın türü açısından
   a) Resmedici vak'a incelemeleri.
   b) Keşfedici vak'a araştırmaları.
   c) Birikimli vak'a çalışmaları.
   d) Kritik olay vak'a çalışmaları.
3. Veri yapısı açısından
   a) Niteliksel vak'a incelemeleri.
   b) Niceliksel vak'a incelemeleri.
   c) Nitel ve nicel içerikli vak'a incelemeleri.

Literatürde son yıllarda çoklu vak'a uygulamalarına dayanan araştırmalar yoğunluk kazanmaya başlamıştır. Bu türünde araştırmacı tek bir kurumda/işletmede değil, herhangi bir örneklem planına dayanmadan birden fazla işletmede/kurumda inceleme yapar. Vak'a türü açısından ele aldığımızda resmedici vak'a incelemelerinin tanımlayıcı, betimleyici (tasvir edici) çalışmalar olduğunu görürüz.[51] Resmedici vak'a analizinde araştırma konusuna ışık tutma ve konuyu aydınlatma amacı güdülür. Resmedici vak'a analizleri pilot araştırma şeklinde yapılır. Keşfedici vak'a incelemeleri ise, kuram oluşturmaya veya belirli bir kuramı test etmeye yönelik olarak yapılır ve genellikle alan araştırması şeklinde gerçekleştirilir. Bazen bu tür vak'a analizlerinde neden-sonuç ilişkilerinin de inceleme konusu yapıldığı görülmüştür. Birikimli vak'a çalışmaları, farklı yer ve zamanlarda yapılan bilgilerin toplanmasına dayalı olarak yapılan çalışmalardır. Daha uzun süreyi ve daha kapsamlı çalışmalar yapılmasını gerektirir. Kritik olay vak'a çalışmaları, belli bir olay veya olgu üzerinde odaklanan incelemelerdir. Bir işletmedeki bir yönetim sorununu, personelin devamsızlık gibi bir davranış sorununu ele alarak bu davranışın nedenlerini inceleyen, araştıran ve yorumlayan çalışmalardır.

Vak'a incelemeleri niceliksel veya niteliksel olabilir ve çoğunlukla da nitelikseldir.[a] Bazı vak'a araştırmaları ise karma bir niteliğe sahiptir. Nitel yorumları ve tanımlayıcı nitelikteki nicel verileri birlikte içerir. Hatta keşfedici vak'a incelemelerinde sonuç çıkarıcı istatistiksel analizleri uygulamak dahi mümkündür.

---

[a] Bir vak'a araştırmasında toplanan verilerin niteliksel ve niceliksel yönü eşit ağırlığa sahipse bu araştırmaları niteliksel araştırma olarak isimlendirmek doğru olmaz. Bu tür çalışmaları sadece "vak'a analizi" olarak isimlendirmek yeterlidir.

Avantaj ve Dezavantajları

Vak'a araştırmalarının kendine özgü avantajları bulunmasına karşılık, önemli ölçüde yetersizlikleri de söz konusudur. Avantajları; bir araştırma projesinin başlangıç aşamasında önemli ölçüde bilgi toplanmasına imkan sağlaması, başka yöntemlerle elde edilemeyecek bazı bilgilerin derlenmesine zemin hazırlaması ve bu araştırmalarla "nasıl" ve "niçin" sorularına daha sağlıklı cevaplar bulunabilmesidir.[52] Dezavantajları ise; bu araştırmalara dayalı olarak neden-sonuç ilişkilerinin kurulamaması veya kurulsa bile bu ilişkilerin daha geniş ana kütlelere genellenememesidir.

Vak'a İncelemesi Prosedürü

Vak'a etüdünde veri toplama işlemi; belge incelemesi, gözlemde bulunma, anket kullanarak alan araştırması yapma, arşiv incelemesi yapma ve görüşme yöntemlerinden yararlanılarak gerçekleştirilir. Vak'a incelemesi yapmak isteyen bir araştırmacı yöntem bilimi açısından aşağıdaki prosedüre uygun olarak hareket eder:

1. Tek veya çok modelli bir yaklaşım belirleme.
2. Katılımcıları belirleme ve seçme.
3. Veri toplama.
4. Geçerlilik ve güvenilirlik çalışmalarını yapma.
5. Veri analizi yapma ve verileri yorumlama.
6. Araştırma raporunu hazırlama.

Vak'a etüdünde veri olarak belgeler, arşiv dokümanları, işletme kayıtları, gazeteler, mülâkatlar, doğrudan yapılan gözlem bilgileri ve işletme uygulamaları kullanılır.

Vak'a Etüdü ve Güvenilirlik

Yin (1994) vak'a etüdünde güvenilirliği sağlamak için "vak'a incelemesi protokolü"nün kullanılmasını önermiştir. Araştırmacı incelediği konuyla ilgili olarak bir veri tabanı oluşturmalı bu veri tabanı içinde veri ve bilgilerin birbiriyle tutarlı olup olmadığını dikkatle incelemelidir. Araştırma raporuna kullandığı ölçüm aracını eklemeli, kimlerle hangi süre içinde görüştüğünü, vak'a çalışmasına katılan kişilerin isimlerini, yaptıkları katkıyı belirtmeli ve kullandığı diğer ikincil kaynakları açıkça ortaya koymalıdır. Böylece vak'a incelemesine ilişkin bütün bilgi ve veri kaynakları tam ola-

rak ortaya konmuş olur. Bu bilgi ve veriler daha sonraki yıllarda başka araştırmacılar tarafından yapılacak teyit etme amaçlı araştırmalara bir baz oluşturacaktır.[53]

Bilim adamı vak'a araştırmasında bir ölçek kullanmışsa bu ölçeğin iç tutarlılığını tek bir kurumda veya birden fazla kurumda çalışan personel üzerinde sınayabilir. Ancak örneklem hacmi yeterince büyük ve kişiler heterojen bir dağılma sahip değilse bu analiz sonucunda elde edilecek yüksek alfa değerleri çok fazla bir anlam ifade etmez. Gerçek iç tutarlılık değerleri; ana kütleyi temsil eden, yeterli bir örneklem hacmine sahip örnek kütlelerden elde edilebilir.

Nitel araştırma birden fazla vak'a üzerinde yapılmışsa her bir vak'adan elden edilen sonuçların benzer olup olmadığına bakılır. Her bir vak'adan elde edilen sonuçlar aynı veya büyük ölçüde benzer çıkmışsa sonuçların güvenilir olduğuna karar verilir. Sonuçlar her bir vak'ada farklı çıkmışsa tartışmalıdır ve daha az güvenilir olarak değerlendirilir.

Vak'a etüdü protokolü. Yin (1994, aktaran Telis, 2003) vak'a incelemesi araştırmalarında önceden yazılı bir "protokol"[a] belirlenmesini güvenilirliği sağlayacak en önemli bir öge olarak görmüştür. Ona göre vak'a incelemesi katı bir biçimde önceden belirlenen protokole uygun olarak yapılmalıdır. Yin, tipik bir vak'a incelemesi protokolünü aşağıdaki gibi belirlemiştir:[54]

1. Vak'a incelemesi projesinin genel çerçevesinin çizilmesi (amacı/amaçları, ele aldığı sorunlar ve araştırılan konular).
2. Alan prosedürleri (ilgili birimde/bölümde/kurumda araştırma yapma gerekçeleri, araştırmayı haklı kılan nedenler, kanıtlar ve bilgi kaynakları).
3. Vak'a incelemesi soruları (araştırmacının veri toplama süreci içinde cevabını almak istediği sorular).
4. Vak'a incelemesi raporu için bir yazım rehberine sahip olunması (çalışmanın planı, nakledici/rivayet edici bir anlatım tarzı).

[a] Vak'a incelemesi protokolü. Vak'a etüdü yapacak araştırmacının vak'a çalışmasını hangi aşamalarda gerçekleştireceğini, hangi sorularla, kimlerle görüşme yaparak ve hangi konularda gözlemde bulunacağını belirlediği yazılı bir plan ortaya koymasıdır. Vak'a incelemesi protokolü nicel araştırmalardaki *araştırma önerisine* benzer, sadece içeriği farklıdır.

Telis'in (2003) Stake ve Yin'den alıntı yapıp bildirdiğine göre vak'a incelemelerinde bir araştırmacı en az altı kaynağı kanıt olarak gösterebilir. Bunlar; belgeler, arşiv kayıtları, doğrudan gözlem kayıtları, katılmalı gözlem kayıtları, mülâkatlar ve fizikî nesneler/eşyalardır.

■ Vak'a incelemesi protokolü örneği.

1. Vak'a çalışmasının amacı.
2. Vak'a etüdünde cevabı aranan sorular.
3. Kişilerin nasıl seçileceği, soruların nerede, hangi süreyle ve nasıl sorulacağı.
4. Kişilerin demografik bilgileri.
5. Her bir mülâkat için kaç dakika ayrılacağı ve kişilere sorulacak sorular. Mülâkat formu örneği. (Duruma göre birinci, ikinci, üçüncü mülâkat formları kullanılabilir).
6. Gözlemde bulunulacak ögeler (ögelerin maddeleştirilmiş listesi).
7. Veri ve bilgileri toplamak için kullanılacak teyp, video kamera ve film çekim cihazları.
8. Bulguların kayıt edileceği formlar, aynen alıntıların tırnak işareti içinde gösterilmesi, bant çözümlerinin nasıl yapılacağı.
9. Yorumlama ve değerlendirmede takip edilecek yöntem. Çoklu yöntem yaklaşımının benimsenmesi, kayıtların aynen değil özet olarak raporlanması.

Vak'a incelemesi protokolünün hazırlanması vak'a çalışmasının güvenilirliğini artırmak için kullanılan temel taktiklerden biridir. Böylece araştırmacı verileri hangi koşullarda, kimlerden ve hangi sorularla toplayacağını belirgin hale getirmiş olur. Belirginlik sınamaya ve test etmeye imkan verir. Araştırmacı protokole uygun olarak sorduğu soruları birden fazla kişi üzerinde test etme ve doğrulama imkanına sahip olur. Protokol, sonuçların sistematik ve analitik bir şekilde raporlanması sonucunu doğurur.

Veri tabanı oluşturma. Güvenilirliği sağlayacak ikinci yaklaşım vak'a etüdü *veri tabanının* oluşturulmasıdır. Araştırmacı vak'a etüdü çalışmaları sırasında topladığı tüm dokümanları organize etmeli, sınıflandırmalı, ayrı

ayrı dosyalamalı ve böylece bir veri tabanı oluşturmalıdır. Veri tabanı iki bölümden oluşur: birincisi orijinal veriler, ikincisi araştırmacının yorum ve değerlendirmeleri. Veri tabanı araştırmacının inceleme planı çerçevesinde dinamik bir şekilde oluşturulur. İnceleme planına göre belgeler ile film ve teyp bantları farklı dosya ve çantalara konabilir. Önemli olan dış bir gözlemcinin kuşku duyduğu konularda veri tabanını inceleme ve değerlendirme şansına sahip olmasıdır.

Kanıtlar zinciri oluşturma. Güvenilirliği sağlayacak üçüncü yaklaşım vak'a etüdü konusuyla ilgili bir kanıtlar zinciri oluşturmaktır.[55] Kanıtlar araştırma sorusunun başlangıç noktasından araştırmanın sonlandırılmasına kadar geçen süre içinde toplanan, kullanılan, konuşulan, gözlenen tüm bilgi ve verileri kapsar. Kanıtlar altı başlık altında toplanmıştır:[56]

1. Belgeler.
2. Arşiv kayıtları.
3. Mülâkat kayıtları.
4. Dolaysız gözlem sonuçları.
5. Katılmalı gözlem sonuçları.
6. Fiziksel nesneler ve eşyalar.

Kanıtlar, dış gözlemcinin vak'a çalışmasında ileri sürülen iddiaların veya görüşlerin nesnelliğini izlemesine olanak sağlar. Vak'a etüdü raporu kanıtlarla birlikte meslektaşlara, danışman öğretim üyesine veya hakemlere gönderilerek bağımsız bir şekilde incelemeleri ve kanaatlerini belirtmeleri istenir. Bağımsız gözlemcilerin görüşleriyle araştırmacının görüşleri örtüştüğü ölçüde çalışmanın güvenilir olduğuna karar verilir.

KRİTİK OLAY İNCELEMELERİ

Kritik olay tekniği John Flanagan tarafından; bireylerin ve örgütlerin belirli durumlarda başarı veya başarısızlıklarına neden olan davranışları gözlemlemek, belirlemek veya işletmede yaşanan kritik olay raporlarını inceleyerek hatalı davranışları analiz etmek amacıyla geliştirilmiştir.[57] Kritik olay tekniğinde araştırmacı, kişilere olaydaki etkili ve etkili olmayan davranışların neler olduğunu sorar ve bu davranışları kendilerinin yaşadıkları örneklerle ilişkilendirmelerini ister. Kritik olay analizi sözlü veya yazılı olarak yapılır. Bu analizde kişilerden aşağıdaki sorulara yanıt vermeleri istenir.

1. Bu olayın ortaya çıkmasına neden olan faktörler nelerdir?
2. Olayda hangi davranışlar doğru ve hangi davranışlar ise yanlıştı?
3. Bu olayda gösterdiğiniz davranışın sonuçları ne oldu?
4. Bu olayda gösterdiğiniz davranış niçin etkiliydi?
5. Sizden hangi davranışı göstermeniz bekleniyordu?

Kritik olay analizi gözlemlere veya tutulan raporlara dayalı bir yorumlama ve değerlendirme çalışmasıdır. Kritik olaylar çoğunlukla içerik analizi yöntemiyle etüt edilir. Araştırmacı metni inceleyerek metnin içinde belirli kategorileri ortaya çıkarmaya çalışır. Bu kategoriler yüzey ve yapı geçerliliğine sahip olmalıdır. John Flanagan ABD'de II. Dünya Savaşı sırasında sık düşerek başarısız olan pilotların durumunu incelemek için bu yöntemi kullanmıştır. Başarılı uçuş yapılan vak'alarla başarısız uçuş yapılan vak'aları ayrı ayrı ele almış ve kayıtları inceleyerek, pilotlarla mülâkatlar yaparak yanlış davranışları belirlemeye çalışmıştır.

Kritik olay çalışmalarında geçerlilik ve güvenilirlik konusuyla ilgili olarak başlıca aşağıdaki sorulara yanıt bulunmaya çalışılır:[58]

1. Geçerlilikle ilgili olarak toplanan vak'alar, beklenen davranış örneklerini temsil ediyor mu?
2. Veri toplama yöntemi toplanan verileri etkiliyor mu?
3. Güvenilirlikle ilgili olarak veriler başka bir kodlayıcı tarafından kodlandığında benzer çıkıyor mu?
4. Makul herhangi bir kodlayıcı, vak'aları beklenen sınıflar veya kodlar içinde gruplandırır mı?

Araştırmacı kritik olay analizinde belirli bir davranışın yanlış veya doğru olduğunu kanıtlamak için yüzlerce kritik olayı inceleyebilir veya raporlarını gözden geçirebilir. Örnekleme alınacak vak'aların incelenen davranışla doğrudan ilgili olması gerekir. Araştırmacı inceleme konularının her biri için belirli sayıda olayı incelemeyi planlayabilir. Buradaki temel sorun kaç kategori olduğu ve araştırmacının her kategoride kaç olayı inceleyebileceğidir. Olay sayısı, yeterince doygunluk verecek kapsamda olmalıdır.

EYLEM ARAŞTIRMALARININ GÜVENİLİRLİĞİ

Eylem araştırması ilk kez 1950'li yıllarda Kurt Lewin tarafından grup davranışını ve sosyal değişimi gözlemek için kullanılmıştır. Bilim adamları bu yöntemi bir araştırma yöntemi olmaktan çok bir yaklaşım biçimi olarak değerlendirmişlerdir.[59] Eylem araştırması, bilim adamının belirli bir grup, kurum veya işletme üzerinde gerçekleştirdiği çok aşamalı bir proje çalışmasıdır. Çok aşamalı süreçte, olay ve olgular incelenirken *spiral döngü* yaklaşımı uygulanır. Yaklaşımda veri toplama, verileri analiz etme ve yorumlama evreleri, sonuçları *değerlendirme* evresiyle birlikte ele alınır. Söz konusu çalışma sırasında bir taraftan sorunlar saptanır ve diğer taraftan saptanan sorunların giderilmesine yönelik sistem, proses veya prosedürler üzerinde bazı değişikliklere gidilir. Değişiklikler izlenerek gözden kaçan veya yan ürün olarak ortaya çıkan yeni sorunlar araştırılır ve ortaya çıkan bu yeni sorunları da gidermek için ek önlemler alınır. Eylem araştırması; sorunları saptama, değişiklik programlarını uygulamaya alma ve duruma göre ortaya çıkan yeni sorunlarla ilgilenme çalışmasıdır. Araştırmacı bu süreç içinde sadece uygulayıcı değil, aynı zamanda elde ettiği sonuçları okurlarına aktaran bir raportör durumundadır.[60]

Eylem araştırması da vak'a araştırmaları gibi sınırlı sayıda birim / bölüm veya kurum üzerinde yapılır, ancak vak'a araştırmalarından farkı bilim adamının sadece gözlemci / değerlendirici olmaması, bunun yanında değişikliği / deneyi / müdahaleyi / testi yapan kişi olmasıdır. O nedenle literatürde bazı yazarlar bu araştırma türü için "Katılmalı Eylem Araştırması" (KEA) terimini kullanırlar.[a] Katılmalı eylem araştırmalarında yeni bir plan, yöntem, prosedür, uygulama veya teknoloji denenir. Araştırmacı hem ölçen hem de süreci uygulayan ve sürece katılan bir kişidir. Araştırmacının uygulamada aktif bir rolü vardır. Eylem araştırmaları profesyonel araştırmacılardan çok okul öğretmenleri, işletme yöneticileri, teknisyenler, danışmanlar gibi pratisyenler tarafından gerçekleştirilir. Eylem araştırmaları da vak'a araştırmaları gibi niceliksel veya niteliksel olabilir, fakat çoğunluksa nitelikseldir. Eylem araştırmalarının pratik bir takım sonuçlar ortaya koyarak akademisyenlerle uygulayıcılar arasındaki açığı kapatma gibi bir takım yararları olmasına karşılık, bu yaklaşımı eleştiren bilim

[a] Eylem araştırmaları için tek bir tür veya uygulama biçiminden söz edilemez. Bazı eylem araştırmalarının katılım olmadan da yapılabileceği belirtilmiştir. Böyle bir uygulamada eylem araştırması vak'a etüdüne benzer. Bazı bilim adamları eylem araştırmalarında uygulayıcının rolünü tam katılımdan hiç katılımın olmadığı bir boyut üzerinde belirlerler. Nitel araştırma literatürü incelendiğinde pek çok eylem araştırmasının aslında vak'a etüdü olarak nitelendirilebileceği görülür.

adamları yöntemin yeterince akademik olmadığı görüşünü dile getirmişlerdir.[61]

Eylem Araştırması Süreci

Eylem araştırması yöntemini uygulamaya karar veren bir araştırmacı bu yaklaşımı yedi temel adımda gerçekleştirir. Bu adımlar aşağıdaki gibi belirlenmiştir:[62]

1. Problemi tanımlayınız.
2. Kuramsal bilgileri, diğer görüş ve verileri toplayınız.
3. Ölçüm araçlarını ve kıyaslama kriterlerini belirleyiniz.
4. Müdahale edilecek yerleri ve müdahale türünü belirleyiniz.
5. Müdahaleyi yapınız.
6. Müdahale sonucunu değerlendiriniz.
7. Problemli verileri yeniden ele alınız.

Bu süreçte birinci adım araştırma sorunun açık ve net bir şekilde ortaya konmasıdır. Nicel araştırmalarda olduğu gibi sorun, "problem cümlesi" olarak belirlenir. Problem cümlesi genel veya oldukça spesifik bir şekilde saptanabilir. İkinci adım, araştırma problemiyle ilgili olarak literatürün araştırılmasıdır. Bilim adamı önceki araştırmaları, konuyla ilgili kaynakları tarayarak konunun ögelerini analitik bir yaklaşımla ele alır. Bu aşamada ayrıca işletmede veya ilgili kurumda konu hakkında bilgi sahibi olan kişilere başvurulur, onların da düşünce ve fikirleri alınır. Üçüncü aşamada araştırmanın varsa hipotezleri ve araştırma soruları belirlenir. Dördüncü aşamada veri ve bilgi toplamak için kullanılacak ölçüm aracı, araçları saptanır. Bunlar anketler, testler, gözlem veya mülâkat teknikleri olabilir. Beşinci aşamada örneklem grubunu belirlemeye yönelik çalışmalar yapılır. Altıncı aşamada araştırma tasarımına karar verilir. Araştırma tasarımı tanımlayıcı, sonuç çıkarıcı veya sadece ilişkileri belirlemeye yönelik olabilir. Bilim adamı, araştırmayı bir *deney serimi* şeklinde de gerçekleştirebilir. Bu yönteme başvurmuşsa deney ve kontrol grupları oluşturarak bu gruplarla çalışır. Bu aşama aynı zamanda müdahalenin yapıldığı evreyi oluşturur. Yedinci aşamada veriler analiz edilir ve araştırma raporu yazılır.

Müdahalenin ortaya koyduğu sonuçların doğru bir şekilde değerlendirilebilmesi için; başlangıç aşamasında süreç içinde ve süreç sonunda ölçümler yapılarak gelişmeler sürekli olarak izlenir. İzlemenin amacı kişiden, programdan ve ara değişkenlerden kaynaklanan bozucu faktörlerin etkisini en aza indirmektir.

Güvenilirlik Değerlendirmesi

Eylem araştırmaları çoğunlukla bilgi birikimine katkı sağlamak amacıyla yapılmaz. Bu tür çalışmalarda güvenilirlik ve geçerlilik kaygıları her zaman ikinci planda kalır. Çünkü bulgular spesifik bir çerçeve için geçerlidir. Araştırmacı ve uygulayıcıların amacı, inceleme yapılan grup veya kurumda pratik bir yararı elde etmek ve hemen bir sonuca ulaşmaktır.[63] Bununla birlikte belirli bir kalite çerçevesinde yapılabilmesi için okurlar bu tür çalışmalarda da geçerlilik ve güvenilirlik koşullarının sağlanmış olmasını arzu ederler.

Eylem araştırmalarının güvenilirliği, toplanan verilerin güvenilirliğiyle ilgilidir. Veri toplama aracı olarak anket, gözlem veya mülâkat yöntemi kullanılmışsa bu yöntem araçları kullanılarak yapılacak daha sonraki ölçümlerde de benzeri sonuçların alınması çalışmanın güvenilir olduğunu gösterir. Ancak ikinci bir ölçüm yapmak her zaman söz konusu olamadığından eylem araştırması sonuçlarının kesinlik derecesi şüphelidir. Lindhult (2003) eylem araştırmalarının kalitesini "ilgililik değeri" ve "doğruluk" terimleriyle açıklamıştır. *İlgililik değeri*, çalışmanın farklı gruplar, amaçlar, örgütler ve bilimsel hedefler için bir anlam ifade etmesidir.[64] İlgililik, *uygunluk* demektir. Araştırma bulgularının günlük pratiklere ve başlangıçta ortaya konulan önermelere uygun olmasıdır. İleri sürülen kanıtlar araştırma iddiası veya önermelerinden bağımsız ise ilgililik koşulu sağlanamamış demektir. Çalışmanın *ilgililik değeri* yükse, araştırma bulguları gelecekte başka gruplara ve düzlemlere de genellenebilir. Bir araştırmacı çalışmanın *ilgililik değerini* belirlemek için aşağıdaki sorulardan hareket edebilir.[65]

1. Araştırma soruma, odaklandığım konuya veya ele aldığım probleme hangi veriler en iyi şekilde cevap verebiliyor?
2. Hangi veriler tekrar niteliğinde?
3. İddiamı destekleyen kanıtlar nelerdir?
4. Hangi bilgilere gerek yok?
5. Hangi bilgiler ilgi çekici, fakat bu çalışmada raporlamam veya ele alamam gerekmiyor?
6. Hangi bilgiler gerçeği yansıtıyor?
7. Hangi bilgiler başka kişilerin görüşlerini yansıtıyor?
8. Hangi bilgilerle aynı görüşte değilim, niçin?

Doğruluk ise, geçerlilik ve güvenilirlik konusuyla ilgilidir. Geçerlilik, araştırmanın "gerçeği" yansıtmasıdır. Güvenilirlik, ölçüm faaliyetlerinin ve kullanılan ölçüm araçlarının iyi veya mükemmel olması anlamına gelir. Ölçüm aracı ve uygulamanın mükemmel olup olmadığını ölçümlerde istikrarlı sonuçlar aldığımız zaman anlayabiliriz. Eylem araştırmasında kalite açısından bilim adamının odaklandığı nokta, *tekrarlanabilirlik* özelliği yerine özgün bir şekilde yapılan *ölçüm ve uygulama sürecinin kendisi* ise güvenilirliği değerlendirmek çok daha zor ve karmaşık bir niteliğe sahip olur. Bu tür niteliksel araştırmalarda ölçüm uygulamasını ve ölçüm aracını kalibre etmek için ikinci bir uygulama yapmak imkansız gibidir.[66]

"Ölçüm ve uygulama süreci" üzerinde odaklanan güvenilirlik çalışmalarında fenomenoloji ve etnografya çalışmalarında takip edilen güvenilirlik yaklaşımlarından yararlanılabilir. Bu çerçevede üçleme yöntemine başvurma, kanıt kütüphanesi oluşturma, meslektaş değerlendirmesi, süreci okuyuculara ayrıntılı bir şekilde açıklama (yansıtıcılık özelliği), kronolojik günlük tutma, dış denetim çalışması yapma ve dış gözlemciye değerletme uygulamalarına başvurulur.

TARİHSEL ARAŞTIRMALARIN GÜVENİLİRLİĞİ

Tarihsel araştırmalar, belirli bir zaman süreci içinde bir kurumda, toplumsal bir olguda, bir toplumda veya bir kişinin, yöneticinin yönetim tarzında ortaya çıkan değişimleri, gelişmeleri sistematik bir biçimde inceleyen ve genel trendi ortaya çıkararak olguyu netleştirmeyi hedefleyen çalışmalardır. Tarihsel araştırmaların kapsadığı zaman aralığı son birkaç yıl, birkaç dönem gibi nispeten dar veya bir iki yüzyılı kapsayacak kadar geniş olabilir. Araştırmacının yaklaşım şekli, son zaman diliminden başlangıca doğru veya bilgi toplanabilen en eski zaman diliminden günümüze doğrudur. Tarihsel araştırma yönteminin tercih edilme nedenleri aşağıdaki gibi sıralanabilir:

1. Genel eğilimi ortaya çıkarmak.
2. Geçmişteki olayların nedenlerini saptamak ve geleceğe ışık tutmak.
3. Olayı aydınlatmak.
4. Eş zamanlı olaylar, kişiler arasındaki ilişkileri veya etkileşimleri görmek.
5. Yeni politikalar belirlemek.
6. Kişilerin, kurumların ve ajansların başarılarını sergilemek.

Tarihsel araştırmalar sadece bir takım gerçeklerin, verilerin veya bulguların kronolojik olarak sıralanmasından veya nakledilmesinden ibaret değildir. Araştırmacı veri ve bulgularına dayalı olarak çıkarımlarda bulunur, nüansları ortaya koyar, olaylar ve insanlar arasındaki bağlantıları kurarak yeni bilgiler üretir.[67] Tarihsel araştırma yapmayı düşünen araştırmacılar öncelikle araştırma konusunu belirlemelidirler. Örneğin, *Türkiye'de Sendikacılığın Tarihsel Gelişimi, Türkiye'de İşçilerin Sahip Oldukları Şirketler, Holding İşletmeciliğinin Gelişimi* gibi başlıklar tarihsel bir araştırma yapıldığını gösterir. İkinci aşamada bilim adamı araştırma hipotezini, araştırma iddiasını, varsayımlarını veya araştırma sorularını belirler. Araştırma soruları kapalı kalan, bilinmeyen ve merak edilen konularla ilgilidir. Üçüncü aşamada veri ve bilgi kaynakları belirlenir.

Tarihsel araştırmaların güvenilirliği büyük ölçüde veri kaynaklarının ve uygulanan yöntemin güvenilirliğine bağlıdır. Tarihsel araştırmaların güvenilirliği birincil ve ikincil kaynaklardaki verilerin tutarlılığı ile ölçülür. Mümkün olduğunca birincil kaynaklara ulaşılmalı ve bu kaynaklar araştırma raporunda okuyucuların bilgisine sunulmuş olmalıdır. Birincil kaynaklar; günlükler, anı yazıları, eski gazeteler, fotoğraflar, el yazısıyla yazılmış mektuplar, haritalar, tutanaklar, kayıtlar, mülâkat çözümleri ve video filmleri gibi belgelerdir. Birincil ve ikincil kaynakların her ikisi de samimiyet ve doğruluk açısından dikkatli bir şekilde incelenmelidir. Birincil kaynak olmasına karşın bir bulgu; önyargılı, politik olarak taraflı veya yeterli bilgi düzeyine sahip olunmadan hazırlanmış olabilir. Dördüncü aşama; veri ve bulguların doğru, objektif ve güvenilir bir şekilde değerlendirilmesi, sentez edilmesi ve yorumlanmasıdır. Bilim adamının bu aşamada çok yönlü bir çalışma yapması gerekir. Birincil kaynaklarda iddiasını doğrulayan bir kanıt bulması güvenilirliği sağlamada tek başına yeterli olmayabilir. Bu nedenle bilim adamı, üçleme yöntemini kullanarak iddiasını değişik kanıtlarla desteklemeye çalışmalıdır. Johonson' (2004) göre, veri ve bulguların güvenilirliği, *pozitif kritik* ve *negatif kritik* yapılarak test edilir.[68] Pozitif kritik, yararlanılan kaynaktaki bilgilerin ve söylemin anlaşılır olmasıyla ilgilidir. "Belirsizlik" ve "mevcudiyet durumu" faktörleri nedeniyle anlaşılırlığı sağlamak her zaman mümkün olmayabilir. Belirsizlik, kaynaktaki ifade ve kavramların ne anlama geldiğinin tam olarak bilinememesidir. Mevcudiyet durumu ise, kavram ve ifadelerin günümüzdeki anlamının geçmişte de var olduğu varsayımına dayanır. Negatif kritik ise, kullanılan kaynaktaki bilgilerin samimi, doğru ve yansız olma durumunun araştırılmasıdır. Ancak bu tür bir araştırmayı yapmak çok daha zordur. Yine, Johonson'a göre (2004) tarihçiler *negatif kritik* yaparken üç farklı

yöntemden yararlanırlar: teyit etme, kaynak gösterme, bağlamlandırma.[69] Teyit etme, yararlanılan belgelerin birbirini doğrulaması; kaynak gösterme, görüşü ileri süren yazarın belirlenmesi ve bağlamlandirma ise olayın nerede, ne zaman ve hangi çerçeve içinde gerçekleştiğinin belirlenmesidir. Beşinci aşama raporlamadır. Bu aşamada yansız bir gözle veri ve bulgular değerlendirmeye alınarak olgu okuyuculara tanıtılır. Tarihsel araştırmanın güvenilirliği nesnel bir üslup kullanmayla ilgilidir. Nesnellik, araştırmacının yazdıkları konusunda partizan bir tutum takınmamasıyla sağlanabilir.

İÇERİK ANALİZLERİNİN GÜVENİLİRLİĞİ

İçerik analizi niteliksel araştırma yöntemlerinden bir diğeridir. Bu yaklaşımda bilim adamı, bir metin veya metinler grubu içinde belirli kelime veya kavramların bulunma durumunu araştırır.[70] Araştırmacı metindeki kavramların sayısına, anlamına ve kavramların birbiriyle olan ilişkisine, kavram ve ifadelerin vurgusuna bakarak metini yorumlar, değerlendirir ve metin hakkında bir hüküm verir. Bu hüküm; metnin kime hitap ettiği, dilinin anlaşılırlığı, vermek istediği mesaj ve yararlılıkla ilgilidir. İçerik analizinin yapıldığı metin bir kitap, bir gazete yazısı, tez ve makale başlıkları, genel olarak gazete yazıları, bir şiir, bir film, bir bölüm, bir makale, bir reklâm ilânı veya reklâm afişi, bir konuşma, diyalog, tiyatro oyunu, bir resim, bir fotoğraf, bir test, bir karikatür, resmî olmayan bir konuşma veya bir tanıtım broşürü olabilir. Trochim'e (2000) göre, içerik analizi nicel olabilir, nitel tarzda hazırlanabilir veya her iki tür incelemeyi içerecek şekilde yapılabilir.[71] Burada önemli olan araştırmacının yönelimidir. Bilim adamı içerik analizi yöntemini uygulamaya karar verdiği zaman keşfedici, tanımlayıcı veya kuramı test etmeye yönelik (sonuç çıkarıcı) bir yaklaşıma sahip olabilir. Keşfedici yaklaşım, niteliksel analiz yöntemini gerektirir. Tanımlayıcı ve teorinin test edilmesine dayanan yaklaşım ise sonuç çıkarıcı istatistikî analiz yöntemleri kullanılarak başarılabilir.

Bilim adamı içerik analizini yapmak için metni önce belirli bir sistem çerçevesinde bölümlendirmeye ve kodlamaya tâbi tutar. Sınıflandırma (kategorize etme); tüm bir metin temel alınarak, başlıklar temel alınarak, cümleler temel alınarak, paragraflar, kişi isimleri, anlamlar, konular, terimler, kavramlar[a] veya sadece kelimeler temel alınarak yapılabilir.[72] İçerik analizini basit bir uygulama olan kelime sayma işleminden farklılaştıran, kodların belirli gruplar halinde sınıflandırılmasıdır. Kelimeler anlam ve

[a] Kavramlar; örtük içeriğe sahip, bir veya birden fazla kelimenin bir araya gelmesiyle oluşan soyut nitelemelerdir.

çağrışım benzerliklerine bakılarak kategorize edilir. Bir metin çok sayıda kavram ve kelimeden meydana gelir. Bu kelimeleri benzerlik / yakınlık - uzaklık / farklılık açısından sınıflandırmak sanıldığı kadar kolay bir iş değildir. Buna karşın yine de belirli ölçütler çerçevesinde kavramlar sınıflandırılarak sayılır ve içerik rakamsal verilere dönüştürülür.

Literatürde içerik analizi yaklaşımları geniş bir grup halinde listelenmiştir. Bu yaklaşımları odaklanma biçimlerine göre sınıflandırabiliriz:

1. Nicel, nitel analizleri içermesine göre.
2. Metin veya konu düzeyinde analiz yapılmasına göre.
3. Tek başına sadece kavramları ele almasına veya kavramlar arasındaki ilişkileri araştırmasına göre.
4. Dizin çıkarma amacına göre.
5. Görünür anlamın veya gizli anlamın araştırılmasına göre.
6. Bilgisayar kullanılarak veya gözle sayılmasına göre.

Nicel – Nitel İçerik Analizleri

Bu yaklaşımda istatistikî analiz kullanılma durumu göz önünde bulundurulmuştur. Kavramlar ve ifadeler arasında istatistiksel ilişkiler araştırılmışsa bu türü *nicel içerik analizi* olarak isimlendirilir. İstatistiksel araştırmaya konu olmayanlara ise *nitel içerik analizi* adı verilir.

Metnin Konusunu Belirlemeye Dayalı İçerik Analizleri

Bu yaklaşımda araştırmacı bir metinde, metinler grubunda, konuşmada, tartışmada, söylevde veya makale ve tez başlıklarında hangi konuların temel alındığını, vurgunun hangi konular üzerinde odaklandığını araştırır. Metinde araştırılan temalar; yazının ana fikri, terimler, ifadeler veya belirli kavramlardır. Örneğin, bir araştırmacı belirli bir dönemdeki tez başlıklarını inceleyerek "iş ve işçi güvenliği" konusu ile "işçi sağlığı" terimlerinin söz konusu başlıklarda ne ölçüde kullanıldığını belirlemeye çalışmıştır. Doolittle'a göre (2001) *konu analizi* temelde iki tür bilgi sağlar: içerik ve içeriğin yapısı. İçerik, metnin ana fikri veya konusu çerçevesinde örgülenir. Yapısal açıdan ele aldığımızda ise konu analizi en azından üç düzeyde inceleme yapılmasını gerektirir: (a) genel başlık düzeyinde, (b) alt başlık düzeyinde, (c) alt başlıkları açıklayan ayrıntılı başlık düzeyinde.[73] Bir araştırmacı en azından üç başlık düzeyinde içerik analizi yaptığı zaman metnin konusunu sağlıklı bir şekilde ortaya çıkarabilir. Sadece tez başlıkla-

rından hareket ederek metnin konusu belirlenmeye çalışıldığında bu yaklaşım büyük ölçüde yetersiz kalır. Çünkü uygulamada bazı tez veya makale başlıkları içeriğinden bağımsız olarak düzenlenebilmektedir. Örneğin, *İşçilerin Sorunlarına Getirilen Palyatif Çözümlerin Geçersizliği* isimli başlıktan bu makalenin iş güvenliğiyle ilgili olduğunu çıkarmak imkansızdır.

Kavramsal – İlişkisel İçerik Analizleri

İçerik analizi yönteminin bir diğer sınıflandırılma şekli kavramlar arasındaki ilişkilere göre yapılmıştır. Buna göre içerik analizi iki şekilde sınıflandırılır: *kavram analizi, ilişki analizi.*[74]

Kavram analizi, bir metindeki terimlerin sayılması ve metnin içerdiği terimler açısından analiz edilmesidir. Terimler, metinde mevcut bulunma durumu veya sıklığı açısından incelenir. Bazı yazarlar tarafından "semantik analizi" olarak da isimlendirilen *ilişki analizinde* ise, belirli terimlerin diğer terimlerle olan ilişkisi veya bir terimle yakından ilgili olan diğer terimler araştırılır. Böylece çok sayıda terim grupları oluşturulur. Amaç, tek bir kelimeyi değil, bir terimle ilgili kelimeler arasındaki anlamlı olan ilişkileri bulmaktır. Örneğin, araştırmacı mülâkat yaptığı kişinin sözlerinden hareket ederek onun stres içinde olup olmadığını anlayabilir. Bunun için stres kelimesinin yanında; kaygı, endişe, zorlanma, baskı, kişinin canının sıkılması, kızma, öfkelenme gibi kelimeleri stres kavramıyla ilişkilendirerek bir grup olarak ele alır. Semantik analizi, bireylerin psikolojik durumlarını saptamaya yönelik olarak, belirli kavramsal yapıları ölçmek ve ortaya çıkarmak için kullanılır. Gottschalk ve Gleser (1969) tarafından geliştirilen semantik analizine ilgi duyan okurlara ilgili literatüre başvurmalarını öneririz.[75]

Görünür Anlam – Gizli Anlam

İçerik analizi ayrıca kelimelerin görünür anlamına ve zımnî (gizli) anlamına bakılarak iki düzeyde gerçekleştirilir. *Görünür anlam düzeyli içerik analizinde* araştırmacı sadece kelimeleri sayar, kelimelerin imâ ettiği anlamla ilgilenmez. Bu yaklaşımda kelimelerin gerçek anlamları üzerinde durulur. *Gizli anlam düzeyli içerik analizinde* ise araştırmacı kelimelerin ve cümlelerin "işaret ettiği" anlamı kodlar.[76]

İçerik Analizi Uygulama Süreci

İçerik analizinin nasıl yapılması gerektiği konusunda belirli bir mutabakat bulunmasa da süreç temel olarak on aşamada gerçekleştirilir:[77]

1. Araştırma probleminin tanımlanması.
2. Metnin veya metinler grubunun belirlenmesi.
3. Örnekleme yönteminin belirlenmesi.
4. Analitik sınıflandırma gruplarının ve gerekiyorsa diğer alt kategorilerin belirlenmesi.
5. Kategori tanımlarının yapılması.
6. Duruma göre kodlama planı ve boş kodlama formunun geliştirilmesi.
7. Kodlama formunun ön testten geçirilmesi ve gerekli revizyonların yapılması.
8. Kodlayıcılar arası güvenilirliğin test edilmesi.
9. Esas kodlama işleminin yapılması.
10. Verilerin analiz edilmesi ve raporun yazılması.

Bir kodlama sisteminin içinde üç farklı şekilde sınıflandırma yapılır: genel terimlere göre sınıflandırma, özel sınıf grupları belirleme ve kuramsal sınıf grupları belirleme. Genel sınıflandırma demografik özellikler temel alınarak yapılır. Özel sınıflandırma mesleklere özgü teknik dille ilgilidir. Kuramsal sınıflandırma ise veri analizi sırasında yeni kavram üretme gereksinimiyle birlikte ortaya çıkar.[78] İçerik analizinde temel gruplandırma kategorileri, ilgili kavramları alacak kadar geniş ve ilgisiz kavramları dışarıda bırakacak kadar dar kapsamda belirlenir.[a] İçerik analizlerinin bir bölümü dereceleme ölçekleri kullanılarak yapılır. Bu yöntemi tercih eden araştırmacılar metindeki belirli cümleleri iki zıt kutbu içerecek şekilde 5 dereceli bir ölçek üzerinde derecelendirirler.

İçerik Analizlerinin Güvenilirliği

İçerik analizinde güvenilirlik; (a) *istikrarlılık*, (b) *yeniden üretilebilirlik* ve (c) *doğruluk* özellikleriyle ölçülür.

İstikrarlılık, kavramları sınıflandırma prosedürünün tutarlı olmasıdır ve değerlendirici içi güvenilirliği gösterir. Aynı araştırmacının daha sonraki incelemelerinde veya kodlama çalışmalarında kategorilerin benzer veya aynı çıkmasıdır.

Yeniden üretilebilirlik ise, başka bir araştırmacı tarafından aynı sistem çerçevesinde kodlandığında kategorilerin ve her bir kategorideki kelime

[a] Eski deyişle, kategoriler "efradını camî, ağyarını manî" olmalıdır (exhaustive and mutually exclusive).

sayısının aynı olmasıdır. İçerik analizinde güvenilirlik sorunları, kelimelerin anlamlarının net olmamasından, kategori tanımları ve kodlama biçimlerinin farklılığından kaynaklanır.[79] Eğer, içerik analizi konusunda uzman olan kişiler birbirinden habersiz olarak kavramları aynı şekilde sınıflandırıyorlarsa ve her bir sınıftaki kelime/kavram sayısı da aynı çıkmışsa yeniden üretilebilirlik sağlanmış demektir. Yeniden üretilebilirlik, değerlendiriciler arası güvenilirliktir. Örneğin, "iş güvenliği" kavramını analiz eden farklı iki araştırmacı metindeki "işçi güvenliği", "işyeri güvenliği", "iş emniyeti", "sosyal güvenlik" kavramlarını birbirlerinden farklı başlıklar altında gruplandırmışlarsa kodlama güvenilirliği düşer. Bunun için araştırmacı her bir kategorinin ne anlama geldiğini hemen yanında açık bir şekilde tanımlayarak belirsizliği azaltmaya çalışmalıdır. Kodlama güvenilirliğini sağlamanın bir diğer yöntemi, "kayıt ve kodlama yönergesi" geliştirmektir. Kodlayıcılar arası güvenilirlik sağlanıncaya kadar bu yönergede gerekli düzeltmeler yapılarak kodlama sisteminde iyileştirmeye gidilmelidir. Kodlamada anlam karmaşasına yol açan, ayırma özelliği olmayan kategoriler iptal edilmelidir. Raporda kategori tanımları yapılmamışsa aynı araştırma başkaları tarafından güvenli bir şekilde tekrar edilemez.

■ Örnek: Kategori tanımı.

1. İş güvenliği: Kaza ve hastalık gibi risklerin ortadan kaldırılması veya en aza indirilmesi suretiyle iş yerinde sağlıklı ve emniyetli bir çalışma ortamının hazırlanması.

Doğruluk, kodlamanın yanlışsız yapılmasıdır. Kodların içerik analizi formlarına veya istatistikî analiz yazılımlarına hatasız bir şekilde girilmesi ve sonuçta doğru hesaplama ile doğru bilgilerin alınmasıdır. Kavramın anlamının genişletilmesi veya daraltılması sonuçta kodlamanın yanlış yapılmasına neden olur.

İçerik analizinde yeniden üretilebilirlik özelliğini saptamak için Cohen Kappa formülünden yararlanılır. Farklı iki hakemin her grupta kaç tane kelime gösterdiği karşılaştırılarak uyuşma durumu saptanır. Rodenburg (2004) Holsti'den aktararak güvenilirliği, *nesnellik*, *sistematik olma* ve *genellik* özellikleriyle açıklamıştır.[80] Nesnellik, araştırmacının kodlamayı nasıl yaptığı, hangi formülleri kullandığı konusunda açık, net ve belirgin olmasıdır. Böylece daha sonraki araştırmacılar da metne aynı yöntemi ve formülleri uygulayarak benzeri sonuçlara ulaşma imkanı elde edebileceklerdir. Sistematik olma, seçilen terim ve kavramların belirli kategorilere alınması veya çıkarılması konusunda hangi kriterlerden hareket edildiğinin

açıklanmasıdır. Genellik ise, içerik hakkında tanımlayıcı nitelikteki bazı bilgilerin sunulmasından öte bulguların kuramsal bir ilgililiğe sahip olma derecesidir. Bilim adamı içerik analizinde iki elin parmakları kadar kavramı/kelimeyi kodlayabileceği gibi bu sayıyı 500'e kadar da çıkarabilir. Ancak kodlanacak kelime sayısı arttığı üçlüde analizin güvenilirliği zayıflar. Neuendorf'a (2002) göre güvenilirlik testleri yapılmadığı sürece içerik analizi sonucunda elde edilen rakamların herhangi bir anlamı yoktur, bu rakamlar değersizdir (aktaran, Lobbard ve d., 2004). Stemler, (2001) içerik analizi yaparken üç tür sorunla karşılaşılabileceğini belirtmiştir. Bunlardan birincisi içerik analizi yapılacak metinlerin (veya örneklemin) sayısının yetersiz olmasıdır. Analizi yapılan metinler ana kütleyi temsil etmiyorsa varılan sonuçların bir anlamı yoktur. İkincisi, metinde kategori sınıflarından herhangi birinin çok düşük veya çok yetersiz ölçüde sayılmasıdır. Yetersiz kayıt varsa bu kategori ölçümden düşürülmelidir. Üçüncüsü ise, bazı metinlerin anlamlarının belirsiz olması nedeniyle kodlanamamasıdır.[81] Böyle olunca yetersiz bir örnekleme dayalı olarak, yetersiz kayıt altında ve yetersiz kodlama çerçevesinde yapılan hesaplamalar çalışmanın güvenilirliğini ve bilimsel değerini düşürür.

SÖYLEM ANALİZLERİNİN GÜVENİLİRLİĞİ

Söylem analizleri kişilerin belirli sosyal ve kültürel çevrelerde dili nasıl kullandıklarını etüt eden incelemelerdir. Bu yaklaşımda tercih edilen iletişim yapılarının sosyokültürel normlar, tercihler ve beklentilerle olan ilişkisi araştırılır ve söz konusu konuşmanın veya yazılı iletişimin yapıldığı sosyal çevrenin özellikleri tanıtılır.[82] Söylem araştırmaları; semantik, fonoloji, sentaks, morfoloji ve pragmatik yaklaşımlarda görüldüğü gibi cümleyi temel alan dar kapsamlı dil analizi çalışmaları değildir. Tam tersine, dil kullanımının daha büyük bölümlerini sosyokültürel bağlam çerçevesinde ele alıp inceleyen bir yaklaşımdır.[83] Söylem analizleri konuşma dili üzerinde odaklanmakla birlikte yazılı metinler üzerinde yapılan incelemeleri de içerir.[84] Bazı yazarlar söylem araştırmalarının ne nitel ve ne de nicel araştırma olarak değerlendirilemeyeceğini söylemişlerdir. Onlara göre söylem araştırmaları sorunlara bilimsel araştırmaya dayalı olarak somut bir cevap vermez. Bir projenin, bir ifadenin veya metnin ontolojik ve epistemolojik kaynakları hakkında bilgi verir.[85] Söylem analizleri ideolojiler, siyasal eğilimler ve inançlar çerçevesinde tanımlayıcı veya eleştirel bir çerçevede yapılabilir. Bilim adamları söylem araştırmalarını aşağıdaki amaçlarla yaparlar:[86]

1. Farklı söylem biçimlerini dil açısından analiz etmek.
2. Ardışık konuşmaları tanımlamak.
3. Konuşma aktivitelerini incelemek.
4. Sözlü veya yazılı ifadeleri tanımlamak.
5. Kişinin zihinsel duruş biçimini saptamak.
6. Bir metnin veya bir konuşmanın arkasındaki gizili güdüleri ortaya çıkarmak.

Söylem araştırması yapan araştırmacılar bu amaçla bir dizi teknikten yararlanırlar. Etnografya araştırmaları, diyalog analizi, spesifik bir konudaki yazı koleksiyonlarının incelenmesi, ayrıntılı video ve teyp araştırmaları bunlardan bazılarıdır. Söylem araştırması için takip edilecek standart bir prosedür belirlenmemiştir. Bu nedenle araştırmacılar bu yöntemi sık uygulayan Jacques Derrida, Michel Foucault, Julia Kristeva, Dell Hymes ve Fredric Jameson gibi post-modern düşünürlerin kuramlarından hareket ederler.[87] Herhangi bir metne, probleme veya duruma uygulanabilen söylem araştırmaları metinle veya sorunla ilgili kesin bir cevap vermez, fakat araştırmacının kişisel görüş ufkunu genişleterek gizli güdüler hakkında belli belirsiz bir bilgi sahibi olmasını sağlar.

Söylem araştırmaları, kişilerin vukufiyeti (iç görüsü) çerçevesinde yapılan yorumlara dayanır. Güvenilirlik ve geçerlilik analizlerinde kullanılan somut bir veri veya rakam bulunmadığından bir kişinin araştırma bulgularının geçerlilik ve güvenilirliği mantığının gücüne argümanlarının tutarlılığına bağlıdır. Bununla birlikte en güçlü tezler dahi karşı görüşlerle çürütülebilir. Bu nedenle analizin geçerliliği, retoriğin kalitesine bağlıdır.[88] Söylem araştırmaları güçlü bir bilimsel temele sahip olmaması nedeniyle her zaman tartışmalara konu olmuş ve güvenilir bir değerlendirme olarak görülmemiştir.

MÜLÂKAT ARAŞTIRMALARININ GÜVENİLİRLİĞİ

Mülâkatlar bireysel görüşmelere dayanan bilgi toplama araçlarıdır. Görüşme; yüz yüze, telefonla veya grup görüşmesi şeklinde gerçekleştirilebilir. Bir araştırmada tek bir yöntem olarak uygulanabileceği gibi niceliksel bir araştırmanın belli bir bölümü şeklinde de düzenlenebilir. Mülâkat araştırması kapsamlı bir literatür taraması çalışmasıyla başlar. Araştırmacı bu süreç içinde konunun sınırlarını, kapsamını, temel kavramlarını, boyutlarını ve önermelerini öğrenir. İkinci aşamada araştırmacı kendi deneyimleri-

ni, araştırma sorularını gözden geçirerek araştırma problemini formüle eder. Üçüncü aşamada ise araştırma problemine uygun bir soru listesi veya anket geliştirilir. Literatürde bu anket formuna "mülâkat protokolü" adı verilmiştir. Mülâkat araştırmasının son ve en önemli aşaması ise toplanan verilerin analiz edilmesidir. Bu aşamada toplanan bilgiler duruma göre kodlanır, gruplandırılır, olaylar ve kişiler arasında ilişkiler kurulur. Literatürdeki bilgilerle mülâkat sonuçları karşılaştırılarak anlamlı ve sistematik bir analiz yapılır.[89]

Uygulanan yönteme göre mülâkat araştırmaları temelde üç değişik şekilde sınıflandırılmıştır:

1. Serbest mülâkatlar.
2. Yarı yapılandırılmış mülâkatlar.
3. Tam yapılandırılmış mülâkatlar.

Aşağıdaki başlıklarda söz konusu mülâkat biçimlerine ilişkin özet açıklamalara yer verilmiştir.

Serbest Mülâkatlar

Serbest mülâkatlar araştırmacının mülâkat yaptığı kişi ile serbest bir etkileşim içinde kurduğu diyaloglara dayanır. İlgili kişi ile konuşmadan önce araştırmacının kafasında hangi soruları soracağına ilişkin bir plan vardır, ancak bu plana çok fazla bağlı kalmaz. Konuşmaların seyrini ilgili kişinin verdiği yanıtlar belirler. Araştırmacı biraz da bu yanıtlara bakarak hangi soruları soracağına karar verir. Araştırmacı, konuşmaları banda alacağından görüşme öncesi ilgili kişiden banda alınmasına izin verdiğini kanıtlamak üzere yazılı bir muvafakat belgesi alır.

Yarı Yapılandırılmış Mülâkatlar

Bu yöntemde araştırmacı ilgili kişi veya kişilerle görüşme yapmadan önce araştıracağı veya soracağı soruları belirler. Katı bir şekilde bu sorulara bağlı kalmasa da bu soruların görüşmesinde kendisine rehberlik etmesini amaçlar. Soruların sorulma sırası kişiden kişiye farklılık gösterdiği gibi ifadelendirme biçimleri de farklıdır. Bir ölçüde serbest bir uygulamaya sahip olmakla birlikte önceden hazırlanmış bir plan çerçevesinde yürütüldüğünden toplanan veriler sistematik bir kümelenmeye sahiptir ve bu nedenle de daha kolay analiz edilir.

Tam Yapılandırılmış Mülâkatlar

Tam yapılandırılmış mülâkatlar sözlü anket uygulamasına benzer. Bu tür araştırmalarda çok az esneklik vardır. Mülâkatı yapan araştırmacı önceden belirlenmiş olan soruların dışına çıkamaz. Bu nedenle birden fazla kişiyle yapılan mülâkat uygulamalarında araştırmacılar arasındaki sapma çok azdır. Mülâkat zamanı en düşük limite indirgenmiştir. Araştırmacılar mülâkat sırasında çoğunlukla basılı soru formlarını kullanırlar.[90] Bu soru formlarının anketlerden farkı soruların cevap şıklarının belirlenmemiş olmasıdır. Büyük ölçüde yapılandırılmış olmasına karşın tam yapılandırılmış mülâkatlarda da araştırmacılar belirsiz olan sorularla ilgili olarak katılımcılara ek açıklamalar yapabilirler.

Mülâkat Bilgilerinin Güvenilirliği

Mülâkat bilgilerinin güvenilirliği, toplanan verilerin kalitesine bağlıdır. Kişilerin araştırılan bilgileri vermesi kendilerini rahat ve serbest hissetmeleri halinde söz konusu olabilir. Kaliteli bilgi elde etmek isteyen mülâkatçı süreç içinde büyük ölçüde dinlemede kalmalı ve kişinin düşünce ve fikirlerini değerlendiren, eleştiren bir yaklaşım tarzından özenle kaçınmalıdır. Mülâkatçının vereceği herhangi bir cevap eleştirel nitelikte, yargılayıcı veya gereksiz yere övgü sözleri içeren bir niteliğe sahip olmamalıdır. Görüşme ne çok hızlı, ne de yavaş bir tempoda sürdürülmelidir. Mülâkatçı, görüşme yapılan kişinin verdiği cevaplardan bazılarını belirsiz veya kapalı bulmuşsa bunları kendisine sormalı, konuyu netleştirmeye çalışmalıdır. Siegel (2003) başarılı bir mülâkat uygulaması için araştırmacılara aşağıdaki önerilerde bulunmuştur:[91]

1. Çoğunlukla dinlemede kalın.
2. Katılımcının ne söylediğini izleyin.
3. Anlamadığınız zaman soru sorun.
4. Konu hakkında daha fazla bilgi edinmeye çalışın.
5. Sorgulamayın, araştırın.
6. Az konuşun ve gerçek soruları sorun.
7. Yönlendirici sorulardan kaçının.
8. Açık uçlu sorular sorun.
9. Katılımcının sözünü kesmeyin, konuşmasını bitirmesini bekleyin.
10. Size sanki herhangi biri imiş gibi konuşmasını isteyiniz.
11. Katılımcıya bir hikaye anlatmasını söyleyiniz.
12. Katılımcının konuya odaklaşmasını sağlayınız.
13. Katılımcıdan ayrıntıları vermesini isteyiniz.
14. Görüşmeyi çok fazla kişiselleştirmeyiniz.

15. Fırsat çıktığında deneyimlerinizi paylaşınız.
16. Hatırlamasını değil, bilgileri yeniden düzenlemesini öneriniz.
17. Katılımcının cevabını pekiştirmeye çalışmayınız.
18. Sempatik olunuz ve esprileri yakalayınız.
19. Sezgilerinizi kullanınız.
20. Mülâkat sorularını dikkatli bir şekilde takip ediniz.
21. Sessizliği hoşgörüyle karşılayınız.

Güvenilirlikte mülâkatın yapılma biçimi kadar soruların niteliği de önemlidir. Araştırmacı mülâkat sorularının yapısal düzenlemesini, soruların uzunluğunu ve soruların karmaşıklık derecesini önceden dikkatli bir şekilde planlamalıdır.[92] Mülâkat bilgilerinin güvenilirliğini artırmak isteyen bilim adamı aşağıdaki faktörleri göz önünde bulundurur:

1. Araştırma sorularının araştırma konusuyla ilgili olması.
2. Mülâkatçının (mülâkatçıların) mülâkat teknikleri konusunda eğitim almış olması.
3. Mülâkatçının (mülâkatçıların) araştırma konusu hakkında bilgili olması.
4. Mülâkat için yeterli zaman ayrılmış olması.
5. Araştırmanın niteliğine göre, mümkün olduğunca birden fazla mülâkatçıdan yararlanma.
6. Mülâkat verilerinin mülâkat yapılan kişiye doğrulatılması.

Mülâkat verilerinin güvenilirliğini sağlamada başvurulabilecek bir diğer yaklaşım yeniden mülâkat yapmaktır. Test-yeniden test uygulamasına benzeyen bu yaklaşımda aynı kişiyle ikinci bir görüşme daha yapılır. Literatürde bu yöntemi uygulayan çok az çalışma yer almıştır.[93] Yeniden mülâkat, psikiyatrik çalışmalarda ve tıbbi tedavi uygulamalarında başvurulan bir yöntemdir.

Mülâkat verilerinin güvenilirliği özellikle araştırmada birden fazla mülâkatçıdan yararlanılması halinde önem kazanır. Çalışmada birden fazla kişi mülâkata katılmışsa bu kişiler görüşme yapılan bireyleri önceden belirlenen kriterleri paylaşarak değerlendirir veya tüm kriterler üzerinden her biri bağımsız olarak değerlendirme yapar. Serbest ve yarı yapılanmış mülâkat biçimine göre tam yapılanmış mülâkat şeklinde yanlılık daha azdır. Mülâkat sırasında yapılan değerlendirme, önceden belirlenen kriterlere puan verme şeklindedir. Daha sonra bu puanlar istatistiksel analiz yazılımı

SPSS'te *two-way mixed-effect* seçeneği ile küme içi korelasyon analizine tâbi tutulur. Birden fazla mülâkatçıdan yararlanılmışsa ve mülâkat sonuçları rakam olarak değerlendirilmişse verilen puanların tutarlılığı "gözlemciler arası değerlendirme güvenilirliği" yöntemiyle saptanır.

Bilgi ve görüş almaya yönelik olarak yapılan mülâkatlarda veri toplama sürecine ait güvenilirliği sağlamak için video veya teyp cihazlarından yararlanılır. Araştırmacının bant çözümlerini mülâkat yaptığı kişiye okutması ve onun yazılı onayını alması güvenilirliği artıracak bir diğer önemli uygulamadır.

GÖZLEM ARAŞTIRMALARININ GÜVENİLİRLİĞİ

Gözlem araştırması, bilim adamının katılımcılara soru sormak, anket uygulamak yerine odak kişi veya kişilerin hareket ve davranışlarını, diğer kişilerle olan etkileşimlerini sistemli bir şekilde gözleyip kayıt altına aldığı, çözümleyip değerlendirdiği bir yöntemdir. Gözlem araştırmalarında kişi ve grupların sözel ve sözel olmayan davranışlarının her ikisi de incelenir. Araştırmacı kişilerin hareketlerini, mimiklerini, yüz ifadelerini, davranışlarını ve sözlerini izleyerek bunlara bir anlam vermeye çalışır. Gözlem, "gerçeklerin veri haline dönüştüğü bir süreç" olarak tanımlanmıştır.[94] Gözlem yöntemi daha çok tıbbî araştırmalarda, psikolojik sorunların teşhisinde, pazarlamada müşteri davranışlarının tahmin edilmesinde, işyerinde personel davranışlarının yorumlanmasında kullanılır.

Niteliği

Gözlem yöntemi odak kişilerin hareket ve davranışlarının, konuşmalarının sistemli bir biçimde ve ayrıntılı olarak izlenmesini gerektirir. Gözlem, doğal ortamında ilgili kişilere fark ettirilmeden, ilgili kişilerin bilgisi dahilinde, fakat hareket ve davranışlarına müdahale edilmeden veya müdahaleyi de içeren katılımlı gözlem şeklinde gerçekleştirilir. Bu özelliği ile, anket ve mülâkat yöntemiyle toplanamayan veri ve bilgilerin elde edilmesine imkan sağlar. Gözlem yönteminden, anket uygulanan alan araştırmalarında ek bilgi toplama aracı olarak, etnografya araştırmalarında ise ana bilgi kaynağı olarak yararlanılır. Etnografya araştırmalarında bilim adamı bir süre incelediği kültürel grupla birlikte yaşayarak onların tüm yaşam biçimlerini (yemeklerini, giyim tarzlarını, törenlerini, inançlarını ve tepkilerini) gözlemlerine dayalı olarak resmeder. Gözlem araştırmaları niteliksel veya niceliksel içerikli olabilir. Bilim adamı araştırmanın başlangıç aşamasında belirli bir hipotezden yola çıkmamış, sadece gözlemlerinin sonucuna göre bir kuram geliştirmeyi hedeflemişse nitel içirikli bir gözlem çalışması ya-

pıyor demektir. Oysa nicel araştırmalarda bilim adamı daha başlangıçta belirlediği bir araştırma hipotezinden hareket eder. Gözlemler, araştırma hipotezine dayalı olarak yapılıyorsa nicel içerikli gözlem araştırmasından söz ederiz.

Türleri

Niteliksel içerikli gözlem araştırmalarını değişik başlıklar altında gruplandırmak mümkündür. Gruplandırma, olguyu daha iyi anlamamıza ve uygun bir araştırma ve veri toplama tasarımı belirlememize yardım eder.

1. Davranışların doğrudan veya dolaylı olarak gözlenmesi.
2. Yapılandırılmış-yapılandırılmamış gözlem.
3. Nitel-nicel içerikli gözlem.

Aşağıdaki paragraflarda başlıca üç ana kategori halinde toplanan bu gözlem türlerine ilişkin açıklamalara yer verilmiştir.

Davranışların doğrudan gözlenmesi. Doğrudan gözlemde bulunma, araştırmacının bizzat kendisinin rol almasını gerektirir. Doğrudan gözlemde bulunma (a) rahatsız etmeden yapılan "pasif gözlem" veya (b) "katılmalı gözlem" şeklinde yapılır.

Rahatsız etmeden yapılan gözlem. Rahatsız etmeden yapılan pasif gözlemde kişiler kendi üzerlerinde gözlem yapıldığının farkında değillerdir. Kendilerine bilgi verilmemiştir ve kuşku uyandıracak herhangi bir davranış da söz konusu değildir. Gözlemci katılımcıları sessizce izler ve anlamaya çalışır. Bu süreçte gözlemci aşağıdaki davranışları sergiler.[95]

1. Katılımcılara soru sormaz.
2. Açıklama yapmaz.
3. Tasarımı veya düzenlemeyi savunmaz.
4. Özür dilemez.
5. Teklif etmez.
6. Tartışmaz.

Bu yöntemin sakıncası araştırmacının yanlı olabilmesi ve sürecin bazı ahlakî kaygılara neden olabilmesidir. Gözlemlenen kişiler bu durumu öğrendiklerinde tepki gösterebilirler. Bir diğer yetersizliği, gözlemlenen değişkenler hakkında kişilere soru sorulamamasıdır. Soru sorulamayınca araştırmacı sadece gördükleriyle yetinmek zorunda kalır.

Katılmalı gözlem . Katılmalı gözlem ise çoğunlukla haberli olarak yapılır ve gözlemde bulunan kişi gözlediği kişilerden biri olur. Sosyal antropolojide yaygın olarak kullanılan bu yaklaşımda kişiler kendi üzerlerinde araştırma yapıldığını önceden öğrenirler.

Bazen katılmalı gözlem gizli bir şekilde de yapılabilir. Bu uygulamada kişiler aralarına katılan kişinin araştırma yaptığını bilmezler ve farkına da varmazlar. Gizli katılmalı gözlemde araştırmacı çok fazla etkileşime girmeden kişilerin konumlarını ve diğerleriyle yaptıkları etkileşimi izler.

Açık katılmalı gözlem yönteminden etkili bir sonuç alınabilmesi için araştırmacının aylarca ve hatta yıllarca ilgili grupla birlikte yaşaması gerekebilir. Kişi bu süre içinde çevresiyle bütünleşir ve onlardan biri haline gelir. Ancak her tür araştırmada uzun süre ilgili kişilerle birlikte olmak gerekmez. Araştırmacı incelediği konuya göre değişik gözlem tasarımları geliştirebilir. Örneğin, araştırmanın niteliğine bağlı olarak her gün birkaç saat veya haftada bir iki gün gözlemde bulunmak yeterli olabilir.

Davranışların dolaylı olarak gözlenmesi. Davranışların dolaylı olarak gözlenmesi video, teyp kullanılarak veya bilgisayarlarda şifreli girişler aracılığıyla yapılır. Bir diğer dolaylı gözlem biçimi, arşiv kayıtlarının taranmasıdır. Bu yöntemde kişiler kendilerinin hangi konuşmalarının veya davranışlarının hangi saatte veya dakikada izlendiğini tam olarak bilemezler.

Yapılandırılmış-yapılandırılmamış gözlem. Gözlem uygulamalarının bir diğer sınıflandırma biçimi önceden hazırlık yapma durumuna göre yapılmıştır. Buna göre gözlem çeşitleri yapılandırılmış ve yapılandırılmamış olmak üzere iki başlıkta toplanır. Nitel gözlemler büyük ölçüde yapılandırılmamış bir niteliğe sahiptir. Yapılandırılmamış gözlemlerin başarısı araştırmacının gözlem kabiliyetine, gördüklerini kayıt altına almasına, gerçekleri yorumlama ve değerlendirme kabiliyetine bağlıdır. Vak'a geliştirme çalışmaları büyük ölçüde *doğal gözlem* yöntemiyle yapılır. Doğal ortamında gözlenmesi zor olan davranışların incelenmesinde ise, yapılandırılmış gözlem yönteminden yararlanılır.

Nicel-nitel gözlem. Nicel gözlem çalışmasında gözlem yapacak kişiler öncelikle özel bir eğitime alınırlar. Bu eğitim sırasında gözlemin kaç saat süreceği, gözlemcilerin hangi davranışlara dikkat etmeleri gerektiği, gözlemlerini nasıl kayıt edecekleri konularında kendilerine bilgi verilir ve denemeler yapılarak yetiştirilmeleri sağlanır. Gözlem sırasında kullanıl-

mak üzere bir *gözlem konuları listesi* oluşturulur. Gözlemciler bu listede belirlenen davranışların / kategorilerin vuku bulma durumunu gözleyerek bu davranışları kayıt altına alırlar ve söz konusu davranışları rakamlarla ifade ederler. Nicel gözlem araştırmasında gözlemci pasif bir izleyici durumundadır. İncelediği olaya veya kişilere müdahale etmediği gibi onları etkileyecek herhangi bir davranışa da girişmez. Daha sonra gözlemciler elde ettikleri rakamlara dayalı olarak matematiksel ve istatistiksel işlem yaparak konuyu yorumlarlar. R. Bales'in (1950) küçük gruplardaki "Etkileşim Süreci Analizi" yaklaşımı tipik bir nicel gözlem çalışmasıdır. Niceliksel gözlem araştırması, bir hipotezi test etmek, değişkenleri ölçmek ve belirli tutum ve davranışları ana kütleye genellemek için yapılır.

Nitel gözlem çalışmalarında ise araştırmacı veya gözlemci gözlediği gruba katılarak onlardan biri haline gelir. Onlarla birlikte değişir; hem onları etkiler, hem de onlardan etkilenir.

Örnekleme Yöntemi

Gözlem yöntemi araştırmacı tarafından belirlenen spesifik bir grup üzerinde ve spesifik davranışlar üzerinde uygulanır. Nitel gözlem yönteminden elde edilen bulgular ana kütleye genellenmediğinden bu grubun incelenen ana kütle açısından "tipik" nitelikte olması önemlidir. Söz konusu grubun ana kütleyi temsil edicilik özelliğinin yüksek olmasına dikkat edilir. Ancak bu değerlendirme yine de görecelidir. Gözlem örneklemi seçilirken *zaman örneklemi*, *durum örneklemi* veya ikisi birlikte göz önünde bulundurulur. Zaman örnekleminde sistematik olarak belirli zaman dilimleri veya rasgele zaman dilimleri seçilir. Durum örnekleminde ise farklı düzlemler, koşullar veya yerler belirlenir. Gözlem yapmadan önce genellikle söz konusu grubun izni alınır ve kendilerine araştırma yapılacağı hakkında bilgi verilir. Araştırmacı bu yaklaşımın gözlenenler üzerinde "Hawthorne etkisi" yaratacağını gözden uzak tutmamalıdır. Pope ve Mays'e göre (2001) gözlem süreci bir "arabulucunun" veya "kolaylaştırıcının" devreye girmesiyle başlar. Bu kişi gözlemciyi gruba tanıtır, yetkililerden izin alınmasını sağlar ve gözlem sırasında ortaya çıkabilecek sorunlarla ilgilenir (aktaran, Catherine ve Nicholas, 2004).[96]

Avantaj ve Dezavantajları

Diğer araştırma yöntemlerinde olduğu gibi gözlem yönteminin de kendine özgü avantaj ve dezavantajları söz konusudur. Avantajları, bazı veri ve bilgilerin doğal ortamında gerçeğine uygun olarak elde edilmesi, verilerin belli bir derinliğe sahip olması, anketler ve mülâkatlarla elde edilmeyecek bilgilerin elde edilmesidir. Gözlem verileri gerçeklere dayanır, kişinin

kendisini beğenilir yapma etkisinden kurtulamadığı "öz-değerlendirme" ölçüm araçlarının yetersizliklerini içermez.

Bu yöntemin zayıf kaldığı yönler ise dikkatli, eğitimli ve sistematik çalışma yapan uzmanlara gereksinim göstermesidir. Bir diğer sakıncası gözlemlerde objektifliğin sağlanmasında yaşanan güçlüktür. Gözlemciler, art yetişimlerinden ve değer yapılarından kolaylıkla etkilenebilirler. Bir başka olumsuz yönü, gözlemin aynı kişilerde tekrarlanma zorluğudur. Daha sonraki gözlemler farklı kişiler üzerinde yapılacağından aynı olgunun iki kez gözlenme olasılığı hemen hemen hiç yoktur. Gözlem; açıkta olan, görünen davranışların saptanmasına imkan sağlar. Kişilerin düşüncelerinin, niyetlerinin, algılarının ve endişelerinin anlaşılmasına imkan vermez. Gözlem verileri geleceğe sınırlı ölçüde yansıtılabilir. Gözlem araştırması, zaman yitirici, pahalı ve bir ölçüde yoğun emek gerektiren bir uygulamadır.

Gözlem Verilerinin Kayıt Edilmesi

Nitel gözlem verilerinin kayıt edilmesi iki şekilde yapılır. Ortam uygunsa kayıtlar doğal düzlemde tutulur. Yöntem ve ortam uygun değilse araştırmacı gözlemlerini akşam kendi odasına çekildiğinde, hatırladığı kadarıyla kayıt altına alır. Gözlem maddelerinin kaydı daha önceden çıkarılmış bir liste çerçevesinde yapılabileceği gibi bu liste daha sonradan da oluşturulabilir. Uzun süreli ve planlı çalışmalarda araştırmacı sahaya çıkmadan önce hangi konuları gözleyeceğini yazılı bir liste haline getirir. Bu liste gün içinde yapılan gözlemler sonucu daha sonraki günlerde genişletilip daraltılabilir. Nicel gözlem verileri ise doğal ortamında eş zamanlı olarak kayıt edilir.

Gözlem Verilerinin Analizi

Gözlem verileri niceliksel veya niteliksel olarak analiz edilir. Verilerin kayıt edilme biçimi verilerin nasıl analiz edileceğini belirler.

Gözlem Çalışmalarının Kalitesi ve Güvenilirliği

Gözlem çalışmalarının kalitesi gözlemcinin kalitesine bağlıdır.[97] Bunun yanında uygulanan veri toplama yaklaşımı da verilerin kalitesini etkiler. Gözlem verilerinin güvenilirliği niceliksel veya niteliksel olmasına göre değişir. Niteliksel verilerin güvenilirliği birkaç faktörden doğrudan etkilenir ve bu faktörler aşağıdaki gibidir:

1. Gözlemcinin tarafsızlığı veya yanlılığı.
2. Gözlemcinin gözlem sürecini yansıtma biçimi.
3. Verilerin doğrulanması.

4. Hawthorne etkisi.
5. Gözlemcinin dikkatliliği.
6. Gözlemcinin eğitimli olması.

Gözlem verilerinin güvenilirliği bilgilerin, temsil edici bir örneklemden ve tarafsız bir gözle alınmasına bağlıdır. Araştırmacı kişisel görüş ve değerlerinden etkilenmişse gözlem verileri yanlıdır. Araştırmacı kendi kişisel eğilimlerinin etkisini azaltmak için mümkün olduğu kadar çok gözlem yapmalı ve olguyu kendi gerçekliği içinde ortaya koymalıdır.

Gözlemci kaliteli bir çalışma yapmak istiyorsa gözlemde bulunma yöntemini, veri toplama ve analiz etme yöntemlerini *yansıtıcı* bir yaklaşımla okuyucularına tanıtmalıdır. Bu amaçla "gözlem protokollerinden" yararlanılır. Gözlem protokolü veri toplamak için kullanılan anket formu, soru listesi veya değerlendirme formudur. Araştırmacı gözlem süreci içinde sahada ne kadar zaman geçirildiğini, bir grubun kaç saat gözlendiğini, gözlemin bildirimli mi yoksa bildirimsiz mi olduğunu, gözlem notlarını nasıl tuttuğunu net bir şekilde araştırma raporunda açıklamalıdır.

Gözlemin kalitesini etkileyen bir diğer öge verilerin doğrulanıp doğrulanmadığı konusudur. Gözlemci doğrulma yapmak için birkaç değişik yöntemden yararlanabilir. Aynı özelliğe sahip birden fazla kişiyi gözleme, başka bir gözlemciyle kendi gözlemlerini karşılaştırma, birden fazla gözlemcinin bulgularını karşılaştırma, aynı kişiyi veya olguyu farklı zamanlarda gözleme, kendinden önceki yazarların gözlem sonuçlarıyla kendi bulgularını karşılaştırma değişik doğrulama yöntemleridir.

Hawthorne etkisi gözlem verilerinin kalitesini etkileyen bir diğer faktördür. Kişiler kendi üzerlerinde araştırma yapıldığını biliyorlarsa gerçek davranışlar yerine araştırmacının beklediği davranışlara veya tam tersi davranışlara yönelebilirler. Bu nedenle gözlemci insan davranışlarında ortaya çıkabilecek Hawthorne etkisini her zaman göz önünde bulundurmalıdır.

Araştırmacı nicel gözlem çalışması yapmışsa böyle bir durumda niceliksel araştırmalar için geçerli olan güvenilirlik analizlerinden yararlanır. Birden fazla gözlemci tarafından toplanan veriler nominal ölçek verisi niteliğinde ise *uyuşma yüzdesi* değeri temel alınır. Toplanan veriler sıralı, eşit aralıklı veya oranlı ölçek verisi niteliğinde ise bu kez *korelasyon katsayılarından* yararlanılır veya *gözlemciler arası güvenilirlik katsayısı* hesaplanır.

HERÜSTİK DEĞERLENDİRMELER

Herüstik değerlendirme, kullanışlılık problemlerinin çözümünde kullanılan ve deneme-yanılma yoluyla en iyi veya en uygun çözümün bulunmasına yardım eden bir değerlendirme veya analiz tekniğidir. Teknik; okullardaki eğitimin değerlendirilmesinde, İnternet'te Ağ kümelerinin değerlendirilmesinde, cep telefonlarının, etkileşimli televizyonların, etkileşimli kioskların, cep bilgisayarlarının, tablet bilgisayarlarının, elektronik adres defterlerinin ve bilgisayar yazılımlarının kalitesinin değerlendirilmesinde kullanılmıştır. Ancak esas itibariyle hayatın her alanında kullanılma özelliğine sahiptir.

Bu bölümde herüstik değerlendirmenin nasıl çalıştığını görmek için bir şirkete ait İnternet'teki Ağ kümelerinin değerlendirme biçimi örnek olarak alınmıştır. Şirket sahibi İnternet'teki Ağ kümesinin kullanışlılığının ve yararlılığının test edilmesini ve bu konuda kendisine bir rapor verilmesini istemiştir. Ağ kümesinin herüstik değerlendirmesi için öncelikle bu değerlendirmeyi yapacak yetenek ve niteliğe sahip 3-5 uzmana ihtiyaç duyulur. Uzmanlar söz konusu Ağ kümesini kullanışlılık ve yararlılık ilkelerine göre değerlendirerek kendi yorumlarını ve değerlendirmelerini bir rapor haline getirirler. Değerlendirmede sadece eksiklikleri değil, nasıl olması gerektiğini de belirleyerek bundan sonraki tasarım için yön ve yol gösterirler.

Herüstik değerlendirme 1990'lı yıllarda Jakob Nielsen ve Rolf Molich tarafından başlatılmıştır. Nielsen, herüstik değerlendirmede kılavuz görevi görmesi için kapsamlı bir değerlendirme listesi geliştirilmiştir. Kullanışlılığı belirlemeye yönelik olarak geliştirdiği 200'den fazla soru daha sonra azaltılarak veya gruplandırmaya gidilerek 10 temel herüstik modele indirgenmiştir. Söz konusu herüstik değerlendirme ilkeleri aşağıdaki gibidir:[98]

1. Sistemin statüsünün görülür olma durumu.
2. Sistem ve gerçek dünya arasındaki denklik.
3. Kullanıcı kontrolü ve özgürlüğü.
4. İstikrarlılık ve standartlar.
5. Hata önleme.
6. Hatırlama yerine tanıma özelliği.
7. Esneklik ve kullanım kolaylığı.
8. Estetik ve minimalci yaklaşım.
9. Hataları teşhis etme ve iyileştirme.

10. Yardım ve dokümantasyon.

Herüstik değerlendirme, bir anlamda resmî olmayan ve sübjektif yargılara dayanan "kullanışlılık analizi" yöntemidir.[99] Değerlendirilen soruna, elektronik araca, Ağ kümesine veya yazılıma göre "herüstik " adı verilen kullanışlılık ilkeleri belirlenir. Bu örnekte Nielsen tarafından yazılımlar için geliştirilen herüstik temel alınmış ve önemli bir değişiklik yapılmadan bu herüstiğin İnternet Ağ kümelerinde de kullanılabileceği öngörülmüştür.

Herüstik değerlendirmede uzmanların her biri belirlenen ilkeler açısından Ağ kümesinin her bir sayfasını veya bütün köprülerini oluşturacakları bir cetvel veya her sayfa için hazırlayacakları ayrı bir tablo üzerinde değerlendirerek görüş ve düşüncelerini açıklarlar. Her sayfada saptadıkları problemleri, hataları veya uygunsuzlukları belirlenen 10 herüstik ilke modeline göre tanımlarlar. Örneğin, "Bu hata, tipik bir şekilde sistem statüsünün görülür olmasıyla ilgilidir." veya "Bu hata, esneklik ve kullanım kolaylığı ilkesini ihlal etmektedir" gibi yorumlar yaparlar.[100] Değerlendirmede kişisel beğeniler üzerinde değil, herüstik üzerinde odaklanırlar[101] Tek kişinin yapacağı değerlendirmeler yeterince güvenilir olmaz. Bu nedenle en az üç uzmanın yaptığı değerlendirmeler bir araya getirilerek birleştirilir. Bu sırada mükerrer görüşler elenerek her bir Ağ sayfası için tek bir problem listesi belirlenir. Daha sonra bu problem listesi 5 veya 3 dereceli bir ölçek üzerinde uzmanlara ortaklaşa olarak değerletilir. Nielsen, kendi orijinal çalışmasında 5 dereceli ölçek formunu kullanmıştır. Bu yaklaşımda 0 değeri kullanışlılık problemi olmadığını, 4 puanı ise kullanışlılık karmaşası bulunduğu anlamına geliyordu.[102] Araştırmacı değerlendirmede üç dereceli yaklaşımı tercih etmişse; 1 puanı problemin önemsiz, 2 rakamı kısmen sorunlu, 3 rakamı ise sorunun ciddi bir karmaşıklığa neden olduğu anlamına gelir. Uzmanların yaptığı değerlendirmelerdeki yüksek uyuşma oranı herüstik değerlendirmenin güvenilir olduğunu gösterir. Uyuşma oranı ,70 veya ,60'ın altında ise değerlendirmenin güvenilirliği düşüktür. Ağ kümesinin iyileştirme çalışmalarına uyuşma oranı yüksek olan maddelerden başlanır ve öncelik bu maddelere verilir.

NİTEL ARAŞTIRMALARIN GÜVENİLİRLİĞİNİ ETKİLEYEN FAKTÖRLER

Nitel araştırmaların güvenilirliği titiz, dikkatli çalışılmasına ve çok yönlü bilgi toplanmasına bağlıdır. Güvenilirliği olumsuz yönde etkileyen faktörler azaltıldığı ölçüde çalışmanın güvenilirliği artar. Aşağıdaki bölümde güvenilirliği etkileyen temel faktörlere değinilmiştir.

Verilerin Kalitesi

Nitel araştırmaların güvenilirliği toplanan verilerin kalitesine bağlıdır. Verilerin güvenilirliği konusunda net bir fikir vermek için veriler; teyp bantları, transkripsiyon metinleri, el yazsı metinleri, video çekimleri şeklinde toplanmalı ve bu kanıtlar okuyucuların incelemelerine açık olmalıdır.

Teyp bantları ve video çekimlerine dayalı olarak yapılan yorumların güvenilirliğini birkaç şekilde test etmek mümkündür. Birincisi, teybin veya videonun birden fazla kişi tarafından dinlenilmesi veya seyredilmesi ve yapılan yorumlar arasında tutarlılık bulunup bulunmadığının araştırılmasıdır. İkincisi, teyp çözümlerin en az birden fazla kişi tarafından incelenmesi ve yorumcular arasında tutarlılık bulunma durumunun araştırılmasıdır. Nitel araştırma yapan kişiler bu tür çalışmalarda güvenilirlikten çok "geçerlilik" konusuna önem vermişlerdir. Bu araştırmacılar, "bir nehirden iki kez geçilemeyeceği" savı ile güvenilirlik konusunu bir ölçüde ihmal etmişlerdir.[103]

Kullanılan Yöntemin Uygunluğu

Nitel araştırmaların güvenilirliği ikinci olarak kullanılan yöntemin uygunluğuna bağlıdır. Sosyal bilimlerde beşerî aktivitenin incelenmesinde farklı felsefe ekolleri ve metodoloji yaklaşımlarının bulunması nedeniyle araştırmacının hangi yöntemi uyguladığını açık bir şekilde söylemesi gerekir. Black'e (1993) göre "sosyal bilimler alanında yapılan araştırmalarda, beşerî aktivite ve etkileşime ait bütün yönlerin ele alınması gerekir"(aktaran Cani, 2003).[104] Bu nedenle, araştırmanının niteliğine göre psikoloji, sosyoloji, antropoloji ve yönetim-organizasyon gibi çok değişik disiplinlere ait yöntemlerden yararlanmak mümkündür.

Nitel araştırtmalarda hiçbir yöntem diğerinden daha iyi değildir. Her bir yöntemin kendine göre avantajları ve dezavantajları vardır. Araştırılan olguya uygun yöntemin seçilmesiyle dezavantajlar büyük ölçüde azaltılmış olmalıdır. Veri toplama yöntemi, veri toplama aracı veya araçlarıyla ilgilidir.

Üçleme Yaklaşımının Kullanılma Durumu

Nitel araştırmalarda veri güvenilirliğini sağlamak için araştırmacının birden fazla yöntem veya yaklaşımla verilerin doğrulmasını yapmasıdır. Üçleme yaklaşımı birçok şekilde uygulanabilir:[105]

1. Veri derleme ve toplamada üçleme.

2. Araştırmacı kullanmada üçleme.
3. Kuramsal temele dayanmada üçleme.
4. Metodoloji seçiminde üçleme.
5. Farklı disiplin bilgilerinden yararlanmada üçleme.

Üçleme yönteminden yararlanmanın amacı kanıtları güçlendirmedir. Kanıtlar birden fazla yöntem ve süreçle desteklenirse inanırlık ve güvenilirlik artar. Üçleme yönteminin tek sakıncası maliyetleri artırması ve bazen zaman problemine yol açmasıdır.

Araştırmacının Deneyimi ve Eğitimi

Nitel araştırmaların güvenilirliği bilim adamının yetkin olmasıyla yakından ilgilidir. Kişinin acemi olması, bir konuyu ilk kez inceliyor olması, konu üzerinde çok fazla bir bilgisinin bulunmaması sonuçları büyük ölçüde etkiler. Araştırmacının deneyimi, onun objektif kalma tutumuyla da ilgilidir. Sorunları yorumlarken olguya kendi görüşlerini katması güvenilirliği zedeler.

Nitel araştırmaların bir bölümünde veri toplamak için birden fazla mülâkatçıdan, gözlemciden veya kodlayıcıdan yararlanılır. Toplanan verilerin güvenilirliğini artırmak için bu kişilerin mülâkat yöntemleri, kodlama formülleri ve gözlem konusunda belirli bir eğitimden geçirilmeleri ve pilot uygulama yaptırılarak sınanmaları gerekir.

Veri ve Bilgilerin Tek Bir Araştırmacının Gözlemlerine Bağlı Olması

Nitel araştırmalar "nakledici" veya "rivayet edici" bir söylem içinde yazılır. Veri ve bilgilerin tek bir araştırmacı tarafından nakledilmesi okurda yeterince güven sağlamaz. Bu nedenle niteliksel araştırmaların güvenilirliğini artırmak için birden fazla gözlemciden, araştırmacıdan veya değerlendiriciden yararlanılması önerilmiştir. Bazı araştırmacılar, niteliksel araştırmalarda birden fazla gözlemcinin veya incelemecinin rol alması halinde yorumların güvenilirliğinin artacağını ileri sürmüşlerdir. Ancak niteliksel araştırmalarda gözlemciler arası değerlendirme güvenilirliği yaklaşımının her zaman iyi sonuç vereceği kuşkuludur. Çünkü değişik gözlemcilerin kendi yaklaşım biçimleri, paradigmaları ve bilimsel art yetişimlerinin farklı olması sonuçta yorumların benzer değil, farklı olması sonucunu da doğurabilir.[106]

Seçilen Örneklemin Uygunluğu

Nitel araştırmalarda araştırma yapılacak çalışma grubu "karar örneklemi" çerçevesinde belirlenir. Tesadüfî değil, bilinçli bir şekilde bir kişi, grup veya örgüt seçilmiştir. Örneklemin doğru seçilmesi, sonuçlar üzerinde doğrudan etkilidir.

Sınıflandırmanın ve Kodlamanın Uygunluğu

Bilim adamları niteliksel araştırmalarda oldukça sık sınıflandırma ve tipolojiler geliştirme yöntemine başvururlar. Geliştirilen sınıflandırma ve tipolojilerin gerçeğe uygunluğu yorumların güvenilirliğini artırır. Sınıflandırmaların gerçekle bağıntısı düşükse taşlar yerine oturmaz ve bu tür sınıflandırmalar kalıcı olmaz. Kavramların kodlanması sınıflandırma sistemleri temel alınarak yapılır.

Kodlama güvenilirliğini sağlamak için aynı sınıflandırma sistemine uygun olarak veriler farklı zamanlarda birden fazla araştırmacı tarafından kodlanır ve benzer değerlerin elde edilip edilmediğine bakılır. Literatürde çoğunlukla *kelime anlamlarına dayalı olarak kodlama* uygulaması yapılmıştır. Bilim adamları *yoruma dayalı kodlama* sistemini genelde uygulamazlar.[107]

Tek Vak'a Üzerinde Çalışma

Tek bir vak'adan elde edilen verilerin güvenilirliğini saptamak zordur. Bunun için araştırmacı daha başlangıç aşamasında olguyu birden fazla vak'a üzerinde test etmeyi bir ilke olarak benimseyebilir.

Bütüncül Yaklaşım Yanılgısı

Bütüncül bir yaklaşımla ele alma yanılgısı, olayları gerçekte olduğundan daha fazla ilişkili olarak görmek ve birbiriyle bağıntılandırmaktır. Araştırmacı verilerdeki ayrık durumları, uç vak'aları görmezlikten gelerek bütünleştirme yoluna başvurur. Bu durum projenin bitimine yakın zamanlarda meydana çıkar. Araştırmacı projenin başlangıç safhasında belli bir kanaate vardığından son zamanlarda gözüne çarpan farklılıkları önemsemez. Bütünleştirme yanılgısından kurtulmak için araştırma projesindeki üyelerin birbirlerinin görüşlerini eleştirmelerine fırsat vermeleri gerekir. Bu konuda yararlanılabilecek bir diğer taktik, ayrık değerleri özel olarak büyüteç altına almak ve bu olgulara başka gözlemlerden elde edilmiş kuramsal açıklamalar getirmektir.[108]

Elit Yanlılığı

Elit yanlılığı, sosyal olgu hakkında bilgi toplarken; çok konuşan, daha bilgili ve daha güçlü olan yerel konuşmacılara ağırlık vermektir. Araştırmacı veri toplarken bilgili gibi gözüken bu kişilere dikkat etmelidir. Bu kişilerin görüşleri geniş kitleyi temsil etmiyor olabilir. Elit yanlılığı özellikle araştırmacının sitede kalmak için yeterli zamana sahip olmadığı durumlarda ortaya çıkar.[109]

Yerlilerin Bilgisine Güvenme Eğilimi

Bazı araştırmacılar bilgi toplarken sadece yerli veya yöre halkından kişilerin söylediklerine önem verirler. Bu kişilerin ne yaptıklarına, nasıl hareket ettiklerine ve kendilerine ne yazdırdıklarına bakarlar. Bunu yaparken objektif gözlemde bulunmayı, diğer kişilerin görüşlerine başvurmayı ve dış gerçekliği gözden uzak tutarlar. Sadece yerli bilgisine güvenme araştırma sonuçlarının yanlı olması sonucunu doğurur.[110]

ALINTI YAPILAN KAYNAKLAR

[1] S. Chang, "Surviving Grounded Theory Methodology [Temelli Kuram Metodolojisi]," <http://www.dis.unimelb.edu.au/staff/graeme/615-610resmeth/Powerpoint%20Slides/5> (29.01.2004).

[2] J.M. Morse ve d., "Verification Strategies for Establishing Reliability and Validity in Qualitative Research [Niteliksel Araştırmalarda Güvenilirlik ve Geçerliliği Sağlamak İçin Uygulanabilecek Geçerleme Stratejileri]," <http://www.ualberta.ca/~iiqm/backissues/1_2Final/pdf/morseetal.pdf> (23.01.2004).

[3] B. Trochim, "Qualitative Validity [Niteliksel Geçerlilik]," <http://trochim.human.cornell.edu/kb/qualval.htm> (23.01.2004).

[4] Jacques de Wet ve Zimitri Erasmus, "Qualitative Research as Rigorous Practice [İhtimam Gerektiren Bir Uygulama Olarak Niteliksel Araştırmalar]," 2003, <http://www.interaction.nu.ac.za/sasa2003/de_wet.htm> (15.02.2004).

[5] B. Trochim, "Qualitative Validity [Niteliksel Geçirlilik]," 2000, <http://trochim.human.cornell.edu/kb/qualval.htm> (05.03.2004).

[6] D. Siegle, "Trustworthiness [Doğruluk]," 2004,<http://www.gifted.uconn.edu/siegle/research/Qualitative/trust.htm> (15.02.2004).

[7] B. Trochim, "Qualitative Validity [Niteliksel Geçirlilik]," 2000, <http://trochim.human.cornell.edu/kb/qualval.htm> (15.02.2004).

[8] E.B. Guevara ve E.P. Mendias, "A Comparative Analysis of the Changes in Nursing Practice Related to Health Sector Reform in Five Countries of the Americas [Amerikanın Beş Eyaletinde Sağlık Sektörü Reformu Çerçevesinde Hemşirelik Hizmetleri Uygulamalarıyla İlgili Değişikliklerin Mukayesili Analizi]," <http://www.scielosp.org/pdf/rpsp/v12n5/14093.pdf> (15.02.2004).

[9] J. Twining, "Naturalistic Inquiry into the Collaboratory: In Search Of Understanding For Prospective Participants [Ortak Çalışmalarda Doğalcı Soruşturma Yaklaşımı]," <http://intertwining.org/dissertation/Chapter1.htm> (15.02.2004).

[10] M. Dereshiwsky, "Evaluating the Credibility of Qualitative Studies [Niteliksel Çalışmaların Kredibilitesini Değerlendirme]," <http://jan.ucc.nau.edu/~mid/edr725/class/makingsense/credibility/reading5-3-1.html> (15.02.2004).

[11] K.D. Kelsey "Validity and Reliability of Qualitative Research [Niteliksel Araştırmalarda Geçerlilik ve Güvenilirlik]," <http://www.okstate.edu/ag/agedcm4h/academic/aged6223/power/7qualva.ppt> (15.02.2004).

[12] Twining, "Naturalistic Inquiry."

[13] T. Oka ve I. Shaw "Qualitative Research in Social Work Sosyal Çalışmalarda Niteliksel Araştırmalar]," <http://pweb.sophia.ac.jp/~t-oka/papers/2000/qrsw/qrsw.html> (05.03.2004).

[14] Siegle, "Trustworthiness."

[15] Trochim, "Qualitative Validity."

[16] Oka ve Shaw "Qualitative Research."

[17] T. Oka ve I. Shaw, "Qualitative Research in Social Work [Sosyal Çalışmalarda Ntel Araştırmalar]," 2003, <http://pweb.sophia.ac.jp/~t-oka/papers/2000/qrsw/qrsw.html> (15.02.2004).

[18] Morse ve d., "Verification Strategies."

[19] B. Roberts, "Biographical Research [Biyografik Araştırma]," < http://mcgraw-hill.co.uk/openup/chapters/0335202861.pdf> (13.02.2004).

[20] Aynı.

[21] "Qualitative Research Methods [Niteliksel Araştırma Yöntemleri]," <http://www.ches.ua.edu/health/CNagy/HHE566/Qualitative.ppt> (26.02.2004).

[22] H.L. Dreyfus, "Phenomenology [Olgu Bilim]," <http://www.csun.edu/~vcoao087/phenom.htm> (22.02.2004).

[23] B. Campbell, "Phenomenology as Research Method [Araştırma Yöntemi Olarak Fenomenoloji]," 1998, <http://www.staff.vu.edu.au/syed/alrnnv/papers/bev.html> (22.02.2004).

[24] C. George Boeree, "Qualitative Methods [Nitel Yöntemler]," 1998, <http://www.ship.edu/~cgboeree/qualmethone.html> (22.02.2004).

[25] R. Turner, "Phenomenology [Olgu Bilim]," <http://www.connect.net/ron/phenom.html> (22.02.2004).

[26] CARP, "What is Phenomenology? [Olgu Bilim Nedir?]," <http://www.phenomenologycenter.org/phenom.htm> (22.02.2004).

[27] ThinkQuest , "Qualitative Research [Niteliksel Araştırma]," <http://t3.preservice.org/T0401367/> (22.02.2004).

[28] Russell J Davey, "Rigorous Sex Research: A Phenomenological Perspective [İhtimamlı Seks Araştırmaları: Fenomenolojik Bir Yaklaşım]," 1999, <http://www.latrobe.edu.au/aqr/offer/papers/RDavey.htm> (22.02.2004).

[29] Aynı.

[30] Colorado State University, "Phenomenology [Olgu Bilim]," 2004, <http://writing.colostate.edu/references/research/observe/com3a6.cfm> (22.02.2004).

[31] B. Bowers, "Varieties and Types of Qualitative Methodologies [Niteliksel Yöntemlerin Çeşit ve Türleri]," <http://tiger.coe.missouri.edu/~wang/portfolio/pages/grounded.htm> (23.02.2004).

[32] H. Kara, "Grounded Theory [Temelli Kuram]," 2002, <http://www.weresearchit.co.uk/theory.htm> (22.02.2004).

[33] "Qualitative Research Methods [Niteliksel Araştırma Yöntemleri]," <http://www.ches.ua.edu/health/CNagy/HHE566/Qualitative.ppt> (26.02.2004).

[34] Chang, "Surviving Grounded."

[35] B.A. Kerlin, "Coding Strategies [Kodlama Stratejileri]," <http://kerlins.net/bobbi/research/nudist/coding/strategies.html> (05.03.2004).

[36] S. Kools ve d., "Hospital Experiences of Young Adults With Congenital Heart Disease: Divergence in Expectations and Dissonance in Care [Konjenital Kalp Hastalığı Olan Genç Yetişkinlerin Hastane Deneyimi]," <http://www.findarticles.com/cf_dls/m0NUB/2_11/91087618/p4/article.jhtml?term=> (23.02.2004).

[37] Barbara M. Kinach, "Grounded Theory as Scientific Method [Bilimsel Yöntem Olarak Temelli Kuram]," <http://www.ed.uiuc.edu/EPS/PES-yearbook/95_docs/kinach.html> (23.02.2004).

[38] psanday@sas.upenn.edu, "What is Ethnography? [Etnografya Nedir?]," <http://www.sas.upenn.edu/anthro/CPIA/METHODS/Ethnography.html> (13.02.2004).

[39] "Qualitative Research Methods [Niteliksel Araştırma Yöntemleri]," <http://www.ches.ua.edu/health/CNagy/HHE566/Qualitative.ppt> (26.02.2004).

[40] D. Garson, "Ethnographic Research [Etnografya Araştırmaları]," <http://www2.chass.ncsu.edu/garson/pa765/ethno.htm> (13.02.2004).

[41] Aynı.

[42] Aynı.

[43] J. Hadley, "The Culture of Learning and the Good Teacher in Japan: An Analysis of Student Views [Japonyada İyi Öğretmen ve Öğrenme Kültürü: Öğrenci Görüşlerinin İncelenmesi]," <http://www.cels.bham.ac.uk/resources/essays/HadleyRes.PDF> (13.02.2004).

[44] Aynı.

[45] A. Gibbs, "Focus Groups [Dak Grupları]," <http://www.soc.surrey.ac.uk/sru/SRU19.html> (13.02.2004).

[46] F.I. Luntz, "Focus Group Research in American Politics [Amerikan Politikasında Odak Grubu Araştırmaları]," <http://www.pollingreport.com/focus.htm> (13.02.2004).

[47] E. McAteer, "Focus Groups [Odak Grupları]," <http://www.icbl.hw.ac.uk/ltdi/cookbook/focus_groups/index.html#endhead> (13.02.2004).

[48] C. Nagy, "Qualitative Evaluation Methods [Niteliksel Değerlendirme Yöntemleri]," <http://www.ches.ua.edu/health/CNagy/HHE566/19> (29.01.2004).

[49] American Statistical Association, "What Are Focus Groups [Odak grupları Nedir?]," <http://www.stat.ncsu.edu/info/srms/surveyfocus.pdf> (13.02.2004).

[50] Luntz, "Focus Group."

[51] D. Viehland, "Case Study Research [Vak'a Etüdü Araştırması]," <http://www.massey.ac.nz/~dviehlan/1> (03.01.2004).

[52] J. Rowley, "Using Case Studies in Research [Vak'a Etüdlerinin Araştırmalarda Kullanılması]," <

[53] Prism, "Constructs of Loyalty [Sadakat Yapıları]," <http://www.prism.com.mt/pdf/PRISM_WP_Loyalty_Constructs.PDF> (03.01.2004).

[54] W. Telis, "Introduction to Case Study [Vak'a Etüdüne Giriş]," <http://www.nova.edu/ssss/QR/QR3-2/tellis1.html> (03.01.2004).

[55] V. Salin, "Case Study Methods in Agribusiness Research and Education [Tarım İşletmeciliği ve Eğitim Araştırmalarında Vak'a Etüdü Yöntemi]," <http://agecon.tamu.edu/faculty/salin/casestudy/Coursenotes.ppt> (26.02.2004).

[56] Aynı.

[57] American Institus for Research, "Critical İncident Technique [Kritik Olay Tekniği]," <http://www.air-dc.org/overview/cit-set.htm> (29.01.2004).

[58] Jane E. Fountain, "A Note on the Critical Incident Technique and its Utility as a Tool of Public Management Research [Kamu Yönetimi Araştırmalarında Kritik Olay Tekniğinin Kullanılması]," 1999, <http://www.ksg.harvard.edu/prg/fountain/citechnique.pdf> (26.02.2004).

[59] H. Bromley vd., "The Qualitative Research and Health Working Group [Niteliksel Araştırma ve Sağlık Çalışanları Grubu]," <http://www.liv.ac.uk/lstm/download/glossary.pdf> (06.03.2004).

[60] "Types of Research Design [Araştırma Tasarımı Türleri]," <http://www.dur.ac.uk/r.j.coe/resmeths/types.doc> (04.01.2004).

[61] E. Lindhult, "The Quality Of Action Research [Eylem Araştırmalarının Güvenilirliği]," <http://www.hh.se/hss/Papers/papers/lindhult.pdf> (06.03.2004).

[62] C. Dimmitt, J. Carey ve Carole Schweid, "Action Research and Evaluation [Eylem Araştırması ve Değerlendirme]," <http://www.umass.edu/schoolcounseling/WelcometoAmherstMassachusetts/ActionResearchandEvaluation.ppt> (05.03.2004).

[63] Richard Watson Todd, "Why Do Action Research? [Niçin Eylem Araştırması Yapanız?]," <http://www.philseflsupport.com/why_ar.asp> (06.03.2004).

[64] Lindhult, "The Quality Of Action."

[65] Stenhouse, "Evaluating Information for Relevance and Validity [Bilgilerin İlgililik ve Geçerlilik Açısından Değerlendirilmesi]," <http://www.stenhouse.com/pdfs/8133ch01.pdf> (06.03.2004).

[66] Aynı.

[67] B. Johonson, "Historical Research [Tarihsel Araştırma]," 2004, <http://www.southalabama.edu/coe/bset/johnson/lectures/ch12.PDF> (07.02.2004).

[68] Aynı.

[69] Aynı.

[70] Writing Center Navigator, "An Introduction to Content Analysis [İçerik Analizine Giriş]," <http://writing.colostate.edu/references/research/content/pop2a.cfm> (07.02.2004).

[71] B. Trochim, "Unobtrusive Measures [Rahatsız Edilmeden Yapılan Ölçümler]," 2000, <http://trochim.human.cornell.edu/kb/unobtrus.htm> (19.02.2004).

[72] Aynı.

[73] Peter E. Doolittle, "Instructional Design for Web-based Instruction [Ağ Temelli Eğitim İçin Tasarım]," <http://edpsychserver.ed.vt.edu/workshops/edtech/pdf/handouts1.pdf> (21.02.2004).

[74] H. Bruce, "Content Analysis [İçerik Analizi]," <http://www.ischool.washington.edu/harryb/courses/LIS570Winter03/contentanalysis.doc> (07.02.2004).

[75] Bu konuda *bk.*, GB Software , "Development of the Scales [Ölçeklerin Geliştirilmesi]," <http://www.gb-software.com/develop.htm> (19.02.2004).

[76] "Content Analysis, What it is, and Where it is Used [İçerik Analizi: Nedir ve Nerede Kullanılır?]," <http://www.tele.sunyit.edu/notes_content_analysis.htm> (07.02.2004).

[77] K. Rodenburg, "Content Analysis of Daily Newspaper From the Front Page [Günlük Gazetelerin Ön sayfalarında İçerik Analizi Yapma]," <http://home.student.utwente.nl/m.ivankovic/Ivankovic_content_%20analysis.pdf> (19.02.2004).

[78] Instructional Support Service, "Content Analysis [İçerik Analizi]," <http://mccoy.lib.siu.edu/projects/psyc/schmeck/iv2.ppt> (19.02.2004).

[79] S. Stemler, "An Overview of Content Analysis [İçerik Analizinin Gözden Geçirilmesi]," <http://pareonline.net/getvn.asp?v=7&n=17> (07.02.2004).

[80] K. Rodenburg, "Content Analysis of Dailey Newspaper From the Front Page [Günlük Gazetelerin Ön Sayfalarının İçerik Analizi]," <http://home.student.utwente.nl/m.ivankovic/Ivankovic_content_%20analysis.pdf> (07.02.2004).

[81] S. Stemler, "An Introduction to Content Analysis [İçerik Analizine Giriş]," <http://www.ericdigests.org/2002-2/content.htm> (19.02.2004).

[82] UCLA Department of Applied Linguistics, "Discourse Analysis [Söylem Araştırmaları]," <http://www.humnet.ucla.edu/humnet/al/textonly/discoursetxt.htm> (14.02.2004).

[83] I. Karasavvidis, "Qualitative Discourse Analysis [Nitel Söylem Araştırmaları]," <http://www.ikaras.org/d_qualitative_analysis.php> (14.02.2004).

[84] Nico Carpentier en Sonja Spee, "Discoursanalyse [Söylem Analizi]," <http://homepages.vub.ac.be/~ncarpent/disc.html> (26.02.2004).

[85] R.A. Palmquist, "Discourse Analysis [Söylem Araştırması]," <http://www.gslis.utexas.edu/~palmquis/courses/discourse.htm> (14.02.2004).

[86] UCLA, "Discourse Analysis."

[87] Palmquist, "Discourse Analysis ."

[88] Aynı.

[89] Del Siegle, "The Long Interview [Uzun Mülâkatlar]," <http://www.gifted.uconn.edu/siegle/research/Qualitative/long.htm> (06.03.2004).

[90] Richard K. Coll ve Richard Chapman, "Choices of Methodology for Cooperative Education [İşbirlikçi Eğitim İçin Yöntem Bilim Tercihleri]," <http://www.apjce.org/volume_1/volume_1_1_pp_1_8.pdf> (15.02.2004).

[91] Del Siegle, "A Phenomenological Approach to In-Depth Interviewing [Derinlemesine Mülâkat Tekniğine Olgu Bilimsel Yaklaşım]," <http://www.gifted.uconn.edu/siegle/research/Qualitative/intervie.htm> (06.03.2004).

[92] P.S. Jensen ve C.Edelbrock "Subject and Interview Characteristics ... [Süjeler ve Mülâkat Özellikleri ...]," <http://www.findarticles.com/cf_dls/m0902/6_27/59037724/p1/article.jhtml> (06.03.2004).

[93] John M. Bushery, J. Michael Brick, Jacqueline Severynse ve Richard A. McGuinness, "How Interview Mode Affects Data Reliability [Mülâkat Şekli Verilerin Güvenilirliğini Ne Şekilde Etkileyebilir?]," <http://srsstats.sbe.nsf.gov/docs/research/5-31.pdf> (06.03.2004).

[94] M.C. Cawley, "Using Observation to Evaluate Extension Programs Genişletme Programlarının Değerlendirilmesinde Gözlem Yöntemini Uygulama]," <http://www.uidaho.edu/extension/observation.ppt> (07.03.2004).

[95] Cawley, "Using Observation."

[96] P. Catherine ve M. Nicholas, "Observational Methods in Health Care Settings [Sağlık Kuruluşlarında Gözlem Yöntemi]," <Observational methods in health care settings> (21.02.2004).

[97] Catherine ve Nicholas, "Observational Methods."

[98] K. Instone, "Site Usability Evaluation [Site Kullanışlılık Değerlendirmesi]," <http://user-experience.org/uefiles/writings/siteeval.html> (14.02.2004).

[99] Sussex University, "A Brief Introduction to Heuristic Evaluation [Herüstik Değerlendirmeye Kısa Bir Giriş]," <http://www.id-book.com/catherb/Introduction.php> (14.02.2004).

[100] Instone, "Site Usability."

[101] Stanford University , "Heuristic Evaluation [Herüstik Değerlendirme]," <http://www.stanford.edu/group/web-creators/heuristics.htm> (14.02.2004).

[102] J. Nielsen, "Severity Ratings for Usability Problems [Kullanışlılık Sorunlarının Değerlendirilmesinde Dereceleme Ölçekleri]," <http://www.useit.com/papers/heuristic/severityrating.html> (14.02.2004).

[103] D. Ratcliff, "Validity and Reliability in Qualitative Research [Niteliksel Araştırmalarda Geçerlilik ve Güvenilirlik]," <http://don.ratcliff.net/qual/validity.html> (20.01.2004).

[104] V. Canı, "Reliability and Validity in Qualitative Research [Niteliksel Araştırmalarda Güvenilirlik ve Geçerlilik]," <http://www.qmuc.ac.uk/psych/RTrek/study_notes/web/sn5.htm> (23.01.2004).

[105] J. Keyton, "Introduction to Qualitative Research [Niteliksel Araştırmaya Giriş]," <http://people.ku.edu/~jkeyton/methods/17> (23.01.2004).

[106] Catherine Pope, Sue Ziebland ve Nicholas Mays, "Analysing qualitative data [Niteliksel Verilerin Analizi]," <http://www.findarticles.com/cf_dls/m0999/7227_320/59110536/p3/article.jhtml?term> (24.01.2004).

[107] S. Purchase ve T. Ward, "Developing Better Definitions Through Analytical Triangulation [Analitik Üçleme Yöntemiyle Daha İyi Tanımlar Geliştirme]," <http://130.195.95.71:8081/www/ANZMAC2000/CDsite/papers/p/Purchas1.PDF> (15.02.2004).

[108] C.D. Cooper, "Not Just A Numbers Thing: Tactics For Improving Reliability And Validity In Qualitative Research [Sadece Sayısal Bir Şey Değil: Nitel Araştırmalarda Geçerlilik ve Güvenilirliği Artırma Taktikleri]," <http://www.aom.pace.edu/rmd/2001forum/methods_article_with_refs.pdf> (15.02.2004).

[109] Aynı.

[110] Aynı.

TEST TÜRÜ BAZINDA GÜVENİLİRLİK ANALİZLERİ

Önceki bölümlerde genel olarak güvenilirlik analizi yöntemleri, güvenilirlik hesaplamalarında kullanılan matematiksel ve istatistiksel işlemler ile araştırma türüne özgü güvenilirlik analizi yöntemleri üzerinde durulmuştu. Bu bölümde ise güvenilirlik analizi yöntemleri *test türü bazında* ele alınmıştır. Testlerin bilgisayar ortamında etkileşimli olarak uygulanmasının güvenilirliği ne şekilde etkileyeceği, hız testlerinde güvenilirlik analizlerinin nasıl yapılacağı, ipsatif ölçeklerde, tanımlama ölçekleriyle yetenek testlerinde uygulanan güvenilirlik analizleri bu bölümde ele alınan başlıca konulardır.

BİLGİSAYAR TEMELLİ TESTLERDE GÜVENİLİRLİK

Günümüzde kağıt-kalem testlerinin matbu formlarının yanında bilgisayar ortamına alınmış olan sürümleri de uygulanabilmektedir. Bunlara "bilgisayar temelli testler" (BTT) adı verilir. Bilgisayar temelli testler bağımsız terminallerde, kurumsal ağ ortamındaki terminallerde veya İnternet ortamında uygulanabilir.

Niteliği

Bilgisayar temelli testler, kağıt-kalem testlerinin elektronik sürümünü oluşturur. Bu testler, kağıt-kalem testlerinde (KKT) olduğu kadar soru içerir ve benzer cevaplama süreleri içinde uygulanır. Cevaplayıcılar eğer bilgisayara yeterince hakim değillerse bu durumda kendilerine işaretleme için kağıt-kalem formundan biraz daha uzun süre verilebilir. Test sonuçları bilgisayarda otomatik olarak hesaplanır ve kişilerin puanları norm grubuyla karşılaştırmalı olarak değerlendirilir.

Bilgisayar temelli testlerin bir bölümü nispeten basit bir düzenleme içinde çalışırken ticarî şirketler tarafından hazırlanan bilgisayar temelli

testler bilgi verme ve raporlama özellikleri açısından çok daha kapsamlıdır. Kapsamlı bir şekilde hazırlanmış bilgisayar temelli testlerde, testin bütünü ve maddelerin her biri için gerekli olan psikometrik analizlerin tamamı bilgisayar ortamında ve programın kendi içinde yapılır. Verileri ayrıca bir istatistik analiz programının içine almaya gerek yoktur.

Ön Koşullar

Bilgisayar temelli testlerin sağlıklı bir şekilde uygulanabilmesi için belirli ön koşulların yerine getirilmiş olması gerekir. Bu koşullar sağlanmadığı durumda sonuçların güvenilirliği kuşkulu hale gelir. Ring'e (1994) göre bu koşullar şunlardır:[1]

1. Test alacak kişiler kağıt-kalem testine göre bilgisayar ortamında daha az avantajlı konuma düşmüş olmamalıdırlar.
2. Bilgisayar ortamındaki olanaklar tam olarak kullanılmış olmalıdır.
3. Test uygulama ortamındaki uygulayıcı kontrolü en üst düzeye çıkarılmış olmalıdır.

Bilgisayarda test uygulayan bazı kişilerin mönüler arasında rahat ve serbest bir şekilde gezinememeleri, geri dönüp yanıtları gözden geçirememeleri, fareyi rahat ve hassas bir şekilde kullanamamaları, tuşlara çekinerek basmaları elde edecekleri puanları olumsuz yönde etkiler. Bu nedenle APA bilgisayar ortamında yapılan testlerde kağıt kalem testlerinde olduğu gibi kişilerin cevaplarını gözden geçirmelerine ve geri besleme almalarına olanak sağlanmasını istemiştir.[2] Test alan kişiler sadece bulundukları ekran üzerinde değil, yanıtını verdikleri sorulardan sonra bir başka ekrana geçtiklerinde dahi geriye dönüp düzeltme yapabilmelidirler. Bu düzeltmenin yapılamadığı durumda kişiler kendilerini rahatsız hissederler. Verilen cevaplarda aynı ekran üzerinde veya daha sonraki ekranlardan geriye dönülerek değişiklik yapılamaması sonuçların güvenilirlik ve geçerliliğini etkiler. Ayrıca kağıt-kalem testlerinde olduğu gibi kişiler gerektiğinde bazı sorulara hiç cevap vermek de istemeyebilirler. Bu gibi durumlarda kişilere zorunlu işaretleme yaptırmak sonuçların güvenilirliğini düşürür.

Bilgisayar temelli testlerde bir diğer önemli konu geri besleme bilgileridir. Katılımcılar uygulama sırasında kaçıncı soruya geldiklerini, kaç dakika zamanları kaldığını her bir ekranda açık bir şekilde görebilmelidirler.

Güvenilirlik Analizleri

Bu testlerin bir bütün olarak veya madde bazında yapılacak güvenilirlik analizlerinde genellikle klasik ölçüm teorisi temel alınır. Bununla birlikte araştırmacı isterse modern test kuramına göre kalibre edilmiş test maddelerinden de yararlanabilir. Bilgisayar temelli testlerde güvenilirlik analizleri kullanılan program içine alınmıştır. Eğer test yazılımı bu hesaplamaları yapacak şekilde üretilmemişse test uygulamasının sonuçları programdan veri matrisi şeklinde alınarak değişik istatistik yazılımlarına yüklenir ve güvenilirlik hesaplamaları söz konusu yazılımlarda yapılır.

Araştırmacılar bilgisayar temelli testlerde şu soruların cevaplarını merak etmişlerdir:

1. Matbu formlar kullanılarak ve el ile işaretleme yapılarak (manüel)[a] uygulanan testlerin güvenilirlik değerleriyle bilgisayar ortamında uygulanan testlerin güvenilirlik değerleri arasında fark var mıdır?
2. Bilgisayarda birden fazla test bir "batarya" halinde uygulandığında acaba güvenilirlik değerleri veya katsayıları bu şekildeki toplu uygulamadan etkilenmekte midir?
3. Karma düzenleme (bilgisayar-matbu form) halinde uygulanan testlerde güvenilirlik sonuçları önemli ölçüde değişiyor mu?

Bilgisayar temelli testlerin güvenilirliğinde klasik ölçüm kuramına uygun olarak alfa iç tutarlılık analizi, yarıya bölme güvenilirliği, paralel formlar güvenilirliği ve test-yeniden test güvenilirliği hesaplamaları yapılabilir. Yapılan araştırmalar testlerin bilgisayar ortamında bir batarya olarak uygulanması halinde bireysel test güvenilirliğinin önemli ölçüde etkilenmediğini ortaya koymuştur.[3] Aynı şekilde karma test uygulaması da benzer sonuçlar vermiştir. Bununla birlikte psikolojik testlere göre nörodavranışsal test sonuçlarının güvenilirliği karma uygulamalarda daha düşük çıkmıştır.[4]

BİLGİSAYAR UYARLI TESTLERDE GÜVENİLİRLİK

Bilgisayar uyarlı testler (BUT), madde-yanıt kuramı temel alınarak kişilerin yetenek düzeylerine uygun soruların sorulduğu ölçüm uygulamalarına dayanır. Araştırmacı kişilerin yeteneklerini ölçerken madde-yanıt kuramı-

---

[a] *Manüel* kelimesi dilimize Fransızcadan "el kitabı" anlamında geçmiştir. Buradaki kullanım biçimi ise "el ile işaretleme" veya "eller kullanılarak" anlamındadır.

na ait 1PL, 2PL ve 3PL gibi değişik modellerden yaralanabilir. Bunun için öncelikle sadece tek bir kavramsal boyutun ölçüldüğü bir soru havuzunun oluşturulması gerekir. Lord (1980) bilgisayar uyarlı ölçümlerde test sorularının tek boyutlu olmasını bir ön koşul olarak görmüştür (aktaran Loerke, 2002).[5] Lord, tek boyutluluk kriterini karşılamak için "kaba" bir prosedür ileri sürmüş ve maddeler arası tetrakorik korelasyon katsayılarının öz değer rakamlarına bakılmasını önermiştir. Green ve arkadaşları (1984) ise yapılacak faktör analizi sonucunda toplam ortak varyansın ,70'ini açıklayan "basit yapının" veya "tek bir faktörün" ortaya çıkması halinde tek boyutluluk koşulunun sağlanmış olacağını ileri sürmüşlerdir. Bu araştırmacılar toplam ortak varyansı ,50'nin üzerinde bir rakamla açıklayan bir faktör çıkması halinde dahi ölçüm aracının tek boyutlu bir yapı olarak değerlendirilebileceğini söylerlerken ,50'nin altındaki toplam varyans değerlerini alt ölçek yapılanması olarak görmüşlerdir.[6]

Bilgisayarların temel alınan ölçüm modelleri çerçevesinde cevaplayıcının yetenek düzeyine uygun soru seçmesi ve sorması gibi programlama özellikleri nedeniyle bilgisayar uyarlı testlerde cevaplandırılan madde sayısı daha azdır. Daha az sayıda madde ile kişinin yetenek ve becerisinin ne olduğuna karar verilir. Bunun için daha önceden bilgisayarlarda kişinin eğitim durumu, yaşı, mesleği dikkate alınarak farklı sayıda soru/ifade kompozisyonları oluşturulur. Bilgisayar uyarlı testlerde iç tutarlılık, yarıya bölme, test – yeniden test yöntemleri değil, madde özellikleri eğrisi yaklaşımı kullanılır. Madde özellikleri kuramındaki "kişi ayırma indeksi"[a] ve "madde ayırma indeksi" değerleri güvenilirlik ölçüsü olarak değerlendirilir.

Uygulama Biçimi

Bilgisayar uyarlı testlerde başlangıç sorusunu verme değişik şekillerde düzenlenebilir. Bazı uygulamalarda alt testler şeklinde ve diğerlerinde ise herhangi bir gruplama yapılmadan kolaydan zora doğru giden bir sıralama yapılmıştır. Maddelerin alt testler halinde zorluk sıralamasına sokulduğu bilgisayar uyarlı testlerde elde mevcut veri yoksa yanıtlayıcı orta derecedeki başarıyı gösteren üçüncü veya dördüncü kademedeki zorluk derecesinden teste başlar. Süreçte arka arkaya üç test sorusuna yanlış cevap vermişse bilgisayar otomatik olarak daha kolay olan bir alt teste geçer. Kolaydan zora doğru giden sıralamanın temel alındığı testlerde ise katılımcı-

[a] Bu değerler, Rasch ölçüm modelinde KR-20 veya Cronbach alfa güvenilirlik değerlerinin eşitidir.

lar eğer yetenek düzeyleri biliniyorsa bu düzeyden; bilinmiyorsa orta güçlük derecesindeki sorudan başlarlar. Doğru yanıt verdikçe daha zor soruları yanlış cevap verdikçe ise daha kolay soruları alırlar. Bilgisayar uyarlı testlerde bir maddeye doğru yanıt verme olasılığı maddenin zorluk derecesine ve kişinin yetenek düzeyine bağlıdır.

- Bir maddeye doğru yanıt verme olasılığı.

$$\text{Olasılık (doğru cevap} \mid \text{yetenek, güçlük)} = \frac{1}{1 + e^{-\text{yetenek ve güçlük}}}. \qquad (12\text{-}1)$$

Literatürde BUT uygulama stratejileri değişik başlıklar altında sırlanmıştır. Bu kitabın temel amacını oluşturmaması nedeniyle bu stratejilerin ayrıntısına girilmeden sadece başlıkları verilmiştir.[7]

1. Çok düzeyli uygulama.
2. İki aşamalı uygulama.
3. Kendinden seçimli uygulama.
4. Piramitsel çok aşamalı uygulama.
5. Tabakalı uygulama.

Yapılan araştırmalar bilgisayar uyarlı testlerde çok az sayıdaki maddenin dahi etkili sonuçlar verebileceğini ortaya koymuştur. Hornke (2000) BUT'da sadece beş maddenin bulunmasının dahi yeterli olabileceğini belirtmiştir. Yaptığı araştırmada 456 maddeli bir havuzdan test seçen programın üzerinde ölçüm yapılan 5912 kişinin yüzde 90'nında 10 veya daha az maddenin etkili sonuçlar verdiğini görmüştür. Ortalama 7,5 maddede sonuç alınmıştır. Hornke cevaplayıcılara her bir madde için üç dakika gibi gereğinden fazla uzun bir zaman vermiş, testin bitimi için süre veya soru sayısı yerine "maksimum bilgi algoritmasını" kullanmıştır.[8] Yapılan başka araştırmalarda da soru havuzundaki maddelerin en fazla %20'sinin sistem tarafından kullanıldığı saptanmıştır.

Bilgisayar uyarlı testlerde "uyarlama süreci" katılımcının önceden belirlenmiş olan belirli bir durdurma kriterini sağlamasına kadar devam eder.

Durdurma kriteri *soru sayısına* göre, *ölçümün standart hatasına* veya ölçümün *kesinlik derecesine* göre belirlenir. Ölçümün standart hatası ,40 veya ,38 gibi bir değer temel alınarak belirlenir. Bu rakam klasik güvenilirlik değerlendirmesinde ,85 güvenilirlik katsayısına denktir.[9] Ölçümün kesinlik derecesi ise madde-yanıt kuramındaki madde bilgi fonksiyonuyla saptanır.

Bilgisayar uyarlı testlerde yüksek yetenekli bireylerle düşük yetenekli bireyler aynı sayıda veya oranda doğru yanıta sahiptirler. Sadece yetenekli bireyler zor soruları düşük yetenekli bireyler ise kolay soruları çözmüşlerdir. Zor soruları çözmeleri nedeniyle yüksek yetenekli bireylerin puanları düşük yetenekli bireylerin puanlarından daha yüksektir.

Amerika Birleşik Devletlerinde AERA, APA ve NCME'nin ortaklaşa yayımladıkları *Standartlar* kitabında, geliştirilen bilgisayar uyarlı testin arkasında yatan mantığın ve sonuçları destekleyici kanıtların belgelenmesi gerektiği belirtilmiştir. Test geliştiricileri; yanlış cevap verildiğinde daha kolay test sorularının hangi mantığa göre seçildiğini, başlangıç noktasının nasıl belirlendiğini, testin nasıl sona erdiğini, test puanlamasının nasıl yapıldığını ve maddelerin açığa çıkmasını nasıl kontrol altına aldıklarını okuyucularına veya testi alan kişilere açıklamalıdırlar.[10]

Bilgisayar uyarlı testlerde ölçümü yapılan kişinin tüm test bilgileri programın veri tabanında tutularak daha sonraki test uygulamalarında karşılaştırma yapmak üzere saklanır. Böylece kişinin birinci test uygulamasıyla daha sonraki test uygulamalarında aldığı puanlar birbirleriyle karşılaştırılabilir.

Son yıllarda bilgisayar uyarlı testlerin çevrimiçi sürümleri yayımlanmaya başlanmıştır. Katılımcılar İnternet ortamına alınan bu testleri uygulayarak kendi durumlarını değerlendirme şansına sahiptirler.

Kişilerin yeteneklerine göre yeni test maddelerinin seçilmesinde veya zor olan maddelerin düşürülmesinde farklı algoritmalardan yararlanılır (Hulin, Drasgow and Parsons 1983; Samejima 1983, aktaran Sabatini).[11] Bu kitabın birincil amacını oluşturmadığından söz konusu algoritmaların ayrıntısına girilmemiştir.

Testin Süresi ve Uzunluğu

Bilgisayar uyarlı testlerde yanıtlayıcı ekranda görülen test sorularını genellikle 60 saniye içinde cevaplandırmak zorundadır. Bilgisayar uyarlı testler kağıt-kalem testlerine ve bilgisayar temelli testlere göre daha kısadır. Böylece katılımcılar daha fazla motive olurlar ve her bir test maddesini daha

dikkatli ve özenli bir şekilde cevaplandırırlar. Ancak bu uygulama aynı zamanda test sonucundan kuşku duyulması gibi duruma da yol açabilir. Kişiler az sayıda madde ile kendilerinin yetenek ve becerilerinin tam olarak açığa çıkarılamayacağını ve yeteneklerinin doğru ölçülmediğini düşünebilirler. Bu nedenle bilgisayar uyarlı testlerde yetenekleri ortaya çıkaracak, kişisel tatmini sağlayacak, fakat aynı zamanda kişileri yormayacak bir düzenleme yapılmalıdır.

Bilgisayar uyarlı testlerde yetenek düzeyleri ne olursa olsun tüm kişiler testi aynı zaman süresi içinde tamamlarlar. Kağıt kalem testlerinde olduğu gibi bazı kişilerin testi erken ve bazılarının ise geç tamamlaması gibi bir durum söz konusu değildir. Bilgisayar uyarlı testlerde kişilere verilen test maddelerinin sayısı ile madde havuzuna alınan madde sayısı aynı değildir. Test maddelerinin deşifre olmaması için havuza aynı konu veya içerikle ilgili çok sayıda madde alınır. Eskiden 100 olarak düşünülen madde sayısı son yıllarda daha da arttırılmıştır. Kısa sürede ve çok sayıda kişiye uygulanan *yüksek nitelikli testlerde* maddelerin açığa çıkma olasılığının bulunması nedeniyle havuza nihaî ölçekteki madde sayısının sekiz veya on katı kadar madde alınır. Wise ve Kingsbury'e göre (2000) günümüzde bilgisayar uyarlı test havuzlarında 1000'den fazla madde vardır ve eğer test sertifikasyon amaçlı olarak kullanılıyorsa madde sayısı 2000'i aşar.[12]

Güvenilirliği

Bilgisayar uyarlı testlerin uzunluğu kağıt-kalem testlerine göre daha kısadır. Kağıt-kalem testleriyle elde edilen aynı güvenilirlik katsayısı bu kez maddelerin yarısıyla elde edilebilir. Çünkü seçilen maddeler test alan kişinin yetenek düzeyine göre belirlenmiştir.[13] Bilgisayar uyarlı testlerde önemli olan sorunlardan biri kullanılan programa göre test bankasındaki bazı maddelerin program tarafından daha fazla seçilmesi ve bu maddelerin gereğinden fazla olarak açığa çıkması veya deşifre olmasıdır. Böyle olunca bu maddelerin testin güvenilirliğini düşürme tehlikesi söz konusudur. Havuzdaki bazı maddeler hiç kullanılmazken bazılarının aşırı açığa çıkması madde seçim sisteminin yeterince sağlıklı çalışmadığının kanıtı olarak görülebilir.

Bilgisayar uyarlı testlerden elde edilen puanın güvenilir olup olmadığını belirlemek için iki yöntemden yararlanılabilir: test-yeniden test yöntemi ve paralel formlar yöntemi.[14] Test-yeniden test yönteminde aynı kişiye test başka bir zaman diliminde ikinci bir kez daha aldırılır. Paralel formlar yönteminde ise aynı zaman diliminde ve aynı soru bankasından çekilmek üzere (daha önce çıkmış olan sorular çıkarılarak) bir başka test daha uygulatı-

lır ve iki test sonuçları arasındaki korelasyona bakılır. Kişilere ait bilgisayar uyarlı test puanları tek başına yorum yapmak için yeterince güçlü değildir. Bu puanları daha sağlıklı bir şekilde yorumlamak için ölçümün standart hatası değerlerini ayrıca hesaplamak gerekir.

İNTERNET ORTAMINDAKİ TESTLERİN GÜVENİLİRLİĞİ

Son yıllarda değişik nitelikteki yetenek, bilgi ve başarı testleri ile tutum ölçekleri İnternet ortamında uygulanmaya başlamıştır. İnternet testleri, kağıt-kalem testlerinin kopyası niteliğinde veya madde-yanıt kuramının temel alındığı bilgisayar uyarlı olarak yapılıyor olabilir. İnternet uyarlı testlerin çok uzaktaki kişilerin sınanmasına imkan vermesi, test sonuçlarının çok çabuk alınması, kişilerin kabiliyetlerinin daha az soruyla çok kısa zamanda saptanması gibi yararları olmasına karşılık test uygulama ortamından kaynaklanan güvenilirlik ve geçerlilik sorunları vardır. İnternet ortamında uygulanan test bir tutum ölçeği ise kağıt-kalem sürümünde olduğu gibi tek boyutlu veya çok boyutlu olması önceden belirlenir.

İnternet Ortamında Test Uygulama Zorlukları

İnternet ortamında uygulanan testlerin uygulama ve yorumlanmasında bir takım zorluklar söz konusudur. İnternet testleri bir yöneticinin nezaretinde ve bilgisayar ortamında uygulanan testlere benzemez. Dikkat edilmesi gereken bir çok faktör kontrol dışındadır.

Uygulama yeknesaklığı yetersizliği. Bu testlerin diğer bilgisayar uyarlı ve bilgisayar temelli testlerden farkı çoğunlukla makinenin başında yönlendirici bir görevlinin bulunmadığı ortamlarda uygulanmasıdır. Bu nedenle test uygulaması sırasında kişinin destek alıp almadığı tam olarak bilinemez. Ayrıca test süresi, soruları çözme sırası, verilen talimatlara uyma derecesi saptanamaz. Testler standart olmayan şartlarda uygulanır.

Yeknesaklık yetersizliği, alınan puanların ne ölçüde gerçeği yansıttığı konusunda kuşku yaratır. Uygulamaların çoğunda kişilerin isimlerini ve demografik bilgilerini saklamaları, gerçek isimleri yerine başka bir isim vermeleri test alan kişilerin testi doldurma konusunda yeterince samimi olmadıkları düşüncesine yol açar. Bazı vakalarda ismi saklama daha gerçekçi bir cevaplama imkanı sağlarken diğerlerinde tam tersi olabilir ve araştırmacının bunu saptaması imkansızdır. İnternet ortamında test alan kişilerin verdikleri demografik bilgilere, telefonla teyit edilmediği sürece, tam olarak güvenilmeyeceğinden bu bilgilere dayalı olarak cinsiyet, yaş, eğitim ve meslek norm değerleri oluşturulamaz.

Bilgisayar yetkinliği sorunları. Bilim adamları ölçüm işleminin kağıt-kalem ortamından bilgisayar ekranına taşınmasıyla zihnin problem çözme kapasitesinde bir takım değişiklikler olabileceği endişesini taşırlar.[15] Bu tür testlerde "Bilgisayara aşina olan kişilerle bilgisayarı hiç kullanmamış veya çok az kullanmış kişilerin bilgi ve yetenek testlerinden alabilecekleri farklı puanlar acaba onların gerçek durumlarını yansıtıyor mu?" sorusu gündeme gelir. Özellikle bilgisayar ortamında uygulanan *hız testlerinde* deneyimli ve deneyimsiz kişilerin aynı ölçüde performans göstermeleri imkansız gibidir. Araştırmalar kullanılan ölçüm aracı türünün özellikle hız testlerinde test sonuçlarının eşitliğini etkilediğini ortaya koymuştur. Böyle olunca farklı uygulama modellerinde testlerin geçerlilik ve güvenilirlik sorunu gündeme gelir.[16] Test alan kişilerin bilgisayar cahili olmaları test sonuçlarının geçerliliğini tehdit eder.

Testin uzunluğu. İnternet ortamında uygulanan testlerin güvenilirliğini etkileyen bir diğer faktör uzunluktur. Uzun testlerin tamamlanma süresi 30 dakikadan bir saate kadar uzayabilir. Ancak bu testlerin dikkatli bir biçimde doldurulduğu veya doldurulacağı konusunda şüpheler vardır. Bu nedenle İnternet testlerinin genellikle 15-20 dakikayı aşmaz istenmez.

İnternet Temelli Testlerin Türleri

Günümüzde artık her türlü test, ölçek ve envanterler İnternet ortamına alınabilmektedir. İnternet temelli testlerin her birinin kendine özgü özel uygulama koşulları vardır. Bu tür testlere güven duyulabilmesi ancak titiz, dikkatli ve özenli bir biçimde hazırlandığı zaman mümkün olabilir.

İnternet testleri ile personel seçimi. İnternet uygulamalarının yaygınlaşması ile işletme yöneticileri personel seçimi uygulamalarında bu araçtan daha fazla yararlanmaya başlamışlardır. Günümüzde büyük ölçekli firmaların hemen tamamı tarama amaçlı, ön seçim amaçlı veya yetenek belirleme amaçlı olarak İnternet test ve ölçeklerini kullanırlar. Özgeçmiş bilgilerinin toplanması, saygınlık testi veya değerlendirmesinin İnternet aracılığıyla yapılması yaygın görülen uygulanmalardır. İnternet ortamında uygulanan BYB testlerinden elde edilen yüksek puanlar yüksek iş başarısıyla[a]

[a] İnternet testleri eğer okul kazanma / giriş testleriyle veya program kazanma / giriş testleriyle ilgili ise geçerlilik kriteri okullarda veya programda alınan derslerin puan ortalamasıdır. Üniversite giriş testlerinin geçerlilik kriteri öğrencilerin sınıf not ortalamaları veya mezuniyet puan ortalamalarıdır.

ve düşük puanlar da düşük iş başarısıyla ilişkili olmadığı sürece bu testlerin geçerli olduğunu söyleyemeyiz. Bu tür testlerde üç konu önemlidir: geçerlilik, güvenilirlik ve işle ilgililik. Güvenilirlik işe alınan kişilere aynı test tekrar uygulandığında benzer puanların alınmasıdır. Test puanlarında istikrarlılık varsa İnternet testi güvenilirdir. İşle ilgililik ise test maddelerinin iş analizi sonuçlarına göre belirlenmesi ve BYB açısından iş ögelerini temsil etmesidir.

Personel seçimi amacına yönelik olarak uygulanan İnternet testleri başvuran adayların öncelikle bir ön testten geçirilmesini gerektirir. Bu ön testte adaylara kendilerini tanımaya imkan sağlayacak demografik içerikli sorular sorulur ve ayrıca bilgisayar ve İnternet deneyimleri anlaşılmaya çalışılır. Bundan sonra çalışacağı işe uygun olarak ilgili kişinin güvenilirliğini, sosyal ilişkilerini, öz güvenini saptayacak ifadelere yer verilir. Testin son bölümü bilgi, yetenek ve beceri konularıyla ilgilidir. İşletme iş için başvuran adayların bilgilerini belirli bir derecelendirme şablonu çerçevesinde değerlendirerek anında sorgulama yapıp sıralamaya sokar. Uygun olmayan adaylara anında otomatik olarak e-posta aracılığıyla olumsuz yanıt verilirken, uygun adaylar ise mülâkata davet edilir.[17] Bu konuda özel olarak geliştirilmiş yazılımların binlerce iş müracaatını 15-20 saniye içinde değerlendirip eleme yaptığını ve mülâkata çağrılacak kişileri belirlediğini İnternet kaynaklarındaki bilgilerden öğrenmiş bulunuyoruz.

- **İnternet temelli kamuoyu araştırmaları.** İnternet ortamında yapılan bir diğer ölçüm işlemi, İnternet kamuoyu araştırmalarıdır (İKA). İnternet kamuoyu araştırmaları kişisel etkileşime girmeden çok sayıda kişiden hızlı bir şekilde veri toplanmasına imkan sağlar. İnternet temelli kamuoyu araştırmaları politik amaçlı, sosyal tercihleri belirlemeye yönelik olarak, sosyal politikaları araştırmaya yönelik olarak, müşteri tatminini belirlemeye yönelik olarak, tüketim kalıplarını ve tercihlerini belirlemeye yönelik olarak yapılabilir. İnternet temelli kamuoyunun araştırmalarında kullanılan ölçüm aracının niteliği önemlidir. Bu araç bir soru listesi, ölçek, indeks veya test olabilir. Görüş ve kanaatleri belirlemeye yönelik olarak oluşturulan soru listelerinin güvenilirlikleri düşüktür. Ölçek, indeks ve testlerin güvenilirlikleri ise bilgisayar temelli testler için geçerli olan kısıtlayıcıların etkisi altındadır.

İnternet temelli zeka ve yetenek testleri. İnsanların zeka testlerine gereğinden fazla ilgi göstermesi sonucunda İnternet ortamındaki zeka testi yapan sitelerin sayısında bir artış ortaya çıkmıştır. Söz konusu Ağ kümelerindeki zeka testlerinin önemli bir bölümü özensiz bir şekilde hazırlanmış

olup; bu tür testler uygulayıcı talimatları, soru sayısı, düzenleniş ve sunuş biçimi açısından yetersiz olduğundan verilen sonuçların güvenilirliği de kuşkuludur. Kersting (2004) zeka testlerini İnternet ortamına alan uygulayıcıların veya bilgisayar yazılımcılarının test sonuçlarını test alıcılarına sunma konusunda yeterli donanıma sahip olmamaları nedeniyle duyarsız kaldıklarını ve bilimsel etik kurallarından uzak bir şekilde hareket ettiklerini bildirmiştir. Psikolojik testlerin yorumlanmasında test alan kişinin durumunun göz önünde bulundurulması gerekirken çevrimiçi testlerde bunu sağlamak imkansızdır. Kersting (2004) İnternet ortamında uygulanan zeka testlerinde aşağıdaki güçlüklerle karşılaşılabileceğini belirtmiştir:[18]

1. Test puanlarının / sonuçlarının güvenliğinin sağlanması.
2. Test malzemelerinin yetkili olmayan kullanımının önlenmesi.
3. Uluslar arası alanda telif haklarının korunması.
4. Test sonuçlarının yeterli bir şekilde açıklanması.
5. Test alıcılarının özel ihtiyaçlarının karşılanması.

İnternet ortamındaki zeka testlerinin bir kısım kuruluşlar veya meslek ahlakı zorunluluğu bulunmayan kişiler tarafından ticarî amaçlarla kullanılması bu testlerin duyarlı, dikkatli ve özenli bir şekilde uygulanmasını önlemektedir. O nedenle İnternet temelli zeka testlerini hazırlayan kurum ve kuruluşların güvenilirliği testin kendisinden daha önemli hale gelmektedir.

İnternet temelli kişilik envanterleri. İnternet temelli kişilik envanterleri, test alan bireylerin kendi kendilerini tanımalarına yarayan araçlardır. Bu tür envanterler danışmanlık, rehberlik ve tanıma amaçlı olarak kullanırlar. Ancak bu testlerin/envanterlerin bir bölümünde geçerlilik ve güvenilirlik analizlerinin yapıldığına ilişkin herhangi bir bilgi bulunmadığından elde edilen sonuçlara tam olarak güvenilemez.

İnternet Temelli Test, Ölçek ve Envanterlerin Güvenilirliği

İnternet ortamında uygulanan test, ölçek ve envanterlerde klasik ölçüm kuramı temel alınmışsa iç tutarlılığı belirlemeye yönelik olarak Cronbach alfa (dereceli ölçeklerde), KR-20 (ikili verilerde), madde-toplam puan korelasyonu, paralel formlar ve dış kriterle karşılaştırarak korelasyon katsayılarını elde etme yöntemleri uygulanır. İnternet testlerinde aynı kişilere test-yeniden test yöntemini uygulamak çok daha zordur. Anonim nitelikteki

İnternet testi uygulamalarında test-yeniden test uygulamaları aynı ana kütleden seçilen farklı gruplarda yapılabilir. Personel seçimi uygulamalarında ve takipli hastalarda ise aynı kişilere test-yeniden test uygulaması yapmak nispeten daha kolaydır. Araştırmalar kağıt-kalem testleri için geliştirilmiş bulunan geçerlilik ve güvenilirlik standartlarına uyulduğunda denetim altına alınmış İnternet testlerinin de aynı ölçüde güvenilir olabileceğini ortaya koymuştur.[19]

KENDİNE REFERANSLI (İPSATİF) TESTLERİN GÜVENİLİRLİĞİ

Kendine referanslı test (KNRT) kavramı ilk kez Cattell (1944) tarafından ileri sürülmüştür. Cattell Latince "kendi" anlamına gelen *ipse* kavramından hareket ederek bir bataryadaki test puanları arasındaki karşılaştırmalara dayanan uygulamalar için "ipsatif test" (kendine referanslı test) teriminin kullanılmasını önermiştir.[20] Kendine referanslı ölçümlerde, bir kişinin, öğrencinin veya personelin özellikleri iki şekilde değerlendirilir:

1. Kendisine uygulanan test bataryasının/ölçek ve alt ölçeklerin puanları arasında karşılaştırmalar yapılarak hangisinden yüksek, hangisinden düşük puan aldığına bakılır.
2. Kendisine uygulanan test bataryasının puanları / ölçeklerin puanları / alt ölçeklerin puanları daha önceki uygulama puanlarıyla karşılaştırılarak bir gelişme veya gerileme olup olmadığı araştırılır.

Birinci türünde, tek bir test uygulamasına dayalı olarak kişilerinin özellikleri veya yetenekleri değerlendirilmeye çalışılır. İkinci türünde ise önce ölçüm yapılır, arkasından kişi belli bir eğitim, yetiştirme, geliştirme veya tedavi programına alınır. Belirlenen programın sonucunda kişiye aynı testler tekrar uygulanarak ne ölçüde bir gelişme olduğuna bakılır. Bireyin kıyas değerleri yine kendi puanlarıdır. Kişi bütün testlerde yüksek, bütün testlerde düşük veya bazı testlerde yüksek ve bazılarında ise düşük puan almış olabilir. Her bir testten alınan toplam puan değişme özelliği bulunmayan "sabit" bir değerdir. Testlerin sonucunda elde edilen toplam puanlar veya ortalama değerler ham puanlardır ve bu puanlar *sabit* olarak isimlendirilir. Sabit değerler kendi başlarına bir anlama sahip değildir, sadece karşılaştırıldıkları grubun değerine göre anlam kazanır.

İpsatif Ölçekler, İpsatif Puanlar ve İpsatif Yorumlar

Bu bölümde olguyu kısa ve özlü bir biçimde ifade etmesi nedeniyle "kendine referanslı test" ifadesi yerine sık aralıklarla ipsatif (ipsative) terimi kullanılmıştır. *İpsatif* kavramı kullanım amacına ve yerine göre birkaç değişik anlama gelebilir ve bunlar aşağıdaki gibidir:

1. İpsatif ölçek.
2. İpsatif puan.
3. İpsatif yorum.

Okuyucunun kullanılan terimler arasındaki bağlantıyı ve geçişimi iyi görmesi için her birini ayrıca ele almakta yarar vardır.

İpsatif ölçek. İkili cevaplama şıklarıyla oluşturulan veya katılımcılara tercih sıralaması yaptırılarak cevaplandırılan ölçek demektir. İpsatif ölçekler daha çok kişilik ve ilgi envanterlerinde kullanılır. Tercih sıralaması yöntemi seçilmişse şık sayısı en çok dört kategoriden oluşur. Daha fazla kategori arasında tercih sıralaması yaptırmak değerlendirici yargılarının güvenilirliğini düşürür. Öte yandan kategori sayısının az tutulması ise karmaşık psikolojik olgunun aşırı bir şekilde basitleştirilmesine neden olur.

İpsatif puan. İpsatif puan elde edilen verilerin niteliğine bağlı olarak birkaç şekilde olabilir.

1. İpsatif ölçeklerden elde edilen puan.
2. Standardize edilmiş normatif ölçeklerden elde edilen fakat karşılaştırma yapılmadan kişi-içi değerlendirme amacıyla kullanılan puanlar.

Genelde, ipsatif puanlar norm grubuna dayanmayan değerler anlamındadır. İpsatif veri, kendine referanslı test puanlarıdır. İpsatif puanlar ya ifadelere getirilen evet/hayır, doğru/yanlış şıklarından birinin zorunlu olarak seçilmesiyle veya cevap şıklarının tercih sıralamasına sokulmasıyla elde edilir. İkincisinde, cevap şıkları "en fazla", "ikinci sırada" ve "en az" şıklarının puanlandırılması suretiyle yapılır. İpsatif ölçeklerde cevap şıkları derecelendirilmek yerine tercih sırasına sokulmuştur. İpsatif ölçeklerde soruların veya ifadelerin cevap şıkları belirlenirken çoğunlukla "zorunlu

tercih" uygulamasına gidilir. İpsatif ölçeklerde yanıtlar tercihin büyüklük derecesini göstermediğinden kişiler arası karşılaştırma yapmaya uygun değildir. İpsatif olmayan ölçeklerde ise tercih şıkları Likert ölçeklerinde olduğu gibi üç, beş veya yedi işaretleme noktası bulunan *dereceleme ölçeği* şeklinde belirlenmiştir. Normatif ölçüm araçlarında dereceleme ölçekleri kullanılır. Literatürde ipsatif ve dereceleme özelliğini göstermek üzere "en fazla" şıkkının kendi içinde 5 ve 4 olarak, "ikinci sırada" şıkkının da 3 ve 2 rakamıyla kodlanmış biçimine "yarı-ipsatif ölçek" adı verilmiştir. Öte yandan bazı test ve envanterlerin ise hem ipsatif hem de normatif sürümleri vardır.

İpsatif yorumlar. Puanlar ister ipsatif ölçeklerden, isterse normatif ölçeklerden elde edilmiş olsun başka kişilerle karşılaştırma yapmadan kişinin kendi verilerine dayalı olarak değerlendirildiğinde bu uygulama *ipsatif yorumlama* tekniği olarak isimlendirilir.

İpsatif Puanlar ve Başarı Tahmini

İpsatif puanların gelecekteki genel başarı veya kişinin gelecekteki iş başarısı hakkında sınırlı ölçüde bilgi verme durumu söz konusudur. Bu nedenle ipsatif veriler iş başarısını tahmin etmek veya personel seçim kararını vermek için kullanılmaz. Çünkü, kişinin bir testten yüksek puan almış olması o özellikte / yetenekte gerçekten güçlü olduğu anlamına gelmez. Bir özellikten elde edilen yüksek puan, kişinin diğer testlere / özelliklere göre o testte daha iyi olduğunu belirler. İpsatif puanlar, bataryadaki / ölçekteki testlere veya alt ölçeklere göre kişinin özelliklerini sıralama özelliğine sahiptir. Bir anlamda kişinin kendi yeteneklerini veya özelliklerini sıralamaya sokmasıdır.

İpsatif Ölçeklerin Kullanım Amacı

İpsatif ölçekler kişiyi tanımaya yönelik araçlardır. Bilim adamları ipsatif ölçekleri bir tartışma zemini yaratması açısından yararlı görmüşler ve bu testlerin daha çok danışmanlık ve kariyer rehberliği alanlarında kullanılmasını önermişlerdir. Bir testin normatif verileri yoksa böyle bir durumda ipsatif değerlendirme yönteminden yararlanılır. İpsatif değerlendirmelerin normatif değerlendirmelere göre daha risksiz olduğu kabul edilmiştir. İpsatif ölçümler sadece kişiye özgü olarak değerlendirilebilir.

Bilim adamlarına göre, normatif verilerin tersine *ipsatif* veriler bireyler arasında karşılaştırma yapmak için kullanılamaz. İpsatif puanlar sadece kişilerin kendi içlerinde her bir testten aldıkları puanlar temel alınarak karşılaştırılabilir. Araştırmacı eğer kişiler arasında karşılaştırma yapmak isti-

yorsa sadece sıralamanın nasıl farklılık gösterdiğini belirlemek için böyle bir yola başvurabilir.

İpsatif Ölçekler ve Personel Seçimi

İpsatif yöntemler kullanılarak hazırlanan kişilik testlerinin tek başına personel seçim kararının verilmesinde kullanılması doğru değildir. Çünkü bu testlerde geçerlilik ve güvenilirlik için belirli normlar temel alınmamış ve iş başarısının tahmin değeri ortaya konmamıştır.

İpsatif Puanların Değerlendirilmesi

Literatürde, ipsatif ölçek puanlarını normatif yüzdelik puanlarına dönüştürmeye yönelik çalışmalar yapılmıştır; ancak zorunlu tercih uygulamasına dayanan kişilik testleri, ilgi envanterleri ve liderlik ölçeklerinin puanları yine de ipsatif nitelikte kalmaya devam etmiştir.[21] Closs (1996) ipsatif değerleri normatif değerlere dönüştürmenin uygun olmadığını ve sonuçların anlamsız puanlar olarak ortaya çıkacağını belirtmiştir (aktaran, Barın, 1995)[22] Çok dereceli ölçeklerde ipsatif değerler eğer norm değerlerine dönüştürülmüşse bu puanlar kişi-içi[a] karşılaştırmaları yapma açısından imkansız hale gelir ve tüm karşılaştırmalar grup puanlarının temel alındığı kişiler arası karşılaştırmalara dayanır.[23] Literatürde, ipsatif puanların normatif değerler olmamasına karşılık McCall *T* puanlarına dönüştürüldüğü görülür. Aslında *T* puanları normatif bir grupta puanların ortalamadan ne ölçüde saptığını göstermesine karşılık sanki ipsatif puanlarmış gibi yorumlanır. Böylece bir kişilik testinde veya yetenek testi bataryasında bir kişiye ait yüksek *T* puanları düşük *T* puanlarıyla karşılaştırılır.[24]

İpsatif Ölçek Türleri

İpsatif ölçümler daha çok kişilik envanterleri, ilgi envanterleri, liderlik envanterleri biçiminde veya "örgütsel kültür" "öğrenme biçimleri" "kişisel değerler", "örgütsel değerler" gibi belli bir kavramsal yapıya ait alt ölçekleri içeren ve şıkları zorunlu tercihe veya sıralamaya dayanan öz değerlendirme ölçekleri şeklindedir. Kuder ilgi envanteri, Jackson mesleki ilgi envanteri, Myers Briggs kişilik envanteri, Wonderlic Kapsamlı Kişilik Profili,[b] WISC zeka testi tipik ipsatif ölçek türleridir. Bilim adamları, kişisel

---

[a] Drummond (1996), ipsatif karşılaştırmaları kişi-içi değerlendirme olarak nitelendirmiştir (aktaran R. P. Brady, "Administrator's Guide [Yönetici Rehberi]," <http: //www.jist.com/ DSI_guide.pdf.> (18.03.2004).

[b] Wonderlic kapsamlı kişilik testi 88 sorulu bir envanterdir. Bu testte bireyin güçlü ve zayıf olduğu yönler duygusal yoğunluk, sezme, kararlılık, motivasyon, iyi izlenim bırakma, güven verme gibi boyutlar altında ölçülür.

değerlerin ölçümünde hiyerarşik bir yapılanmaya sahip olması nedeniyle ipsatif tekniklerin kullanılmasını önermişlerdir. Kişiler kendi değer yapılarını ancak karşılaştırmalar yaparak sıraya sokabileceklerinden değer yapılarının ölçümünde ipsatif tekniklerden yararlanmak daha doğrudur.

İpsatif Veriler ve Güvenilirlik

İpsatif veriler üzerinde genelde korelasyon analizleri ve faktör analizi yöntemleri uygulanamaz. İpsatif veriler, eğer seçeneklerin tercih sıralamasına sokulmasıyla (rank order) elde edilmişse söz konusu rakamlara dayalı olarak korelasyon analizi ve faktör analizi yöntemini uygulamak imkansızdır. Veriler *Evet-Hayır* gibi zorunlu tercihe dayanan ikili bir yapıya sahipse keşfedici ve teyit edici faktör analizi ancak her bir maddeye gelen yanıtlar normal dağılım özelliği gösteriyorsa uygulanabilir. Maddelere gelen yanıtlar normal dağılım özelliği göstermiyorsa faktör analizi yapmadan önce bu verileri "dönüştürme" yöntemleriyle *normalleştirmek* veya diğer özel prosedürleri uygulamak gerekir. Bununla birlikte normalleştirme de son çözüm değildir, çünkü bu yaklaşımların da kendine özgü başka sorunları çıkarması söz konusu olabilir.[25] Uygulamada ipsatif ölçek verileri üzerinde faktör analizi yapılamadığından öne sürülen boyutların güvenilirlik ve geçerliliği kuşkuludur. İngiltere Psikoloji Derneği üyelerine personel seçimi uygulamalarında ipsatif ölçekleri kullanmamalarını önermiştir.[26] Korelasyon analizi ve faktör analizi yöntemleri uygulanamayınca bu verilerin güvenilirlik ve geçerliliği tartışmalıdır. İpsatif veriler sadece ilgili kişi için geçerlidir.

Literatürde bu verilerin güvenilirlik ve geçerliliği ile ilgili bulgular çelişkilidir. Bazı araştırmacılar ipsatif ölçeklerin iç tutarlılığının normatif ölçeklere göre daha yüksek ve abartılı olduğunu bildirmişlerdir.[27] Diğer araştırmacılara göre ise, normatif ve ipsatif ölçümlerin her ikisi de dış kriter değişkeniyle yapılan korelasyon analizlerinde benzer güvenilirlik katsayıları vermiştir.[28] Bu çerçevede Bartram (1996) ise, bataryadaki ölçek/test sayısı az ve ölçekler/testler arasındaki korelasyon katsayıları yüksekse ipsatif güvenilirlik değerlerinin normatif güvenilirlik değerlerinden daha düşük olduğunu bildirmiştir (aktaran, Cambell vd., 2003).[29] Baron (2004) ipsatif ölçümlerde bataryadaki test veya ölçek sayısının 30 veya daha yüksek olması ve testler arasındaki korelasyon katsayısı ortalamalarının düşük olması halinde sağlıklı bir güvenilirlik analizi yapılabileceğini belirtir.[30] İpsatif verilerin güvenilirliği ile ilgili araştırma bulguları kullanıcılara tam bir netlik sağlamamaktadır. Literatürün bu bölümüne ilişkin araştırma bulguları yetersizdir. Cornwell ve Manfredo'ya göre ipsatif öl-

çekler psikometrik değerlendirmeler için uygun değildir ve ipsatif puanlarla herhangi bir teori de test edilmez (aktaran, Bower, 2004).[31]

İpsatif Ölçeklerin Avantaj ve Dezavantajları

Kendine referanslı test uygulamalarının en önemli avantajı "sistematik yanıtlama eğilimini" azaltmasıdır. Bu tür testleri uygulayan kişiler şıklar arasında tercih veya sıralama yapmak durumunda olduklarından *içeriğe aşina* olma ve *sosyal beğenilirlik* gibi faktörlere göre yanıt verme eğiliminden nispeten uzak hareket ederler. İpsatif ölçeklerdeki zorunlu tercih formatının, normatif ölçeklerin doğasında bulunan *yanıt yanlılığı* olgusunu bir ölçüde giderdiği bildirilmiştir.[32] Katılımcılar, sosyal beğeni sağlamak için hep olumlu dereceleri işaretleme gibi bir eğilim içinde olmazlar.

KİŞİLİK ENVANTERLERİNİN GÜVENİLİRLİĞİ

Kişilik envanterleri bireylerin değişik nitelikler açısından özelliklerini belirlemeye yönelik olarak uygulanır. Kişilik envanterlerinin geliştirilme nedeni büyük ölçüde literatürdeki "özellik kuramından" kaynaklanır. Raymond B. Cattell, Allport, Eysenck ve Murray gibi bilim adamları kişiliğin "özellik" ve "faktörlerden" oluştuğunu öne sürmüşlerdir. Özellikler kuramına göre bireyler kısa zaman diliminde değişmeyen oldukça kalıcı kişilik özelliklerine sahiptirler. Kişilik özellikleri kalıcı ise, bunları ölçüp bir bireyin kişiliği hakkında yorum yapma imkanına sahip olabiliriz. Bununla birlikte kişilik özellikleri kültürler arasında farklılık gösterir. Belirli bir kültüre özgü olarak saptanan kişilik özellikleri başka kültürler için geçerli olmayabilir. Literatürde yer alan *çok faktörlü kişilik envanterlerinin* bir çoğunun bu açıdan dikkatli bir şekilde uyarlanması ve incelenmesi gerekir. Kişilik envanterlerinin kültürler arasındaki uygulanabilirliği güvenilirlik-geçerlilik analizleri ve standardizasyon çalışmalarıyla sağlanır.

Kişilik Envanterlerinin Niteliği

Kişilik testlerinde/envanterlerinde çok sayıda madde bulunur. Bir kişilik envanterlerindeki madde sayısı 200 ilâ 600 arasında değişir. Cevaplama süresi de yarım saten 1,5 saate kadar uzanabilir. Kişilik envanterleri zaman sınırlaması getirmeden uygulanır. Kişilik envanterleri bir ülkede çok sayıda kişiye uygulanarak içerdiği faktörler açısından o ülke insanlarına özgü ulusal normlar geliştirilir. Kişiliğin belirli yönünü ölçmeyi amaçlayan envanterleri ise 20-30 kadar daha az sayıda madden oluşur ve bu envanterler genellikle 10-15 dakikada tamamlanır. Kişilik envanterlerinin geliştirilmesinde "mantıksal içerik" veya "kuramsal içerik" stratejisinden hareket edi-

lir. Mantıksal içerik yaklaşımında araştırmacı gözlemlerine ve mantığına dayalı olarak ölçüm boyutlarını ve her bir boyutun altında toplanan vasıfları (nitelikleri) belirler. Kuramsal içerik stratejisinde ise daha önceden yapılmış araştırma bulgularından ve geliştirilen kuramsal görüşlerden yararlanılır.

Literatürde kişilik kuramları beş başlık altında toplanmıştır: psikoanalitik kişilik kuramları, özellik kuramları, hümanistik / fenomonolojik kişilik kuramları, davranışsal/sosyal öğrenme kuramları, bilişsel öğrenme kuramları. Araştırmacı hangi kuramı temel almışsa geliştireceği kişilik envanterini de bu kuramın çerçevelediği bilgilere dayalı olarak geliştirir.

Kişilik Envanterlerinin Türleri

Kişilik envanterleri klinik teşhis amaçlı, bireyi genel olarak tanıma veya iş barısı hakkında tahmin yürütme amaçlı olarak kullanılabilir. Kullanım amacı aynı zamanda kişilik envanterinin oluşturulma biçimini belirler. Klinik amaçlı kişilik envanterleri; kişilik bozukluklarını, depresyonu ve psikotik kişilik özelliklerini ortaya çıkarmaya yöneliktir. İş çevrelerinde daha çok başarı üzerinde etkisi olduğuna inanılan kişilik özellikleri araştırılır. Bu envanterlerin bir bölümü kişiliği bir bütün olarak değerlendirirken diğerleri kişiliğin sadece belirli bir boyutunu ölçmeye yönelik oluşturulmuştur. Örneğin zihinsel sertlik, kararlılık, savunma biçimleri, öz saygınlık, iletişim, motivasyon testleri bu gruba girer. Literatürde çok sık kullanılan genel amaçlı kişilik envanterleri aşağıdaki gibidir:

1. Klinik amaçlı kişilik envanterleri.
   a. The Minnesota Multiphasic Personality Inventory (MMPI).
   b. Rorschach Inkblot Test.
   c. Personality Assessment Inventory.

2. Normal kişilik envanterleri
   a. California Psychological Inventory (CPI).
   b. Edwards Personal Preference Schedule.
   c. Myer's-Brigg's Type Indicator.
   d. 16 PF kişilik testi.
   e. 15FQ +
   f. The Big Five Personality.
   g. Five Factor Model of personality.
   h. Self-Directed Search.
   i. Eysenck Personality Inventory.

3. Kişiliğin belirli bir boyutunu ölçen envanterler.
   a. Rotter Locus of Control.
   b. Rosenberg Self-Esteem Scale.
   c. Açık veya kapalı biçimde düzenlenen saygınlık testleri.
   d. Type-A Type-B Personality Test.

Kişilik envanterleri sonuçların standardize edilmiş olmasına göre *objektif kişilik envanterleri* ve *izdüşümlü kişilik envanterleri* olmak üzere ayrıca iki grupta incelenir. Objektif kişilik testleri/envanterleri MMPI, MMPI-II NEO Kişilik Envanteri ve MBTI Myers-Briggs Tip Göstergeleri gibi geçerlilik ve güvenilirlik analizleri yapılmış ve belirli gruplar için norm değerleri çıkarılmış olan testlerdir. İzdüşümlü kişilik testleri/ envanterleri ise Rorschach [ro'şa] mürekkep lekesi testi, TAT Konu Algılama Testi, Kinetic Aile Üyeleri Çizim Testi, Rotter Cümle Tamamlama Testi gibi geçerlilik ve güvenilirlikleri tartışmalı olan ölçüm araçlarıdır.

Kişilik Envanterleri ve Personel Seçimi

Son yıllarda personel seçimi sürecinde kişilik envanterlerinden de yararlanılmaya başlanmıştır. Kişilik envanterleri seçim kararının verilmesinde temel öge olarak değil, yardımcı öge olarak kullanılır. İş adamları ve yöneticiler kişilik envanterlerinden daha çok yerleştirme ve bireyi tanımak amacıyla yararlanırlar. Bunun yanında halen çalışan kişilere uygulandığında eğitime alınacak kişileri belirleme, çalışanlardan bazılarını müşteri hizmetlerine yöneltme, kişilere kariyer danışmanlığı yapma ve kişilerin kendilerini tanıyarak etkinliklerini artırma amacı güdülür.

Kişilik Envanterlerinin Güvenilirliği

Kişilik envanterlerinin güvenilirliği benzer koşullarda benzer sonuçların alınmasıyla sağlanır. Kişilik envanterlerinin hazırlanış biçimi güvenilirliğini etkiler. İpsatif ölçekler şeklinde hazırlanan kişilik envanterlerinin güvenilirliğini saptamak zordur. İpsatif ölçekler şeklinde hazırlanan kişilik envanterlerinde sadece test-yeniden test güvenilirliği yapılabilir. Derecelemeölçekleri kullanılarak hazırlanan kişilik ölçeklerinin güvenilirliği her bir faktör veya alt boyut düzeyinde Cronbach alfa, yarıya bölme yöntemi, faktör analizi, test-yeniden test, paralel formlar yöntemleri kullanılarak belirlenir. Kişilik envanterlerinde test-yeniden test güvenilirlik katsayısının en az ,70 olması gerekir. Bilim adamları kişilik envanterlerinde test-yeniden test güvenilirlik analizi yapılabilmesi için geçmesi gereken süreyi altı aya kadar uzatmışlar, daha kısa sürede güvenilirlik rakamlarının yapay bir şekilde yüksek çıkacağını iddia etmişlerdir. Araştırmacı güvenilirlik

analizi için test-yeniden test yönteminden yararlanmayı düşünüyorsa ilk uygulamada kişilere örneğin bir ay sonra ikinci bir test daha uygulayacağını belirtir, fakat bunu yaparken aynı testi uygulayacağını söylemez. Kişilik envanterleri geliştirilirken güvenilirlik ve geçerlilik rakamlarının sağlıklı olarak elde edilmesi için kişilere ad ve soyadları yazdırılmaz. Çünkü kişiler kendi üzerlerinde bir değerlendirme yapılacağını düşüneceklerinden sosyal beğenirlik faktörün göre cevap verme eğilimi içine girebilirler.

Kişilik Envanterlerinde Güvenilirliği Etkileyen Faktörler

Bireylerin kendi kendilerine doldurdukları kişilik envanterlerinin en önemli sakıncası bazen belirli faktörlerin etkisiyle alınan sonuçların gerçeği tam olarak temsil etmemesidir. Kişilik envanteri sonuçları başlıca üç faktörden etkilenir: belli bir şekilde yanıtlama eğilimi, sosyal olarak arzu edilen cevapları verme eğilimi ve yanıltma.

Yanıltma. Anket alan kişilerin kasıtlı olarak araştırmayı yapan kişileri yanıltacak cevaplar vermeleridir. Yapılan bir çok araştırmada cevaplayıcıların işaretlemelerini bilinçli bir şekilde gerçek durumlarını gizleyerek yaptıklarını ortaya koymuştur.[33] Bunun için araştırmacılar kişilik envanterlerine verilen cevapların güvenilirliğini artırmak için test içinde kişinin yalan söyleyip söylemediğini belirlemeye yönelik olarak test soruları koyarlar. Test sorularına verilen tutarsız yanıtlar anket formunun iptal edilmesine neden olur. Test sorularının sorulmadığı durumlarda kişinin verdiği yanıtların tutarlı ve geçerli olduğunu saptamak çok zordur. Araştırmacılar kişilik envanterlerinde bireyin yanıltıcı cevap vermelerinin önüne geçmek için üç farklı yaklaşımdan yararlanırlar: (a) test maddelerini yanıltmaya imkan vermeyecek şekilde yazmak, (b) testin içine tuhaf cevapları yakamaya imkan verecek geçerlilik ölçekleri almak, (c) cevaplayıcının yaptığı işaretlemeleri kontrol ederek yanıltma derecesini belirlemeye çalışmak.[34] Araştırmacılar iki tür soruda kişilerin yanıltıcı cevap verme konusunda güçlük çekeceklerini belirtmişlerdir. Bunlardan birincisi, doğruluğu kanıtlanabilir cevaplar ve ikincisi ise neyi ölçtüğü tam olarak belli olmayan belirsiz içerikli maddelerdir.

Sosyal beğenirlik faktörüne göre cevap verme. Kişilik envanterlerinin ve ilgi envanterlerinin değerlendirilmesinde dikkat edilmesi gereken bir diğer husus test uygulayan kişinin kendi zayıf yönlerini gizleme eğilimi içinde olması ve güçlü yönlerini ise abartmasıdır. İnsanlar daha çok sosyal beğenirlik faktörlerine göre yanıt verme eğilimi içinde olurlar. Bunun yanında az sayıda kişide ise tevazu eğilimi yüksek olabilir. Bu kişiler kendi-

lerini olduğundan daha düşük değerlikli değerlendirirler. Düşük değerlikli puanlar verme ayrıca yetenek ve becerilerinden tatmin olmama, çok daha yüksek beklentilere sahip olmayla da ilgili olabilir. Bu kişiler oldukça yüksek yetenek ve becerilere sahip olmalarına karşılık bunları yetersiz görerek kendilerini zayıf değerlendirirler. Bu davranış tevazu ile ilgili değil, doyumsuzlukla ilgilidir.

Belli bir şekilde yanıtlama eğilimi. Kişilerin kendileriyle ilgili olarak gerçek düşüncelerini değil, belli bir eğilimi yansıtacak şekilde cevap vermeleridir. Kişilerin olumsuz tutuma sahip olmaları veya çok iyimser bir tutuma sahip olmaları "cevaplama eğilimi" yaratır. Araştırmacı cevaplama eğilimini ortadan kaldırmak veya önlemek için test maddelerini pozitif ve negatif anlamlı olarak dengeli bir şekilde oluşturabilir. Normal gruplarda yanıtlama yanlılığı doğuran maddeler gözden geçirilerek bunların elenmesi yoluna başvurulabilir.

Yalan söyleme, saklama, abartma, tevazu ve doyumsuzluk kişilik puanlarının güvenilirliğini etkiler. Test geliştirici kişilik testleri uygulama yönergesinde test alan kişileri bu konuda uyarmalı ve kendilerini saklamadan ve abartmadan, büyük ölçüde tatminsizlik duyguları yaşamadan değerlendirmelerini istemelidir.

GÜÇ VE HIZ TESTLERİNDE GÜVENİLİRLİK

Testler, yanıtlama süresi açısından üç grupta değerlendirilir: Hız testleri, güç testleri ve melez testler.

Güç Testleri

Kişilerin zihinsel olarak güçlü olma durumunu ölçen, cevaplandırılması bir ölçüde zor olan testlerdir. Bu grupta testi alan kişiler zaman baskısı altında değillerdir. Bununla birlikte bazı test soruları o kadar zordur ki kişiler bu test maddelerine hiç cevap veremeyebilirler. Ancak cevaplayıcılara test sorularını çözmeleri için yeteri kadar süre verilmiştir. Bu süre tüm katılımcıların tüm test sorularını çözebilecekleri kadar uzundur. Ancak uygulamada mutlak anlamda güç testi yoktur. Tüm güç testleri bir ölçüde hızlandırılmıştır. Bunlara *hızlandırılmış güç testleri* adı verilir. Normal güç testlerinde katılımcıların %90-%95'inin testi "tamamladıkları süre" optimum yanıtlama süresi olarak belirlenir. Testi *tamamlama* ile *doğru olarak tamamlama* farklı olgulardır. Güç testlerinin optimum cevaplama süresi zaman içinde yapılacak diğer test uygulamalarındaki zaman kullanım değerleri de göz önünde

bulundurularak revize edilir ve yeniden düzenlenir. Güç testlerinde klasik kuramdaki güvenilirlik analizlerini yapma konusunda herhangi bir sorun yoktur. İç tutarlılık, test-yeniden test ve paralel formlar yönteminden herhangi biri uygulanabilir.

Bir ölçüm uygulamasında; önce güç testleri, daha sonra hız testleri ve en son psikomotor testleri uygulamaya alınır. Güç testlerinin uygulandığı vak'alarda kişiler testi eğer tam zamanında bitirememişlerse kendilerine ekstra bir iki dakika daha zaman verilebilir. Ekstra zaman verilmesi test sonuçlarının güvenilirliğini olumsuz yönde etkilemez. Tipik güç testleri örnekleri aşağıdaki gibidir:

1. Aritmetik muhakeme ve problem çözme testi.
2. Belirli kodları akılda tutmayı gerektiren bellek testi.
3. Üç boyutlu uzay ilişkileri testi.
4. Kelime hazinesi zenginliği testi.
5. Bilgiyi ölçen diğer testler.

Bu ölçüm araçlarında sorular genellikle kolaydan zora doğru dizilir. Amaç test alan kişilerin soruları ne kadar hızlı çözdüğünü görmek değil, çözüp çözemediğini belirlemektir. Aritmetik muhakemenin dışında kalan aritmetik işlem ve hesaplama yeteneği büyük ölçüde hıza bağlı olduğundan güç testi olarak değil, hız testi olarak değerlendirilmiştir.

Hız Testleri

İkinci gruptakiler *hız testleri* olarak adlandırılır. Hız testleri güç testlerine oranla daha kolay sorulardan oluşur, fakat cevaplandırma süresi kısıtlanmıştır. Cevaplandırıcıların hiç birisinin belirlenen süre içinde testi tamamlayamaması halinde *pür hız testinden* söz ederiz. Cevaplandırıcıların büyük çoğunluğunun test maddelerinin tamamını süresi içinde bitirememesi ise *kısmî hız testi* olarak adlandırılır. Uygulamada, testler pür hız veya pür güç testi şeklinde değil, kısmen hızlandırılmış bir şekilde uygulanır. Hız testlerinde bulunan 50, 80 veya 100 kadar sorunun her biri için kişilere, temel alınan hızlandırma faktörüne göre, 1 ilâ 5 saniye arasında bir süre verilir. Testteki madde sayısının çokluğu nedeniyle duruma göre kişiler testteki tüm maddeleri yanıtlayamamış olabilirler. Basit ve kolay bir şekilde cevaplandırılan hız testi türlerini aşağıdaki gibi sıralayabiliriz:

1. Farklı olanı bulmaya yönelik grafik / çizim temelli genel yetenek testleri.
2. Nesne (obje) eşleştirme testleri.
3. Basit matematiksel dört işlem testleri.
4. Dikkat testleri.
5. Farklılık ve benzerlikleri bulma testleri.
6. Yazım (imlâ) yanlışlıklarını bulma yeteneği testleri.
7. Kodlama hızı testleri (rakamları belirli hücrelere işaretleme).
8. Verilen bir kelimenin eş veya zıt anlamlısını bulma testi.
9. Verilen bir harften önce gelen harfi bulma testi.
10. Sözel akıcılık testi (verilen harflerden hareket ederek belirli kategorilerde kelime üretme).
11. Cümle tamamlama testi.
12. Sessiz veya sesli okuma testi.
13. Telaffuz testi.
14. Psikomotor testleri (kalemle kağıt üzerine işaret koyma testi, yerleştirme, döndürme, takma, çıkarma, birleştirme ve dağıtma testleri).

Bir testin hız testi olup olmadığını belirleyen faktör büyük ölçüde zorluk derecesidir. Herhangi bir test, katılımcıların büyük çoğunluğu tarafından 15-20 saniyeden daha kısa sürede cevaplandırılabilen maddelerden oluşmuşsa bu testte muhtemelen hız faktörü çalışıyor demektir. Bir test hız faktörüne göre uygulanmadığında elde edilen puanlar eğer homojen nitelikte çıkıyorsa böyle bir durumda testin hızlandırılmasına karar verilir. Bilim adamları bir testin hız testi mi yoksa güç testi mi olduğuna karar vermek için korelasyon analizlerinden yararlanmışlardır. Güç ve hız testi olarak uygulandığında bir ölçüm aracının puanları arasındaki korelasyon katsayısı yüksek çıkıyorsa testin güç testi olduğuna karar verilir.

Hız testlerinde güvenilirlik araştırmalarının tarihsel gelişimi. Hız testlerinin güvenilirliğiyle ilgili ilk çalışmalar Anastasi ve Drake (1954) tarafından yapılmıştır. Bilim adamları hızlandırılmış testlerin güvenilirliğini tahmin etmek için dört yöntem önermişlerdir (aktaran Donlon, 1980):[35]

1. Spearman-Brown, tek-çift yöntemi.
2. Gutman L_4, tek-çift yöntemi.
3. Gutman L_4, ayrı bir şekilde zamanlandırılmış yarılar yöntemi.

4. Cronbach ve Warrington'un alt sınır güvenilirliği yöntemi.

Hız testlerinin güvenilirliğiyle ilgili bir diğer çalışma Gulliksen (1950) tarafından yapılmıştır. Gulliksen kısmen hızlandırılmış testlerin güvenilirliğini hesaplamak için birkaç yöntem önermiştir. Bu yöntemlerde paralel form değerlerine ihtiyaç bulunmuyordu. Pür ve kısmî hızlandırılmış testlerin güvenilirliği için formüller önermiş ve "teşebbüs edilmemiş puanların sayısına dayalı" olarak aritmetik ortalama ve standart sapma değerlerini belirlemeyi içeren bir hesaplama yöntemi geliştirmiştir.[36] Bu hesaplama, teşebbüs edilmemiş maddelerin standart sapmasının teşebbüs edilmemiş maddeler artı yanlış bir şekilde cevap verilmiş veya boş bırakılmış maddelerin sayısına olan oranı şeklinde belirleniyordu. Oran, *pür güç* testlerinde sıfır ve *pür hız* testlerinde ise 1,0 olarak çıkıyordu.[37] Daha sonraki yıllarda Rasch, (1960) ve Van der Ven, (1969) yüksek derecede hıza sahip testlerde "gerçek puan" modellerini geliştirmek için *Poisson süreç modellerini* kullanmışlardır.

Madde-yanıt kuramının gelişmesine paralel olarak 1980'li yıllardan sonra hız testlerinin güvenilirlik değerlendirmesinde lojistik parametre değerleri dikkate alınmaya başlamıştır. Bejar (1985), Davey (1990) ve Douglass (1981) madde-yanıt kuramı çerçevesinde hız testlerinin güvenilirlik araştırmalarını yapmışlardır. Roskam, (1987) ve Van Breukelen, (1989) zaman sınırlı testlere tek parametreli Rasch lojistik modelini uygulamışlardır.[38] Bilim adamı, verileri analiz etmek için madde-yanıt kuramını temel almışsa madde parametreleri hızdan büyük ölçüde etkilenir ve parametreler doğru olmayan bir şekilde tahmin edilir. Bu amaçla test sonundaki yanıtsız bırakılan test maddelerinin parametre tahminlerini geçersiz hale getirmesini önlemeye yönelik olarak belirli stratejiler geliştirilmiştir.[39] Fakat, madde-yanıt kuramına dayalı hız testlerinin güvenilirlik analizleri önemli ölçüde karmaşık formülasyonlara dayalıdır. Bu nedenle de literatürde yaygın bir uygulamaya sahip değildir. Kitapta söz konusu modellerin teknik ayrıntılarına girilmemiş yapılan araştırmalara sadece işaret etmekle yetinilmiştir. Okuyucu hız testlerinin güvenilirliklerini belirlemeye yönelik araştırmaların tarihsel gelişimine yakından ilgi duyuyor ise bu konuda Donlon'un kapsamlı bir şekilde yaptığı çalışmaya başvurabilir.

Hız testleri ve güvenilirlik analizi yöntemleri. Hız testlerinde her tür güvenilirlik analizi uygulanmaz. Yapay biçimde yüksek güvenilirlik katsayıları verdiğinden; hız testlerinde yarıya bölme, Cronbach alfa ve KR-20 gibi iç tutarlılık analizi yöntemlerini uygulamak doğru değildir. Yanıtlama girişiminde bulunulmamış maddeler yapay bir şekilde yüksek tutarlılık / homo-

jenlik gösterdiğinden hesaplanan iç tutarlılık katsayıları da yüksek çıkar. Bu tür testlerde güvenilirlik analizleri için;

1. test-yeniden test,
2. alternatif formlar,
3. eşdeğer formlar[a] veya
4. her bir yarı için iki farklı zaman diliminin belirlendiği Rulon yarıya bölme

yöntemi uygulanır.[40] Hız testlerinde katılımcıların bir çok maddeyi işaretlememiş olmaları sonuçta bilgi kaybına yol açar. Bilim adamı bu sorunu yenmek için bir kriter geliştirmek zorundadır. Örneğin güvenilirlik analizine alabilmek için test sorularının %40'ının, %60'ının yanıtlanmış olması gibi bir ölçüyü kriter olarak kabul edebilir. Tutum ölçeklerinde cevapsız yanıtlara medyan değeri atanabilirken bilişsel hız testlerinde cevapsız maddelere sıfır değeri girilir.[41] Öte yandan hız testlerinde güç testlerinin tersine katılımcılara ekstra zaman verilmesi test sonuçlarını geçersiz hale getirir.[42]

Genel yetenek testleri ve hız faktörü. Cronbach'a göre modern genel yetenek testlerinin uygulanmasında şu yöntem takip edilir. Önce test maddeleri zorluk sırasına sokulur ve daha sonra katılımcıların neredeyse büyük çoğunluğunun testi tamamlayabildikleri bir süre belirlenir. Daha sonra test ilgili kişilere uygulanır. Cronbach İkinci Dünya Savaşı'ndan önceki yıllarda genel yetenek testlerinin büyük ölçüde hızlandırılmış bir biçimde uygulandığını ifade etmiştir. Hız testlerinin farklı ırk, etnik köken ve cinsiyet gruplarında farklı sonuçlar vermesi ve geçerliliğinin soruşturulabilir bir nitelik kazanması üzerine ABD'de Ulusal Araştırma Konseyi'nin (National Research Council) ılımlı ölçüde hızlandırılmış güç testi uygulamasına geçtiği bildirilmiştir.[43] Ree, diğer şartlar eşit olmak koşuluyla bir testteki hız faktörünün azaltılmasının, testin *g* ile yüklü olma durumunu düşüreceğini ve dolayısıyla çok sayıda işte iş başarısı tahmin etmek için testin zayıf hale geleceğini belirtmiştir (aktaran Peterson).[44]

Hız testlerinde sürenin belirlenmesi. Hız testlerinde testin tamamlanma süresini belirlemek için aritmetik ortalama ve standart değerlerinin sınırlı ölçüde yararı olması nedeniyle hesaplamalar daha çok frekans/yüzde dağı-

[a] Literatürde *paralel formlar* terimi yerine geçmek üzere bazı yazarlar *eşdeğer formlar* tabirini kullanma eğilimi içinde olmuşlardır.

lımlarına veya yüzdelik dilimlerine bakılarak yapılır. Örneğin, 50 kişiye uygulanan bir testten Tablo 12-1'deki veriler elde edilmiş olsun.

Tablo 12-1. Yüzdelik Dilimlerine Göre Hız Kriterinin Belirlenmesi

Süre	2 d.	2,5 d.	3 d.	3,5 d.	4 d.	5 d.	5,5 d.	6 d.	6,5 d.
Adayların yüzdesi / yanıtladıkları madde sayısı	%10	%20	%30	%40	%50	%60	%70	%80	%90
A testi	12	14	18	25	27	32	34	35	40[a]
B Testi	14	20	22	26	32	36	42	48	50

[a] Testteki toplam madde sayısı 40'tır.

Tablo 12-1'deki örnekte yanıtlanan madde sayısı temel alındığında; maddelerin %75'ini (30 madde) katılımcıların %60'nın cevaplandıracağı tahmin edilir. Belirlenen sürede maddelerin %75'ine cevap verilmişse bu test hız testi değil melez bir testtir. Belirlenen sürede maddelerin %75'inden azına (örneğin, 27'isine) cevap verilmişse bu test, hız testidir. Hız testlerinde testin optimum süresine değişik açılardan inceleyerek karar vermek daha doğru olur. Bunun için değişik bir düzenleme Tablo 12-2'deki gibi yapılır.

Tablo 12-2. Hız Kriterinin Belirlenmesinde Hızlılık İndeksi Varyansı

Testler	Testin %100'ünü tamamlayan adayların oranı	Testin %75'ini tamamlayan adayların oranı	Hızlılık indeksi varyansı[a]	Adayların %80'nin ulaştığı madde sayısı	Toplam madde sayısı
A testi	96,4	98,5	,03	35	40
B Testi	63,2	95,1	,01	42	50

[a] Hızlılık indeksi varyansı: Ulaşılamayan madde sayıları varyansının doğru madde sayıları varyansına olan oranı. Sonucu ,25'ten yüksek olan değerler ölçüm aracının hız testi olduğunu gösterir.

Tablo 12-2'deki sonuçlar A ve B testlerinin her ikisinin de büyük ölçüde güç testi olduğunu göstermektedir.

Yanıtsız bırakılan sorular. Hız testlerinde katılımcıların yanıtsız bıraktıkları sorular birkaç şekilde değerlendirilebilir. Test soruları eğer bilgiyi ölçmeye yönelik ise katılımcının en son ulaşabildiği soru numarası dikkate alınır. Arada boş bırakılan maddeler görmezlikten gelinir. "Katılımcı kaçıncı

soruya kadar gelmiştir?" sorusuna yanıt aranır. Ancak hız testi yanıtlarının niteliği; kısa, basit ve çok fazla düşünmeyi gerektirmeyen türden cevaplar ise böyle bir durumda aralarda boş bırakılan sorular da dikkate alınır. Bu kez, ulaşılan son test maddesine değil, boş bırakılan soru sayısına bakılır ve boş bırakılan maddelere sıfır değeri girilir. Böyle bir durumda ulaşılamayan madde sayısı, boş bırakılan madde sayısıdır.

Testin hızlılığını değerlendirme. Hızlılık, testin tamamlanma süresini gösterir. Literatürde bilgi ve yetenek testlerinin hız faktörünü belirlemek için değişik modeller önerilmiştir. Hız faktörünün belirlenmesi, tek test uygulamasına veya çoklu test uygulamasına göre değişir.

Tek test uygulamasına göre hız faktörü. Hız testi analizleri ya tek bir uygulamaya dayalı olarak veya deneysel maniplasyon sonucunda belirlenir. Amerika Birleşik Devletleri'nde "Eğitim Testleri Dairesi" (Educational Testing Service – ETS) başparmak kuralı olarak tek grup uygulamalarında hız faktörünü aşağıdaki kurallara göre belirlemiştir:[45]

1. Bir testi alan cevaplayıcıların belirlenen sürede %100'ünden azı test maddelerinin en azından %75'ine ulaşmışsa ve testi alan cevaplayıcıların %80'ininden azı belirlenen süre içinde testi bitirmişlerse bu test hız testidir. Eğer belirlenen sürede cevaplayıcıların %80'inden fazlası testi tamamlamışsa güç testi olarak isimlendirilir.
2. Hızlılık indeksi varyansı (ulaşılamayan madde sayıları varyansının doğru bir şekilde işaretlenen madde sayıları varyansına olan oranı).
3. Katılımcıların %80'inin belirlenen süre içinde kaç test maddesini yanıtlamış oldukları (ETS'nin daha sonraki yıllarda geliştirdiği üçüncü kriter).
4. Stafford'un (1971) hızlılık katsayısı.
5. Belirlenen süre içinde cevaplandırma teşebbüsünde bulunulan maddelerdeki doğru yanıt yüzdesi.

Bir testte katılımcıların eğer %80'i belirlenen süre içinde son maddeyi de yanıtlamışlarsa bu test hız testi sayılmaz. Bazı yazarlar ise test maddelerini temel almışlar ve test maddelerinin %75'in altında işaretlenmesi halinde bu testleri hız testi olarak görmüşlerdir. Buna göre bir testin hız testi sayılabilmesi için belirlenen süre içinde katılımcıların testi tamamlama oranını %80'in altındaki bir rakama çekmek veya test maddelerinin işaretlenme

oranının %75'in altında olmasına dikkat etmek gerekir. Aynı şekilde *varyans indeks değeri* 0,15'in altında ise bu testler de hız testi olarak değerlendirilmez. Hız testi olarak değerlendirilebilmesi için varyans indeks değerinin 0,25'in üzerinde olması gerekir. Varyans değeri 0,16-0,25 arasındaki testler orta derecede hızlandırılmış *melez test* grubunda değerlendirilir. Ancak bu indis değerleri geçici rakamlar olarak görülmelidir. Gerçek hızlandırılmış test değerleri; testin türüne, uygulandığı çevreye ve amacına göre belirlenir.[46]

Çoklu test uygulamasına göre hız faktörü. Çoklu test uygulamaları yönteminde ise test kişilere hızlandırılmış ve hızlandırılmamış koşullarda verilerek elde edilen sonuçlara bakılır. Sonuçlar araştırmacıya bir testin hızlandırılması veya hızlandırılmaması gerektiğini söyler. Cronbach ve Warrington (1951) bu yöntemle uyumlu olarak *tau* kavramını önermişlerdir. Tau, paralel formlar için süreli ve süresiz test uygulamaları arasındaki –zayıflığı yenme formülüyle düzeltilmiş– korelasyon katsayısının karesidir.[47] Tau, süreli ve süresiz test uygulamaları arasındaki güvenilir zaman limitinin oranı hakkında bilgi verir. Ancak bu bilgi, pratik bir sonuçla ilgili değildir, araştırmacıya sadece bilimsel bir veri sağlar.

Hız faktörünü belirlemede kullanılan diğer yöntemler. Bir testin hızlılık faktörünü belirlemek için bilim adamı başka yöntemlerden de yararlanabilir. Bunlardan biri Rindler tarafından önerilmiştir. Rindler testin hız faktörüne uygun ve hız limiti getirilmeden uygulanmasını önermiş ve bu uygulamanın sonucunda üç değerle çalışılabileceğini belirtmiştir: (a) normal zaman diliminde elde edilen doğru cevap sayısı, (b) tamamlanmış testlerdeki doğru cevap sayısı, (c) testi tamamlamak için gereken süre. Rindler daha sonra elde edilen bu puanların korelasyon ve regresyon analizine tabi tutulacağını bildirmiştir.[48]

Uygulama biçimi. Hız testleri kişilere uygulanmadan önce gerekli açıklamalar yapılarak kişiler bilgilendirilmelidir. Hiçbir açıklama yapmadan veya örnek test sayfasında söz konusu testin bir *hız testi olduğu* belirtilmeden böyle bir test uygulamak doğru değildir. Kişiler farklı olmayan koşullarda yarışmak durumunda kalabilirler. Böyle olunca da test sonuçlarının güvenilirliği kuşkulu hale gelir. Bunu önlemek için sadece test kitapçığına değil, aynı zamanda cevap formuna da testin bir hız testi olduğu yazılmalıdır.

Hız testlerinin sakıncası. Hız testlerinde hız faktörü nedeniyle ortalama %50 veya %60 oranında doğru yanıt verilebilmesi kişileri rasgele işaretleme

yapmaya sevk eder. Böyle olunca rasgele iki veya üç soruyu doğru yanıtlayan bir kişi haksız bir şekilde bir çok kişinin önüne geçebilir. Bu nedenle hız faktörünün tayininde dikkatli olmak gerekir. Ölçülen zihinsel veya kavramsal yapı hızla ilgili değilse hızlandırma işlemine başvurulmaz. Örneğin bilim adamları, öğrencilerin bilgisini ölçmeye yönelik olarak yapılacak bilgi / başarı testlerinde hız faktörünün kullanılmasını doğru bulmamışlardır. Bu tür sınavlarda zaman kısıtı yaşayan öğrencilerin rasgele işaretleme yaptıkları ve bilgilerini tam olarak ortaya koyamadıkları ifade edilmiştir.[49] Hız testlerinde testin gereğinden fazla uzun olması bazı kişilerde psikolojik reaksiyonlara neden olabilir. Bu kişiler sıkılarak testi bırakma, rasgele işaretleme yapma eğilimi gösterirler. Yapılan bazı araştırmalarda hız testlerinde yoğun pratik yapan kişilerin test puanlarındaki açığı kapattıkları bulunmuştur.[50]

Hız testlerinin bir diğer sakıncası farklı etnik ve cinsiyet gruplarında farklı sonuçlar alınabilmesidir. Amerika Birleşik Devletleri'nde yapılan bir araştırmada SAT testinin matematik bölümünün cevaplandırılmasında kız öğrencilerinin zaman yoğunluklu algoritma stratejilerini kullanamamaları nedeniyle dezavantajlı durumu düştükleri görülmüştür.[51] Bu kişilere ekstra bir zaman dilimi daha verildiğinde öğrenciler puanlarını önemli ölçüde iyileştirmişlerdir.

Melez Testler

Üçüncü gruptaki testler verilen sürenin niteliği açısından melez bir yapıya sahiptir. Bu testler hem güç, hem de belirli ölçüde hız faktörünü içerirler. Bu nedenle bazı kaynaklarda bu ölçüm araçları "hızlandırılmış güç testi" olarak isimlendirilmiştir. Bir testin belirlenen süre içinde katılımcılarının %80 veya %90'ı tarafından tamamlanması onun melez yapıya sahip bir test olduğunu gösterir. ABD'de Psikoloji Hizmetleri Şirket yetkilileri bir testin hızlandırılmış olduğuna "hızlılık indeksi değerine" (HİD) bakarak karar verirler. Hızlılık indeksi değeri doğru yanıtlanan soru sayısı ile teşebbüs edilen madde sayısı arasındaki korelasyon katsayısının ,90'dan büyük olmasına göre belirlenir. Hesaplanan HİD katsayısı ,90'dan büyükse bu testler hızlandırılmıştır. Katsayısı ,90'dan düşük olan testler ise melez olarak isimlendirilir.[52] Kişilere eğer bilişsel yetenek test bataryası uygulanıyorsa bu bataryadaki testlerin bir kısmı güç ve bir kısmı ise hız testi niteliğindedir. Gulliksen'e göre (1950) zihinsel yetenekler güç ve hızın her ikisini de içerir (aktaran Tindal, 1999).[53] Böyle bir durumda melez nitelikteki bir test bataryasından söz etmek gerekir. Günümüzde saf bir biçimde güç testi bulunmadığından %5 veya % 10 oranında hızlandırılmış testler yine de genel bir terim olarak *güç testi* olarak isimlendirilir. Güç testlerinin belirli bir ölçüde hızlandırılması bu testlerin geçerliliğini bozmaz. Melez testlerde hızlandırma faktörü

ne hız testleri kadar yüksek ne de güç testleri kadar düşüktür. Melez testlerde optimum hız aşağıdaki faktörlere göre belirlenir:

1. Test edilen ana kütlenin özelliği.
2. Ölçülen kavramsal yapının özelliği.
3. Testin kullanım amacı.

Hız testlerinin farklı etnik kökene sahip kişilerde, erkeklere oranla kadınlarda bir ölçüde düşük sonuçlar verdiği bildirilmiştir. Ayrıca hızlandırılmış güç testlerinin hızlandırılmamış güç testlerine göre hız testleriyle daha yüksek korelasyon rakamlarına sahip olması nedeniyle melez testlerin geliştirilmesi için güçlü bir kanıt elde edilmiştir. Ree (1993) bir testin uygulanmasında hız faktörünün değiştirilmesiyle ölçülen yapının da önemli ölçüde değişebileceğini belirtmiştir (aktaran, Peterson).[54] Bu nedenle bilim adamları kavramsal yapısı net olmayan ölçümlerde testin hızlandırılmış bir biçimde ve hızlandırılmadan uygulanıp iki ölçüm sonuçları arasındaki ilişkilerin incelenmesini önermişlerdir. Hızı azaltmanın her zaman "farklı bir yapının ölçülmesi sonucunu doğuracağı" iddiasından uzak olarak, bu gibi durumlarda hızdan farklı *dominant bir boyutun* ortaya çıkacağı (*g* faktörünü veya başka bir boyutu ölçen) bildirilmiştir. Hızlandırılmış ve hızlandırılmadan uygulanan testlerin sonucunda aynı testin paralel olmayan iki ayrı versiyonu söz konusu olabilir.[55]

YETENEK TESTLERİNİN GÜVENİLİRLİĞİ

Literatürde yetenek testleri dört grup altında sınıflandırılır: bilişsel yetenek testleri, psikomotor yetenek testleri, fiziksel yetenek testleri, duyusal yetenek testleri. Bu bölümde ise, söz konusu dörtlü sınıflandırmaya beceri (yatkınlık) testleri ile bilgi testleri de ilave edilerek yetenek testlerinin güvenilirliği konusu altı alt başlık halinde incelenmiştir.

Bilişsel Yetenek Testleri

Bilişsel yetenek testleri (sözel, sayısal, mekanik, düzlemsel muhakeme, sözel akıcılık, dikkat vb. yetenekleri ölçen testler), (a) okullarından yeni mezun olmuş kişilerin yeteneklerini ölçmeye yönelik olarak veya (b) kendi branşında üç, beş sene çalıştıktan sonra belirli bir uzmanlık düzeyine gelmiş olan kişilerin kabiliyetlerini ölçmeye yönelik olarak uygulanır. Testler, hangi ana kütleye yönelik olarak uygulanıyorsa o ana kütlede iç tutarlılık, para-

lel formlar, yarıya bölme güvenilirliği ve test-yeniden test güvenilirlik analizleriyle sınanmalı ve ondan sonra kullanılmalıdır.

Bilişsel yeteneklerin türleri. Literatürde bilişsel yetenekler değişik başlıklar altında sıralanmıştır. Genelde *sayısal* ve *sözel* başlıkları altında sıralanan bu yetenekler kendi içinde ayrıca değişik alt bölümlere ayrılır.

1. *Zihinsel keskinlik*: hızlı öğrenme, hüküm verme yeteneği, problem çözme ve muhakeme yeteneği.
2. *Meslek terimlerine hakimiyet.* Meslek terimlerini iyi bilme ve kullanma.
3. *Genel kelime hazinesinin geniş olması.* Kişinin günlük dilde kullandığı kelime hazinesinin geniş olması ve bu kelimeleri rahat bir şekilde kullanabilmesi. Literatürde bu yeteneğe aynı zamanda sözel akıcılık adı verilmiştir.
4. *Bellek.* Kişinin hatırlama ve hafızada tutma gücünün yüksek olması.
5. *Sayısal yetenek.* Kişinin rakamsal değerlerle kolayca işlem yapabilmesi, problem çözebilmesi. Kişilerin tamsayılarla ondalıklı sayılarla ve kesirli sayılarla dört işlem yapma yeteneğini ölçer.
6. *Mekanik ilgi.* Kişinin teknik konulara ilgi duyması, teknik problemleri çözmesi.
7. *Sözel yetenek ve sözel muhakeme testleri.* Sözel yetenek testleri yazılı malzemeleri okuma, anlama ve yorumlama yeteneğini ölçer. Sözel muhakeme ise kelimeler ve kavramlar arasında anlamlı bir ilişki veya bağ kurma ve akıl yürütme konusuyla ilgilidir.

Sekiz on kadar yetenek veya yatkınlık testlerinin bir araya getirilmesiyle oluşturulan test bataryaları *g* faktörünü ölçmeye yönelik olarak, dört beş testten oluşturulan test bataryaları belirli bir yetenek alanını ölçmeye yönelik olarak ve bir iki testten oluşturulan test bataryaları ise spesifik bir yetenek alanını ölçmeye yönelik olarak oluşturulur. Bu testler birbirinden bağımsız olmakla birlikte yapılan araştırmalar testler arasındaki korelasyonların oldukça yüksek olduğunu göstermiştir. Literatürde sık kullanılan bilişsel yetenek testlerinden bazıları aşağıdaki gibidir:

1. *Otis Self Administering Tests.* Yönetim dışı işler için.
2. *Wonderlic Personnel Test.* Elli maddeli ve 12 dakika limitli genel yetenek testi.
3. *Wechsler Adult Intelligence Scale.* Yetişkinler için genel yetenek ve zeka testi.

Bilişsel yetenek testlerinde uygulanan güvenilirlik analizleri. Bilişsel yetenek testlerinde çoğunlukla test-yeniden test, yarıya bölme ve paralel formlar yöntemi uygulanır.

*Test-yeniden test yöntemi.*Bilişsel yetenek testlerin tümünde test-yeniden test yöntemi uygulanabilir. Bunun için testin niteliğine göre aradan en az iki hafta ile bir ay kadar bir sürenin geçmesi gerekir. Test-yeniden test yöntemi bataryanın toplam puanları temel alınarak veya bataryadaki her bir testin toplam puanları temel alınarak iki düzeyde hesaplanabilir. Birinci uygulamayla ikinci uygulama sonuçları arasındaki korelasyon katsayısı, ρ (xx) güvenilirlik katsayısını verir. Güvenilirlik katsayısı, iki uygulamaya ait puanlar arasındaki varyansın önemli ölçüde değişmediği varsayımına dayanır.

Yarıya bölme yöntemi. Bilişsel yetenek testleri tutum ölçeklerindeki gibi belirli ifadelere dayanmaz. Bu nedenle bu ifadelerin kavramsal yapıyı doğru temsil etmesi olgusundan çok geliştirilen maddelerin kendi içinde bir bütünlük gösterip göstermediğine bakılır. Bunun için sözel, sayısal, mekanik, düzlemsel muhakeme, sözel akıcılık, dikkat testleri kendi içinde ikiye bölünerek iki yarısı arasındaki korelasyonlar incelenir. Ancak bu testlerin kolaydan zora doğru sıralanan güç testi niteliğinde olmaması gerekir. Eğer maddelerin güçlük dereceleri eşit ise yarıya bölme yöntemi doğrudur. Güç testlerinde ise tek-çift madde uygulaması ile yarıya bölme yöntemi uygulanabilir. Yetenek testlerinin güvenle uygulanabilmesi için güvenilirlik katsayılarının ,80'inin üzerinde olması gerekir.

Paralel formlar yöntemi. Bilişsel yetenek testlerinde sık uygulanan bir diğer yöntemdir. Bir bilişsel yetenek testinin A ve B formları olmak üzere en az iki sürümünün geliştirilmesini ve bu sürümlerin çok sayıda kişiye aynı zaman diliminde "birbirinden bağımsız iki ayrı form halinde" kısa zaman aralığıyla (aynı gün veya bir iki hafta içinde[a]) uygulanmasını gerek-

[a] Paralel formların aynı gün içinde uygulanması daha çok yarıya bölme güvenilirliğine benzer. Aradan iki hafta gibi bir zaman geçtiğinde ise bağımsız bir uygulama olur. Bu özel-

tirir. A ve B formları arasındaki korelasyon katsayısının yüksek olması testlerin güvenilir olduğunu gösterir. Paralel formlar yönteminde korelasyon katsayısı en ,75 olmalıdır. Paralel formlarda her bir madde, benzer veya değişik kelime ve ifadeler kullanılarak yeniden üretilir. Bir testin gerçek anlamda paralel bir diğer sürümünü oluşturmak için test sonuçlarının aritmetik ortalama ve varyans değerlerinin eşit olması gerekir.

Pratik hayatta her zaman paralel test geliştirmek güç olduğundan paralel test yerine çoğunlukla alternatif testler uygulanır. Alternatif testlerde gözlem puanlarının aritmetik ortalama ve varyans değerlerinin yaklaşık olarak eşit olduğu varsayılır. Paralel veya alternatif test uygulamalarında yetersiz veya uygun olmayan madde örneklemesi, hata varyansının artmasına neden olur.

Yatkınlık Testlerinin Güvenilirliği

Yatkınlık testleri bilişsel yetenek testlerinden farklıdır. Yatkınlık testlerinde mekanik yetenek, büro işleri yeteneği, dil kullanma, klavye kullanma, müzik, güzel sanatlar, at veya köpek yetiştiriciliği, sürücülük gibi *spesifik yeteneklerin* ölçülmesi durumu söz konusudur.[a] Yatkınlık testlerinde daha çok uygulamaya dayanan beceriler ön plandadır. İş çevrelerinde sık kullanılan yatkınlık testlerinden bazılarının isimleri aşağıdaki gibidir:

1. *Minnesota Clerical Test.* Büro işlerinde gerekli olan hız ve doğruluğu ölçer.
2. *Revised Minnesota Paper Form.* Mekanik yetenek ve estetik yönelimi ölçer.
3. *Bennett Mechanical Comprehension Test.* Mekanik işlemler hakkındaki bilgileri ölçer.
4. *Computer Competence Tests.* Bilgisayar terminolojisi ve uygulamalarıyla ilgili yeteneği ölçer.
5. *Dil yeteneği.* Kişinin bir yabancı dili okuyup anlaması, çeviri yapabilmesi ve konuşabilmesiyle ilgili yetenektir.

---

liği ile test-yeniden test uygulamasına benzemekle birlikte kullanılan formlar aynı değil farklıdır.

[a] Literatürde yatkınlık testleriyle yetenek testleri arasındaki ayrım kesin çizgilerle belirlenmemiştir. Bazı yazarlar yatkınlık-yetenek testleri kavramlarını aynı anlamda kullanırlar. Ancak bu kitapta uygulamaya dönük beceri testleri yatkınlık testleri olarak isimlendirilmiştir.

İş çevrelerinde birbiriyle ilgili birkaç tane yatkınlık testi bir araya getirilerek yatkınlık test bataryası oluşturulabilir. Örneğin büro işleri test bataryası bürolarda kullanılan birden fazla yeteneği ölçen bir test grubundan oluşur.

Yatkınlık testlerinin güvenilirliğini belirlemek için kağıt-kalem testi şeklinde hazırlanmışsa iç tutarlılık analizi ve paralel formlar yöntemlerinden, aletli test şeklinde hazırlanmışsa test-yeniden test yöntemlerinden yararlanılır. Ancak yatkınlık testlerinde test-yeniden test güvenilirliği yöntemini uygulama önerilmemiştir. Çünkü yatkınlık testlerinde pratik yapma ile kişinin yeteneklerini geliştirmesi söz konusu olabileceğinden elde edilecek güvenilirlik rakamları gerçeği yansıtmayabilir.[56] Eğer test-yeniden test yöntemi uygulanacaksa aradan geçen sürenin kısa tutulmasında ve az sayıda pratik yapılmasında yarar vardır. Zaman içinde sonuç değerlerinin önemli ölçüde dalgalanma gösterdiği yatkınlık testleri için daha çok yarıya bölme güvenilirlik analizinin kullanılması önerilmiştir.

Psikomotor Testlerinin Güvenilirliği

Psikomotor (devinimsel) testler, zihinsel ve bedensel yetenekleri belirli bir karışım içinde ölçen araçlardır. Bu tür testlerde kas koordinasyonu, parmak becerisi, el-göz koordinasyonu gibi zihinsel değerlendirme ve fiziksel hareketleri birlikte içeren yetenekler önemlidir.

Psikomotor alan. Psikomotor yetenekler psikomotor alanla ilgilidir. Psikomotor alan, değişik bilim adamlarınca farklı biçimlerde tanımlanmış ve sınıflandırılmıştır. Psikomotor yetenekler; (a) izleme, (b) taklit etme, (c) uygulama ve (ç) alışkanlık geliştirmeyle kazanılır.[a] Bir araştırmacı bilişsel alanla psikomotor alan arasında net bir ayırım yapamıyorsa aşağıdaki sorulardan hareket edebilir:[57]

1. Hız bir faktör mü?
2. Herhangi bir araç, alet araç, cihaz kullanılıyor ve işlem yapılması gerekiyor mu?
3. Yapılan faaliyete ilişkin bir başarı değerlendirmesi söz konusu olabilir mi?

[a] Devinimsel becerilerin kazanılması konusunda literatürde değişik sınıflandırma modelleri vardır. Bu konunun ayrıntısına girilmemiştir.

Bu sorulardan herhangi birine verilecek *evet* yanıtı o alanın psikomotor yeteneklerle ilgili olduğunu gösterir.

Fleishman (1972) geniş bir yelpazede yer alan çok sayıdaki bilişsel-devinimsel hareket üzerinde yaptığı korelasyon ve faktör analizi sonuçlarına dayalı olarak test edilebilecek 11 farklı psikomotor beceri belirlemiştir: nişan alma, el-kol kımıldamazlığı, kontrol hassasiyeti, parmak becerisi, el becerisi, çoklu kas koordinasyonu, hız kontrolü, reaksiyon zamanı, tepki yönelimi, kol hareketlerinin hızı, bilek-parmak hızı (aktaran Vervys).[58] Little (2004) ise "devinimsel alanı" altı temel grupta toplamıştır: refleks hareketleri, temel beden hareketleri, tüm duyu organlarından gelen sinyalleri algılama hareketi, fiziksel yetenekler, ince maharet gerektiren yetenekler, beden dili ile konuşma hareketleri.[59] Literatürde devinimsel beceriler; *genel beceriler* ve *hassas beceriler* olmak üzere iki grupta toplanmış ve kadınların hassas beceri grubunda erkeklerden daha mahir oldukları bulunmuştur. Literatürde yer alan genel ve hassas becerileri ölçen psikomotor test türleri aşağıdaki gibidir:

1. *Genel hareket yeteneğini ölçen testler.* Genel hareket becerileri "genel vücut hareketleri" ile ilgilidir. Yürüme, oturma, eğilme, kaldırma, ellerini kullanma, monte etme, sökme gibi hareketleri içerir.
   a. Stromberg Dexterity Test.
   b. Minnesota Rate of Manipulation Test.
2. *Hassas hareket yeteneğini ölçen testler.* Eli hassas bir şekilde kullanmayı gerektiren becerilerdir. Küçük nesneleri el ile tutmayı ve parmakları kullanarak bunlar üzerinde işlem yapmayı gerektirir.
   a. The O'Conner Tweezer Dexterity Test.
   b. The Purdue Pegboard.
   c. Crawford Small Parts test.
3. *Kombine testler.* Genel vücut hareketlerini ve eli birlikte kullanmayı gerektiren görevlerde kullanılan testlerdir.
   a. The Pennsylvania Bi-Manual Worksample.
   b. Hand-Dexterity tests.

Psikomotor yetenekler belirli mesleklerde daha çok ön plana çıkar. Örneğin, hemşirelerde, ameliyat yapan hekimlerde, diş hekimlerinde, itfaiyecilerde, araba sürücülerinde, astronotlarda, uçak pilotlarında yapılan görevlerin bir bölümü devinimsel yeteneklerle ilgilidir.

Psikomotor yetenek testlerinin güvenilirliği. Bilim adamı psikomotor yetenek testlerinin güvenilirlik analizinden önce bu testlerin işle olan ilgililiğini sağlamak zorundadır. Devinimsel yetenek testleri iş gereklerinin bir bölümüyle ilgili olmak zorundadır. Devinimsel yetenekler bilişsel ve devinimsel alanın her ikisiyle de ilgili olduğundan diğer bilişsel testlerle birlikte kullanılır. Fleishman psikomotor yetenek testlerinin ilgili diğer özel yetenek testlerinin güvenilirliğinden daha düşük olduğunu bulmuştur. Yaptığı araştırmalarda psikomotor yetenek testlerinin güvenilirliğinin ,70 ilâ ,88 arasında değiştiğini bulmuştur. Güvenilirlik rakamlarının düşük olmasının nedeni olarak pratik yapma etkisi üzerinde durulmuştur. Devinimsel yetenek testlerinin güvenilirlik analizi için, kağıt-kalem testi - aletli test olması ve ayrıca testte hız faktörünün bulunup bulunmadığı araştırılır. Kağıt-kalem testi şeklinde hazırlanmış testlerde iç tutarlılık analizleri, aletli testlerde ise test-yeniden test ve gözlemciler arası değerlendirme tutarlılığı güvenilirlik analizi yöntemleri uygulanır. Güvenilirlik analizleri devinimsel alanın her birini ayrı ölçen testlerin her biri için tekrarlanır. Devinimsel test bataryasının genel veya ortalama güvenilirlik değerini bulmaya gerek yoktur. Bilim adamı bunun için mutlaka bir değer vermek istiyorsa medyan değerinden hareket edebilir.[60]

Psikomotor yeteneklerin ölçümünde her test deneklere belirli bir süre içinde arka arkaya üç kez uygulanır ve puanların güvenilirliği iki tekrarın sonucuna göre analiz edilir. Birinci uygulama "deneme" niteliğindedir. Deneme analiz sonuçları ölçüme katılmaz. Psikomotor testlerinin önemli bir bölümü hızlandırılmış olan testlerdir. Bu nedenle hız faktörü bulunan psikomotor testlerinde hız testleri için uygun olan güvenilirlik analizleri yapılır.

Fiziksel Yetenek Testlerinin Güvenilirliği

Fiziksel yetenek (performans) testleri kaldırma, çekme, itme, taşıma gibi konulardaki güç ve dayanıklılığı ölçer. Askerlerin, itfaiyecilerin, kamyon sürücülerinin, atletlerin, koruyucuların, ağır yükleri kaldırmayı ve taşımayı gerektiren görevlerde çalışan işçilerin seçiminde fiziksel yetenek testlerinden yararlanılır. Personel seçiminde fiziksel yetenek testlerinin kullanılmasının iş kazalarının azalmasına da katkı yaptığı bildirilmiştir. Bu tür işlerde personelin aşağıdaki yeteneklere sahip olması araştırılır:

1. *Statik güç*: Çekme, itme, taşıma.
2. *Patlayıcı güç*. Gücün bir anda ve çok yüksek yoğunlukta kullanılması.
3. *Vücut koordinasyonu*. Vücudun bir bütün olarak değişik yönlerde insicamlı bir şekilde hareket ettirilebilmesi.
4. *Kardiyovasküler zindelik*. Kas-damar sisteminin sağlıklı ve zinde olması.
5. *Denge*. Bisiklete binme, çok dar bir zemin üzerinde yürüme ve dayanıklılık gösterme yeteneği.

Kanada İnsan Hakları Komisyonuna göre, fiziksel değerlendirme testlerinde en azından üç kriterin karşılanıp karşılanmadığı araştırılmalıdır.[61]

1. Testler iş ögeleri temel alınarak belirlenmiş olmalıdır.
2. İş öğeleri için gerekli olan emniyetli, güvenli ve güvenilir bir başarı kapasitesi veya kriteri önceden tanımlanmış olmalıdır.
3. Testler kişinin emniyetli, etkili ve güvenilir bir performans gösterme durumunu ortaya çıkarma özelliğine sahip olmalıdır.

Fiziksel yetenek testlerinin güvenilirliğini saptamak için gözlemciler arası değerlendirme, gözlemci içi değerlendireme ve test-yeniden test güvenilirliği yöntemleri uygulanır.

Fiziksel yetenek testlerinin güvenilirliği kadar önemli bir diğer husus bu testlerde başarılı olmak için kriter olarak kabul edilen kesim noktasının ne olacağıdır. Kesim noktasını çok yüksek tutmak personel bulma ve ölçme işini zorlaştırırken düşük tutma, insanların güvenliğini tehlikeye atabilir. Fiziksel yetenek testlerinde kesim noktasını belirlemek için değiştirilmiş Angoff yönteminden, beklenti tablolarından, Taylor-Russel tablolarından, norm veya kriter değerlerinden yararlanılabilir. Norm veya kriter değerlerinden yararlanmak isteyen araştırmacıların bunun için en az 30 kişilik bir örneklem üzerinde araştırma yapmaları gerekir. Fiziksel yetenek testlerinde geçme puanı saptanırken testin geçerliliği, olumsuz etki faktörü, güvenlik, iş başarısı, verimlilik ve yasal düzenlemeler göz önünde bulundurulur. Bir iş için birden fazla fiziksel testi içeren bir bataryadan yararlanılmışsa bataryadaki test puanları ile iş başarısı arasında yapılacak

regresyon analizi sonucunda yüksek R^2 değerleri elde edilmiş olmalıdır. Yüksek R^2 değerleri testin aynı zamanda geçerli olduğunu gösterir.

Bilgi Testlerinin Güvenilirliği

Bilgi testleri, daha çok okullarda uygulanan Matematik, Fen Bilgisi, İngilizce ve Türkçe gibi derslerin kavranma ve öğrenme durumunu saptamaya yönelik olarak uygulanan sınavlarla ilgilidir. Bu sınavların her biri değişik yöntemlere göre uygulanan testlerden oluşur. Testler açık uçlu kompozisyon soruları, çoktan seçmeli sorular, kısa yanıtlar, paragraf şeklinde uzun yanıtlar, doğru-yanlış seçenekleri şeklinde düzenlenmiş sorular şeklindedir.

Bilgi testlerinin güvenilirliğinde maddeler için iç tutarlılık analizi, test-yeniden test, paralel formlar güvenilirliği ve ölçümün standart hatası değerleri kullanılır.[62] Cronbach alfa iç tutarlılık güvenilirliği ölçümün standart hatasının saptanmasına yönelik olarak hesaplanır. Bilgi testlerinde eğer sorular sadece çoktan seçmeli sorular şeklinde değil de karma soru düzenlemeleri şeklinde sorulmuşsa alfa güvenilirlik katsayısı düşük değerlikli bir katsayısıdır. Karma özellikli sınav sorularında iç tutarlılık güvenilirliği için Feldt-Raju indeks değerlerinin kullanılması önerilmiştir.[63]

Duyusal Yetenek Testlerinin Güvenilirliği

Duyusal yetenekler beş duyu organının hassasiyeti ile ilgilidir. İş hayatında en çok koklama, görme ve işitme duyusunun yeterliliği araştırma ve inceleme konusu olmuştur. İşitme duyusu kişinin çevresinden gelen sesleri algılaması ve aynı zamanda ayırt etmesiyle ilgilidir. İşitme duyusunun keskinliği, çevredeki normal konuşma seslerinin duyulması, cılız seslerin duyulması, fısıltıların duyulması ve belirsiz vücut seslerinin duyulması gibi giderek artan bir hassasiyet çerçevesinde belirlenir. Görme yeteneği, bireyin çevresindeki küçük nesneleri dahi görerek algılaması ve ayırt etmesiyle ilgilidir. Örneğin, kişinin bilgisayar ekranındaki yazıları yarım metre uzaktan okuyabilmesi, 6 metre uzağındaki bir kişiyi tanıyabilmesi, nesnelerin renk, renk doygunluğu, düzenleme, şekil ve tasarımlarını ayırt edebilmesi görme keskinliği olarak belirlenir. Koklama duyusu, bireyin çevresindeki uçucu gazları algılayabilmesi çevresel gazlarla parfüm gibi kişilerden kaynaklanan gazlar arasında ayrım yapabilmesidir.

Duyusal yetenek testlerinin güvenilirliği test-yeniden test yöntemi, gözlemciler arası değerlendirme güvenilirliği, deney ve kontrol gruplarında yapılan ölçümlerin karşılaştırılması suretiyle yapılır.

TANIMLAMA (RUBRİK) PUANLARININ GÜVENİLİRLİĞİ

Tanımlayıcı değerlendirme puanları (tanımlama puanları)[a] daha çok okullarda öğrencilerin yaptıkları çalışmaların kalitesini belirlemek için kullanılır. Öğretmenler yapılan işin niteliğini belirlemek için önce en düşük ve zayıf başarı durumunu tanımlayan "bir açıklama" veya "bir tanım" geliştirirler. Daha sonra bu açıklamalar kademe kademe daha iyi başarı durumunu gösterecek şekilde genişletilir. Yapılan açıklamaların her birine belirli bir puan verilir. Puanlar sıfırdan başlayıp beşinci veya yedinci dereceye kadar devam eder. Tanım puanlarının güvenilirliği, kategorilerin net bir şekilde açıklanmış veya tanımlanmış olmasına bağlıdır. Kategoriler arasında belirgin bir farklılaşma olmalı ve bu farklılaşma kategoriler arasında belirli bir dengeye sahip olmalıdır. Bazı kategori aralıkları diğerlerinden daha geniş olmamalıdır. Araştırmacı değerlendirme işlemine girişmeden önce her bir kategoriyi iyi tanımlayan örnekler geliştirebilir. Bu örnekler daha sonraki değerlendirme işleminde çipo görevi görür.

Tanım puanlarının güvenilirliğinde iki yöntemden yararlanılır. Değerlendirici içi güvenilirlik ve değerlendiriciler arası güvenilirlik.

Değerlendirici İçi Güvenilirlik

Değerlendirici içi güvenilirlik değerlendirmeyi yapan araştırmacının, öğretmenin veya gözlemcinin zaman içinde ne ölçüde tutarlı puanlama yaptığını belirlemeye yöneliktir. Öğretmenler değerlendirme sürecinde bir süre sonra yorulup daha dikkatsiz bir değerlendirme yapmaya başlayabilirler. Eğer test sorularının / başarı belgelerinin sayısı fazla ve değerlendirme tanımları ayrıntılı ise öğretmenler çok çabuk yorulurlar. Aynı öğretmen eğer başka bir zaman diliminde aynı testi veya başarı dosyasını değerlendirdiğinde farklı bir puan veriyorsa sonuçların güvenilirliği düşük çıkar. Puanlar arasındaki tutarsızlık öğrencilerin performansındaki değişiklikten değil değerlendirme koşullarından kaynaklanır. Değerlendirme tanımları önceden geliştirilmekle birlikte araştırmacı değerlendirme sürecinde bu tanımlarda istikrarlılığı sağlamak amacıyla bazı değişiklikler yapabilir.[64]

---

[a] Tanımlayıcı değerlendirmelere ait puanlar (rubrics) önceden belirlenen bir şablondur. Öğretmenler yapılan çalışmayı bu şablona bakarak değerlendirirler. Yazımda sözel akıcılık sağlaması için uzun bir şekilde "tanımlayıcı değerlendirme puanları" ifadesi yerine kavram kısaltılmış ve sadece *tanımlama puanları* sözcükleri kullanılmıştır.

Değerlendiriciler Arası Güvenilirlik

Değerlendiriciler arası güvenilirlik öğrencinin başarı dosyasının birden fazla öğretmen tarafından "tanım puanları" dikkate alınarak değerlendirilmesi ve bu değerlendirme sonucunda verilen puanların benzer çıkmasıdır. Değerlendirmede öğretmenler kendi kişisel yargıları yerine tanım puanlarını dikkate alırlar. Tanım puanları değerlendirme farklılıklarını bütünüyle ortadan kaldırmasa bile büyük ölçüde azaltma niteliğine sahiptir.[65]

PERSONEL SEÇİM TESTLERİNİN GÜVENİLİRLİĞİ

Personel seçim testlerinin güvenilirliği başlıca iki alanda incelenir. Birincisi test sonuçlarının istikrarlılığına ve ikincisi ise test alan kişilerin iş performans puanlarının istikrarlılığına bakılarak.

Test Sonuçlarının İstikrarlılığı

Test sonuçlarının istikrarlılığı test-yeniden test, paralel formlar ve iç tutarlılık (yarıya bölme, Cronbach alfa, KR-20, madde analizi) güvenilirlik analizleriyle yapılır. Personel seçimi uygulamalarında, test verilerek seçilen kişilere aradan 15 gün veya bir ay gibi bir süre geçtikten sonra aynı test tekrar uygulandığında sonuçlar benzer çıkmışsa testin güvenilir olduğu sonucuna varılır. Ancak aradan geçen süre içinde bir çok faktör test sonuçlarını etkileyebilir. Test alan kişinin psikolojik veya fizyolojik durumundaki değişiklikler, çevresel faktörlerdeki değişiklikler, test formunda yapılacak değişiklikler veya paralel nitelikte başka bir formun uygulanması, personel seçim değerlendirmesini yapan kişilerin, gözlemcilerin değişmesi yeniden test sonuçlarını etkileyebilir. Bu nedenle test sonuçlarının test-yeniden test güvenilirliği yapılırken kontrol edilmesi zor olan tesadüfî ölçüm hatalarının göz önünde bulundurulması ve mümkün olduğu ölçüde azaltılmaya çalışılması gerekir. Test sonuçlarının istikrarlığı ayrıca paralel formlar yöntemi veya iç tutarlılık analizleriyle de (yarıya bölme, Cronbach alfa, KR-20) sınanabilir.

Performans Puanlarının İstikrarlılığı

Performans puanları, psikolojik testlerle işe alınan kişilerin iş başında gösterdikleri başarıların dönemsel olarak değerlendirilmesiyle elde edilir. Kişilerin performans puanlarının istikrarlılığı aynı zamanda test sonuçlarının güvenilirliği konusunda önemli bir kanıttır. Performans puanları da değerleyicilerin psikolojik durumlarından büyük ölçüde etkilenir. Bu nedenle performans puanlarından hareket ederek güvenilirliği kesin ölçülerle tespit

etmek çok zor olmakla birlikte sonuçlar arasındaki korelasyon katsayıları değerleyicilere bir fikir verebilir.

Personel Seçim Testlerinin Güvenilirliğini Artırma

Personel seçim testlerinin güvenilirliği iki şekilde artırılabilir. Personel seçiminde kullanılan test maddeleri işi büyük ölçüde temsil ediyorsa ve başarı puanlarının değerlendirilmesinde yanlılık büyük ölçüde azaltılmışsa.

Testlerde içerik hatasının azaltılması test bataryasındaki testlerin işin içeriğiyle yakından ilgili olmasını artırır. Test bataryasına işte önemli olan, kritik olarak değerlendirilen, işteki başarıyı %60 veya %80 oranında etkileyen hareketler, davranışlar veya özellikler alınır. Bataryada nispetten önemsiz olan, işi yaklaşık olarak temsil eden, işle ilgililiği düşük olan testlerin veya test maddelerinin alınması ölçüm hatasını ve sonunda testin / bataryanın güvenilirliğini düşürür. Kişinin bazı özel yeteneklerini saptamaya yönelik testler bataryaya alınabilir, fakat bataryadaki testler arasındaki standart puan farklılıkları arttıkça bataryanın güvenilirliği düşer.

Performansı değerleyen üstlerin kendi aralarındaki değerleme farklılıkları da değerlemenin güvenilirliğini büyük ölçüde azaltır. Önemli ölçüde değerleme farkının bulunduğu maddelerde ek çalışmalar yapılarak resmin daha gerçekçi olarak belirlenmesi gerekir.

Personel Seçimi Amacıyla Kullanılacak Testlerin Güvenilirlik Puanlarının Değerlendirilmesinde Dikkat Edilecek Konular

İnsan kaynakları yöneticileri psikometrik test araçlarını eğer dış danışmanlık firmalarından temin etmişlerse öncelikle bu testlerin "teknik el kitaplarında" yer alan güvenilirlik araştırma sonuçlarını incelemeli ve hangi tür güvenilirlik analizi yapıldığını belirlemelidirler. Test geliştiren kişi ve kurumlar bu testleri ticarî olarak pazarlamadan önce geçerlilik ve güvenilirlik analizlerini yapmış olmalıdırlar. Güvenilirlik analizleri testin türüne göre değişiklik gösterebilir. Bazı testler için yarıya bölme güvenilirliği önemli iken diğerleri için KR-20, yarıya bölme, Cronbach alfa, test-yeniden test güvenilirliği veya gözlemciler arası değerlendirme güvenilirliği ön plana çıkabilir. Bir testin güvenilirlik katsayısının ,70'in üzerinde olması genelde kabul edilebilir bir ölçüdür. Ancak bu değer, sabit bir rakam olarak görülmemelidir. Testlerin asgarî güvenilirlik katsayıları da testin türüne göre değişir. Bunun yanında testler sadece güvenilirlik katsayılarının yüksek olmasına bakılarak belirlenmez. Güvenilirliğin yanında geçerlilik ve gruplar arasında ayırım yapma özelliğinin bulunmaması da aranması gereken özellikler arasındadır. İnsan kaynakları yöneticileri satın aldıkları testlerin

güvenilirlik ve geçerlilik katsayıları yanında "ölçümün standart hatası" değerlerini de incelemelidirler. Ölçümün standart hatası değeri test alan bir kişinin aldığı puanın hangi değerler arasında oynayabileceği hakkında bir fikir verir. Ölçümün standart hatası test puanlarının doğruluğunu belirleme ölçüsüdür. Bu değer küçük çıktığı ölçüde ölçüm sonuçlarının doğru olduğu anlamına gelir. İnsan kaynakları yöneticileri, test teknik el kitaplarında güvenilirlik analizleri yapılırken ölçümde hangi tür tesadüfi ölçüm hatalarının söz konusu olduğuna ilişkin bilgilerin ve söz konusu ölçüm hatalarının nasıl kontrol altına alındığına ilişkin verilerin raporlanıp raporlanmadığını incelemelidirler. Test teknik el kitabında güvenilirlik analizi yapılan örnek kütlenin demografik özelliklerine ilişkin bilgiler (yaş, cinsiyet, eğitim durumu, meslek, kıdem, deneyim) tam olarak verilmiş olmalıdır. Her bir demografik özelliğe sahip grubun/dilimin örneklem büyüklüğü, söz konusu demografik grubun ana kütle içindeki dağlımı göz önüne alınarak belirlenmelidir. Testler için eğer norm değerleri çıkarılmışsa güvenilirlik katsayıları her bir yaş dilimi, meslek grubu veya eğitim düzeyi için ayrı ayrı hesaplanmalıdır. Bu konudaki temel kriter, satın alınan veya satın alınması düşünülen testin kendi alanında uzmanlaşmış personel seçim örgütlerince belirlenmiş personel seçim testlerinin taşıması gereken standartlara sahip olup olmadığıdır.

MÜLÂKAT PUANLARININ GÜVENİLİRLİĞİ

Mülâkat, iki kişinin bilgi almak ve vermek üzere yaptıkları konuşmalar ve görüşmelerdir. Mülâkat puanlarının güvenilirliği mülâkat sorularından değil, mülâkat yapan ve yapılan kişinin konumlarından etkilenir. Mülâkat yapılan kişinin cinsiyeti, yaşı, gösterişli olması, aksanı ve uyumlu olması mülâkatı yapan kişileri etkileyebilir. Aynı şekilde mülâkatçı da mülâkat yapılan bireyi olumlu veya olumsuz yönde etkileyebilir. Mülâkat puanlarının güvenilirliğini belirlemek için birden fazla mülâkatçıdan ve önceden belirlenmiş tanımlama şemaları ve puanlarından (rubrics) yararlanılır. İki mülâkatçının verdiği puanlar arasında benzerlik varsa mülâkat değerlendirmesi güvenilirdir. Verilen puanlar önemli ölçüde birbirinden farklı ise mülâkat puanlarının güvenilirliği düşüktür.

Mülâkat puanlarının güvenilirliğini değerlendirmede yararlanılacak bir diğer yöntem mülâkatın aynı kişi tarafından farklı zamanlarda tekrarlanmasıdır. Eğer her iki uygulamada da benzer puanlar verilmişse mülâkat puanlarının güvenilir olduğuna karar verilir.

Mülâkat puanlarının güvenilirliğini etkileyen diğer faktörler mülâkatın uzunluğu ve mülâkatın yapılanma biçimidir. Mülâkatın süresi uzadıkça

değerlendirme güvenilirliği artar. Yapılan bir araştırmada ortalama mülâkat süresi 38,95 dakika ve mülâkatın standart sapma değeri ise 25,79 olarak bulunmuştur. Bu süre içinde kişilere 4 ilâ 34 arasında soru sorulmuştur. Ortalama soru sayısı 16,50 ve standart sapma değeri ise 8,71 olarak bulunmuştur.[66] Mülâkat yapılandırılmamış veya sorulacak sorular önceden belirlenmemişse mülâkat bilgilerinin güvenilirliğini değerlendirmek zorlaşır. Tam yapılandırılmış mülâkat uygulamalarında ise değerlendiricilerin verdikleri puanların birbiriyle tutarlı olma durumu daha iyi araştırılır. Literatürdeki araştırma bulguları yapılandırılmış mülâkat uygulamalarını destekler niteliktedir. Wagner (1949) tüm mülâkatların belirli bir amaç çerçevesinde gerçekleşmesi için yapılandırılmış formlar kullanılarak yapılmasını önermiştir. Mayfield (1964) tatmin edici güvenilirlik rakamlarının elde edildiği tüm mülâkat çalışmalarının "yapılandırılmış mülâkat" olduğunu belirtmiştir. Son yıllarda Huffcutt ve Arthur (1994) tarafından yapılan araştırmalarda da geçerlilik ve güvenilirliğin yapılandırılmış mülâkat uygulamalarında daha yüksek çıktığı bulunmuştur (aktaran, Campion).[67]

Personel Seçim Mülâkatlarının Güvenilirliği

Günümüzde iş hayatında çok sayıda personel seçim mülâkatı yapılır. Bu mülâkatların önemli bir bölümü geçerli ve güvenilir değildir. Çünkü, yöneticiler yapılan çalışmaların çoğunda görüşmecileri sübjektif yargılarına göre değerlendirirler. Yapılan mülâkatlar nadiren bütünüyle işle ilgili olarak sorulan sorulara dayanır. Mülâkatı yapan kişiler mülâkat teknikleri geçerlilik ve güvenilirlik konusunda özel bir eğitim almadıklarından ve gözlemlerini sistematik bir biçimde değerlendirmediklerinden vardıkları yargılarda öznel olmaktan kurtulamazlar. Personel seçim çalışmalarının güvenilirliğini sağlamak için yapılandırılmış mülâkat sürecinde aşağıdaki önlemler alınmalıdır.[68]

1. Ölçülebilir değerlendirme kriterleri belirlenir.
2. İş için gerekli olan bilgi, yetenek ve becerilerin listesi çıkarılır.
3. Belirli bir iş için kişilere sorulacak soruların listesi oluşturur. Çok sayıda soru sorularak mülâkat süresi uzun tutulur.
4. Tüm kişilere aynı sorular sorulur. İzleme ve teşvik etme soruları mümkün olduğu kadar kısıtlanır. Adayla tartışılmaz ve adayın sorularına cevap verilmez.
5. Her bir soruya verilmesi gereken veya verilebilecek standart cevap belirlenir ve bu cevap karşılaştırma kriteri olarak kullanılır.
6. Sorulara verilen yanıtların puanlama formatı belirlenir.

7. Değerlendirmenin panel grubu çerçevesinde yapılması sağlanır.
8. Mülâkatı yapacak kişiler eğitilerek puanlamanın nasıl yapılacağı hakkında kendilerine bilgi verilir.
9. Puanlamada ayrıntılı bir dereceleme ölçeği kullanılır.
10. Mülâkatçıların verdikleri puanlar istatistiksel olarak toplanır. Klinik değerlendirme olarak adlandırılabilecek soru bazında yatay toplama işlemine girişilmez.
11. Mülâkat sistemi belirli sayıda kişiyle görüşüldükten sonra tekrar gözden geçirilir.

Yapılandırılmış mülâkat tekniğinde gözlemciler arası değerlendirme güvenilirliğinin ,80'in üzerinde olması sağlıklı bir mülâkat tekniğinin uygulandığını gösterir.

Mülâkat Verilerinin Kalitesini Etkileyen Faktörler

Mülâkat verilerinin kalitesi başlıca iki faktörden etkilenir: Gözlemci yanlılığı ve gözlemci etkisi. Gözlemci yanlılığı, gözlemcinin değerlerinden, eğitim geçmişinden, kişisel tercihlerinden ve deneyimlerinden kaynaklanır. Bir diğer etken, belirli şekildeki tepkilere benzer puanlar verilerek "yanıt yanlılığı" olgusunun ortaya çıkmasıdır. Gözlemciler mülâkat yapılan kişiyi tanıyorlarsa, kişinin baskın bir özelliği açık bir şekilde ortaya çıkmışsa yine gözlemci yanlılığı söz konusu olabilir. Gözlemci etkisi ise, mülâkatı yapan kişinin/kişilerin mülâkat yapılan düzlem veya kişi üzerindeki etkisidir.[69] Mülâkatçının otoriter bir tutum takınması, mülâkat yapılan kişiyle sorgulanır bir tarzda görüşülmesi mülâkatçı etkisini gösterir.

İLGİ ENVANTERLERİNİN GÜVENİLİRLİĞİ

İlgi envanterleri, daha çok meslekî kariyer danışmanlığı yapma ve uygun işe yerleştirme amacıyla kullanılan ölçüm araçlarıdır. Belli bir süreye dayalı olmadan, kağıt-kalem araçları kullanılarak veya bilgisayar ortamında uygulanır. Bir ilgi envanterinin doldurulma süresi yarım saatten bir saate kadar uzayabilir. Personel seçimine yönelik olarak nadiren kullanılır. Bununla birlikte iş hayatında bazen yetenek ve beceri testleriyle birlikte kullanılarak kişi hakkında daha fazla bilgi edinilmeye çalışılır. Araştırmacılar, kişilerin değişik alanlarındaki ilgilerini ortaya çıkarmak amacıyla çok sayıda ilgi envanterleri geliştirmişlerdir. Bunların bir kısmı yaşam tarzına ait ilgiler iken diğerleri meslek tercihini belirlemeye yöneliktir. Günümüzde gelişmiş ülkelerde 40 bine yakın mesleki kariyer ve 300'e yakın akademik branş vardır. Meslekî ilgi envanterleri genel

branş vardır. Meslekî ilgi envanterleri genel gruplama veya kümeleme felsefesine dayalı olarak insanların ilgisini ortaya çıkarmayı hedefler. Bazı meslekî ilgi envanterleri ise daha spesifik alanların içindeki alt bölümlerle ilgilidir. Örneğin, işletmecilik branşının alt bölümlerine duyulan ilgiyi ortaya çıkarmak için geliştirilen envanterler bu kapsamda değerlendirilir. Yapılan araştırmalar ilgi envanterlerinin orta yaştaki kişiler arasında oldukça istikrarlı sonuçlar verdiğini, yirmi yaşın altındaki gençlerle üniversiteden yeni mezun olan gençlerde ise istikrarlılığın daha düşük olduğunu ortaya koymuştur.[70] İlgi envanterleriyle meslekî başarı ve eğitim başarısı arasındaki ilişkinin ,10 gibi düşük bir değer olduğu bulunmuştur (Schmidt ve Hunter, 1996, aktaran PSCC).[71]

İlgi envanterleri belirli bir kuramsal çerçeveye, kavramsal çatıya dayalı olarak veya işlemsel şemaya dayalı olarak geliştirilir. Örneğin, Jackson Meslekî İlgi Envanteri Holland'ın altı meslekî kişilik tipi temel alınarak geliştirilmiştir. Holland herkesin bir "iş kişiliğine" sahip olduğuna inanıyordu ve söz konusu iş kişilik tiplerini altı grup içinde sınıflandırmıştı: gerçekçi iş kişiliği, araştırıcı iş kişiliği, sanatçı iş kişiliği, sosyal iş kişiliği, girişimci iş kişiliği, konvensiyonel (ayrıntılı iş özellikleri, büro ortamı, düzen) iş kişiliği. İşlemsel şemalarda ise pratik hayattaki boyutlar veya temel görünümler, temel iş grupları dikkate alınır.

Genel meslekî ilgi envanterleri katılımcıların eğitim düzeyini ve deneyimlerini dikkate almadan hazırlanır. Bu nedenle bu testlere yanıt veren kişiler ilgilerini belirlemeye çalışırken daha çok geçmişteki deneyimlerini ve yaşantılarını göz önünde bulundururlar. İlgiler sadece yetenek ve yaşantılara bağlı olunca gerçek durumu yansıtmaktan uzaklaşır. İlgi envanterini uygulayan araştırmacı bu durumu saptamışsa söz konusu kişi için envanterde yer almayan başka boyutların da kişinin ilgi alanı içinde olabileceğini göz önünde bulundurmalı ilgileri kullanılan ölçekle sınırlandırmamalıdır. Envanter sonuçlarının daha sağlıklı ve doğru olarak yorumlanması ilgili kişinin kültürel çevresinin ve sosyal koşullarının ve yaşantısının birlikte ele alınmasını gerektirir.[72] İlk yıllarda envanterler belirli bir aktiviteye karşı bireyin duyduğu "hislere" dayandırılırken daha sonraki yıllarda ilgiler "deneme" aktiviteleriyle belirlenmeye çalışılmıştır. Ancak deneme aktivitelerinin çok uzun zaman alması nedeniyle günümüzde daha çok dereceleme ölçekleri ve ipsatif ölçekler kullanılmaya başlanmıştır.

Bir etkinlikten hoşlanma söz konusu etkinliğin hangi tür bir çevre içinde gerçekleştirildiğine, aktivitenin karmaşıklık derecesine ve aktivitenin sıklık derecesine bağlıdır. Aktivite tekrarlandığı uygun çevre içinde düzenlediği ölçüde hoşlanma ve ilgi duyma durumu daha belirgin olarak ortaya çıkar.

İlgi envanterleri değişik şekillerde düzenlenmiştir. Kontrol listeli ilgi envanterleri, resimli ilgi envanterleri ve sözel envanterler bunların başlıcalarıdır. Resimli ilgi envanterleri ilköğretimin birinci kademesindeki çocuklara uygulanır.

İlgi envanterlerinin yapılandırılma biçimi güvenilirliğini etkiler. İlgi envanteri zorunlu tercihe dayalı ipsatif ölçek şeklinde oluşturulmuşsa bu ölçüm araçlarının güvenilirliğini saptamak çok zordur. Ölçüm aracı dereceli ölçek şeklinde oluşturulmuşsa her bir boyut kendi içinde önce iç tutarlılık analizine tabi tutulabilir. Bunun için Cronbach alfa ve yarıya bölme güvenilirlik analizleri yapılır. Fakat daha doğru olan veya daha sağlıklı sonuç veren yöntem aradan on beş gün, bir ay, üç veya altı ay gibi bir zaman geçtikten sonra (hangisi uygulanabilir nitelikte ise) güvenilirlik için test-yeniden test yönteminin uygulanmasıdır. İlgi envanterlerinde test-yeniden test güvenilirliği eğer ,60'ın altında çıkmışsa ölçülen özeliğin çok çabuk değiştiğine, veya kişinin ilgilerinin somut bir şekilde belirginleşmediğine karar verilir.

İlgi envanterlerinin güvenilirliği kadar önemli olan bir diğer konu güvenilirlik analizlerinin yapıldığı örnek kütlenin niteliğidir. Güvenilirlik analizleri yapılan örnek kütlenin büyük olması, geniş bir mesleki tabana oturması, verilerin geniş bir sosyoekonomik ve demografik ana kütleden elde edilmiş olması güvenilirlik rakamlarını daha anlamlı hale getirir. İlgi envanterinin üniversitede okuyan öğrencilere, iş hayatında çalışan kişilere, yöneticilere, büro elemanlarına uygulanmasıyla test daha sağlam bir zemin üzerinde oturur. Belirli hedef kitlelere yönelik norm değerlerinin oluşturulmasıyla ilgi envanterinin norm temelli olarak yorumlanması mümkün olur. Literatürde sık kullanılan ilgi envanterlerinden bazıları aşağıdaki gibidir:

1. Carnegie Interest Inventory.
2. Kuder Preference Record.
3. Kuder Occupational Interest Survey.
4. Strong Interest Inventory.
5. Campbell Interest and Skills Survey.
6. Business Career Inventory.
7. Jackson Vocational Interest Survey.
8. Career Assessment Inventory.

9. Self-Directed Search.
10. College Major Interest Test.
11. Guilford-Zimmerman Interest Inventory.
12. California Occupational Preference Survey.
13. Ohio Vocational Interest Survey.
14. JOB-O.
15. Career Occupational Reference System.
16. Harrington O'Shea Career Decision Making System.
17. Occupational Interest Profile.

Bilim adamları ilgi envanterlerinde cinsiyet faktörünün önemli bir konu olduğunu belirtmişlerdir. Gottfredson (1981) kızların erkeklere göre daha dar meslekî tercih portföyüne sahip olduğunu ve bu özelliğin cinsiyet rollerine ilişkin sosyalleşme olgusundan kaynaklandığını ifade etmiştir (aktaran, Farmer).[73] Bazı araştırmacılar cinsiyetle ilgili maddelerin ölçekten çıkarılması gerektiğini savunurlarken diğerleri kromozom eksikliğine yol açacağından bu maddelerin ölçekte kalması gerektiğini ifade etmişlerdir. Amerika Birleşik Devletlerinde "Ulusal Eğitim Enstitüsü" ilgi envanterlerinde cinsiyet yanlılığını azaltmaya yönelik olarak bir rehber veya kılavuz geliştirmiştir.

YABANCI DİLDEN UYARLANMIŞ TESTLERİN GÜVENİLİRLİĞİ

Araştırmacı bir ölçüm aracını üç şekilde elde edebilir: (a) tüm maddelerini gözlem ve literatür taramasına dayalı olarak kendi geliştirebilir, (b) yabancı dilde yayımlanmış "bir ölçüm aracını" çevirerek Türkçeye uyarlayabilir, (c) yabancı dilde hazırlanmış "tek bir ölçeği temel alarak" veya yabancı dilde hazırlanmış "birden fazla ölçekten yararlanarak" kendine ait yeni bir ölçüm aracı oluşturabilir. Araştırmada hangi yöntem uygulanırsa uygulansın her üç yöntemde de geçerlilik ve güvenilirlik analizlerinin yeniden yapılması gerekir.

Yabancı Dilden Çevrilerek Uyarlanan Ölçüm Araçlarının Güvenilirliği

Yabancı dilden *aynen alınarak* veya *küçük değişiklikler yapılarak* çevrilen ölçüm araçlarının kullanılabilmesi için orijinal yazarından izin alınması gerekir. Ölçüm aracının yazarı veya geliştiricisi referans göstermek şartıyla kullanım serbestisi vermişse izin almaya gerek yoktur. Uluslar Arası Test Komisyonu – UATK, (International Test Commission – ITC) uyarlanmış testlerin uygun bir şekilde çevrilmesi ve kullanılması için halen devam eden bir çalışma içindedir. Hambleton (1994) bu çalışmadan yararlanarak test uyarlama çalışmalarına yardımcı olması için 22 maddeden oluşan bir rehber geliştirmiştir.[74] Bilim adamı, yaptığı araştırmada yabancı dilden çevirerek kullandığı test ve ölçekler hakkında öncelikle gerekli olan ön bilgileri verir. Testin daha önce kimler tarafından kullanıldığını, geçerlilik ve güvenilirlik analizi sonuçlarının ne olduğunu açıklar. Kullanılan ölçüm aracıyla ilgili yayımlanmış makaleler varsa bunların neler olduğu konusunda bilgi verir. Bir ölçüm aracının yabancı dilden Türkçeye uyarlama sürecinde atılması gereken başlıca iki adım vardır.[75]

1. Yapı, kavram ve dil eşitliğini sağlamak.
2. Ölçüm aracının psikometrik özelliklerini değerlendirmek.

Uyarlanan testlerde birincil derecede önemli olan konu *yapı* eşitliğidir. Ölçülmek istenen yapı, çeviri yapılan kültüre yabancı ise ölçüm sonuçları havada kalır. Araştırmacılar arasında yaygın olan bir mit (hurafe), "kavramsal yapıların evrensel olduğu" düşüncesidir. Örneğin, *Dönüşümsel Liderlik Ölçeği*'nde "idealleştirilmiş etki" boyutunu kısaca "karizma" olarak çevirebiliriz, ancak anketi dolduran kişiler bu boyut altında toplanan ifadelere verdikleri yanıtlarda eğer başka bir boyutun veya kavramsal yapının etkisi altında kalmışlarsa ölçüm sonuçları yanlıştır veya geçersizdir. Yapı eşitliği her iki kültüre ve ölçüm konusuna aşina olan uzmanlar tarafından saptanır. Yapı eşitliği konusunda herhangi bir tereddüt varsa söz konusu yapı veya boyut araştırmacı tarafından yeniden tanımlanır.

Kavram ve dil eşitliği ise, çift çeviri yöntemiyle sağlanır. Kavramlar ve terimler bütün kültürlerde aynı anlama gelmez. Anlam benzerliği olmasına karşılık nüans farklılıkları söz konusu olabilir. Başarılı bir çeviri veya test uyarlaması, nüans farklılıklarının da ustaca sergilenmesini gerektirir. Bunun için ölçüm aracının Türkçe çevirisi tekrar yabancı dile çevrilmeli ve söz konusu yabancı dilden metin ikinci bir kez daha Türkçeye tercüme

edilmelidir. Böylece metin kaynak dilden iki kez Türkçeye çevrilmiş olacaktır. Çevrilen maddelerin kalitesi üç faktör açısından değerlendirilir:[76]

1. Çevirinin doğruluğu ve cümlelerin açık olması.
2. Kullanılan kelimelerin zorluk derecesi.
3. Kullanılan dilin akıcılığı.

Çeviri yapan kişilerin yabancı dili sadece çok iyi bilmeleri yeterli olmaz, bu kişiler aynı zamanda çevri yaptıkları kültürü de çok iyi tanıyor olmalıdırlar.[77] Hambleton ve Patsula, (1999) çeviri yapacak kişilerde en azından dört kriterin bulunması gerektiğini belirtmişlerdir: (a) her iki dili de çok iyi biliyor olmak, (b) her iki kültürü de yakından tanıyor olmak, (c) test edilen veya ölçülmeye çalışılan konu hakkında yetkin biri olmak, (ç) madde yazma konusunda uzman olmak (aktaran Sireci).[78]

Çevirisi yapılan sözel içerikli ölçeklerde bazı ifadelerin, deyimlerin ve terimlerin diğer ülkelerde tam karşılıkları bulunmayabilir. Bu gibi durumlarda bu ifadelerin ya ölçekten çıkarılması veya yerlerine yakın anlamlı başka ifadelerin oluşturulması gerekir. Bilim adamı yabancı dilden çevirdiği ölçüm araçlarında eğer bazı değişiklikler yapmışsa bu değişikliklerin testin hangi bölümleriyle veya maddeleriyle ilgili olduğunu, değişiklik sonucunda güvenilirlik rakamlarında ne gibi değişiklik gözlemlendiğini yorumlamalıdır. Bilim adamı başarılı bir uyarlama çalışması için mümkün olduğu kadar birden fazla kişiden veya uzmandan yararlanmalıdır. Böylece gözden kaçan çeviri hataları yakalanmış ve kültüre uygun daha başarılı bir adaptasyon çalışması yapılmış olacaktır. Bu açıdan bireysel çeviri yöntemi yerine panel çeviri yöntemi daha yararlıdır.

Uyarlanan ölçüm aracının psikometrik özellikleri ise, pilot araştırma yapılarak testin faktöriyel yapısının, geçerlilik ve güvenilirlik analizlerinin çıkarılmasıyla ilgilidir. Literatürdeki faktör yapısı ile araştırmacının ampirik araştırma sonucunda elde ettiği faktöriyel yapı uyuşma göstermeyebilir. Bilim adamı, araştırmasında kendi topladığı verilere dayanan faktöriyel yapıyı temel alır. Geçerlilik ve güvenilirlik analizleri de ayrıca yapılmalıdır. Böylece literatürdeki geçerlilik ve güvenilirlik analizi sonuçlarıyla Türkiye'deki geçerlilik ve güvenilirlik analizi sonuçlarını karşılaştırmak mümkün olur.

Test geliştirme çalışmalarında olduğu gibi test uyarlama çalışmalarında da süreklilik esastır. Araştırmacı Türkçeye uyarladığı bir test veya ölçeği pilot araştırmanın sonunda esas araştırmada tekrar sınayacak ve elde ettiği

sonuçları yeniden gözden geçirecektir. Uyarlanmış bir testin daha iyi bir noktaya gelmesi belki birkaç kez uygulanıp gerekli iyileştirme çalışmaları yapıldıktan sonra mümkün olabilir. Uyarlanmış testlerde orijinal madde sayısının üç katı kadar madde geliştirilmez. Sadece maddeler üzerinde belirli revizyonlar yapılır.

Yabancı Dilde Yayımlanmış Ölçüm Araçlarından Yararlanarak Yeniden Geliştirilen Testlerin Güvenilirliği

Bilim adamı test geliştirirken literatürdeki araştırmalardan ve daha önce geliştirilmiş olan ölçüm araçlarından da yararlanabilir. Bu uygulama aynen çeviri olarak değerlendirilmediğinden söz konusu çalışmalarda çift çeviri koşulu aranmaz. Kültürler arası bir karşılaştırma yapılmayacaksa diğer test ve ölçeklerden yararlanarak orijinal bir çalışma ortaya koymak daha doğrudur. Araştırmacı serbest çeviri ile bir veya birden fazla ölçüm aracından yararlanabilir. Bu konuda iki yaklaşımdan biri uygulanır. Birinci yaklaşımda *tek bir ölçüm aracı* temel alınarak bu ölçüm aracındaki boyutlara dayalı olarak madde geliştirilir. Maddelerin bir bölümü orijinal ölçüm aracından alınırken önemli bir bölümü araştırmacı tarafından geliştirilir. Bu yaklaşım "çevirerek uyarlama" olmadığından araştırmacı nihai ölçekte olmasını arzu ettiği madde miktarının en az üç katı kadar madde geliştirmek zorundadır. Belirli bir ölçekten/testten hareket edilmiş fakat testin yapısı bozularak çok daha fazla madde geliştirilmiştir. Araştırmacıların sık yaptıkları yanlış, geliştirdikleri ölçek "aynen çeviri" niteliğinde olmadığı halde temel aldıkları ölçekteki madde sayısı kadar madde ile yola koyulmaları ve faktör analizini bu maddeler üzerinden yapmalarıdır.

İkinci yaklaşımda ise test geliştirilirken belirli bir kavramsal yapıyı ölçen birden fazla ölçekten yararlanılır. Faktör sayısı, bu ölçekler genel olarak değerlendirildikten sonra belirlenir. Faktörler ve her bir faktörün altındaki göstergeler değişik ölçeklerden alınmış olabilir. Bu uygulamada da araştırmacı nihai ölçekte/alt ölçekte bulunmasını arzu ettiği madde sayısının en az üç katı kadar madde ile yola koyulmalıdır. Herhangi bir ölçeğin madde sayısını kavramsal alan evreni olarak belirlemek doğru değildir.

ALINTI YAPILAN KAYNAKLAR

[1] G. Ring, "Computer Administered Testing [Bilgisayara Uyarlanmış Testler]," 1994, <http://www.aset.org.au/confs/iims/1994/qz/ring1.html> (20.03.2004).

[2] Aynı.

[3] K.A Campbell, D.S. Rohlman, D. Storzbach, LM. Binder, W.K. Anger, C.A. Kovera, K.L. Davis ve S.J. Grossmann, "Test-retest Reliability of Psychological and Neurobehavioral Tests Self-administered by Computer. [Bilgisayara Uyarlanmış Psikolojik ve Nöro-davranışsal Testlerin Test-yeniden Test Güvenilirliği]," 2003, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=9971880&dopt=Abstract> (14.03.2004).

[4] Aynı.

[5] R.B. Loerke, "The Psychometric Benefits of Soft-linked Items [Yumuşak Bağlantılı Maddelerin Psikometrik Yararları]," <http://www.findarticles.com/cf_0/m0FCG/3_29/91707800/p3/article.jhtml?term=> (03.04.2004).

[6] Loerke, "The Psychometric Benefits."

[7] Nükhet Çıkrıkçı-Demirtaşlı, Türk Psikoloji Bülteni, "Psikometride Yeni Ufuklar: Bilgisayar Ortamında Bireye Uyarlanmış Test," 1999, <http://www.psikolog.org.tr/bulten/13/13_yeniufuk.htm> (03.04.2004).

[8] Lutz F. Hornke , "Item Response Times in Computerized Adaptive Testing [Bilgisayar Uyarlı Testlerde Madde Yanıt Kuramı]," 2000, <http://www.uv.es/psicologica/articulos1y2.00/hornke.pdf> (03.04.2004).

[9] Aynı.

[10] James B. Olsen, "Guidelines for Computer-Based Testing [Bilgisayar Uyarlı Test Uygulamaları İçin Rehber]," <http://www.isoc.org/oti/articles/0500/olsen.html> (20.03.2004).

[11] John P. Sabatini, "Using Technology for Assessment in Adult Learning [Yetişkin Öğreniminin Değerlendirilmesinde Teknolojinin Kullanımı]," <http://www-tcall.tamu.edu/archives/hopey/08.pdf> (20.03.2004).

[12] Steven L. Wise ve G. Gage Kingsbury, "Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program [Bilgisayar Uyarlı Test Programları Geliştirme ve Uygulamada Karşılaşılan Sorunlar]," <http://www.uv.es/psicologica/articulos1y2.00/wise.pdf> (20.03.2004).

[13] Sabatini, "Using Technology."

[14] Educational Testing Services, "Standards [Standartlar]," <http://ftp.ets.org/pub/corp/standards.pdf> (04.04.2004).

[15] O. Wilhem, M. Witthöft ve A. Gröbler, "Comparisions of Paper and Pencil and İnternet Administrated Ability and Achievement Tests [Kağıt-Kalem ve Bilgisayar Uyarlı Yetenek ve Başarı Testlerinin Karşılaştırılması]," <> (27.03.2004).

[16] Wilhem, Witthöft ve Gröbler, "Comparisions of Paper."

[17] Paul W. Brooks, "Internet Assessment: Opportunities and Challenges [İnternet Değerlendirmesi: Fırsatlar ve Tehditler]," <http://www.ipmaac.org/conf00/brooks.pdf> (27.03.2004).

[18] Karen Kerstıng, "How do You Test on the Web? Responsibly [Ağ'da Sorumlu Bir Şekilde Nasıl Test Uygulayabilirsiniz?]," 2004, <http://www.apa.org/monitor/mar04/test.html> (27.03.2004).

[19] Kerstıng, "How do You Test."

[20] W. Chan, "Analyzing Ipsative Data in Psychological Research [Psikolojik Araştırmalarda Kendine Referanslı Verilerin Analizi]," <http://wwwsoc.nii.ac.jp/bsj/B30_1_6.pdf> (14.03.2004).

[21] R. J Cohen, P. Montague, L. S. Nathanson ve M. E. Swerdlik, "Personality Assessment [Kişilik Değerlendirmesi]," <http://www.indiana.edu/~educy520/sec5982/week_3/cohen88.pdf> (14.03.2004).

[22] H. Baron, "Strengths and Limitations of Ipsative Measurement [İpsatif Ölçümlerin Gücü ve Sınırlılıkları]," *Journal of Occupational and Organisational Psychology*, Aralık 1995.

[23] Baron, "Strengths and Limitations."

[24] G.P. Gruber, "Standardized Testing and Employment Equity Career Counselling [Standardize Edilmiş Testler ve Kariyer Danışmanlığında İstihdam Eşitliği]," <http://www.psc-cfp.gc.ca/centres/employment_equity/eecco/pdf/standardized_e.pdf> (14.03.2004).

[25] Robin D. Froman, "Elements to Consider in Planning the Use of Factor Analysis [Faktör Analizini Kullanmayı Planlarken Göz Önünde Bulundurulması Gereken Ögeler]," <http://www.snrs.org/members/SOJNR_articles/iss05vol02.pdf> (19.03.2004).

[26] Bienvenue a Psychtech International, "The myths and Realities of Psychometric Testing [Psikometrik Testlerde Mitler ve Gerçekler]," <http://www.psytech.co.uk/france/myth.htm> (14.03.2004).

[27] Campbell vd., "Test-retest Reliability."

[28] Aynı.

[29] Aynı.

[30] Baron, "Strengths and Limitations."

[31] R. Bower, "Results [Sonuçlar]," 2004, <http://randy.bower.com/dissertation/Results.html> (14.03.2004).

[32] Constance J. Jones, "Developmental Paths of Psychological Health From Early Adolescence to Later Adulthood [Yetişkinlikten Yaşlılığa Kadar Geçen Sürede Psikolojik Sağlığın Gelişim Yolları]," <http://www.apa.org/journals/pag/pag152351.html> (14.03.2004).

[33] Reagan D. Brown ve Robert J. Harvey, "Detecting Personality Test Faking with Appropriateness Measurement: Fact or Fantasy? [Kişilik Test Yanıltmacalarını Ölçümün Uygunluğu Açısından Değerlendirme: Gerçek mi Hayal mi?]," <http://harvey.psyc.vt.edu/Documents/BrownHarveySIOP2003.pdf> (11.04.2004).

[34] Brown ve Harvey, "Detecting Personality."

[35] T.F. Donlon, "An Annotated Bibliography of Studies of Test Speededness [Hız Testi Çalışmalarına İlişikin Açıklamalı Bibiliyografya]." Educational Testing Service, 1980.

[36] Aynı.

[37] N.G. Peterson, "Review of Issues Associated with Speededness of GATB Tests [GATB Testlerindeki Hızlılıkla İlgili Sorunların Gözden Geçirilmesi]," 1993, <http://www.onetcenter.org/dl_files/Speed_GATB.pdf> (13.04.2004).

[38] Measurement and Research Department Reports, "A Logistic Model For Time Limit Tests Zaman Kısıtı Getirilen Testler İçin Lojistik Model]," <http://download.citogroep.nl/pub/pok/reports/Report92-1.pdf> (12.04.2004).

[39] A. S. Cohen, J.A. Wollack, D.M. Bolt ve A.A. Mroch "Test Spededness [Test Hızlandırması]," < http://wiscinfo.doit.wisc.edu/exams/cohen_wollack_bolt_mroch02.pdf> (25.01.2004).

[40] D. Culları, "Psychological Testing [Psikolojik Testler]," <http://www.lvc.edu/psychology/courses/testing.html> (18.10.2002).

[41] A. Yu, "Using SAS for Item Analysis [Madde Analizi İçin SAS'ı Kullanma]," <http://seamonkey.ed.asu.edu/~alex/teaching/assessment/alpha.html> (25.01.2004).

[42] hr-guide.com, "Administering Assessment Instruments [Değerlendirme Araçlarının Yönetimi]," <http://www.hr-guide.com/data/G365.htm> (25.01.2004).

[43] Peterson, "Review of Issues."

[44] Aynı.

[45] ETS, "Major Field Test [Temel Alan Testi]," <http://ftp.ets.org/pub/corp/hea/overview2003.pdf> (25.01.2004).

[46] ETS, "Major Field Test."

[47] Peterson, "Review of Issues."

[48] Aynı.

[49] TerraNova, " Technical Quality [Teknik Kalite]," <http://www.ctb.com/media/mktg/terranova/other_media/tech_quality/superior_process.pdf > (28.03.2004).

[50] M. Gross, "Fairness in Employment Testing [İstihdam Testlerinde Eşitlik]," <http://web.mit.edu/~gross/Public/autor/fairness-summary.doc> (27.03.2004).

[51] ETS, "Effect of Fever Quoestions per Section on SAT I Scores [SAT I Testinin Her Bölümünde Daha Az Soru Sorulmasının Etkileri]," <http://www.ets.org/research/dload/RR-03-08.pdf> (12.4.2004).

[52] J.A. Weiner, "Computerizing Cognitive Ability Assessments [Bilişsel Yetenek Değerlendirmelerinin Bilgisayarlaştırılması]," <http://www.psionline.com/PDFLibrary/SIOP_2002_CBT_TechIssues.pdf> (13.04.2004).

[53] Gerald Tindal, "A Summary of Research on Test Changes [Test Değişiklikleri Konusundaki Araştırmaların Özeti]," 1999, <http://www.ihdi.uky.edu/msrrc/Word%20Docs/Tindal&Fuchs1.doc> (28.03.2004).

[54] Peterson, "Review of Issues."

[55] Aynı.

[56] Laljit Sidhu, "Concepts in Psychological Assessment [Psikolojik Değerlendirmede Kavramlar]," <http://www.nldontheweb.org/concepts.htm> (02.04.2004).

[57] HCC, "Learning Domains [Öğrenme Alanları]," <http://honolulu.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/m-files/m-domain.htm> (16.04.2004).

[58] C. Verwys, "Tests of Special Abilities [Özel Yetenek Testleri]," <http://www.rpi.edu/~verwyc/Chap9tm.htm> (13.04.2004).

[59] Julie K. Little, "Psychomotor Domain [Devinimsel Alan]," <http://itc.utk.edu/~jklittle/edsmrt521/psychomotor.html> (16.04.2004).

[60] R.C. Gardner, "The Attitude/Motivation Test Battery: Technical Report [Motivasyon Test Bataryası Teknik Raporu]," <http://publish.uwo.ca/~gardner/amtb4e.htm> (16.04.2004).

[61] BC Puplic Service Agency, "Assessment Methods: Physical Tests [Değerlendirme Yöntemleri: Fiziksel Testler]," <http://www.hrtoolkit.gov.bc.ca/staffing/staffing_steps/assess_methods/physical_tests/Physical_Tests.htm> (03.04.2004).

[62] New York State Department of Education, "2001 New York State Grade 4 Mathematics Statewide Assessment [2001 New York Eyaleti 4. Sınıfların Matematik Dersi Puanlarının Değerlendirilmesi]," <http://www.google.com.tr/search?q=cache:M326yj_C5MsJ:www.emsc.nysed.gov/ciai/testing/assesspubs/G4marpt01.PDF+conditional+sem+response+item&hl=tr&ie=UTF-8&inlang=tr> (14.08.2003).

[63] Aynı.

[64] Barbara M. Moskal ve Jon A. Leydens, "Scoring Rubric Development: Validity and Reliability [Tanım Puanları Geliştirme: Geçerlilik ve Güvenilirlik]," <http://pareonline.net/getvn.asp?v=7&n=10> (27.03.2004).

[65] Aynı.

[66] J.B. Campion, "A Review of Structure in the Selection Interview [Seçim Mülakatlarında Yapının Gözden Geçirilmesi]," <http://www.ipmaac.org/files/campion.pdf> (10.04.2004).

[67] Campion, "A Review of Structure".

[68] J. Sullivan, "Interview Information [Mülâkat Bilgisi]," <http://online.sfsu.edu/~johns/Mgt614/IntvwInfo.htm> (10.04.2004).

[69] Jeffery Oescher, "Educational Research [Eğitimsel Araştırma]," <http://employees.oneonta.edu/bischopj/Chapter%2007.ppt> (10.04.2004).

[70] M. Mills, "Psychological Test Report [Psikolojik Test Raporu]," <http://bellarmine.lmu.edu/faculty/mmills_fp/Testing/smpl-rpt-c.htm> (10.04.2004).

[71] PSCC, "Interest Inventories [İlgi Envanterleri]," <http://www.psc-cfp.gc.ca/centres/employment_equity/eecco/interest_e.htm> (10.04.2004).

[72] Aynı.

[73] H.S. Farmer, "Gender Differences in Adolescent Career Exploration 8Yetişkinlerin Kariyer Araştırmasında Cinsiyet Farklılıkları]," <http://www.ericdigests.org/1996-3/gender.htm> (10.04.2004).

[74] ITC, "ITC Projects [ITC Projeleri]," <http://www.intestcom.org/itc_projects.htm#ITC Guidelines on Adapting Tests> (18.04.2004).

[75] Medical Outcomes Trust, "SAC Instrument Review Criteria [SAC Aracı Gözden Geçirme Kriteri]," <http://www.qolid.org/public/34sacrev.htm> (10.04.2004).

[76] Ronald K. Hambleton ve Liane Patsula, "Increasing the Validity of Adapted Tests [Uyarlanmış Testlerin Geçerliliğini Artırma]," 1999, <http://www.testpublishers.org/Documents/journal0114.pdf> (18.04.2004).

[77] Stephen G. Sireci, "Guidelines for Adapting Certification Tests for Use Across Multiple Languages [Uyarlanmış Sertifika Testlerinin Farklı Kültürlerde Kullanılması İçin Rehber]," <http://www.cesb.org/Guidelines%20for%20Adapting.htm> (18.04.2004).

[78] Aynı.

ÖLÇÜMÜ ETKİLEYEN FAKTÖRLER VE ÖLÇÜMÜN İYİLEŞTİRİLMESİ

Tercih edilen modele, uygulanan ölçüm aracının türüne ve ölçüm çalışmasının gerçekleştirilme sürecine göre toplanan verilerin değişik sayıda faktörden etkilenmesi söz konusudur. Başarı testlerini etkileyen faktörler gözlemci değerlendirmelerini etkileyen faktörlerden ve bunlar da kişisel özdeğerlendirme anket formlarını etkisi altına alan psikolojik ve sosyal koşullardan farklıdır. Bilim adamı, ölçüm çalışmasının kendi özel koşullarını göz önünde bulundurarak ölçüm sonuçlarını "bulandıran" faktörlerin etkisini en alt düzeye çekmeye çalışmalıdır. Ölçüm sırasında araya giren etkenlerin verilerdeki saflığı bozmasına "bulaşma etkisi" ve söz konusu etkenlere de "gürültü" adı verilir. Verilerdeki "gürültünün" yoğunluğu ne kadar fazla ise yapılan çalışma o denli daha az güvenilir olarak değerlendirilir. Bu bölümde önce verilerde gürültüye neden olan faktörler ele alınmış ve daha sonra ölçümün ve veri kalitesinin iyileştirilmesine yönelik önlemler, süreçler üzerinde durulmuştur.

ÖLÇÜMÜ ETKİLEYEN FAKTÖRLER

Klâsik test kuramında, belirli sayıda faktör bir testin/ölçeğin veya gözlemci değerlendirmesinin güvenilirliğini etkiler. Uygulanan güvenilirlik yöntemleri bazı hatalardan daha fazla etkilenirken diğerlerinden daha az etkilenir. Ölçümde hataya yol açan faktörler; olgunlaşma, reaksiyon, taşıma etkisi, alan örneklemesi, alanın türdeşliği ve puanlayıcı etkisi başlıklarında sıralanabilir. Test-yeniden test güvenilirlik analizlerinde olgunlaşma, reaksiyon ve taşıma hata varyansı daha çok etkili olurken, paralel formlarda olgunlaşma ve alan örneklemesi faktörleri sonuçları önemli ölçüde etkiler. Yarıya bölme yönteminde ise alan örneklemesi faktörü nedeniyle sonuçlar belirli ölçüde hata içerir. Alfa yönteminde alan örneklemesi ve alanın türdeşliği, gözlemciler arası değerlendirmede puanlayıcı etkisi sonuçları bir ölçüde çarpıtma özelliğine sahiptir. Klâsik test kuramına göre uygulanan

güvenilirlik yöntemleri ve hata faktörleri Tablo 13-1'de özet olarak verilmiştir.

Tablo 13-1. Güvenilirlik Yöntemleri ve Hata Faktörleri

Hata faktörleri	*Test-Yeniden test*	*Paralel formlar*	*Yarıya bölme*	*Alfa iç tutarlılık*	*Gözlemciler arası tutarlılık*
Olgunlaşma	°	°	×	×	×
Reaksiyon	°	◊	×	×	×
Taşıma etkisi	°	◊	×	×	×
Alan örneklemesi	×	°	°	°	×
Alanın türdeşliği	×	×	×	°	×
Puanlayıcı etkisi	×	×	×	×	°

Not. Simgeler ve açıklamaları: ° = güvenilirlik tahmin değerine hata varyansı dahil edilmiştir, × = hata varyansı dahil edilmemiştir, ◊ = güvenilirlik tahmin değeri hata varyansını kısmen içerir.

Kaynak. National Chung-Cheng University, "Reliability [Güvenilirlik]," <http://psy.ccu.edu.tw/testroom/ Reliability.doc> (09.10.2002).

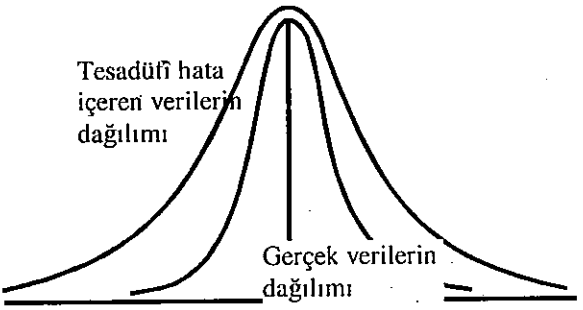
Ölçümü etkileyen faktörleri daha geniş bir biçimde tasarımla ilgili faktörler, ölçüm aracıyla ilgili faktörler, ortamla ilgili faktörler, kişilerle ilgili faktörler ve uygulama süreciyle ilgili faktörler başlıkları altında inceleyebiliriz.

Tasarımla İlgili Faktörler

Klâsik test kuramında güvenilirliği etkileyen faktörlerin başında araştırma tasarımı gelir. Araştırmanın tasarımında ise, örneklem hatası ve örneklem büyüklüğü önemli bir rol oynar.

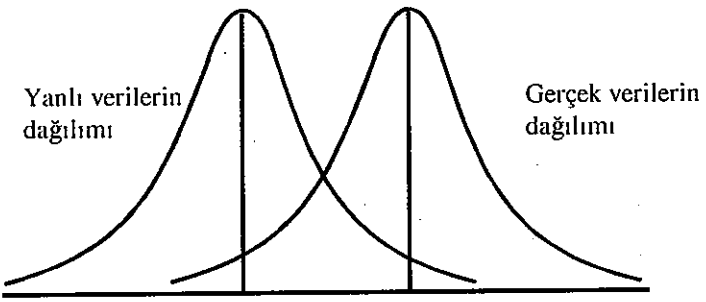
Örneklem hatası. Örnekleme ait sonuçların ana kütleyi temsil etme konusundaki yanılma payını gösterir. Örneklem hatası, ancak ana kütleden çok sayıda örneklem seçilmesi ve bu örneklemlerde ölçüm yapılması halinde tam olarak belli olabilir. Örneklem hatası %3 veya %5 gibi belirli oranlarla ifade edilir. Örneklem hatasının iki temel kaynağı vardır: tesadüfî hata ve sistematik hata. Tesadüfî hata, bilim adamının kontrol etme konu-

sunda güçlük çektiği etkenlerden kaynaklanır ve örneklem sonuçlarıyla gerçek sonuçlar arasındaki farklılığı belirtir. Tesadüfî hata içeren verilerin aritmetik ortalama değeriyle hata içermeyen gerçek verilerin aritmetik ortalama değerleri arasında genelde önemli bir farklılık yoktur fakat her iki veri dizisinin dağılımı veya varyansı farklıdır (*bk.*, Şekil 13-1). Sistematik hatada ise gerçek verilerin aritmetik ortalaması hata içeren verilerin ortalamasından farklı bir konumdadır (*bk.*, Şekil 13-2). Sistematik hata daha çok, araştırmanın tasarımından kaynaklanır. Örneklemin yanlı olarak seçilmesi veya örneklemin homojen bir niteliğe sahip olması sistematik hataya yol açar.



Şekil 13-1. Tesadüfî hata içeren verilerin dağılımı.

Kaynak. B. Trochim, "Measurement Error [Ölçüm Hataları]," <http://trochim.human.cornell.edu/lectureshows/1> (25.01.2004).



Şekil 13-2. Sistematik hata içeren verilerin dağılımı.

Kaynak. B. Trochim, "Measurement Error [Ölçüm Hataları]," <http://trochim.human.cornell.edu/lectureshows/1> (25.01.2004).

Medya kuruluşlarında "hata marjı" veya "hata payı" adı da verilen örneklem hatası değişik nedenlerle ortaya çıkar. Bunlardan birincisi *örneklem çerçevesinin* doğru bir şekilde belirlenmemiş olmasıdır. Ana kütlenin daraltılmış biçimi olan örneklem çerçevesi, temsil edici bir şekilde belirlenmediği zaman seçilen örnekleme ait veriler yanlı bir özellik gösterir ve sonuçta güvenilirlik rakamları düşük çıkar. Örneklem çerçevesinde her birey eşit seçilme şansına sahip olmalıdır. Örneğin, bir araştırmacı tekstil sanayinde çalışan işçilerin işe yabancılaşmalarını belirlemeye yönelik bir ölçek geliştirmeye çalışırken zaman yetersizliği ve izin alma konusundaki güçlükler nedeniyle pilot araştırmasını sadece tekstil sanayindeki tek bir holdinge ait üç firmanın işçileri üzerinde yapmıştır. Seçilen örneklem, temsil edici bir örneklem çerçevesine dayandırılmadığından sonuçlar belli bir yöne eğilimli olarak çıkmış ve sonuçta ölçeğin alfa güvenilirlik rakamları düşük elde edilmiştir. Buradaki sorun, ölçeğin kendisinden değil, örnekleme yöntemi ve örneklem çerçevesinden kaynaklanmıştır. Ölçüm, testlerde norm geliştirmeye yönelik olarak yapılıyorsa norm grubunun yaş, deneyim ve eğitim gibi faktörler açısından uygulanması düşünülen hedef kitleyi temsil etmesi gerekir.

Örneklem hatasının oluşma nedenlerinden ikincisi ana kütlenin veya evrenin açık ve doğru bir şekilde tanımlanmamasıdır. Bazen araştırmacılar ana kütleyi ölçüm çalışması süreci içinde belirlemeye çalışırlar. Böyle olunca, belirsiz bir ana kütle yaklaşımından hareket edilerek belirli bir ana kütle çerçevesi oluşturulmaya çalışılır. Örneğin, bir araştırmada geliştirilen ölçek önceleri sadece işçilere uygulanırken daha sonra yöneticilere de uygulanmasına karar verilmiş ve örnek kütlenin tanımlanmasında bulguların bir kısmı işçilere ve diğer kısmı yöneticilere yönelik olarak gerçekleştirilmiştir.

Üçüncüsü, seçim hatasıdır. Araştırmaya katılan kişilerin gönüllüler arasından seçilmesi, yöneticinin direktif vermesi nedeniyle bazı kişilerin araştırmaya zorla katılmaları, anketlerin kolay cevap alınacak birimlerde uygulanması seçim hatasını ortaya çıkarır. Bu tür uygulamalarda örneklem ana kütleyi temsil etme özelliğine sahip değildir. Örneklem verileri ana kütle parametrelerini tam olarak yansıtmadığından bulguların ve dolayısıyla sonuçların güvenilirliği kuşkuludur.

Ana kütle açık bir şekilde tanımlanmamışsa ve buna bağlı olarak örneklem çerçevesi sağlıklı bir şekilde belirlenmemişse, araştırmada temsil edici bir örnekleme çalışması yapılmamışsa bilim adamının "hata marjının ±%2 gibi düşük bir değer olduğu" iddiası anlamsızdır. Çünkü, elde edilen sonuçlar ana kütle parametrelerini sağlıklı bir şekilde yansıtmaz. Ana kütle-

nin tamamında yapılacak ölçümlerde sonuçlar ±%2'den daha yüksek çıkar. Sağlıklı bir güvenilirlik analizi yapabilmek için; (a) hata marjı kabul edilebilir bir örneklem büyüklüğüne ulaşılmalı, (b) katılımcılar tesadüfi örnekleme yöntemine göre belirlenmeli ve (c) örneklem birimleri ana kütleyi temsil edecek şekilde saptanmalıdır. Elde edilen hata marjının ancak bu koşullarda bir anlamı olabilir. Örnekleme hatasının düşük olduğu bu tür araştırmalarda verilerin standart hatası da düşük çıkar. Örneklem hatasının oranını tespit etmek her zaman mümkün olmayabilir, ancak bilim adamı hata marjını belirlemeye yönelik bir çaba içinde olmalıdır. Alan araştırmalarında hata marjı, ana kütle varyansı ve seçilen örneklemin büyüklüğünden hareket edilerek hesaplanır (*bk.*, Eşitlik 13-1).

- Hata marjı için basit bir formül.

$$HM = \frac{\sqrt{n}}{n}. \qquad (13\text{-}1)$$

HM = Hata marjı.
n = Örneklem büyüklüğü.

Örneğin 380 kişilik bir öğrenci grubu üzerinde yapılan araştırmada öğrencilerin %70'i üniversite kütüphanelerinin "açık raf sistemiyle" çalışmasından yana olduklarını bildirmişler, %30'u ise "kapalı raf sistemini" savunmuşlardır. Hesaplama sonucunda hata marjı ,05 çıktığından gerçek dağılım, %70 ±%5 ve %30 ±%5 olarak gerçekleşir. Hata payı *güven aralığında* olduğu gibi belirli bir güvenilirlik düzeyinde ana kütle oranının hangi değerler arasında değişebileceğini gösterir. Sosyal ve davranışsal bilimlerde genelde %95 güvenilirlik düzeyinde çalışıldığından hata marjı da çoğunlukla bu çerçevede hesaplanır (*bk.*, Eşitlik 13-2).

- Hata marjı için ayrıntılı formül.

$$hm = z\sqrt{\frac{p(1-p)}{n-1}}. \qquad (13\text{-}2)$$

hm = Hata marjı.
z = %95 güvenilirlik düzeyinde z değeri (1,96).
n = Örneklem büyüklüğü.
p = Ölçülmek istenen özelliğin örnek kütlede saptanan oranı.

Örneklem büyüklüğü. Bir ölçümde örneklem büyüklüğü arttıkça hata marjı azalır. Ancak örneklem hatası doğrusal bir azalış göstermez. Örneklemdeki kişi sayısı arttıkça hata marjı çok daha az oranda azalır. Sosyal içerikli alan araştırmalarında genellikle 1000-1500 cevaplayıcıyla çalışılır. Çünkü bu büyüklüklerdeki hata marjı %3,1 ilâ %2,5 arasında değişir (*bk.*, Tablo 13-2). İronik bir gerçek olarak vurgulamak gerekir ki, örnek kütle büyüklüğünün artması her zaman araştırma sonuçlarının daha doğru olacağı anlamına gelmez. Dikkatli bir şekilde planlanmış, temsil edici bir örnekleme sürecine dayandırılmamışsa tek başına örneklem büyüklüğü sağlıklı bir sonuç vermez. Ayrıca temsil edicilik özelliği sağlanmış olsa bile bu kez *örneklem dışı hataların* artma olasılığının bulunması nedeniyle "tam sayıma" dayanan ölçümler yerine "örnekleme yöntemi" daha güvenilir veya daha doğru sonuçlar verir. Çünkü, *tam sayımda* örneklem hatası giderilirken örneklem dışı hata oranı artar. Tam sayım araştırmalarında sistematik hata yapma olasılığı sıfırlanırken tesadüfî hata yapma olasılığı büyük ölçüde artar.

Ölçümlerde belirlenebilecek minimum örneklem büyüklüğü 3'tür, çünkü 2 kişiden oluşan örneklem büyüklüğünde kötü ölçüm, iyi ölçümden ayırt edilemez.[1] Bir çalışmada 30 veya daha az sayıda kişiden oluşan ölçüm grupları "küçük örneklem" olarak adlandırılır. Otuz ilâ 100 arasındaki örneklem büyüklükleri "orta büyüklükteki" örneklem grubu ve 100'den büyük olanlar ise kaba ölçülerle "büyük örneklem" olarak nitelendirilir.

Tablo 13-2. Örneklem Büyüklüğü İle Örneklem Hatası Arasındaki İlişkiler

Örneklem büyüklüğü	%95 güven aralığında örneklem hatası
10	± %31,0
50	± %13,9
100	± %9,8
200	± %6,9
1000	± %3,1

Kaynak. G.T. Fong, "The Relationship Between Sample Size and Sampling Error [Örneklem büyüklüğü ve Örneklem Hatası Arasındaki İlişkiler]," <http://www.arts.uwaterloo.ca/~gfong/psych101/samperr.html>

Tablo 13-2'den de görüleceği gibi örneklem büyüklüğünün artmasıyla birlikte %95 güvenilirlik düzeyinde hata oranları önemli ölçüde azalmak-

tadır. Belirli bir hata payında örneklem büyüklüğünü belirlemek isteyen araştırmacılar bunun için geliştirilmiş bulunan formüllerden yararlanırlar.

Bilim adamı, modern ölçüm kuramlarından madde-yanıt modelini temel almışsa bu kez örneklem hatası formülleri yerine seçilen model için araştırmalarla belirlenmiş örneklem büyüklükleriyle çalışmalıdır. MYK modellerinde örneklem büyüklüğünün araştırılan gizli değişkenin ana kütlede normal dağılım özelliği gösterecek bir büyüklükte olması gerekir. Normal dağılım özelliği gösteren ana kütlede gizli değişken, 0 aritmetik ortalama ve 1 standart sapma değerine sahiptir. Baker (1998) madde-yanıt modeliyle ilgili olarak yaptığı simülasyon çalışmalarının sonucunda 50 madde ve 500 katılımcıyla BILOG isimli yazılımda parametre özelliklerinin mükemmel bir şekilde belirlenebileceğini göstermiştir (aktaran, García-Pérez, 1999).[2] Rasch modelinde ise gizli değişkene ait parametre özelliklerinin dağılımı hakkında herhangi bir tahmin yapılmaz ve bu konuda bir varsayım da ileri sürülmez. Maksimum olasılık hesaplamasında, gözlem hatalarının beklenen değerler çerçevesinde aşağı yukarı normal dağılacağı varsayılır.[3] Bu nedenle Rasch modelinde 100 gibi çok daha küçük örnek kütle hacimleriyle çalışılabilir.

Ölçüm Aracıyla İlgili Faktörler

Güvenilirliği etkileyen faktörlerden bir diğeri kullanılan ölçüm aracıyla ilgilidir. Bu kitapta ele alınan klâsik veya modern kalibrasyon çalışmalarından geçirilmemişse hataların önemli bir bölümü ölçüm aracının kendisinden kaynaklanır. Güvenilirliği etkileyen ölçüm aracıyla ilgili etkenler; testin gereğinden fazla uzun olması, test maddelerindeki ifadelendirme yanlışları, hileli sorular ve test maddelerinin homojen olmaması gibi konuları kapsar.

Testin/ölçeğin gereğinden fazla uzun olması. Bir ölçek teorik olarak ne kadar uzun ise, ölçekte ne kadar fazla madde varsa güvenilirliği o kadar artar. Ancak bunun anlamı, güvenilirliği artırmak için yüzlerce maddeden meydana gelen bir test/ölçek geliştirmek değildir. İlave maddeler ancak orijinal maddeler kadar kaliteli ve iyi ise güvenilirlik artar, aksi halde düşük kaliteli maddeler güvenilirliğin düşmesine neden olur. Klâsik test kuramına göre bir ölçekte veya testte arzu edilen güvenilirlik düzeyinde kaç tane madde olması gerektiğine Spearman-Brown formülü kullanılarak karar verilir. Gereğinden fazla uzun olan bir test, cevaplayıcıda yorgunluğa ve tesadüfî hatanın artmasına neden olur. Güvenilirlik nedeniyle bir test gereğinden fazla madde içeriyor ve bu nedenle yanıtlama süresi uzuyorsa

böyle bir durumda testteki madde sayısını artırmak yerine güvenilirlikten taviz vererek kısaltma yoluna başvurmak daha uygundur.[4]

Modern test kuramlarından madde-yanıt modelinde ise *uzunluk* güvenilirliğin garantisi değildir. Bir test veya ölçek, hedeflenen *test bilgi fonksiyonunu* çok daha az madde ile de sağlayabilir. Bunun için madde-yanıt modelinde hedeflenen test bilgi fonksiyonunu sağlayacak mümkün olduğu kadar az sayıda madde belirlenmeye çalışılır.

Maddelerin ifadelendirme yanlılıkları. Ölçek/test maddelerinin ifadelendirilme biçimi veya maddelerin anlaşılırlığı ölçeğin güvenilirliğini etkiler. Test maddeleri belirsiz veya anlaşılmaz ise, yanıtlara ilişkin bazı ip uçlarını içeriyorsa, yanlı bir şekilde yazılmışsa, cevaplayıcıları gizli bir şekilde yönlendirmeyi hedeflemişse güvenilirlik düşük çıkar. Madde yanlılığı cinsiyet ayrımcılığı nedeniyle; kültürel, etnik veya dinî ayrımcılık nedeniyle ya da maddenin belirli bir sınıfın özelliklerini yansıtması nedeniyle ortaya çıkabilir. Madde, ülke genelinde yaygın olarak kullanılmayan sözcükleri veya teknik uzmanların anlayabileceği bazı özel kelimeleri içeriyorsa "dil yanlılığına", genel kültüre özgü olmayan uygulamaları veya anlamları içeriyorsa "kültürel yanlılığa" sahiptir. Maddelerin gruplara özgü yanlılığını saptamak için diferansiyel madde analizi ve diğer klâsik madde analizleri yapılarak iyi ve kötü maddeler açığa çıkarılır. Maddeler yazılırken basit bir dil kullanılarak belirtilmek istenen görüş direkt bir şekilde ifade edilir.

Maddelerin kavramsal yapıyla doğrudan ilgili olup olmadığı ifadelendirme biçimine dayanır. Maddelerin kavramsal yapısıyla ilgisi faktör analizi yapılarak belirlenir. Maddeler anlamları, içeriği ve kapsamı itibariyle kavramsal yapıyla ilgili olmalıdır. Bilim adamının bir maddenin kavramsal yapıyla ilgili olduğunu düşünmesiyle uygulayıcıların algılamaları farklı olabilir. Faktörü temsil etme oranı düşük olan maddeler belirsizdir. Bunların ölçekten çıkarılması veya üzerinde yeniden düşünülerek iyileştirme yapılması gerekir.

Okullarda uygulanan çoktan seçmeli testlerde "Hangisi değildir?" örneğinde olduğu gibi negatif cümle yapılarının kullanılması karışıklığa neden olur ve sonuçta hata sayısının artmasıyla birlikte testin güvenilirliği düşer. Çoktan seçmeli testlerde, madde kökleri pozitif cümle yapılarına göre yazılmalıdır.[5] Yine, çoktan seçmeli sorularda cevap şıkları mümkün olduğunca birbirine benzer uzunlukta olmalı, bazı şıkların çok uzun ve bazılarının bir iki kelime olduğu düzenlemelerden kaçınılmalıdır. Cevap şıklarında; "asla", "her zaman", "yukarıdakilerin hepsi", "yukarıdakilerin hiç birisi değil" gibi bilgi niteliği taşımayan tanımlamalar kullanılmamalıdır.

Soru kökü, devrik cümle yapıları kullanılarak yazılmamalı ve cümle uzunluğu yedi kelimeden kısa, 17 kelimeden uzun olmamalıdır.

Araştırmacı eğer Likert ölçeği veya Likert tipi bir ölçek geliştiriyorsa ifadelerin yazımında belirli kurallara dikkat etmesi gerekir. Bu kurallardan bazıları aşağıdaki gibidir:

1. İfadeler gereğinden fazla uzun ve aynı zamanda gereğinden daha kısa olmamalı, fikri özlü bir biçimde ve tam olarak ortaya koymalıdır. Üç dört satır uzunluğundaki ifadeler okuyucuları usandırır.
2. Bir maddede sadece tek bir görüş dile getirilmelidir.
3. Ölçekteki ifadelerin yaklaşık yarısı olumlu ve diğer yarısı ise olumsuz içeriğe sahip bulunmalıdır.
4. Olumlu ve olumsuz ifadeler bir olumlu ve bir olumsuz şeklinde sıralanarak değil, ölçek içinde rasgele dağılmış olmalıdır.
5. Çok boyutlu ölçeklerde yine ifadeler rasgele dağılmış olmalıdır. İfadeler okuyuculara belirli başlıklar altında sunulmamalıdır.
6. Olumsuz içerikli ifadeler yazılırken "işimi sevmiyorum", "işimden memnun değilim" şeklinde olumsuz yüklemli cümlelerden değil "işimden nefret ediyorum" veya "işimi sıkıcı buluyorum" şeklindeki olumlu yükleme sahip, fakat içeriği olumsuz olan cümlelerden yararlanılmalıdır.
7. İfadelerde argo kelimelerden, klişe sözcüklerden, teknik dilden ve kişileri rahatsız edebilecek kelimelerin kullanımından kaçınılmalıdır.
8. İfadeler konuya, ölçüm alanına, kavramsal alana odaklanmış olmalı; değişik konulara dağılmamalıdır.
9. İfadeler birbirini tekrarlayan nitelikte veya büyük ölçüde aynı anlama gelecek kelimelerden oluşturulmamalıdır.
10. İfadeler basit bir dille ortaya konmalı, edebî, estetik veya sanatsal yazım biçiminden sakınılmalıdır.
11. İfadelerde yazım (tapaj) hatası bulunmamalıdır.
12. İfadelerde; *genellikle, çoğunlukla, daha çok, zaman zaman, neredeyse* gibi konuşma dilinde kullanılan sıklık tanımlamalarından kaçınılmalıdır.

13. "Teknik konularda cahil halkın değil, ilgili uzmanların görüşlerine önem veririm" örneğinde olduğu gibi ifadelerde *yönlendirici, mesaj yüklü veya şartlandırıcı* görüşlerin sunulmasından kaçınılmalıdır.
14. İfadelerin kendilerine anket uygulanacak tüm katılımcılar için geçerli olmasına dikkat edilmelidir.
15. İfadeler; cinsiyet, etnik köken, ırk, dinî inanış, mezhep veya ideoloji temelinde kişileri sorgulayan bir içeriğe sahip olmamalı ve kişiler söz konusu ifadelerden dolayı rahatsızlık hissetmemelidirler. Araştırmacı kültürel farklılıklar konusunda duyarlı olmalıdır.
16. İfadeler, ölçeğin etiketleri göz önünde bulundurularak (beyan edici, sıklığı gösterici, değerlendirici vb. gibi) yazılmalıdır.
17. Toplam ifade sayısı mümkün olduğunca 8, 10, 12, 18, 20 ve 22 gibi çift rakamlı olarak belirlenmelidir.
18. İfadeler arasında 12 nokta boşluk bırakılmalı ve birbirlerinden net olarak ayrılmaları sağlanmalıdır.
19. İfadelerin etiketleri ve dereceleme puanları ölçeğin sağ tarafında ayrılacak bölümde gösterilmelidir.
20. İfadeler tablo içine yazılmış olsa bile tablo çizgileri gizlenmeli, ölçek temiz ve net bir görünüme sahip olmalıdır.
21. İfadelerin yazımında vurgu yapmak için koyu siyah yazım biçiminden kaçınılmalıdır.
22. İfadeler değişik nedenlerle küçük veya büyük punto büyüklükleriyle değil, 10-12 punto büyüklüğüyle yazılmalıdır.
23. İfadelerin yazımında anket formunun dört bir tarafından en az 2,5 cm boşluk bırakılmalıdır.

Test maddelerinin homojen olmaması. Test maddelerinin karmaşık konularla ilgili olması testin farklı özellikleri ölçtüğü anlamına gelir. Heterojen maddeler testin güvenilirliğini düşürür, bu maddeler çoğunlukla tutarsızdır. Maddeler homojen olursa testin güvenilirliği daha yüksek çıkar. Örneğin *Sözcük Bilgisi Testi*'nin güvenilirliği; sözcük bilgisinin yanında dil bilgisi ve paragraf bilgisini de ölçen "karma bir testten" daha güvenilirdir. Test maddelerinin türdeşliği maddeler arası tutarlılık, yarıya bölme yöntemi ve paralel formlar yöntemiyle sınanırken esas olarak bütün maddelerin aynı özelliğe birlikte "asılıp asılmadığına" bakılır. Basit kavramsal

yapıların ölçümünde maddelerin homojen bir şekilde belirlenmesi nispeten kolayken karmaşık kavramsal yapılarda homojen nitelikli madde oluşturmak oldukça zordur. Bu nedenle karmaşık kavramsal yapılarda alt ölçekler veya testçikler oluşturularak çalışılır.

Güçlük derecelerinin uygun bir dağılıma sahip olmaması. Ölçüm bir başarı testi şeklindeyse tüm maddelerin çok zor veya çok kolay olması, güvenilirliği düşürür. Orta derecede zorluğa sahip maddeler ise güvenilirliği arttırır. Maddelerin zorluk derecesi için ,15-,85; ,20-,80 veya ,30-,70 değişim aralıkları önerilmiştir. Aralığın genişliği veya darlığı sorulardaki seçenek sayısına göre değişir. Seçenek sayısı arttığı ölçüde daha geniş bir erim; daraldığı ölçüde daha dar bir erim belirlenir. Bunun yanında maddelerin zorluk derecelerinin dağılımı araştırmanın amacına göre saptanır. Bilim adamının amacı testi alan kişilerin yetkinlik / yeterlilik düzeylerini belirlemek ise güçlük indeksinin ,35'in altında olması hedeflenir. Tam tersine testin amacı kişiler arasındaki farkı ortaya çıkarmak veya kişileri farklılaştırmak ise, maddelerin güçlük indeks değerlerinin ,30 ilâ ,70 aralığında normal dağılım gösterecek şekilde yayılması arzulanır. Sınıf ortamında yapılan başarı testlerinde güçlük indeks değerlerinin geniş bir dağılıma sahip olması önemlidir. Ancak yüksek başarıyı hedefleyen okullar güçlük indeks değerlerini dar bir erimde belirlemiş olabilirler. Başarı testlerinde maddeler belirli bir bölgede veya noktada toplanmamalı değişik yetenek düzeylerini ortaya çıkaracak şekilde geniş bir erime sahip bulunmalıdır. Bir test, güçlük derecesi sadece yüksek olan maddelerden oluşturulursa kişilerde hayal kırıklığı yaratırken çok kolay maddelerden oluşturulduğu durumda ise can sıkıntısına neden olur. Heterojen dağılıma sahip bir sınıfta test maddelerinin %60'nın ,30-,50 güçlük indeks değerine sahip olmasının testin gücünü arttıracağı belirtilmiştir.[6] Test uygulanacak sınıf veya kişiler homojen nitelikte ise bu kez güçlük indeks değeri geniş bir ranj aralığı yerine kısıtlanarak belirlenir. Maddelerin güçlük indeks değerlerinin şans düzeyinin altına düşmesi istenmez. Şans düzeyi dört şıklı sorularda ,25 ve beş şıklı sorularda ise ,20'dir. Çoktan seçmeli testlerde maddelerin güçlük indeks değerlerinin aritmetik ortalamasının ,60 ilâ ,80 arasında olması gerektiği belirtilmiştir. Doğru-yanlış şeklinde cevaplandırılan testlerde ise ortalama güçlük indeks değeri için ,75 oranı önerilmiştir. Optimum güçlük düzeyini belirlemek için Eşitlik 13-3'deki formül kullanılır.

$$OGD = ,5 + ,5\left(\frac{1}{a}\right). \qquad (13\text{-}3)$$

OGD = Optimum güçlük düzeyi.
a = Seçenek sayısı.

Değişik maddelerin güçlük indeks değerlerinin aritmetik ortalaması alınarak bulunacak bu değerlerde düşük rakam daha zor bir testi, yüksek rakam ise nispeten daha kolay olan bir testi tanımlar.[7] Güçlük indeks değeri ancak büyük örneklemlerle çalışıldığı zaman anlamlıdır. Bir okulun 25 kişilik sınıfında öğrencilere uygulanan test sonuçlarına bağlı olarak elde edilen maddelerin güçlük indeks değerleri büyük ölçüde çarpıktır ve gerçeği yansıtmaz.[8] Güçlük indeks değerinin gerçeğe daha yakın olması için bu değer, aynı sınıftaki tüm şubelerin öğrencileri birleştirilerek hesaplanmalı ve örneklem sayısı en azından 100'ün üzerinde olmalıdır.

Soruların hileli olarak hazırlanması. Hileli sorular, yazılı olarak ifade edilen başarı testleriyle ilgilidir. Ölçeklerde ve psikometrik testlerde hileli değil, yetersiz soru olur. Stanley ve Hopkins (1990) hileli maddeleri yanıtın, kolaylıkla gözden kaçırılabileceği önemsiz bir kelimeye bağlı olması veya olumsuz anlamın/yüklemin kafa karıştıracak bir şekilde kullanılması olarak tanımlamışlardır (aktaran Roberts, 2003).[9] Yine Thorndike, Cunningham ve Hagen (1991) hileli soruyu; "cevabı maddenin içeriğiyle hayatî derecede ilişkili olmayan" biçiminde ele almışlardır (aktaran Roberts, 2003).[10]

Hileli sorular yanıtlayıcıları tuzağa düşürerek daha fazla yanlış yapmalarına neden olur. Bu tür sorularda sadece zayıf öğrenciler değil iyi öğrenciler de yanılırlar. Sorularda harf, rakam, kelime düzenlemelerine dayanan ve bilginin içeriyle ilgili olmayan yanıtlar hileli soru olarak nitelendirilir. Araştırmacılar maddeleri mümkün olduğu kadar açık, birden fazla anlama gelmeyecek şekilde ve çözülebilecek tarzda hazırlamalıdırlar.

Maddelere ağırlık verilmesi. Çoktan seçmeli başarı testlerinde maddelere ağırlık verilmek istenmesinin nedeni testin ayırt edicilik özelliğini artırmaktır. Ancak yapılan araştırmalarda ağırlık verme uygulamasının testin güvenilirlik ve geçerliliğini arttırmadığı bulunmuştur. Maddelere ağırlık verme sadece kompozisyon şeklinde cevaplandırılan kısa testlerde

istenen sonucu sağlayabilir. Çoktan seçmeli testlerde maddelere sadece 0 ve 1 şeklinde ağırlık verme ideal bir uygulama olarak görülmüştür.

Ortamla İlgili Faktörler

Ölçüm bir başarı testi (bilgi veya yetenek) şeklindeyse test uygulanan ortam testin güvenilirliğini etkiler. Odanın ısısı, ışıklandırması, gürültü seviyesi, yazı yazılan ortam, test odasında başkalarının bulunması, gözlemcilerin test alan kişilerin dikkatini dağıtacak ölçüde fısıldaşarak konuşmaları test puanlarında hata varyansını artırır. Ortam, sadece başarı testlerinin değil, tutum ölçeklerinin yanıtları üzerinde de etkili olur.

Kişilerle İlgili Faktörler

Ölçüm sonuçları; soruları tasarlayan kişilerden, soruları soran kişilerin konumlarından, soruları/ifadeleri yanıtlayan kişilerin sosyal ve psikolojik konumlarından etkilenebilir. Birincisine tasarımcı etkisi, ikincisine mülakatçı etkisi ve üçüncüsüne ise yanıtlayıcı etkisi adı verilir.

Tasarımcı etkisi. Ölçümü yapan bilim adamının veya araştırmacının önyargılı olarak soruları, ifadeleri veya ölçüm maddelerini belirli bir görüşü empoze edecek şekilde oluşturmasıdır. Bu tür ifadeler "yönlendirici soru" olarak nitelendirilir. Bilim adamı yanıtlayıcıları "ikna etme" gibi bir çaba içinde olmamalıdır. Araştırmacının dünya görüşü, ideolojisi, dinî inançları ve yaşam biçimi sadece kendisine aittir. Bilimsel ölçüm; yansız, tarafsız ve nesnel olmayı gerektirir. Bilim adamı ne kendisini ne de başkalarını aldatacak veya yanıltacak bir tutum içinde olamaz.

Mülakatçı/anketçi/gözlemci etkisi. Bir araştırmada eğer anketçilerden yararlanılmışsa bu kişilerin standart uygulama yapabilmeleri için dikkatli bir şekilde eğitilmeleri ve kendilerine lüzumu halinde başvuracakları "uygulama yönergelerinin / talimatlarının" verilmesi gerekir. İki mülakatçının / anketçinin tutum ve tavırları aynı olmayabilir. Bu nedenle bazen iki farklı mülakatçı aynı kişiyle mülakat yaptıklarında farklı sonuçlar alırlar. Bu durum verilerin güvenilirliğini düşürür. Soruların sorulma şekli, anketçinin yaklaşım biçimi kişilerin verecekleri yanıtları etkiler. Kişiler genelde rizikolu gördükleri belirli nitelikteki sorulara yanıltıcı cevap verme eğilimi içindedirler. Ölçüm uygulamasından önce anketçiler kişilerin muhtemel tepkileri ve soruları konusunda eğitilerek takınacakları tutum konusunda bilinçlendirilmelidirler.[11]

Mülakatçı veya anketçi soruların soruş biçiminde önyargılı veya yönlendirici bir yaklaşıma sahip olmamalıdır. Araştırma konusuyla ilgili ol-

madığı sürece ölçüm yaptığı kişinin sosyoekonomik durumunu, değerlerini ve yaşam biçimini sorgulamamalı sadece ölçüm konusuyla ilgilenmelidir.

Bir araştırmada eğer gözlemcilerden yararlanılmışsa bu kişilerin sayısının artmasıyla birlikte güvenilirliğin daha fazla etkilenmesi söz konusudur. Gözlemci sayısı arttıkça sübjektif faktörler işin içine daha fazla gireceğinden hata oranı artacaktır. Öğrencilerin sınav kağıtlarının belirli bir sayının üzerinde farklı öğretmenler tarafından değerlendirilmesi, bilim projelerinin çok sayıda hakem tarafından değerlendirilmesi ölçüm sonuçlarındaki hata oranını artırır.

Ölçüm bağımsız gözlemciler veya değerlendiriciler tarafından yapılıyorsa bu kişilerde görülebilecek yanlılık sonuçların güvenilirliğini etkiler. Örneğin, başarı değerlendirme uygulamalarında her bir personel en az iki üst tarafından belirli kriterler çerçevesinde değerlendirilir. Değerlendiricilere özel bir eğitim verilmemişse çeşitli ölçüm hatalarıyla karşılaşılır. Bunlardan birincisi *merkeze yönelme* eğilimidir. Yöneticiler personele orta derecedeki puanları verme eğilimi içinde olurlar. İkincisi, *hâle* etkisidir. *Hâle etkisi* yöneticilerin / değerlendiricilerin değerlendirdikleri kişi veya nesnenin bariz bir özelliğinden olumlu veya olumsuz biçimde etkilenmeleridir. Üçüncüsü *gevşek* değerlendirmedir. Bazı gözlemciler değerlendirmelerinde yüksek puan verme eğilimi içinde olurlar. Dördüncüsü, *katı* değerlendirme yaklaşımıdır ki kişilere düşük puanlar verme eğilimini yansıtır. Beşincisi, *yakınlık* hatasıdır. Ölçüm aracında birbirine benzer maddeler eğer alt alta sıralanmışsa değerlendiriciler bu maddelere benzer puanları verirler. Altıncısı, *gözlemcinin yanlılığından* kaynaklanan hatalardır. Yanlılık haksız bir şekilde destekleme veya haksız bir şekilde karşı olma anlamındadır. Bazen değerlendiriciler sevdikleri kişilere yüksek puanlar verme eğilimi içinde olurlarken sevmedikleri kişilere karşı ise ön yargılı davranıp düşük puanlar verirler. Gözlemcilerin eğitimle bilinçlendirilmediği durumlarda gözlemci hatalarıyla karşılaşmak kaçınılmazdır.

Yanıtlayıcı etkisi. Ölçüm sonuçlarını, yanıtlayıcılar da doğru olmayan cevaplar vererek etkileyebilirler.Yanıtlayıcıya müdahale anlamına geleceğinden bu tür hataları önlemek çok zordur. Yanıtlayıcılar değişik beklentileri, korkuları, endişeleri, kızgınlıkları ve içinde bulundukları ruh halleri nedeniyle gerçeği olduğundan farklı bir biçimde yansıtabilirler. Yanıtlayıcı etkisinden kaynaklanan hataları en alt düzeye düşürmek için ölçümün en uygun koşullarda ve gönüllü katılım ilkesine bağlı olarak yapılması gerekir. Zorlama, uygun olmayan koşullar ve ölçüm sonuçlarının ne şekilde kullanılacağının bilinmediği durumlarda kişiler gerçek düşünce ve tutumlarını gizleme eğilimi içinde olurlar. Yanıtlayıcı etkisi bazen mülakatçı

etkisiyle birlikte çalışır. Yanıtlayıcılar anket uygulayan veya ölçüm yapan kişinin yeterliliğinden, güvenilirliğinden kuşku duymuşlarsa ölçümü ciddiye almama eğilimi içinde olurlar. Yanıtlayıcı etkisini azaltmak için ölçüm yapılacak kişilere duruma göre birkaç gün önceden bilgi verilmeli, kişiler katılım konusunda kendilerini özgür hissetmeli ve kendilerine yeterince zaman tanınmalıdır.

Ölçüm amacının yanıtlayıcılara bildirilmesi konusu üzerinde bilim adamı dikkatle düşünmelidir. Bazen bu amacın bildirilmesi sonuçların güvenilirliğini düşürürken bazen de tam tersine daha doğru yanıtlar alınmasına yol açar. Test alan kişiler kendi kazançlarına veya zararlı çıkmalarına neden olabileceğini düşündükleri ölçüm çalışmalarında gerçek davranışlarını gizleme eğilimi içinde olurlar. Bu tür araştırmalarda ölçüm amacının soyut ve genel olarak ifade edilmesi daha uygundur. Ölçüm, eğer kişinin kendi durumuyla doğrudan ilgili değilse araştırma amacının açık bir şekilde açıklanması daha gerçekçi yanıtlar alınmasını sağlar.

Ölçüm yine bir başarı testi şeklinde ise, test alan kişilerin konumu / durumu test puanlarındaki hata varyansını artırabilir. Bunlar; test alan kişinin uykusuz olması, endişeli olması, almış olduğu ilaçların kendisini etkilemesi, duygusal olarak test sonuçlarına güvenmemesi ve bu nedenle testlere karşı alaysı bir tutum takınması gibi faktörlerdir.

Yanıtlayıcı etkisiyle ilgili bir diğer faktör, ölçülen özelliğin araştırma yapılan örnek kütlede homojen bir dağılıma sahip olmasıdır. Örneklemdeki kişiler "sigara içme" davranışı konusunda hep aynı görüşlere sahip iseler güvenilirlik rakamları düşük çıkar. Ölçülen özellik örnek kütlede kısmen heterojen ve farklı bir dağılıma sahipse güvenilirlik rakamları yüksek elde edilir.

Uygulama Sürecinde Ortaya Çıkan Faktörler

Ölçüm uygulamasının yapılış biçimi, ölçüm için ayrılan süre, ölçüme veya araştırmaya katılması gereken kişilerin tavırları da ölçüm sonuçları üzerinde etkili olur. Bilim adamı planlama aşamasında olduğu kadar yürütme aşamasında da hata faktörlerinin etkisini kontrol altında tutmaya çalışmalıdır.

Yetersiz katılım. Uygulama sürecinde ölçüm sonuçlarını etkileyen önemli faktörlerden biri yetersiz katılımdır. Örneklem grubu içinde oldukları halde kişilerin araştırmaya / ölçüme katılmak istememeleri sonuçları etkiler. Katılmama olgusu değişik nedenlerden dolayı ortaya çıkabilir:

1. Kişinin bulunamaması.
2. Kişinin yanıt vermeyi tehir etmesi.
3. Kişinin yanıt vermek istememesi.
4. Kişinin kısmen yanıt vermesi.
5. Kişinin anketi iade etmemesi.
6. Kişinin mazereti nedeniyle ölçüme katılamaması.

Yetersiz katılım örneklem hatasının yüksek çıkmasına, ana kütleyi temsil kabiliyetinin azalmasına ve dolayısıyla sonuçların güvenilirliğinin düşmesine neden olur.

Testin kısıtlanmış süre içinde uygulanması. Testin veya ölçeğin doldurulma ve cevaplandırılma hızı veya tahsis edilen süre testin güvenilirliğini etkiler. Bu nedenle *hız testlerinde* iç tutarlılık güvenilirliği yerine test-yeniden test güvenilirliği veya paralel formlar güvenilirliğine önem verilir. Hız testlerinde cevaplandırma süresi genelde pilot araştırma sırasında temsil edici örneklem üzerinde yapılan çalışmalara dayanır. Hız testlerinde hiç kimse testi zamanı içinde tamamlayamaz veya çok az kişi testi süresi içinde tamamlar. Hız testlerinde sınav/test kaygısı, test uygulama talimatlarının açık olmaması, test sonuçlarının kritik kararlara konu olması gibi faktörler verilerde büyük ölçüde gürültüye neden olur. Böyle bir ortamda test-yeniden test ve paralel form güvenilirlik çalışmaları farklı sonuçlar verir.

Grubun homojen veya heterojen olması. Konuyu klâsik test kuramı ve madde-yanıt kuramı açısından irdeleyebiliriz. Klâsik test kuramında ölçeğin veya testin uygulandığı grubun benzer özelliklere sahip olması test puanlarının benzer olması sonucunu doğurur. Homojen gruplarda bireyler belirli yanıtları verme eğilimi içinde olurlar. Örneğin, kişiler 7 dereceli bir ölçekte 4 ve 4 puanın altında hiç işaretleme yapmamış olabilirler. Bu nedenle homojen gruplarda standart hata yüksek çıkar. Heterojen gruplarda ise, standart hata gözlem puanları üzerinde çok daha az etkilidir. Yetenek, bilgi testlerinde, tutum ölçeklerinde ölçüm yapılan grubun heterojen olması gerekir. Örneğin, matematik yeteneği ölçülüyorsa sınıfın yetenek dağılımı açısından heterojen bir dağılım göstermesi gerekir ki elde edilen sonuçların güvenilirliğinden söz edilebilsin. Sadece kriter referanslı testlerde homojen gruplarla çalışılır.

Madde-yanıt kuramında ise, ölçümler örneklemden bağımsızdır. Özellikle yetenek, zeka, başarı ve bilgi ölçümlerinde maddeler başarı oranını değişik yetenek düzeylerine göre ortaya koyduğundan örneklemin homojen veya heterojen olması üzerinde durulmaz. Önemli olan ölçüm çalışmalarının oldukça büyük örneklemlerde yapılmış olması ve her bir maddenin "parametrelerinin", "madde özellikleri eğrisinin" ve "madde bilgi fonksiyonu"nun ortaya çıkarılmasıdır. Ancak küçük örneklemlerde çalışıldığı zaman grubun homojen veya heterojen olması sonuçları etkiler.

Şans/tahmin faktörünün etkisi. Başarı testleri değişik şekillerde uygulanır. Kimi uygulamalarda test alan kişilere belirli sayıdaki yanlış yanıtın doğru yanıtı iptal edeceği açıklaması yapılırken diğerlerinde "tahminen işaretleme olgusunu" önlemek için *düzeltme formülünün* uygulanacağı bildirilir.[a] Böylece kişiler tahmine dayalı işaretlemeyi çok daha dikkatli bir şekilde yapacaklar veya hiç bu yola başvurmayacaklardır. Ancak test alan grupta kimlerin tahminen işaretleme yaptığı tam olarak saptanamaz. Tahminde bulunarak yapılan işaretlemeler sonuçların güvenilirliğini olumsuz yönde etkilediği için sınırlı yararı olan *tahmin düzeltmesi* formülü kullanılır (*bk.*, Eşitlik 13-4). Klâsik test kuramında tesadüfen işaretleme olgusu madde-yanıt kuramında ele alındığı kadar önemsenmemiştir.

■ Tahmin düzeltmesi formülü.

$$TD = \frac{D - Y}{n - 1}. \qquad (13\text{-}4)$$

TD = Tahmin düzeltmesi.
D = Doğru cevapların sayısı.
Y = Yanlış cevapların sayısı.
n = Her bir maddedeki/sorudaki seçenek/cevap şıkkı sayısı.

---

[a] Test uygulaması sırasında kişilere düzeltme formülü uygulanacağının veya yanlış soruların doğru soruları gidereceğinin bildirilip bildirilmemesi tartışmalı bir konudur. Bazı bilim adamları bildirmeme davranışının etik olmadığını düşünürlerken diğerleri bildirmenin yönlendirmeye neden olacağı görüşündedirler. Normal koşullarda testi alan kişilere "yanıtını bilmediğiniz soruları tahmin ederek cevaplandırmaya çalışmayınız. Bir sorunun yanıtını hiç bilmiyorsanız o soruyu boş bırakınız. Eğer sorunun iki veya üç şıkkını elimine edebiliyorsanız soruyu boş bırakmak yerine cevaplandırmaya çalışınız" açıklaması yapılır.

Bu formülde boş / yanıtsız bırakılan sorular değerlendirmeye alınmamıştır. Çünkü yanıtsız bırakılan sorular yanlış olarak değerlendirilmez. Bilim adamı eğer yanıtsız bırakılan soruları da değerlendirmeye almayı düşünüyorsa bu kez Eşitlik 13-5'teki formülü kullanır.

$$TD = \frac{(Y + B)}{n} + D \,. \qquad (13\text{-}5)$$

Formülde B simgesi boş bırakılan soru sayısını gösterir. Sınıf ortamında yapılan testlerde bir maddeyi testi alan kişilerin %5'inden fazlası boş bırakmışsa, birden fazla şıkkı işaretlemişse veya hem boş bırakıp hem de birden fazla şıkkı işaretlemişse o madde teste alınmaz.[12] Yapılan araştırmalar tahmin düzeltmesi formülünün sınıf ortamında yapılan bilgi ve başarı testleri için çok uygun olmadığını göstermiştir. Çünkü test alan kişilerin tahminen işaretleme yapma davranışları benzer değildir. Bazıları daha çok tahmin yöntemine başvurmuşken diğerleri kendilerine verilen talimatlara tam olarak uymuş bulunabilirler. Bu nedenle son yıllarda test şirketleri düzeltme formülünü uygulamaktan vazgeçmeye başlamışlardır.

ÖLÇÜMÜN İYİLEŞTİRİLMESİ

Ölçümü iyileştirme çalışmaları seçilen modele göre *klâsik test kuramı* çerçevesinde veya *madde-yanıt kuramı* (MYK) çerçevesinde ele alınır. Bu bölümde önce klâsik test kuramına göre yapılabilecek iyileştirme çalışmaları üzerinde durulmuş ve daha sonra MYK çerçevesindeki madde kalibrasyon çalışmalarına değinilmiştir.

Klâsik Test Kuramına Göre İyileştirme

Klâsik test kuramında iyileştirme çalışmaları esas olarak "madde" üzerinde değil, "test" üzerinde odaklanır. Kuşkusuz maddeler iyileştirilmeden test de iyileştirilemez. Fakat maddelerin iyileştirilmesi temelde testin iyileştirilmesi amacına yöneliktir.

Madde ifadelerinin gözden geçirilmesi. Madde ifadelerinin gözden geçirilmesi başarı testleriyle ölçeklerde farklı şekillerde yapılır. Bilgi ve başarı testlerinde "açıklık" ve "anlaşılırlık" olgusu öne çıkarken tutum ölçeklerinde maddelerin kavramsal yapıyı temsil etmesi ve aynı zamanda *şişkin özgünlüğe* yol açmaması önemlidir. Ölçüm aracı bir tutum ölçeği,

bir indeks veya yanıt ölçeği niteliğinde ise öncelikle anlaşılırlık önem kazanır. Cümlelerin kısa olması, uygun kelimelerin seçilmesi, görüşlerin kurallı cümle yapısı içinde sunulması kavramayı ve algılamayı hızlandırır. Test maddelerinin yazımında teknik ifadelerin ve yabancı kelimelerin kullanılmasından kaçınılmalıdır. Madde içeriklerinin kişileri incitici, rahatsız edici, çirkin ifade ve kelimeler içermesi onların tepkisel yanıtlar vermelerine neden olur. Öte yandan maddelerin kendi içinde "çift yargı" veya "çift soru" içermemesi, cümlenin çift olumsuz ifadeye sahip olmaması gerekir. Soruların "... olmaması aşağıdakilerden hangisi değildir?" şeklinde çift olumsuzluğu içerecek şekilde yazılması cevaplayıcıları şaşırtacağından elde edilen sonuçların geçerlilik ve güvenilirliği zayıflar. Çoktan seçmeli sorularda madde kökünün de negatif cümle yapısı yerine pozitif cümle yapısıyla yazılması önerilmiştir (*bk.*, Tablo 13-3).

Tablo 13-3. Değişik Test Uygulamalarında İfadelerin Gözden Geçirilmesi

Bilgi, başarı, yetenek testleri	Tutum ölçekleri, kişilik ve ilgi envanterleri
• Açıklık	• Tek boyutla ilgili olması
• Anlaşılırlık	• Şişkin özgünlüğe yol açmaması
• Tek soru	• Kültürel uygunluk
• Çift olumsuzlama yapılmaması	• Tek yargı / düşünce
• Tuzak ifadelerden kaçınma	• Negatif ifadelerin açık olması

Tutum ölçekleri, olgusal ölçekler, kişilik ve ilgi envanterlerinde maddelerin sözel olarak ifade edilmesinde dikkat edilmesi gereken bir diğer kural maddenin mümkün olduğunca kısa olarak yazılmasıdır. Uzun cümlelerden oluşan ifadeler cevaplandırıcıları usandırır ve onların bazen belirtilmek istenen düşünceyi yanlış anlamalarına neden olur.[13]

Maddelerin ölçüm aracının niteliğine uygun bir biçimde sıralanması. Kullanılan ölçüm aracının güvenilirliğini arttırmak için soru veya ifadelerin sıralanış biçimine de dikkat etmek gerekir. Cevaplayıcıların yanlılığı bazen soruların sıralanış biçiminden kaynaklanır. Soruların/ifadelerin sıralanış biçimi testin niteliğine göre değişiklik gösterir.

Kağıt ve kalem araçları kullanılarak yapılan *başarı ve bilgi testlerinde* sorular kolaydan zora doğru sıralanır.[14] Böylece cevaplayıcılar kendilerine güven geliştirmiş olurlar. Yine cevaplayıcıların ilgisini çekecek sorulara

ilk sıralarda yer verilir.[a] Tutum ölçeklerinde, kişilik ve ilgi envanterlerinde alt boyutlar veya faktörler önem kazandığından bir boyuta ait maddeler arka arkaya gelmemeli karmaşık bir sıra içinde düzenlenmelidir. Ancak tek bir kavramsal boyutu ölçen ifadelerin karmaşık bir sıra içinde sunulmasıyla birden fazla kavramsal boyutu ölçen anketlere ait maddelerin hepsinin karmaşık bir sıra içinde sunulması aynı değildir. Çok boyutlu ve alt ölçek şeklinde düzenlenmiş anketlere ait maddeleri tek bir anket içinde karıştırmak doğru değildir. Tek maddeli yanıt ölçeklerinde ise sorular mantıksal bir sıraya sahip olmalı; soruların karmaşık, kişisel ve duygusal içerikli olanlarına anket formunun son bölümünde yer verilmelidir.

Test ve ölçeklerde maddeler sıralanırken bir sorunun yanıtının ondan sonra gelen sorunun cevabını etkilememesine dikkat edilir.

Pilot araştırma yapma. Güvenilirliği arttırmak için araştırmacı öncelikle bir pilot araştırma yapmalı ve bu araştırmanın sonucuna göre iyileştirilmesi gereken maddeleri belirlemelidir. Pilot araştırma süreci içinde hata varyansına neden olan faktörler araştırılır. Anketin anlaşılma durumu değerlendirilir ve bu aşamada uygulanacak yönteme, cevaplama biçimine, cevaplama süresine karar verilir. Pilot araştırma sırasında test maddelerinin anlaşılırlık, geçerlilik ve güvenilirlik analizleri yapıldığından maddelerin kalibrasyonu büyük ölçüde *pilot araştırma örneklemine* bağlıdır. Kalibrasyon örneklemiyle operasyon örneklemi birbirinden farklı ise test sonuçları da önemli ölçüde farklı çıkar. Örneğin, bir araştırmacı geliştirmiş olduğu *Örgütsel Stres* ölçeğinin kalibrasyon çalışmalarını üniversitedeki araştırma görevlileri üzerinde sınayarak yapmış ve sonuçta yüksek güvenilirlik katsayıları elde etmişti. Daha sonra bu ölçeği tekstil sektöründeki yöneticilerde uygulamak istiyordu. Esas araştırma uygulaması sırasında yöneticilere bu testi uyguladığında tam tersi bir durumla karşılaşmış ve güvenilirlik katsayıları düşük çıkmıştı. Pilot araştırma örneklemiyle asıl çalışma (operasyon) örnekleminin ayrı olması test sonuçlarının güvenilirliğini düşürmüştü.

Bilim adamı pilot araştırma sonunda yaptığı değişikliklerin istenen sonucu verip vermediğini görmek için ikinci bir pilot araştırma daha yapabilir. Bilimsel araştırma amacıyla geliştirilen ölçek ve testlerde ikinci pilot araştırma her zaman mümkün olmasa da ticarî amaçla geliştirilen *yüksek ödüllü test* ve ölçeklerde kişilerin yaşamları, kariyerleri veya meslekî ge-

[a] Bilgisayarlı test uygulamalarında bu ilkeye uyulmayabilir. Bilgisayarlı uygulamalarda ilk soru orta zorluktadır. Doğru yanıt verildikçe sorular zorlaşırken yanlış soru verildikçe bu kez daha kolay sorular sorulur.

lişmeleriyle ilgili kritik kararlar söz konusu olacağından sağlıklı sonuçlar alınıncaya kadar pilot araştırma çalışmalarına devam edilir.

Örneklemdeki cevaplayıcıların dağılımı. Örneklemdeki cevaplayıcılar belirli ölçüde heterojen bir dağılıma sahip olduğu oranda ölçeğin/testin güvenilirliği artar. Katılımcılar türdeş olduğu ölçüde güvenilirlik azalır. Bir araştırmacı geliştirmiş olduğu *Pozitif Davranış İndeksi* isimli ölçeğini sadece bir bankanın genel merkezinde çalışan personele uygularsa güvenilirlik katsayısı muhtemelen yüksek çıkmayacaktır. Çünkü bir kurumdaki çalışanların büyük bir kısmı tavır takınma açısından birbirlerinden etkilenmişlerdir. Cevaplayıcıların heterojen bir dağılıma sahip olması için pilot araştırma da dahil olmak üzere örneklem hacminin belirli bir büyüklüğe sahip olması ve ana kütleyi temsil etmesi gerekir. Asgari örneklem büyüklüğü için 1:5, 1:6, 1:10 gibi *madde sayısı – kişi oranları* belirlenmiştir. Hem pilot araştırma ve hem de operasyon örnekleminde bu oranların takip edilmesi ölçümün güvenilirliğini artırır.

Örneklemdeki cevaplayıcıların dağılımını etkileyen bir diğer faktör örnekleme yöntemidir. Tesadüfî örnekleme yöntemine göre *küme örneklemesi* yöntemi seçilmişse muhtemelen cevaplayıcılar birbirlerine benzer özelliklere sahip olacaklardır. Küme örneklemesi yapılmamış bile olsa "homojen özelliklere sahip bir ana kütleden" tesadüfi yöntemle örnek kütle belirlenmiş olması türdeşlik özelliği nedeniyle sonuçları değiştirmez. Örneğin İşletme Fakültesi öğrencilerinden tesadüfî olarak seçilecek bir örnek kütle üzerinde *İletişim Biçimleri Ölçeği* uygulandığında sonuçların varyansı muhtemelen düşük çıkacaktır. Eğer sonuçlar kriter grup yerine "genel grupta" homojen bir yığılım ortaya çıkarmışsa bu tür araştırmaların bilimsel değeri düşüktür. Çünkü çıkan sonuçlar ölçeğin güvenilirliği hakkında fazla bir bilgi vermez. Yanlı bir eğilimi yansıtır. Bu tür araştırmalarda ana kütleyi veya örneklem çerçevesini tanımlarken sınırları anlamlı bir biçimde geniş tutmak gerekir.

Diferansiyel madde fonksiyonunun incelenmesi. Diferansiyel madde fonksiyonu (DMF), bir maddenin değişik etnik, kültürel ve cinsiyet gruplarında farklı bir şekilde çalışması anlamına gelir. Klâsik test kuramında geliştirilen ölçeğin veya testin DMF özelliği gösterip göstermediğini belirlemek için *P* güçlük indeksi değerlerinden veya ki-kare analizinden yararlanılır.

Güçlük indeks değeriyle diferansiyel madde fonksiyonunun araştırılması. Güçlük indeks değerlerinden yararlanmak için her bir madde farklı iki grubun üyelerine uygulanır ve her bir grupta maddelerin güçlük indeks değerle-

ri hesaplanır. Daha sonra odak grubun indeks değerleri referans grubun indeks değerlerinden çıkarılarak madde bazında *güçlük indeks değerleri farkı* bulunur.[15]

■ Cinsiyet faktörü temel alınarak erkekler ve kadınlar üzerinde yapılan bir ölçümde DMF özelliğinin araştırılmasına ilişkin bir örnek: 1. ve 2. maddelerin odak ve referans grubundaki güçlük indeks ve cinsiyet farklılığı (CFRK) değerleri aşağıdaki gibi elde edilmiştir.

$1P_o = ,45$ $\quad$ $2P_o = ,63$
$1P_r = ,35$ $\quad$ $2P_o = ,60$

$1_{CFRK} = 1P_o - 1P_r$. $\quad$ $2_{CFRK} = 1P_o - 1P_r$.
$1_{CFRK} = ,45 - ,35$. $\quad$ $2_{CFRK} = ,63 - ,60$.
$1_{CFRK} = ,10$. $\quad$ $2_{CFRK} = ,03$.

İkinci aşamada, odak ve referans grubunun her biri için testteki maddelerin ortalama güçlük indeks değerleri saptanır.

$\overline{X}_O = 6,12/12 = ,51$. (Odak grup için ortalama güçlük derecesi)
$\overline{X}_R = 4,21/12 = ,35$. (Referans grubu için ortalama güçlük derecesi)

Üçüncü aşamada cinsiyet faktörüne göre odak ve referans gruplarının ortalama güçlük derecelerinin farkları alınır ($\overline{X}_{CGFRK} = \overline{X}_O - \overline{X}_R$). Bundan sonra madde farkları, grup farklarından büyük olan maddeler işaretlenerek iyileştirilmesi gereken DMF özelliğine sahip maddeler olarak belirlenir. Bu yöntem gruplar arasındaki "ana etki" farklılığını temel alır. Maddelerdeki farklılık ana etkiden büyükse yöntem, bu maddeleri zayıf olarak değerlendirir.Yaklaşımda, maddelerde meydana gelebilecek hata varyansı dikkate alınmaz. Maddeler aynı gruptaki kişilere tekrar uygulanmış olsaydı muhtemelen sonuçlar farklı çıkardı. Bu yöntem bir barometre gibi değerlendirilebilir. Barometredeki rakamlar yükseldikçe DMF'nin gerçekliği konusunda daha güçlü bir baskı veya stres ortaya çıkacaktır.

Ki-kare hesaplamasıyla diferansiyel madde fonksiyonunun araştırılması. Ki-kare analizi farklı iki grupta yapılan ölçümler arasında anlamlı bir farklılık bulunup bulunmadığını belirlemek için kullanılır. Ki-kare analizi test

bazında değil, her bir madde için ayrı yapılır ve analiz sonucunda grupları farklılaştıran maddelerin elenmesi yoluna başvurulur.

■ Örnek:

Bir testte 1 numaralı maddede odak grubunda yer alan 500 kişiden 245'i doğru yanıt vermişlerdir. Doğru yanıt oranı %45'tir. Yanlış yanıt verenlerin oranı ise %55'tir. Bilim adamı odak grubunda testin ayrımcılığa yol açıp açmadığını görmek istemektedir. Odak ve referans grupları arasında fark yoksa doğru / yanlış yanıt verme oranları arasında da fark olmaması gerekir. Referans grubunda 312 kişi arasında ölçüm yapılmış ve bu kişilerin 175'i doğru yanıt vermişlerdir. Doğru yanıt verme oranı %56 ve yanlış yanıt verme oranı ise %44'tür. Araştırmacı burada iki grubun yüzde oranları arasındaki farklılığın anlamlı olup olmadığını belirlemeye çalışacaktır. Doğru yanıt verme açısından %45 ilâ %56 arasındaki %11'lik farkın anlamlı olup olmadığı ki-kare testi ile analiz edilir. Ki-kare formülü Eşitlik 13-6'daki gibidir.

■ Ki-kare formülü.

$$x^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]. \qquad (13\text{-}6)$$

f_o = Gözlenen frekans.

f_e = Beklenen frekans.

Σ = Odak grubunda doğru yanıt veren kişilerin gözlenen frekansından beklenen frekansın çıkarılması ve karesinin alınarak beklenen frekansa bölünmesi + odak grubunda yanlış yanıt veren kişilerin gözlenen frekansından beklenen frekansın çıkarılması ve karesinin alınarak beklenen frekansa bölünmesi.

Odak ve referans grupları arasında hiçbir fark olmasaydı, odak grubu için doğru yanıtlara yönelik beklenen frekans (f_o), referans grubundan elde edilen doğru yanıtlara ait yüzdenin odak grubundaki örnek kütle büyüklüğüyle çarpımına eşit olurdu. Yanlış yanıtlara yönelik beklenen frekans (f_o) ise referans grubundaki yanlış yanıtlara ait yüzdenin odak grubundaki örnek kütle büyüklüğüyle çarpımına eşittir. Buna göre doğru yanıtlar için beklenen frekans, 500*0,56 = 280 ve yanlış yanıtlar için beklenen frekans, 500*0,44 =

220'dir. Değerleri ki-kare formülünde yerine koyduğumuzda aşağıdaki sonuçlar elde edilir.

$$x^2 = \left[\frac{(245-280)^2}{280} + \frac{(255-220)^2}{220}\right], \quad (13\text{-}7)$$

$$x^2 = \left[\frac{(-35)^2}{280} + \frac{(35)^2}{220}\right], \quad (13\text{-}8)$$

$$x^2 = \left[\frac{1225}{280} + \frac{1125}{220}\right], \quad (13\text{-}9)$$

$$x^2 = [5{,}56 + 5{,}11], \quad (13\text{-}10)$$

$$x^2 = 10{,}67. \quad (13\text{-}11)$$

Hesaplanan ki-kare değeri, 2x2 tablosundan elde edilen 1 serbestlik derecesinde ve %5 anlamlılık düzeyinde tablo değeri 3,841'den yüksek olması nedeniyle "odak ve referans grubunun maddeye doğru yanıt verme oranları arasında fark yoktur" hipotezi ret edilir ve maddenin DMF özelliğine sahip olduğu söylenir.

Ki-kare analizi yöntemi, diferansiyel madde fonksiyonunu herhangi bir farklılığın "ana etkiden" daha büyük olup olmadığını belirlemek için kullanılır. Bu yöntem, maddeler arasındaki DMF farklılığından çok gruplar arasındaki gerçek farklılığı ortaya koyar.

Cevaplama oranın artırılması. Güvenilirliği artırmanın bir diğer yöntemi anketlerin geri dönüş/yanıtlama oranını artırmaktır. Belirli bir hipotezi test eden sonuçların genellenebilmesi için belirlenen örneklemdeki kişilerin en az %70'i anketi yanıtlamış olmalıdır. Böylece gelen cevaplardaki değişkenlik daha geniş bir aralığa dağılmış olur.[16] Bazı bilim adamları alan araştırmalarında cevaplama oranını %50 olarak belirlerken diğerleri bu oranı %30'a kadar düşürmüşlerdir. Ancak 300 kişilik bir örnek kütlenin %30'u ile 3000 kişilik bir örnek kütlenin %30'u aynı değildir. Üç yüz kişilik bir örnek kütlede %30'luk cevaplama oranı muhtemelen kabul edilebilir bir rakam olarak görülmeyecek ve toplanan veriler güvenilir bulunma-

yacaktır. Kabul edilebilir cevaplama oranı, posta anketlerinde, telefonla yapılan araştırmalarda, İnternet ortamında ve yüz yüze yapılan anket uygulamalarında farklı rakamlarla ifade edilmiştir. Yeni Zellenda'da Massey Üniversitesi pazarlama departmanı tarafından yapılan bir araştırmada 1991-1997 yılları arasında posta anketi ile yapılan araştırmalarda geri dönüş oranının %66 ilâ %71 arasında değiştiği bulunmuştur. Fakat, başka ülkelerin bu konudaki araştırma sonuçları örneğin, ülkemiz için baz veri olarak kullanılamaz. Çünkü ülkemizde insanlarımızın bu tür anketlere yanıt verme oranları çok daha düşüktür. Kabul edilebilir bir cevaplama oranın ne olması gerektiği konusunda bilim adamları arasında henüz tam bir mutabakat oluşmamıştır. Bilim adamı yaptığı araştırmanın bilimsel genellenebilirlik durumunu göz önünde bulundurarak bu soruya yanıt vermelidir. Düşük cevaplama oranı ölçek verilerinin güvenilirliğini düşüreceğinden söz konusu oranı arttırmak için ilave çalışmaların yapılmasına ihtiyaç vardır. Bu konuda Dillman, Christensen, Carpenter ve Brooks (1974) tarafından yapılan bir araştırmada ek çalışmalarla yanıt sayılarının Tablo 13-4'deki gibi arttırılabileceği bulunmuştur (aktaran, Fesenmaier, 2002).[17]

Cevaplama oranını artırmanın bir diğer yöntemi, örneklem büyüklüğünü anketlerin tahmin edilen geri dönmeme oranı kadar çoğaltmaktır. Böylece seçilen örneklem büyüklüğüne daha yakın değerler elde edilebilecektir. Geri dönmeme tahmini önceki araştırma sonuçlarına, ülke insanlarının alışkanlıklarına veya pilot araştırma sonuçlarına bakılarak belirlenir.

Tablo 13-4. Cevaplama Oranının Zaman İçinde Yapılacak Girişimlerle Artırılması

	Süre	*Cevaplama oranı*
Postayla gönderilen ilk anket	1. hafta	23,8
Posta kartıyla hatırlatma yapılması	2. hafta	42,0
İkinci bir anket daha gönderme	3. hafta	59,0
Üçüncü bir anket daha gönderme	7. hafta	72,4

Kaynak. D. Fesenmaier, "Response Rate [Yanıt Oranı]," ty, <http://www.tourism.uiuc.edu/itn/etools/eGuides_survey_responserate.htm> (01.12.2002).

Yanıtları tutarsız olan anketlerin elenmesi. Öğrencilere yapılan anketlerde, bir fabrikada işçilere kendi istekleri dışında zorla yapılan anketlerde cevaplayıcıların gerçek görüş ve düşüncelerini saklayarak, araştırma-

cıyı yanıltmak isteyebilecekleri ihtimali gözden uzak tutulmamalıdır. Bu nedenle her bir anketin baş tarafına, uygulama kılavuzuyla ilgili bölüme verilen cevapların tutarlı olmasına özen gösterilmesi konusunda bir uyarı cümlesinin yazılmasında yarar vardır. Bilim adamı ayrıca her bir ankete gelen yanıtları tek tek gözden geçirmeli, bir ankette çelişkili sayılabilecek madde sayısı %20'yi aşmışsa bu anketleri değerlendirme dışı bırakmalıdır. Tutarsız yanıt incelemesi her tür ölçek için uygun değildir. Ölçüm konusu, kişilerin yalan söylediklerini veya açık bir şekilde tutarsız yanıtlar verdiklerini ortaya koyacak bir nitelikte ise bu yönteme başvurulur. Örneğin bir öğrencinin "Tüm ödevlerimi tam olarak ve eksiksiz bir şekilde hazırlarım." ifadesine olumlu yanıt vermesi ile başka bir maddedeki "Benim için ödevler hiç önemli değildir." ifadesine aynı şekilde olumlu yanıt vermesi ölçeğin/anketin dikkatsiz bir şekilde doldurulduğu anlamına gelir.

Eksik verilerin kodlanması. Eksik verilerin bir bölümü cevaplayıcılardan, diğer bölümü ise verileri bilgisayara giren kodlayıcılardan kaynaklanır. Eksik veriler; hızlı okuma nedeniyle bazı maddelerin atlanılması, unutma, kasıtlı olarak cevap vermeme, birden fazla işaretleme yapılması, nasıl cevap verileceğinin bilinememesi, test cihazında veya sonuçları kayıt eden cihazda arıza ortaya çıkması, verileri bilgisayara tanımlarken eksik girme gibi nedenlerle ortaya çıkar. Eksik veri sayısının ölçüm sonuçlarını etkileyecek miktarda olup olmadığını belirlemek için aşağıdaki yöntemlere başvurulur:[18]

1. Eksik veri içeren değişken "0 = eksik", "1 = tam" şeklinde iki grup halinde kodlanır ve başka bir değişkenle *t*-testi uygulanır.
2. Eksik veri içeren değişken "0 = eksik", "1 = tam" şeklinde iki grup halinde kodlanır ve başka bir değişkenle kendisi arasında ki-kare analizi uygulanır.
3. Eksik veri içeren değişken "0 = eksik", "1 = tam" şeklinde iki grup halinde kodlanır ve başka bir değişkenle kendisi arasında korelasyon analizi yapılır.

Çıkan sonuçların anlamlılık düzeyi dikkate alınarak, anlamlı bir farklılık yaratan kaç tane değişken olduğuna bakılır. Araştırmacı güvenilirliği artırmak için eksik verileri iyileştirmeye yönelik değişik yöntemlere başvurabilir. Bu yöntemlerin bir bölümü anket formu ve değişkenlerin iptal edilmesiyle ilgiliyken diğerleri eksik verilerin yerine yeni değerlerin atan-

masını gerektirir. Eğer az sayıdaki ankette (örneklemin %10'u kadar) bazı değişkenlere cevap verilmemişse bu anketler analize alınmaz. Bir başka yöntem cevap verilmeyen değişkenlerin analize alınması ve toplam puan yerine ortalama puanlar ile çalışılmasıdır. Bilim adamı, anket bölümlerini çıkarmak, değişkenleri düşürmek yerine eksik verilere yeni değerler atama yoluna da başvurabilir. Atama yöntemlerinin başlıcaları aşağıdaki gibidir:

1. Araştırmacı, eksik veriyi sağlıklı bir şekilde tahmin edebiliyorsa kendi tahminini kullanır. Bunu yaparken önceki ve sonraki sorularla benzeri diğer maddelere gelen yanıtları inceler.
2. Eksik verilerin yerine yuvarlama yapılarak aritmetik ortalama değeri, medyan veya duruma göre mod değeri yazılabilir. Ancak bu yöntemin bir çok yetersizlikleri vardır. Bu uygulama varyansı azaltabilir. Bu nedenle sonuçta maddeler arasındaki korelasyon azalır. Yöntem maddeler arasındaki ilişkileri dikkate almadığından çok değişkenli ilişkileri çarpıtır.[19]
3. Az sayıdaki eksik verinin yerine ölçek değerleri temel alınarak rasgele bir değer atanabilir.
4. Bir diğer yaklaşım, regresyon değerinin ikâme edilmesidir. Bu değer aynı satırdaki diğer değerler ve korelasyon matrisi dikkate alınarak belirlenir. Tahmin etmekten ve aritmetik ortalama değerlerini kullanmaktan daha güvenilir bir yöntemdir.
5. Çoklu atama yöntemi başka bir yaklaşımdır. Bu yöntemde verilerdeki dağılım temel alınır.
6. Bir diğer yöntem Mokken tarafından önerilen *tek parametreli lojistik model* (TPLM)[a] yaklaşımıdır. Bu modelde boş değişkenlerin yerine birden fazla değeri atamak mümkün olabilmektedir.

Eksik verilerin iyileştirilmesi konusuna kitabın "Girdi Kalitesinin Değerlendirilmesi İçin Veri Taraması" başlıklı dördüncü bölümde ayrıntılı olarak değinilmişti. Okurların bu bölüme tekrar başvurmalarında yarar vardır.

---

[a] Tek parametreli lojistik model (One parameter logistic model). Madde-yanıt kuramında sadece maddenin güçlüğünü temel alan hesaplama yaklaşımı.

Verilerin dağılım özelliğinin incelenmesi. Klâsik test kuramına göre güvenilirlik analizi yapılacak veriler birinci ve ikinci ölçümlerden, testin birinci ve ikinci yarısından, alternatif formlardan veya farklı gözlemcilerden elde edilmiş olabilir. Araştırmacı öncelikle bu verilerin normal dağılım eğrisine uygun olup olmadığını, çarpıklık ve basıklık katsayılarını incelemelidir. Çarpıklık ve basıklık katsayılarının normal bir dağılım için sıfıra yakın çıkması gerekir. Herhangi bir ifadeye ait veriler büyük öçlükte *çarpık* ise güvenilirlik konusunda kuşkular uyanacaktır. Verilerin, ölçekteki tüm maddelere ait puanların normale yakın bir dağılım göstermesi istenir. Ölçekteki bir iki vak'anın çıkarılması ile verilerin normal dağılım özelliği artıyorsa bu yönteme başvurulabilir, ancak bu işlem dikkatli bir şekilde yapılmalıdır. Çünkü çıkarılan vak'alar eğer bir veya birden fazla faktör üzerinde "etkili" olma özelliğine sahipse ölçeğin tahmin gücünün zayıflama tehlikesi ortaya çıkar.

Araştırmacı verilerin dağılım özelliğini incelemek için istatistiksel analiz programlarının histogram grafiklerinden, gövde-yaprak grafiklerinden, Q grafiğinden, kutu grafiğinden ve iki madde arasındaki ilişkileri görmek için ise *nokta dağılım grafiğinden* yararlanır.

Ortalama ve standart sapma değerlerinin incelenmesi. Maddelerin dağılım özelliklerinin incelenmesinden sonra ortalama, standart sapma ve varyans değerlerine ilişkin değerlendirmeler yapılır.

Eğer maddenin aritmetik ortalaması diğer maddelerinkinden büyük ölçüde farklılık gösteriyorsa bu madde ölçekten çıkarılır. Ortalama puanları "ayrık bir biçimde" çok yüksekse veya çok düşükse tavan-taban etkisi söz konusudur. Bu maddeler yapıyı ya çok yüksek veya çok düşük derecelerle ölçmeye çalışıyor demektir. Bu maddelerin 5 veya 7 dereceli ölçekteki dağılımları dengeli değildir ve bu nedenle problemli maddeler olarak değerlendirilir. Bu maddeler çıkarıldıktan sonra tekrar hesaplatma yapılarak alfa değerindeki değişmeler gözlenmelidir.

Bir maddenin standart sapması küçük ise o maddenin değişkenliği çok azdır ve herkes benzer şekilde cevaplama yapmış demektir. Bu şekilde ayırt etme gücü düşük olan maddeler üzerinde yeniden durulmalıdır. Katılımcıları farklılaştırmayan maddeler ölçeği zayıflatır. Standart sapma ile yakından ilgili bir diğer gösterge varyanstır. Standart sapma örneklem verileri hakkında bilgi verirken, varyans ana kütle değerlerini daha iyi yansıtır. Maddelerin varyans değerleri sıfır veya sıfıra çok yakın ise, bu maddeler de ölçekten çıkarılır. Çoktan seçmeli sorularda maddelerin varyans değerlerinin ,16'dan büyük olması hedeflenir.

Ayrık değerlerin elenmesi. Araştırmada ölçüm değerleri dikkatle incelenerek ayrık değerlerin elenmesi veya iyileştirilmesi yoluna başvurulur. Tutum ölçeklerinde maddelere ait ayrık değerler çok fazla anlamlı olmayabilir. Ancak test toplam puanlarındaki ayrık değerler önemlidir. Psikoteknik ölçümlerde örneklem grubunda görülebilecek bir kaç tane çok yüksek veya çok düşük puan ölçümün güvenilirliğini büyük ölçüde etkiler. Bu nedenle söz konusu puanlar ölçüm verilerinden çıkarılarak çarpıklık ve basıklık katsayıları yeniden incelenmelidir. Bir diğer yöntem ayrık verileri analizden çıkarma yoluna başvurmadan bu değerleri daha kabul edilebilir değerlere çekmektir. Psikoteknik testlerde asgarî bir barem puan belirlenerek bu puanın altında kalan kişiler araştırmaya/ölçüme katılmayabilirler. Bu uygulama psikometride sınır/kesim değerinin üstünde kalma ilkesine uygundur. İstatistiksel analiz programı SPSS'te ayrık değerleri kutu-bıyık grafiğinde görmek mümkündür. Grafikte kartiller arası değerin 1,5 katından fazla olanlar *ayrık değer*, 3 katından fazla olanlar ise *uç değerler* olarak kabul edilir. Grafikte ayrık değerler küçük "o" harfiyle, uç değerler ise küçük yıldız simgesiyle gösterilmiştir.

Çoklu gözlemde/ölçümde bulunma. Gözlem/ölçüm sayısı arttıkça toplanan verilerin güvenilirliği artar. Bir kişiye psikometrik bir test uygulandığında, tek bir test yerine birden fazla ölçüm yapılmasıyla rakamlar arasındaki sapmalar azalır. Test-yeniden test uygulamalarını oldukça istikrarlı rakamlar elde edinceye kadar sürdürebiliriz. Öte yandan değerlendiricilerin/gözlemcilerin sayısının da belirli bir miktara kadar arttırılmasıyla daha tutarlı ve daha doğru sonuçlar elde edilir.

Birden fazla ölçek kullanma. Araştırmacı belirli bir psikolojik yapıyla ilgili olarak ölçüm yapıyorsa aynı konuda birden fazla ölçek bularak bu ölçekleri aynı zaman diliminde veya farklı zaman dilimlerinde uygulamak suretiyle sonuçların benzer çıkıp çıkmadığını araştırabilir. Birden fazla ölçek verilerinden benzer sonuçlar elde edilmesi geliştirilen ölçeğin/testin güvenilir olduğu konusunda oldukça güçlü bir kanıttır.

Puanları standartlaştırma. Araştırmacı alt ölçekler yerine tüm ölçeğin veya değişik bilişsel testlerden oluşan bir test bataryasının genel güvenilirlik katsayısını elde etmek istiyorsa ölçeklerin/testlerin ham puanlarını öncelikle standart puanlara çevirmelidir. Değişik standart puan yöntemleri arasında en sık MacCall *T* puanları kullanılır. Bunun için ham puanlar ortalaması 50 ve standart sapması 10 olan *T* puanlarına çevrilmeli ve güvenilirlik analizi bu puanlara dayandırılmalıdır. Ancak bu yöntemin uygula-

nabilmesi için bataryadaki testlerin aynı özelliği ölçmesi gerekir. Farklı özellikleri saptamaya çalışan testler için *birleşik alfa güvenilirliğini* ölçmeye gerek yoktur. Ayrıca bir kişinin psikometrik testlerden elde ettiği toplam başarı puanı bazen testlerin ölçüm özelliği dikkate alınarak yapılan ağırlıklı puanlara göre de hesaplanabilir.

Madde sayısının artırılması. Ölçekteki/testteki madde sayısı belirli bir sayıya kadar arttıkça ölçeğin veya testin güvenilirliği arttığından düşük güvenilirlik rakamlarında testteki madde sayısının yeterliliği soruşturulur. Uzun testler kısa testlere göre daha güvenilirdir. Örneğin, tutum ölçeklerinde üç maddeli testlere göre 10-15 maddeli testler daha güvenilirdir. Çünkü bu testlerde verilerdeki ayrık değerler ölçeğin toplam puanını daha az etkiler. Bir tutum ölçeğinde kaç tane madde olması gerektiği faktör analizi sonuçlarına bakılarak belli olur. Ancak üç veya dört maddeden daha az olmamalıdır.[a] Bilgi testlerinde spesifik bir konuyla ilgili madde sayısı en az altı olmalıdır. Bilişsel yetenek testlerinde madde sayısı; testin süresi, güvenilirlik katsayısı, testin norm veya kriter referanslı olarak kullanılma durumuna göre belirlenir. Testlerde madde sayısının artmasıyla birlikte güvenirliğin artması *Spearman-Brown Kehanet Formülü* ile hesaplanarak tahmin edilebilir (*bk.*, Eşitlik 13-12).

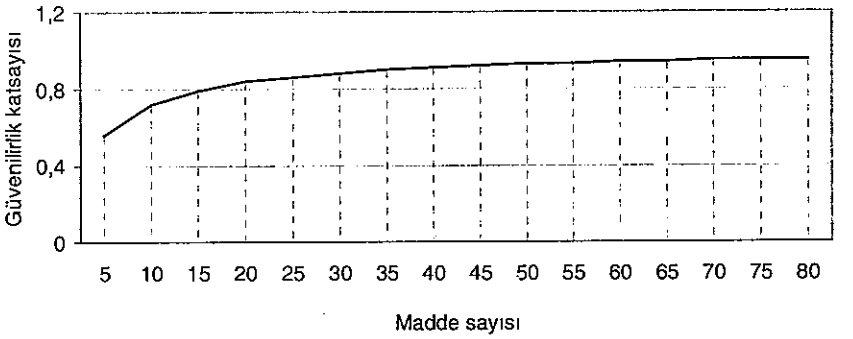
$$r_{yeni} = \frac{K(r_{eski})}{1 + (K-1)(r_{eski})}. \qquad (13\text{-}12)$$

Bu formülü çalıştırabilmek için iki bilgiye ihtiyaç vardır. Birincisi, hali hazırda hesaplamış olduğumuz güvenilirlik katsayısıdır. İkincisi ise testin veya ölçeğin uzunluğunu veya madde sayısını etkileyecek faktör sayısıdır.

---

[a] Bir ölçekte, belirli bir "özelliği" ölçen veya ortaya çıkaran madde sayısı konusunda bilim dalları arasında farklılıklar olabilmektedir. Örneğin pazarlama araştırmalarında bir özellik "somut olarak kavranabilecek" bir şekilde ise birden fazla madde sorulmasına gerek olmadığı belirtilmiştir. *Somut özellik*, hakemlerin bir maddenin ne olduğu veya anlamı konusunda fikir birliği içinde olmalarıdır. Drolet ve Morrison (2001) somut olarak kavranabilecek özelliklerde birden fazla madde sorulmasının gereksiz olduğunu, bu uygulamanın tesadüfî hatayı azaltmadığını tersine ölçüm konusunun ötesinde başka özelliklerin işin içine girmesine neden olduğunu bildirmişlerdir. Bu konuda *bk.*, J.R. Rossiter, "Scale Develpment in Marketing [Pazarlama Araştırmalarında Ölçek Geliştirme]," <http://130.19 5.95.7:80 81/ WWW/ANZMA C2001/anzmac/Special%20Session/ SPECIAL % 20SESSION%20V.pdf.>

Faktör sayısı *K* harfiyle gösterilmiştir. Bu simgeyi "artırılmak istenen kat" olarak da tanımlayabiliriz. Araştırmacı ölçekte tek bir faktör yerine iki faktör olması halinde veya madde sayısını "iki kat kadar" artırmak istemesi halinde madde sayısının ikiye katlanacağını düşünebilir. Böyle bir durumda "olası" güvenilirlik katsayısını hesaplamak için *Spearman-Brown* formülü çalıştırılır ve yeni bir değer elde edilir. Ancak bu değer, tahminî bir göstergedir. Gerçek uygulamada güvenilirlik katsayısı bu değerin üstünde çıkabilir veya daha düşük bir değer olabilir. Şekil 13-3'de deneme amaçlı olarak çizilen grafikte beş maddeden oluşan bir ölçeğin güvenilirlik katsayısı r =,56 iken madde sayısının artmasıyla güvenilirlik katsayısının da arttığı görülür. Bununla birlikte Şekil 13-3'teki grafik, özellikle 25 madden sonra güvenilirlik katsayısında önemli bir artış olmadığını ortaya koymaktadır.



Şekil 13-3. Madde sayısının artmasıyla birlikte güvenilirlik katsayısının artması.

Kaynak. J. Parkes, "Some Concrete Examples of Reliability Issues [Güvenilirlik Sorunlarıyla İlgili Bazı Somut Örnekler]," <http://www.unm.edu/~parkes/574/reliability.pdf> (25.01.2004).

Yapılan araştırmalar Cronbach alfa güvenilirlik katsayısının 18 maddeye kadar arttığını bu sayıdan sonra ise önemli bir artış göstermediğini ortaya koymuştur.[20] Bununla birlikte Winer (1986) az sayıda madden oluşan bir test, yüksek güvenilirlik katsayısına sahip olsa bile araştırmacının yanıltıcı bir güvenlik hissine kapılmadan bu rakamı dikkatle değerlendirmesi gerektiğini belirtmiştir. Ona göre nispeten kısa sayılabilecek bir testte güvenilirlik katsayıları çok yüksek veya çok düşük çıkmışsa test/ölçek yeni-

den metodolojik bir incelemeye tâbi tutulmalıdır (aktaran Torabi ve Ding, 2003).[21]

Testteki madde sayısını testin niteliğine göre düşünmek gerekir. Bir tutum ölçeğindeki madde sayısı, bir bilişsel yetenek testindeki madde sayısı, okullarda sınıf ortamında yapılan matematik sınavındaki madde sayısı ve bir sertifikasyon veya liselere, üniversiteye giriş sınavındaki madde sayısı farklıdır. Sertifikasyon veya yüksek ödüllü test sınavlarında test maddeleri alan örneklemesi yöntemine göre değişik alanlardaki yetkinliği belirleyecek şekilde saptanır. Bilgi, yetenek ve beceri alanının genişliğine göre test maddelerinin sayısı 100 ilâ 200 arasında değişir. Hatta daha geniş yetkinlik alanlarının ölçülmek istenmesi halinde madde sayısı daha da artırılabilir. Bir ölçekteki/testteki madde sayısını belirlerken aşağıdaki faktörler göz önünde bulundurulur:

1. Testin doldurulması için katılımcıların zaman ayırabilecekleri optimum süre.
2. Testi alanların veya ölçeği dolduranların katlanabilecekleri yorgunluk.
3. Kaliteli ölçek maddesi hazırlamak veya kaliteli test oluşturmak için ayrılabilecek süre.
4. Yeni test maddesi oluşturmanın maliyeti.
5. Ölçüm yapılacak yetkinlik alanının genişliği.

Görüldüğü gibi madde sayısının artırılması diğer faktörlerle birlikte ele alınıp değerlendirilmesi gereken bir konudur.

Örneklem büyüklüğünü artırma. Güvenilirliği iyileştirmenin bir diğer yöntemi uygun örneklem büyüklüğü ile çalışmaktır. Örneklem büyüklüğü arttıkça güvenilirlik tahmin değeri de artar. Ölçekteki madde sayısı ile örneklemdeki vak'a sayısı arasında olumlu yönde bir ilişki vardır. Büyük örneklemlerde nadir görülen özellikleri veya olguları daha doğru bir şekilde tahmin etme olasılığı ortaya çıkar. Örneklem hacmi yeterince büyük değilse, toplanan veriler araştırılan faktör yapısına iyi uymaz.

Yüksek lisans öğrencileri, örneklem büyüklüğünü belirlerken geliştirdikleri ölçeğin güvenilirlik rakamlarından çok hipotez testlerini göz önünde bulundururlar. Ayrıca istatistik kitaplarında örneklem büyüklüğünün hesaplanmasıyla ilgili değişik formüller ve rakamlar sunulmuş olması on-

ları bir ölçüde şaşırtır. Genel eğilim "tek bir kriter" belirleyip bu kriter çerçevesinde örnek kütle büyüklüğünü saptama şeklinde gerçekleşir. Oysa konu karmaşıktır ve biraz daha ayrıntılı bir inceleme yapılmasını gerektirir. Örneğin, alfa güvenilirlik hesaplamasında örneklem hacmi çok küçükse test güç kaybeder ve güvenilirlik katsayısının güven aralığı geniş çıkar. Örneklemin çok büyük olması ise kaynak israfına neden olur. Örneklem büyüklüğü konusunda Fleiss (1986) 15-20 kişilik bir grubun yeterli olacağını bildirirken Nunnally ve Bernstein (1994) ölçüm kuramının yüksek dozda örneklem hatası içeren çalışmalara tolerans gösteremeyeceğini ve bu nedenle 300 veya daha büyük rakamlardaki örneklem büyüklüğü ile çalışılmasını önermişlerdir (aktaran Banet).[22] Herhangi bir ölçümde, uygun örneklem büyüklüğü; (a) testin sahip olması arzulanan güce, (b) etki büyüklüğüne, (c) testin duyarlılık derecesine ve (ç) arzulanan alfa düzeyine bağlıdır. Bu nedenle örneklem büyüklüğünü basit bir şekilde "100'ün üstünde veya 300'ün üstünde olması iyidir" gibi belirli baş parmak kuralları-na bağlamak doğru değildir.[23]

Test uygulayıcılarının veya anketçilerin eğitilmesi. Ölçüm uygulamalarında güvenilirliği iyileştirmenin bir diğer yöntemi testi verecek teknisyenleri; testin uygulama biçimi, testin süresi, aşamaları ve sonuçların kayıt edilmesi konusunda bilinçlendirmek ve eğitmektir. Eğitimin temel konusu *yeknesaklık* ve *standardizasyondur*. Aynı şey tutum ölçekleri ve indekslerle veri toplama yöntemi için de geçerlidir. Burada anketçiler özellikle uygulama biçimi konusunda eğitilirler. Anketçiler düşünme, algılama ve uygulama açısından benzer tutumlara sahip olmalıdırlar. Bunun için kişiler mümkün olduğunca aynı eğitim branşından seçilirler. Sosyal nitelikle bir araştırma için mühendislik eğitimi görmüş anketçilerden yararlanmak doğru değildir.

Test standartlarına uyma. Testlerde hata varyansını azaltmak için test uygulama biçiminin testi alan tüm kişiler için standart hale getirilmesi gerekir. Test tüm kişilere konforlu şartlarda verilmeli, kişilerin dikkatlerini dağıtacak faktörler, etkenler ortadan kaldırılmalıdır. Testi alan kişiler aşırı derecede yorgun, uykusuz olmamalı testi alma konusunda istekli bulunmalıdırlar. Test; kişilere zorla, istekleri dışında uygulanmamalıdır. Test talimatları ve açıklamaları tüm katılımcılara standart bir şekilde okunmalı, beklenmedik sorulara testi uygulayan teknisyenler kendi yargılarını kullanarak kısa ve makul cevaplar vermelidirler.[24]

Kontrol ve deney gruplarından yararlanma. Araştırmacı ölçüm sonuçlarının güvenilirliği konusunda daha sağlıklı bilgiler elde etmek için kontrol ve deney gruplarından yararlanabilir. Kontrol grubu ölçüm yapılan grubun dışındaki genel gruplardır. Bunlar genelleme yapılacak odak gruplardan farklı bir sonucun çıkması doğal olan gruplardır. Eğer bu gruplarda da sonuçlar farklı çıkmıyorsa ana kütleye ilişkin olarak yapılan genellemeler güvenilirliğini yitirir.

Çeldiricilerin çalışıp çalışmadığının kontrol edilmesi. Kullanılan test çoktan seçmeli bilgi testi niteliğinde ise çeldiricilerin (yanlış şıkların) çalışma durumu dikkatli bir şekilde analiz edilir. Hiçbir cevaplayıcı tarafından seçilmeyen çeldiriciler değiştirilir veya iptal edilir. Çünkü bu şıklar iyi öğrencilerle zayıf öğrencileri ayırt etme özelliğine sahip değildir. Çeldiricilerin başarılı, vasat ve başarısız öğrenciler tarafından eşit sayıda tercih edilmiş olması halinde bu şıklar incelemeye alınır. Fakat esas önemli olan vasat ve başarısız öğrencilerin çeldiricileri daha yüksek sayıda/oranda işaretlemiş olmalarıdır. Başarısı düşük gruptaki kişilerin çeldiricileri daha yüksek oranda seçmesi $+Ç$ simgesiyle; başarısı yüksek gruptaki kişilerin çeldiricileri daha yüksek oranda seçmesi ise $-Ç$ simgesiyle gösterilir. Öte yandan bir soruya ait çeldiricilerin eşit sayıda/oranda tercih alıp almadığı önemli değildir, çünkü çeldiricilere değişik sayıda öğrenci yakalanabilir.[25] Sayının veya oranın bizatihi kendisinden çok çeldiricilerin yüksek ve düşük başarılı gruplardaki dağılımı önemlidir. Yüksek ve düşük başarılı gruplarda $Ç$ oranları eşit çıkmışsa bu şıkların gözden geçirilmesi gerekir.

■ Örnek: Çeldirici analizi sonuçları.

İstavrit balığı daha çok hangi denizde bulunur?

	Başarısı yüksek grup	*Başarısı düşük grup*
Karadeniz	%15	%35
Akdeniz	%6	%12
Marmara*	%48	%12
Ege	%12	%24

Testte eğer iki doğru cevap varsa doğru şıklara gelen yanıtlarla yanlış şıklara gelen yanıtlar arasındaki orantıya bakılır. Başarılı öğrencilerin doğru şıklara verdikleri doğru yanıt oranları yanlış şıklara verdikleri doğru

yanıt oranlarından yüksek olmalıdır. Maddelerin güçlük indeks değerleri için eğer ,20-,80 aralığı baz alınmış ve soruların yanıtları 5 şıklı olarak oluşturulmuşsa çeldiricilerin her biri ,05 oranında tercih edilmiş olmalıdır. Bunun için, en düşük güçlük oranı değeri temel alınarak bu değer yanlış şık sayısına bölünür ve böylece bir çeldiricinin alması gereken yanıt oranı bulunur (,20/*d*). Formüldeki *d* simgesi *yanlış şık sayısını* gösterir. Bu rakam, beş şıklı sorularda 4, dört şıklı sorularda 3 ve üç şıklı sorularda ise 2'dir.

Madde-yanıt kuramında çeldirici analizinin mantığı biraz daha farklıdır. Bu kuramda çeldiricilerin güçlü olması maddelerin zorluk düzeyini artırır. Madde-yanıt kuramında yüksek yetkinlik düzeyine sahip olan kişilerin/öğrencilerin doğru yanıtı seçme olasılığı yüksek iken yetkinlik düzeyi düşük olan kişilerin/öğrencilerin doğru yanıt verme olasılıkları düşüktür. Bu kişiler daha yüksek oranda çeldiricilere yönelirler. Modern ölçüm kuramında bu incelemeye "opsiyon analizi" adı verilmiştir.[26] Madde-yanıt kuramına göre tek parametreli ölçüm modelini analiz eden yazılımlardan RUMM'un çoktan seçmeli sorular için karmaşık nitelikte çeldirici analizleri yaptığı bildirilmiştir.

Çok kolay ve çok zor olan bazı maddelerin elenmesi. Klâsik test kuramının temel alındığı analizlerde test maddeleri çok kolay olmaları nedeniyle katılımcıların büyük çoğunluğu tarafından yanıtlanmışsa bu maddeler ayırt edicilik özellikleri bulunmadığından testten çıkarılır ve yerlerine nispeten daha zor olanlar koyulur. Aynı şekilde bir testin çok zor olan çok sayıda maddeden oluşması da istenmez. Ancak çok zor ve çok kolay olan maddeler bütünüyle test dışında tutulmaz. Sadece bu maddelerin yoğunluklu olarak testin içinde bulunmasına izin verilmez. Güçlük ve kolaylık açısından maddelerin testin içinde normale yakın bir dağılıma sahip olması amaçlanır.

Test maddelerindeki zayıflığın iyileştirilmesi. Zayıflık, ölçek veya testin katılımcılara uygulaması sırasında ortaya çıkan ölçüm hataları nedeniyle maddeler arasındaki ilişkileri belirlemeye yönelik olarak yapılan korelasyon analizi sonucunda korelasyon katsayılarının beklenenden daha düşük çıkmasıdır. Zayıflık tek başına ölçüm hatalarından kaynaklanmaz. Ölçek derecelerinin 7'den 3'e düşürülmesi, sürekli nitelikteki parametrik verilerin nanparametrik ikili veri haline dönüştürülmesi de sonuçta korelasyon katsayılarında zayıflığa neden olur. Bunun için *Zayıflığı Düzeltme*

Formülü uygulanır.[a] Zayıflığı düzeltme formülü, Spearman (1904) tarafından önerilmiştir. Amacı ölçüm sonrası elde edilen korelasyon değerlerinde bir düzelteme yapmaktır. Fakat, güvenilirlik katsayısının puanlardaki/verilerdeki ölçüm süreci hataları yerine ölçeğin kendisinden daha fazla etkilendiği düşünülüyorsa zayıflığı düzeltme formülünü uygulamanın bir anlamı yoktur. Zayıflığı düzeltme, Eşitlik 13-13'deki formülle hesaplanır.

$$r_{t_1t_2} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}} \qquad (13\text{-}13)$$

r_{t1t2} = Birinci ölçümle ikinci ölçüm sonuçları arasındaki korelasyon katsayısı.
r_{11} = Birinci ölçümün güvenilirlik katsayısı.
r_{22} = İkinci ölçümün güvenilirlik katsayısı.

Modern Test Kuramına Göre İyileştirme

Modern test kuramı, madde-yanıt modeli üzerinde odaklanmıştır. Günümüzde madde-yanıt modeli sadece başarı testlerinde değil kişilik envanterleri ve tutum ölçeklerinde de kullanılır. Madde-yanıt modelinde maddelerin iyileştirilmesi, klâsik test kuramında olduğu gibi ana kütle–örneklem verilerine bağlı değildir. Test maddeleri bir kere kalibre edildikten (*a*, *b* ve *c* parametreleri hesaplandıktan) ve maddelerin bilgi fonksiyonları [$I_i(\theta)$] belirlendikten sonra bu maddeler her tür *örneklem–ana kütle* grubu için eşit ölçüde başarılı bir şekilde uygulanabilir. Bu nedenle madde-yanıt modeli sadece farklı örneklem grupları için değil, aynı zamanda değişik ana kütleler arasındaki farklılıkları görmek isteyen araştırmacılar için de cazip

[a] İstatistikî analiz programı SPSS'te zayıflık analizi yapmaya yönelik özel bir test bulunmamaktadır. Yapısal eşitlik modelini değerlendiren AMOS ve LISREL gibi programların gizli değişkenleri ortaya çıkarmak için oluşturulan ölçeklerdeki bazı değişkenlere cevap verilmemesi gibi zayıflıkları başarıyla ele aldığı bildirilmiştir. Zayıflığın derecesi büyük ölçüde bazı değişkenlere/maddelere cevap verilmemesinden veya yanlı olarak sadece belirli şıklara cevap verilmesinden kaynaklanır. Zayıf ölçeklerde varyans (değişenlik) düşüktür, verilerin dağılımı sağa veya sola çarpıktır, ayrıca veriler dikkati çekecek bir şekilde basık veya sivri bir dağılım özelliği gösterir. Yaklaşık olarak normal dağılım özelliği gösteren veri yapılarında ölçeğin zayıflığı ortadan kalkar. Varyansı büyük olan maddeler ölçeğe daha fazla katkıda bulunur. Bu nedenle varyansı düşük cevap şıkları, ölçeğin tek bir yönünde yoğunlaşmış maddeler *zayıf* olarak değerlendirilir.

bir yöntem olarak değerlendirilmiştir.[a] Bu kuramda, maddelerin *yanlılık incelemesi* ve *paralel maddeler eşitlik* değerlendirmesi maddelerin içerdiği koşullu standart hata değerlerine bakılarak yapılır. Standart hata değerleri ise testteki madde sayısı ile testi alan kişi sayısına ve puanların dağılımına bağlıdır.

Modern test kuramında gizli özelliği/yeteneği ölçmek için oluşturulan maddeler test oluşturma süreci içinde iyileştirilir. İlk aşamada test oluşturma ile iyileştirme farklı iki süreç değildir. Kalibre edilmiş test havuzu oluşturulduktan sonra ise iyileştirme yeni maddelerin havuzdaki maddelerin düzeyine çıkarılma işlemi anlamına gelir. Test oluşturma / iyileştirme süreci ayıklamayı, değiştirmeyi ve duruma göre madde havuzundan yeni test maddesi almayı veya bazı kişilerin test puanlarının ölçümden çıkarılmasını gerektirir. İlk kez yapılan bir test geliştirme faaliyetinde iyileştirme / oluşturma süreci aşağıdaki adımlarda gerçekleştirilir.

Pilot araştırma yapılması. Bu süreçte pilot araştırma veya ön test yapılarak veriler klâsik test kuramındaki analizlere tâbi tutulur. Zayıf maddeler klâsik madde analizi yöntemiyle ayıklanır. Maddeler toplam puanla iki serili korelasyon analizine tâbi tutularak, güçlük indeks değerleri ve ayırt etme özellikleri açısından değerlendirilir. Pilot araştırma sonunda güçlük indeks değerleri açısından geniş bir dağılıma sahip olan ve aynı zamanda ayırt etme indeks değerleri ,30'dan büyük olan maddeler belirlenmiş olur.

Pilot araştırma verilerine bağlı olarak yapılacak bir diğer analiz ölçeğin/testin tek boyutluluğunu belirlemeye yönelik olarak faktör analizi yönteminin uygulanmasıdır. Faktör analizi sonucunda eğer birden fazla boyut ortaya çıkmışsa test, ya alt testlere bölünerek ele alınır veya ilgili maddeler bir araya getirilerek küçük test grupları (testçikler) oluşturulur.

Model uyuşumunun saptanması. Maddelerin parametre tahminlerine dayanan madde-yanıt kuramının sağlıklı ve güvenilir sonuçlar verebilmesi

[a] Örneklemden bağımsız olmakla birlikte uygulamada madde bankaları genele ait değil, belirli yaş grupları veya sınıflar için ayrı ayrı belirlenmiştir. Belirli bir sınıfa ait kalibre edilmiş madde bankasında bir ders için 100-150 kadar soru bulunur.

için model-veri uyuşumunun[a] ve ölçülmek istenen boyutun/boyutların veriler tarafından doğrulandığı bilgisinin ortaya konması gerekir. Modelin verilere uygun olduğu *apriori* olarak belirlenemez. Bilim adamı model-veri uyuşumunu görmek için tek parametreli modeller yerine daha az kısıtlayıcı varsayımlara dayanan *genel modellerden* hareket etmişse (2PL ve 3PL gibi) model uyuşumu kolay sağlanır, fakat bu yaklaşım 1000, 2000 gibi büyük örnek kütlelerle çalışılmasını gerektirir. Bilim adamı eğer küçük örneklemlerle çalışıyorsa tek parametreli modelleri seçmesi daha uygun olur. Seçilen ölçüm modelinin önceki kuramsal bilgilere veya ampirik araştırma sonuçlarına dayandırılması güvenilirliği artırır.[27]

Modeller, araştırmacının gözlemlerine dayalı olarak veya literatürde yaptığı incelemelere bağlı olarak oluşturduğu soyut kavramlar, ölçüm objeleri, ön kabuller veya ilişki ağlarıdır. Madde-yanıt kuramında modeller tek boyutlu, çok boyutlu veya çok yüzeylidir. Ayrıca tek boyutlu modeller de kendi içinde parametre özellikleri dikkate alınarak tek parametreli, iki parametreli veya üç parametreli olarak belirlenebilir. Tek parametreli Rasch modelinin kendi içinde dahi değişik ölçüm modelleri vardır. Yanıt tipleri temel alındığında ise, ikili–çok dereceli; tek yanıtlı–çok yanıtlı–kısmî kredili ölçüm modellerinden söz edilir. Ölçüm modelleri genelde matematiksel denklemlerle ifade edilir. Gerçek hayattan elde edilen veriler hiçbir zaman geliştirilen matematiksel modellere tam olarak uymaz. Bu nedenle modelin verilere olan uygunluğu "modelin güçlü" olmasıyla tanımlanır. Model güçlü ise, bazı maddelerde görülen uygunsuzluklar modeli sarsmaz. Tek parametreli Rasch ölçüm modelinin[b], model gereklerinden sapmalara karşı oldukça güçlü olduğu bildirilmiştir.[28] Madde-yanıt kuramında bilim adamı eğer uyuşmazlıkla karşılaşmışsa üzerinde çalıştığı mo-

---

[a] Madde-yanıt kuramında, *model-veri* uyuşumu mu yoksa *veri-model* uyuşumu mu konusu tartışmalara neden olmuştur. Bütün tanımlayıcı istatistiksel tekniklerde olduğu gibi MYK'de de geliştirilen model verilere uygun hale getirilmeye çalışılır. Öncelik model üzerindedir. Örneğin, bilim adamı "parametre cimriliği" ilkelerinden hareket eder ve bir model verilere uygun gelmemişse bu kez diğer modeli deneyerek hangi modelin verilere daha uygun olduğunu bulmaya çalışır. Bu süreçte verilerin niteliği her zaman ikinci planda kalır. Fakat bu yöntemin sakıncası kötü verilerin iyi bir teorinin reddedilmesine neden olmasıdır. Rasch yönteminde ise verilerin modele ne ölçüde uygun olduğuna bakılır. Öncelik veriler üzerindedir. Rasch, bilime giden yolun verilerin modele uygun hale getirilmesinden geçtiğini belirtmiştir. Literatürde bu olguya *veri perspektifi* yaklaşımı adı verilir. Konuyla ilgili daha fazla bilgi için *bk.*, Institute for Objective Measurement, "Do Bad Data Refute Good Theory? [Kötü Veri İyi Teoriyi Kovar mı?], <http://www.rasch.org/rmt/rmt114h.htm> (25.01. 2004).

[b] Bilimsel dergi editörlerinin bir bölümü Rasch ölçüm modelleri için "madde-yanıt kuramı" teriminin kullanılmasını istemezler.

delde değişiklik yaparak daha iyi uyuşma gösteren başka bir modele geçer. Rasch yönteminde uyuşmazlıkla karşılaşılmışsa model değiştirilmez; uyuşmazlığa neden olan kişi veya maddeler ölçümden çıkarılarak parametre özelliği yeniden hesaplatılır. Hambleton ve Swaminathan (1985) model-veri uyuşumu analizlerinde uyuşum ölçüsünün üç tür kanıta dayandırılması gerektiğini belirtmişlerdir:[29]

1. Veri seti için önerilen modele ait varsayımların geçerli olduğunun kanıtlanması.

   a. Tek boyutluluk.
   b. Testin hız testi olmaması.
   c. Tahminin en alt düzeyde kalması (1PL ve 2PL için).
   d. Bütün maddelerin eşit ayırt etme gücüne sahip olması (1PL için).

2. Modele ait beklenen özelliklerin gerçekleşme derecesinin saptanması.

   a. Madde parametre tahminlerinin değişmezliği.
   b. Yetenek parametre tahminlerinin değişmezliği.

3. Modele ilişkin tahminlerin doğruluk derecesi.

   a. Tahminlerin doğruluk derecesi için artık değerlerin incelenmesi.
   b. Model-veri uyuşmazlığı halinde tatmin edici başka bir MYK modelinin seçilmesi.

Model uyuşumu grafik yöntemlerle veya istatistiksel hesaplamalarla belirlenir. Grafik yöntemlerde uyuşma grafiği (fit plot) ile Levine ve Williams yöntemi uygulanır. Uyuşma grafiğinde madde özellikleri eğrisi (MÖE) geçerlilik örnekleminden[a] elde edilen ampirik madde özellikleri eğrisi (AMÖE) veya teorik madde özellikleri eğrisiyle (TMÖE) karşılaştırılır. Söz konusu grafik BILOG isimli yazılımla çizilir. Bu kitabın amacının dışında kaldığından konunun ayrıntılarına girilmemiştir.[30]

İstatistiksel olarak ise her bir madde için ki-kare uyuşma istatistiği hesaplanır. Ki-kare uyuşma istatistiğinin uygulanabilmesi için testte en az 20

[a] Geçerlilik örneklemi, analizi yapılacak verilerin başlangıçta iki gruba ayrılması ve bunlardan birinin kalibrasyon örneklemi ve diğerinin ise geçerlilik örneklemi olarak belirlenmesidir. Geçerlilik örneklemi, model-veri uyuşumunun sağlanıp sağlanmadığını görmek amacıyla kullanılır.

madde olmalı ve örneklem hacmi ise, 100'ün üzerine çıkmalıdır.[31] Drasgow, Levine, Tsien, Williams ve Mead'e göre (1995) çeşitli ki-kare değerleri arasında karşılaştırma yapabilmek için düzeltilmiş ki-kare değerini kullanmak daha doğrudur (aktaran, Stark, 2003).[32] Düzeltilmiş ki-kare değeri için, normal ki-kare değeri örneklem büyüklüğü ve onun serbestlik derecesine göre yeniden hesaplanır. Hesaplama sonucunda düzeltilmiş ki-kare / serbestlik derecesi oranı 3'ten küçük ise modelin verilerle iyi bir şekilde uyuştuğuna karar verilir.[33] Uyuşma istatistiği küçük örneklemlerde istikrarsız sonuçlar verdiğinden bu örneklemlerde maddeler oldukça büyük uyuşma değeri ortaya koyabilir. MYK modellerinde ki-kare analizleri sonuçlarıyla uyuşmayan maddelerin testten düşürülmesi gerekir.[34]

Rasch ölçüm modelinde veri-model uyuşumu artık değerlere bakılarak belirlenir. Rasch modelini analiz eden yazılımlardan elde edilen artık/kalan (residual) istatistiği, model-veri uyuşum derecesini gösterir. Bunun için çıktılardaki "ağırlıklandırılmış uyuşum istatistiği" (INFIT) değerleri dikkate alınır. Maddelerin standardize edilmiş uyuşum istatistik değerleri eğer 2'nin üzerindeyse bu maddeler uyuşumsuz olarak saptanır.[35] Uyuşum değerleri 0,8 ilâ 1,2 arasında ise bu maddeler iyi uyuşan göstergeler olarak değerlendirilir.

Model uyuşumu MODFIT istatistiksel analiz programıyla veya Excel'de model-veri uyuşumu analiziyle hesaplanabilir. Ayrıca model-veri uyuşumunu model türüne göre analiz eden daha spesifik yazılımlar da söz konusudur. Örneğin 3PL modeline göre analiz yapan Resid isimli yazılım (Rogers, 1994) bunlardan birisidir.[36] Rasch yaklaşımında model-veri uyuşumu için Winsteps, RUMM ve ConQUEST adlı yazılımlar kullanılır. BILOG ve LOGIST isimli yazılımlarla ise her üç modele ilişkin model-veri uyuşum analizini yapmak mümkündür.

Diferansiyel madde, diferansiyel deste ve/veya diferansiyel test fonksiyonlarının hesaplanması. Diferansiyel madde fonksiyonu (DMF), daha önce belirtildiği gibi, bir testteki maddelerin değişik gruplarda farklı şekillerde çalışma özelliğini yansıtır. DMF bazen gruplar arasında ölçülen özellikle/yetenekle ilgili gerçek farklılığı yansıtırken bazen de ölçüm alanının yeterliliğinin dışında yapay farklılıkları temsil eder. Gerçek farklılıkta maddenin geçerliliği bozulmamışken yapay farklılıkta maddenin yanlı olduğundan söz edilir.[37] Bir maddenin yanlılığını belirlemek için DMF analizi yapmak gerekli olmakla birlikte yüksek DMF değeri her zaman maddenin yanlılığını göstermez. Bir maddenin DMF özelliğine sahip olduğu saptandıktan sonra maddenin içeriği uzmanlar tarafından incelenerek duruma göre testte kalmasına veya testten çıkarılmasına karar verilir.[38]

Büyük ölçekli herhangi bir ölçme ve değerlendirme çalışmasında testin DMF değerlendirmesinin yapılması *Eğitim ve Psikolojik Test Standartları*'nda (Standards for Educational and Psychological Testing – AERA, APA, NCME, 1985) temel bir gereklilik olarak belirlenmiştir.[39]

Klâsik test kuramında DMF'nin araştırılmasında değişik gruplara ait ortalama puanlar arasındaki farklılık veya yanlılık ön plana çıkarken madde-yanıt kuramında grup ortalama puanlarından hareket edilmez. Madde-yanıt kuramının terminolojisi kendine özgüdür. MYK'de *diferansiyel madde fonksiyonu*, aynı veya benzer teta değerine sahip odak ve referans grubu[a] üyelerinin bir maddeye doğru yanıt verme veya bir özelliği onaylama olasılıkları arasındaki farklılık anlamına gelir. Odak ve referans gruplarında madde özellikleri eğrileri birbirinden önemli ölçüde farklı çıkmışsa maddenin DMF'ye sahip olduğu söylenir. Farklılık belirli bir grup madde üzerinde gözlenmişse *diferansiyel deste fonksiyonu* (DDF) olarak isimlendirilir. Konuyla ilgili bir diğer kavram *diferansiyel test fonksiyonudur*. Diferansiyel test fonksiyonu (DTF), her bir grup için madde özellikleri eğrilerinin toplamları arasındaki farktır. Günlük hayatta kararlar test sonuçlarına bakılarak verildiğinden DTF, diferansiyel madde fonksiyonundan daha önemlidir. Bilim adamı test maddelerinin farklı gruplarda değişik bir şekilde çalışıp çalışmadığını görmek için DMF / DDF / DTF analizlerinden yararlanır. Diferansiyel madde / deste / test fonksiyonu pilot araştırma verileri üzerinde yapılır. Pilot araştırma veya ön araştırma sırasında bir maddeye bir grup diğer gruptan farklı bir biçimde yanıt veriyorsa o maddenin yanlı olabileceğini söyleriz. Testte tek bir maddenin yanlı olması sorun oluşturmaz. Birkaç yanlı maddenin bulunması ise test sonuçları üzerinde fark edilebilir bir sonuç yaratır, fakat yanlı madde sayısı arttıkça test artık kişileri âdil olmayan bir biçimde ölçmeye başlar. Bu nedenle testteki yanlı maddelerin ayıklanması veya testten çıkarılması gerekir. Fakat bu konuda dikkatli olunmalıdır. İyi yapılandırılmış bir testte eğer çok sayıda DMF içeren madde varsa bu maddelerin testten çıkarılmasıyla testin yapısal geçerliliği düşebilir. Bu gibi durumlarda hem yapısal geçerlilik azalır hem de maddelerin parametre tahmin değerlerinin gücü zayıflar.

Diferansiyel madde fonksiyonunun ortaya çıkarılması. DMF'nin belirtileri, ölçüm yapılan odak grubun örnek alınan referans gruba göre daha iyi

---

[a] Odak ve referans grubu tanımlamaları izafîdir. İncelemeye alınan iki gruptan herhangi biri odak grubu olarak tanımlanabilir. Literatürde daha çok azınlık grupları odak, çoğunluk grupları ise referans grubu olarak nitelendirilmiştir. Cinsiyet faktörünün incelendiği bir araştırmada grupların büyüklükleri eşitse kendilerine haksızlık yapıldığı düşünülen grup odak ve karşılaştırma yapılan grup ise referans grubudur.

(veya daha kötü) sonuçlar alındığının gözlenmesi halinde ortaya çıkar. Ancak bu belirtilerin anlamlı olabilmesi için ölçüm yapılmadan önce her iki grubun, arka plandaki yetenek/bilgi/özellik düzeyi açısından tam olarak eşitlenmesi gerekir. Eşitleme işlemi için genellikle gruplarda kişilerin testten elde ettikleri toplam puanlar dikkate alınır. Lise ikinci sınıftaki kız öğrencilerinin lise üçüncü sınıftaki erkek öğrencilerle karşılaştırılması veya toplam puanları farklı olan kişilerin birbirleriyle karşılaştırılması anlamsızdır. Bir maddenin denkleştirilmiş odak ve referans gruplarında klasik yöntemdeki *p* güçlük indeks değerleri farklı çıkmışsa DMF'nin ilk belirtisi elde edilmiştir. Daha kapsamlı bir inceleme için klâsik analizler yerine MYK kapsamındaki analiz yöntemleri ile (ki bu yöntemde teta değerleriyle bir maddenin odak ve referans gruplarında daha zor veya daha kolay olup olmadığına bakılır) kontenjan tablolarıyla yapılan modelleme yaklaşımları ve regresyon analizi yöntemleri uygulanır.

Diferansiyel madde fonksiyonunu türleri. Madde-yanıt kuramında DMF Mellenbergh (1982) tarafından tanımlandığı şekliyle *biçimli* ve *biçimsiz* olmak üzere iki grupta değerlendirilir (aktaran, Dickinson ve Coates, 2003).[40]

Bir grupta bir maddeye doğru yanıt verme olasılığı tüm yetenek düzeylerinde diğer gruptan daha yüksek ise buna *biçimli DMF* adı verilir. Biçimli DMF için grupların beta katsayıları karşılaştırılır. Maldonado ve Greenland'a göre iki grup arasında ,10'dan fazla bir farklılık varsa verilerde biçimli DMF olduğuna karar verilir (aktaran, NACC).[41] Biçimli DMF bütün yanıtlar üzerinde etkilidir.

Biçimsiz DMF ise, sadece bazı yanıtları etkiler. Örneğin, bir grupta bir maddeye doğru yanıt verme olasılığı bütün yetenek düzeylerinde diğer gruptan daha yüksek değildir.[42] Bazı yetenek düzeylerinde yüksek iken diğerlerinde düşük olabilir. Maddelerin biçimsiz DMF özelliğine sahip olup olmadığını belirlemek için anlamlılık testi yapılır ve sıfır hipotezi grupların yetenek düzeyleri arasında fark yoktur şeklinde belirlenir (H_0: $b = 0$; H_1: $b \neq 0$). Eğer b parametre değeri sıfıra eşit değilse sıfır hipotezi ret edilir ve maddede *biçimsiz DMF* bulunduğuna karar verilir.[43] Biçimli ve biçimsiz DMF tanısı, daha çok lojistik regresyon analizleri sonucu belirlenir. Bu analizleri el ile yapmak zor olduğundan bu amaçla hazırlanmış bulunan EZDIF, LINKDIF, SIBTEST, RUMM, QUEST, WINSTEPS, Bilog MG gibi istatistiksel analiz programlarından yararlanılır.

Diferansiyel madde fonksiyonunu tespit etmeye yönelik analiz yöntemleri. Diferansiyel madde veya diferansiyel test fonksiyonunu teşhis etmek

için değişik prosedürlerden ve değişik istatistikî analiz yöntemlerinden yararlanılabilir. Bu yöntemler üzerinde bilim adamları tam bir fikir birliği içinde değildirler. Yöntemlerden bir bölümü madde-yanıt kuramı çerçevesinde çalışırken diğerlerinde lojistik regresyon ve kontenjan tabloları temel alındığından söz konusu analizler bu amaçla yazılmış özel istatistikî analiz programlarıyla test edilir. Örneğin, bir sınıflamada verilerin parametrik ve nanparametrik olmasına göre analiz yöntemleri aşağıdaki gibi ele alınmıştır.[44]

- Diferansiyel madde fonksiyonu prosedürleri.

  Parametrik
  1. Lord ki-kare analizi.
  2. Olasılık oranı testi.
  3. İşaretlenmiş ve işaretlenmemiş alanlar yöntemi.

  Nanparametrik
  1. SIBTEST.
  2. Mantel-Haenszel.

- Diferansiyel test fonksiyonu prosedürleri.

  Parametrik
  1. Raju'nun DFIT Yöntemi.

  Nanparametrik
  1. SIBTEST.

DMF analizini yapmak için literatürde en çok lojistik regresyon analizi ve Mantel-Haenszel ki-kare prosedürü kullanılmıştır. Bunun dışında araştırmacılar son yıllarda DMF analizlerini yapacak yeni ve daha ayrıntılı yöntemler geliştirmeye başlamışlardır.

Lojistik regresyon analizi. Lojistik regresyon analizi literatüre Swaminathan ve Rogers (1990) tarafından tanıtılmıştır ve bu prosedür *logit* kavramına dayanır (aktaran, Chiang ve Lam, 2004).[45] Logit, bir maddeye doğru olarak yanıt verme olasılığının yanlış olarak yanıt verme olasılığına olan oranın logaritmik değeridir.

Tarihsel olarak lojistik regresyon analizleri hep ikili veriler üzerinde yapılmıştır. Lojistik regresyon, odak ve referans grubundaki üyelerin belirli bir ölçüm kriteri çerçevesinde bir maddeye doğru yanıt verme olasılığını

test etmeye dayanır. Bağımlı değişken, genellikle ölçek/test maddesinin başarılı/başarısız şeklinde kodlandığı değerlerden oluşur. Lojistik regresyon analizinde bağımsız değişkenler ise ikili veya sürekli veri niteliğindedir. DMF analizinde odak ve referans grubu tanımlaması bağımsız değişkendir. Lojistik regresyon analizini SPSS'te yapmak mümkündür. Bu tekniği uygulamak isteyen araştırmacılara istatistiksel analiz kitaplarına başvurmaları önerilir.

Lord ki-kare istatistiği. Lord ki-kare istatistiği Eşitlik 13-14'deki formül çerçevesinde hesaplanır.[46]

$$\chi^2 = v_i \sum^{-1} v_i . \qquad (13\text{-}14)$$

Formüldeki v_i i'ninci maddenin odak ve referans gruplarında tahmin edilen parametre değerleri arasındaki farklılığı simgeleyen vektördür. $Sigma_i$ ise, madde parametre tahminlerinin varyans-kovaryans matrisini simgeler. Analizde ,05 anlamlılık düzeyi ve 3 serbestlik derecesinde "DMF yoktur" hipotezi için gözlem değerleri kritik değerle karşılaştırılır. Lord ki-kare analizini yapmak için aşağıdaki adımlar atılır:[47]

1. BILOG yazılımı kullanılarak odak ve referans grubunda maddenin parametreleri ve kovaryansları hesaplanır.
2. Metrik birliği sağlamak için odak grubunun parametre değerleri referans grubunun parametre değerlerine dönüştürülür. Maddelerin parametre puanlarını eşitlemek için bu amaçla hazırlanmış bulunan EQUATE99 isimli yazılımdan yararlanılır.
3. Dönüştürülmüş parametre ve kovaryans değerlerinden hareket edilerek Lord ki-kare istatistiği yapılır.
4. Analiz sonucuna göre DMF maddeleri belirlendikten sonra, DMF içermeyen maddeler için metrik eşitleme işlemi yeniden yapılır.
5. Bundan sonra 2. ve 4. maddeler arasındaki işlemler DMF içermeyen aynı maddeler elde edilinceye kadar tekrarlanarak devam eder. Çünkü, bazen DMF içeren maddeler çıkarıldıktan sonra başka bir maddenin DMF içermesi söz konusu olabilir.

Lord ki-kare analizinde görülen tekrarlamalı süreçlerin ITERLINK isimli yazılımdan yararlanmak suretiyle yapılabileceği, yazılımda *tekrarlamalı bağlama yönteminin*[a] 2PL ve 3PL modellerinde başarıyla uygulandığı belirtilmiştir.[48]

Mantel-Haenszel ki-kare prosedürü. Mantel ve Haenszel (1959) tarafından geliştirilen ve Holland ve Thayer (1988) tarafından kolej giriş testlerinde madde yanlılığını belirlemek için uyarlanan bu uygulama, küçük örneklemlerle çalışılabilmesi nedeniyle pek çok araştırmacı tarafından tercih edilmiştir, fakat yöntemin zayıf tarafı biçimsiz DMF'ye karşı duyarlı olmamasıdır.[49] MH prosedüründe odak ve referans gruplarının her ikisi de kendi içinde gösterilen performansa göre iki ayrı alt gruba bölünür.[b] Denkleştirilmiş iki gruptaki performans daha sonra *olasılık oranı* değerleri dikkate alınarak karşılaştırılır. Olasılık oranı, bir gruptaki kişilerin bir soruya doğru yanıt verme olasılığının diğer gruptaki üyelerden yüksek olup olmadığını belirler. Bilim adamları olasılık oranını, daha kolay yorumlayabilmek için "delta ölçeği" adı verilen bir indeks değerine dönüştürmüşlerdir. DMF istatistiği, delta ölçeği üzerindeki *farklılık* olarak bilinir. Bu farklılık literatürde kısaca MH F-DMF değeri olarak tanımlanmıştır. MH F-DMF değeri eksi sonsuzdan artı sonsuza uzanır ve 0 değeri DMF bulunmadığını gösterir. MH F-DMF 1,00'in anlamı maddenin analize alınan odak grubundaki üyeler için referans grubundaki üyelere göre 1 delta değeri kadar daha zor olduğudur.[50] MH prosedürü ki-kare istatistiğine dayanır. Ki-kare istatistiği için iki nominal değişken belirlenir (*bk.*, Tablo 13-5). Bunlardan birincisi gruplama değişkenidir ki literatürde bu değişkenin düzeylerini tanımlamak için "tabaka" terimi kullanılmıştır; ikincisi ise kişilerin başarılı ve ya başarısız olduklarını gösteren değişkendir. Ki-kare analizi sonucunda olasılık değeri ,05'ten düşük çıkmışsa sıfır hipotezi reddedilerek grupların doğru ve yanlış yanıtları arasında istatistiksel olarak anlamlı bir farklılık olduğuna karar verilir. Diğer bir deyişle söz konusu madde yanlıdır.

---

[a] Linking. Maddeler arasında eşitlik veya denklik sağlama işlemi.

[b] Bazı yazarlar, grupları denkleştirme işleminde yetenek düzeylerini dikkate alarak belirli puan aralıklarında üç ilâ beş düzeyden oluşan tabakalar yapılabileceğini belirtmişlerdir. Böylece her tabakadaki bireylerin yetenekleri birbirine eşitlenmiş olur.

Tablo 13-5. Mantel-Haenszel Ki-kare Analizi İçin Veri Tablosu

Grup	*Doğru (1)*	*Yanlış (0)*	*Toplam*
Referans	A_j ($P_{Rj}1$)	B_j ($P_{Rj}0$)	
Odak	C_j ($P_{Fj}1$)	D_j($P_{Fj}0$)	
Toplam			Tj

Not. Aj = Gözlem sayısı, P_{Rjl} = Örtük (gizli) olasılık.

MH testinde olasılık oranıyla ilgili sıfır hipotezi, olasılık oranı = 1 (DMF yoktur) şeklinde belirlenir.[51] Olasılık oranının 1 olması, gruplar arasında cinsiyet, ırk, kültür gibi faktörlerin (hangisi inceleniyorsa), elde edilen puanlar üzerinde herhangi bir etkisinin olmadığını gösterir. Oranın 1'den büyük olması örneğin, cinsiyet faktörü söz konusuysa, maddenin cinsiyet ayrımcılığına yol açtığını ifade eder. Diferansiyel madde fonksiyonu etkisinin büyüklüğünü belirlemek için *olasılık oranı* değerinin (odds-ratio) iyi bir gösterge olduğu belirtilmiştir. Olasılık oranının 3,6 çıkmış olması test sonuçlarının odak grubundaki kişiler için 3,6 kat daha zor olduğu anlamına gelir. Analiz sonucunda elde edilen yüksek MH alfa değerleri eşit yeteneğe sahip odak ve referans gruplarında maddenin farklı bir şekilde çalıştığını gösterir. Peteren (1987), Mantel-Haenszel istatistiği sonucunda test maddelerini üç kategori içinde sınıflandırmıştır (aktaran, Habing, 2003):[52] Bu sınıflandırmada, A maddeleri "çok az DMF içerenler", B ve C "çok fazla DMF içerenler" ve D puanları "DMF'nin miktarını ölçen" değerler olarak belirlenmiştir. A ve D'de DMF yoktur, D puanı 1'den küçükse istatistiksel olarak anlamlı sayılabilecek bir DMF yok demektir. B ve C şeklinde sınıflandırılan maddeler daha sonra odak ve referans grubundan kişilerin de yer aldığı panel grubunda yeniden değerlendirmeye alınır.[53] Bu değerlendirmenin sonunda söz konusu maddeler testte bırakılır veya testten çıkarılır. Mantel-Haenszel istatistiğinde pozitif değer DMF'nin odak grubunun, negatif değer ise referans grubunun aleyhine olduğunu gösterir.

MH testini SPSS'te yapmak için *k*x2x2 veri matris tablosundan yararlanılır. Bunun için birinci değişken karşılaştırma yapılan grupları temsil etmek üzere 1 ve 2 şeklinde, ikinci değişken ise maddede başarılı ve başarısız olanları göstermek üzere 1 ve 0 şeklinde kodlanır. SPSS'te olasılık oranı tablosundaki değerler; sıfır hipotezi ret edildiği, gruplar arasında istatistiksel olarak anlamlı bir farklılık bulunduğu zaman kullanılır. Amacı fark-

lılığın büyüklüğünü ortaya koymaktır. MH testini yapmak için araştırmacılar bu amaçla geliştirilmiş özel istatistiksel analiz yazılımlarından da yararlanabilirler. Bunlardan biri Fidalgo (1994) tarafından geliştirilen ve iki aşamalı MH prosedürünü uygulayan MHDIF isimli yazılımdır.[54] Bu yazılımın birinci aşamasında ki-kare değerlerinin hesaplandığı ve ikinci aşamasında ise %5 düzeyinde istatistiksel olarak anlamlı olan maddelerin çıkarılmasıyla geri kalan maddeler üzerinde eşleştirme çalışmaları yapıldığı bildirilmiştir.[55]

SIBTET. Shealy ve Stout (1993) tarafından geliştirilen Sibtest, (Simultaneous Item Bias) nanparametrik nitelikte istatistiksel bir prosedürdür.[56] Bu prosedürle hem bir maddedeki veya maddeler grubundaki DMF'nin miktarı tahmin edilmekte hem de DMF testi yapılmaktadır. Sibtest'in regresyon düzeltme tekniğini kullanarak şişkin çıkan Tip I hatasını kontrol ettiği bildirilmiştir. Yöntemin kernel düzleştirme metodunu kullanarak hem çapraz DMF'yi incelemek için ve hem de lokal DMF'yi analiz etmek için etkili bir araç olduğu bildirilmiştir. Prosedür çok dereceli maddeler kadar, çok boyutlu ölçekler için de kullanılabilmektedir.[57] Çok boyutlu ölçeklerde "hedef yetenek" boyutunun dışında yardımcı veya ikincil öneme sahip boyutlar da değerlendirmeye alınır. DMF özelliği saptanan maddelerin bazen ikincil boyutu ölçtüğü saptanmıştır.

DMF değerlendirmesinde Tip I hatasının kontrol edilmesi. Diferansiyel madde fonksiyonunun değerlendirilmesinde Tip I hatasının kontrol edilmesi gerekmektedir. Test maddelerinin üretilmesi maliyetli olduğundan bir maddenin gereksiz yere iptal edilmesi kaynakların israf edilmesi anlamına gelir. DMF'ye neyin veya hangi faktörlerin yol açtığını araştıran araştırmacılar Tip I hatasının büyük ölçüde karışıklığa neden olduğunu bulmuşlardır. Bu nedenle DMF araştırmalarında alfa hatasının kontrol altına alınması önemli bir hedef olarak görülmüştür.

DMF sonuçlarının değerlendirilmesi. Bir madde/test herhangi iki gruba uygulandığında gruplar arasında az da olsa belirli düzeyde bir farklılık ortaya çıkar. Örneğin, test maddesi bir gruba göre biraz daha zor ve diğer gruba göre ise biraz daha kolay değerlendirilmiş olabilir. DMF analizinin en önemli yönü, gruplar arasındaki farklılıkların önemsiz veya aslına çok benzeyen türden bir kalibrasyon farklılığına dayanıp dayanmadığını belirlemektir.[58] Du'ya göre (1995) DMF için aşılması gereken üç engel vardır.[59] Bunlardan birincisi DMF'nin istatistiksel olarak anlamlı olup olmadığıdır. Bunu belirlemek için bazen iki maddenin kalibrasyon değerlerini

sadece gözle taramak dahi yeterlidir. Örneğin, odak grupta 10 üye varsa ve ikili bir maddenin güçlük derecesinin 1,0 logit daha zor olduğu söylenmişse ayrıca anlamlılık testi yapmaya gerek yoktur. Çünkü 10 kez yapılacak tekrarlamada 1,0 logit'in standart hatası 0,6 logit çıkacak ve bu değer de t <1,96'dan küçük olduğundan farklılık açık bir şekilde belirlenmiş olacaktır. Ancak, diferansiyel madde fonksiyonu asimetrik olduğundan istatistiksel olarak anlamlı çıkmaması gerçek anlamda DMF olup olmadığı konusunda araştırmacıyı kuşku içinde bırakır. İstatistiksel anlamlılık testinin, kendi bünyesindeki zayıflık nedeniyle araştırmacıya istediği bilgiyi tam olarak temin edemeyeceği bildirilmiştir.[60] Du'ya göre aşılması gereken ikinci engel, DMF'nin gerçekten önemli sonuçlara yol açıp açmadığının incelenmesidir. Örneğin bir madde 10.000 erkek ve 10.000 bayana uygulandığında aralarında bayanlar lehine ,1 logit fark bulunduğu saptanmış bulunsun. Acaba böylesine çok küçük bir logit değeri için ayarlama veya uyarlama yapmak gerekir mi? Du, bu sorunun yanıtı için ,1 logitin standart hatasının istatistiksel olarak ,3 logit olması nedeniyle ayarlama yapmaya gereksinim varmış gibi gözüktüğünü söylemektedir. Fakat, DMF'nin tam hedef madde üzerindeki puan etkisi ,3 logit iken hedef maddenin uzağındaki etkisi ,1 logit olduğundan söz konusu küçük farklılıklar grup performansları hakkındaki düşüncelerimizi değiştirmez. Böyle bir durumda odak grubunda yer alan kişilerin en fazla %3'ü sadece bir puan cezalandırılmış olur. Üçüncü engel, "DMF'nin gerçek mi yoksa tesadüfî / arızî mi olduğu" sorusudur. Tek bir istatistiksel test, gerçek ve tesadüfî etki arasındaki farkı ortaya koyamaz. Bir parayı üç kere attığımızda üçünde de yazı gelmesi o paranın yanlı olduğunu göstermez. Bunun için odak grup çeşitli şekillerde yarılara bölünerek her bir alt grupta DMF'nin devam edip etmediğine bakılır. Eğer maddenin DMF özelliği devam ediyorsa farklılığın gerçek olduğu sonucuna varılır, devam etmiyorsa farklılık tesadüfîdir veya sadece bir tür örneklem kazasıdır.[61]

DMF yazılımları. Diferansiyel madde fonksiyonunu ölçmek için değişik yazılımlardan yararlanılır. Bunlardan biri DIFCOMP isimli programdır. DIFCOMP iki boyutlu madde-yanıt kuramı çerçevesinde hazırlanan bir testteki tek bir madde için veya belirli sayıdaki madde grubu için DMF analizi yapar. Yazılım odak ve referans grubu için iki boyutlu madde-yanıt fonksiyonlarını ortaya koyar. DIFSIM, çok boyutlu testlerde DMF analizini yaparken SIBTEST'in diferansiyel madde ve diferansiyel deste fonksiyonunu analiz ettiği bildirilmiştir. SIBTEST diferansiyel madde fonksiyonu ve diferansiyel deste fonksiyonunu çok boyutlu DMF modeline göre hesaplar. Aynı grupta yer alan diğer iki yazılım Crossing SIBTEST ve

Polytomous SIBTEST adlarıyla bilinir.[62] MS-DOS altında birbirinden bağımsız olarak çalışan bu yazılımlar son yıllarda yoğun talep üzerine birleştirilerek tek bir paket halinde Windows ortamına alınmış ve DIFPACK olarak isimlendirilmiştir. Bir diğer yazılım BILOG-MG3 adını taşıyan madde-yanıt kuramı analiz programıdır. Lojistik regresyon ve MH prosedürleri ise, SAS isimli istatistiksel analiz programının PROC FREQ adlı modülü kullanılarak yapılabilir. Ayrıca bu analizlerin bir bölümünü SPSS ortamında da yapmak mümkündür. Lojistik regresyon tekniğini uygulamak isteyen araştırmacıların yararlanabilecekleri bir diğer program DIFdetect adını taşır. DIFdetect'in grup tanımlamasını ikili veri haline getirmeden de çalışabildiği, ikili veri niteliğinde olmayan demografik değişkenlere bağlı olarak da analiz yapabildiği bildirilmiştir. Bu yazılımlar hakkında daha fazla bilgi edinmek isteyen okurlara İnternet kaynaklarına başvurmalarını öneririz.

Baz ölçüm değerlerinin belirlenmesi. Baz ölçüm sonraki ölçümlere temel oluşturacak ilk ölçümdür. Pilot araştırma çalışmasının yapılması, model uyuşumunun belirlenmesi ve DMF analizinden sonra belirli bir sayıya indirilmiş olan maddeler bu kez MYK çerçevesinde "kalibrasyon örnekleminde" sınanır. Uygulama sonucunda her bir maddenin *madde özellikleri eğrisi* çıkarılır ve ayrıca bu maddeler Birnbaum paradigması[a] çerçevesinde incelenerek bilgi fonksiyonu yüksek olanlar belirlenir. Madde özellikleri belirli niteliğe sahip olanların ve madde bilgi fonksiyonu yüksek olan maddelerin seçilmesiyle baz ölçüm temelinde test kalibre edilmiş olur. Kalibrasyon çalışmasının sonucunda elde edilen maddelere *kalibre edilmiş madde bankası* adı verilir.

Madde özelliklerinin / parametrelerinin tahmin edilmesi. Madde özellikleri veya parametreleri seçilen ölçüm modeline göre belirlenir. Ölçüm modelleri çoğunlukla tek parametreli, iki parametreli veya üç parametreli olarak belirlenir.

Tek parametreli modellerde madde özelliklerinin belirlenmesi. Tek parametreli (1PL – Rasch) modelinde izlem grafiği ile birlikte *madde haritası* kullanılır. Rasch yönteminde maddelerin *ayırma özelliklerinin* eşit

[a] Birnbaum paradigması: Arzulanan özelliklere sahip bir test oluşturmak için 1968'de Birnbaum tarafından önerilen "madde bilgi fonksiyonu" kavramı ve bu fonksiyonun yorumlanması.

olduğu kısıtlaması nedeniyle maddeler sadece güçlük özelliğine, *b*[a] göre teta boyutu üzerinde sıralanır. Bilim adamı bu aşamada maddeleri belirli bir kesim puanına göre mi yoksa ortalama etrafında mı dağılacağına karar verir. Eğer maddelerin bir kesim puanı etrafında dağılmasını istiyorsa bu puanı temel alarak maddelerin bu puanın etrafında dağılmış olanlarını alır. Örneğin, –1,0 logit yetenek düzeyini kesim puanı olarak belirlemişse maddelerin güçlük derecesi açısından bu puan çerçevesinde dağılmasını isteyecektir. Yetenek düzeyi, kesim puanı yerine bir aralık şeklinde de belirlenebilir. Örneğin, başka bir araştırmacı yetenek düzeyini –,50 ilâ +1,50 arasında veya –2,0 ilâ +2,0 arasında belirleyebilir. Böyle bir durumda maksimum bilgi düzeyi 2,5 olur. Ancak tahmin hatası nedeniyle, hesaplanan teta değerleri *gerçek teta* değerlerini tam olarak yansıtmaz. Büyük örneklemlerde ve hatta uzun testlerde dahi ampirik izlem eğrisi, "gerçek" yanıt fonksiyonundan biraz daha farklı çıkabilir.[63]

Tek parametreli modellerde maddelerin güvenilirliği tahmin ve/veya şans faktörüne bağlı olarak işaretleme yapmanın en alt düzeye indirilmesine bağlıdır. Eğer bu konuda katılımcıları etkilemek mümkün değilse diğer modelleri düşünmek daha doğru olur.

Tek parametreli modellerde madde yanlılığı, madde güçlük değerlerine bakılarak belirlenir. Rasch yönteminde bir maddenin güçlük tahmin değeri gruplar arasında yarım logitlik bir kayma göstermişse (–2,5 ilâ +2,5 değerleri arasında %10'luk bir değişimi yansıtır) bu kayma dikkat çekici olarak değerlendirilir. Wright ve Douglas'a göre (1975) kayma yarım logitten daha az ise bu kaymanın ölçüm verilerinin doğruluğu üzerindeki etkisi çok azdır (aktaran Lober, 2003).[64]

2PL modelinde parametre belirlemesi. İki parametreli modellerde maddelerin güçlük özelliğinin yanı sıra ayırt etme özellikleri de dikkate alınır. Her bir maddenin iki indeks değeri vardır: güçlük değeri ve ayırt etme değeri. Bu modelde madde özelliklerinin eğimi farklılaşabilir. İki parametreli Birnbaum modeli Rasch modelinde görülen log-lineer model (logaritmik doğrusal) olarak isimlendirilemez, çünkü bu modelde logaritmik yanıt olasılıkları model parametrelerinin doğrusal fonksiyonu değildir. Bu yöntem sayesinde kişilerin yeteneklerine ilişkin tahmin değerleri incelenirken güçlük indeksi yerine ayırt etme indeksi *a*, daha fazla önem kazanmıştır. Bir

[a] Literatürde güçlük özelliği genellikle *b* harfiyle gösterilmiştir. Ancak, Rasch modeli üzerinde odaklanan yazarlar bu özelliği göstermek için farklı simgeler kullanmışlardır. Örneğin, bu yazarlar güçlük indeksi için delta δ, kişi/yetenek parametresi için ise beta β_j simgesini kullanırlar.

maddenin kalitesi hakkında karar verirken güçlük indeksinden çok maddenin ayırt etme özelliğine bakmak gerekir. Ayırt etme değeri diğer faktörler eşit olması koşuluyla gizli özellik θ hakkında daha fazla bilgi verir. Bu nedenle yetenek testlerinde iki ve üç parametreli modeller daha fazla kullanılır.

İki parametreli modellerde bir dizi maddenin içinde *a* parametresi yüksek olan maddeler dikkate alınır. Ayırma parametresinin dikliğine göre test özellikleri eğrisi maksimum bilgi fonksiyonunu farklı yetenek düzeyinde sağlıyor olabilir. Maddenin ayırt etme özelliği arttıkça bilgi fonksiyonu da artar. Araştırmacı teta boyutu üzerinde çeşitli yetenek düzeylerine sahip yeni maddeler ilave ederek veya çıkararak test bilgi fonksiyonundaki gelişmeleri izler. En yüksek bilgi fonksiyonunu veren maddelerden oluşan set güvenilirliği en yüksek olan testtir.

üç parametreli ölçüm modellerinde parametre belirlemesi. Üç parametreli modellerde *güçlük* ve *ayırt etme* özelliklerine ilave olarak yeni bir faktör daha incelemeye alınır. Bu faktör *c* simgesiyle gösterilen şans/tahmin parametresidir. Bir ve iki parametreli modellerde izlem eğrisinin alt ucu veya başlangıç noktası sıfıra sabitlenmişken üç parametreli modellerde izlem eğrisinin başlangıç noktası etkili minimum değeri verecek şekilde sıfırdan farklı bir değerdir. Soruların doğru yanıtlarını bilmeyen kişiler boş bırakmak yerine tahminde bulunacaklarından izlem eğrisinin başlangıç noktası hiçbir zaman sıfır olmaz. Kullanılan test bir başarı testi yerine bir kişilik testi dahi olsa sosyal beğenirlik gibi nedenlerle kişiler hiç bir zaman bütün yanıtlara sıfır başlangıç noktası yaratacak bir şekilde cevap vermezler.[65] Grafikte *c* noktası olasılık boyutunu belirli bir noktada keser. Tahmin parametresi *c*'nin 0 olması hiç tahmin yapılmadığı anlamına gelir. Üç parametreli ölçüm modelleri başarı testlerinde daha çok kullanılır. Örneğin, iki dereceli olarak kodlanan başarı (bilgi ve bilişsel yetenek) testlerinin kalibrasyonu üç parametreli lojistik model (3PL) çerçevesinde yapılır. Yapılan çok sayıda araştırma 3PL modelinin bilişsel testlerle çok iyi uyuştuğunu göstermiştir.[66] Bilişsel olmayan kişilik ve ilgi envanterleri gibi testlerde ise hangi modelin temel alınması gerektiği konusunda bilim adamları tam bir mutabakat içinde değillerdir. Likert ölçekleri gibi çok dereceli maddelerin kalibrasyonu için ise *Samejima Derecelendirilmiş Yanıt Modeli* kullanılır (Samejima Graded Response Model).

Madde özellikleri eğrilerinin incelenmesi. Madde parametreleri, rakamsal değerlerin yanında görsel olarak "madde özellikleri eğrisi" veya bir başka adlandırmayla "izlem grafiği" ile incelenir.

Rasch ve 1PL modelinde madde özellikleri eğrisi. Rasch ve 1PL modelinde madde özellikleri eğrileri birbirlerine paralel bir seyir izler. Grafikte sadece maddelerin güçlük *b* değerleri dikkate alınarak izlem grafiği çizilmiştir. Aynı grafiğin üzerinde birden fazla maddenin izlem eğrisi gösterilebilir. Araştırmacı tarafından belirlenen delta aralığındaki maddeler güvenilir olarak kabul edilir.

İki parametreli modellerde madde özellikleri eğrisi. İki parametreli modellerde madde özellikleri birbirlerine paralel değildir. İzlem eğrileri birbirini kesebileceği gibi eğimleri de farklı olabilir. Eğimi dik olan izlem eğrileri ayırt etme gücü yüksek olan maddeleri temsil eder. Ancak maddenin güvenilirliği sadece izlem eğrisiyle değil aynı zamanda test bilgi fonksiyonuyla birlikte değerlendirilerek belirlenir.

Üç parametreli modellerde madde özellikleri eğrisi. Üç parametreli modellerde madde özellikleri birbirlerine paralel olmadığı gibi izlem eğrisinin başlangıç noktası da 0'dan farklıdır. İki parametreli modellerde olduğu gibi maddenin güvenilirliği test bilgi fonksiyonu göz önünde bulundurularak saptanır.

Test bilgi fonksiyonunun incelenmesi. Bir test çok sayıda maddeden oluşur ve testteki her bir madde kendi bilgi fonksiyonuna sahiptir. Yerel olarak bağımsız bir dizi maddenin bilgi fonksiyonu Eşitlik 13-15'deki formülle belirlenir.

$$I(\theta)=\sum_i I_i(\theta)\,. \qquad (13\text{-}15)$$

Maddelerin bilgi fonksiyonlarının toplamı testin bilgi fonksiyonunu oluşturur. Belirli bir θ düzeyinde test bilgi fonksiyonu yüksek çıktığı ölçüde test kişileri iyi ayırt ediyor demektir ve bu testlerde ölçüm hatası daha küçüktür.

Standart hata. Madde-yanıt kuramında maddelere ait standart hata ve madde bilgi fonksiyonu terimleri klâsik test kuramındaki güvenilirlik kavramına benzer.[67] Madde bilgi fonksiyonuna ait yüksek rakamlar ölçümde kesinliğe sahip duyarlı değerleri gösterir. Daha yüksek bilgi fonksiyonu; eğer maddeler daha yüksek ayırt etme *a* değerine ve daha küçük *c* paramet-

re değerine sahipse elde edilir. Şans parametresi *c*, düşük ve ayırt etme parametresi *a*, büyük olan maddelerin bilgi fonksiyonu eğrileri yüksek olur. Harvey'e (2003) göre, klâsik test kuramında ölçümün standart hatasının güvenilirlik katsayısıyla ters yönde ilişkili olmasına benzer şekilde, madde-yanıt kuramında da belirli bir yetenek düzeyiyle θ ilgili standart hata tahmin değeri, madde bilgi fonksiyonuyla ters yönde ilişkilidir.[68]

Madde-yanıt kuramında madde seçimi, maddelerin çıkarılması veya iyileştirilmesi maddenin standart hatası (MSH), testin standart hatası (TSH), madde bilgi fonksiyonu (MBF), test bilgi fonksiyonu (TBF) ve diferansiyel madde özellikleri (DMF) belirlenerek yapılır. Bilim adamı hedeflediği *test bilgi fonksiyonu* ve *test standart hata* değerlerini dikkate alarak oluşturmak istediği ölçeği/testi şekillendirir. Madde-yanıt kuramında teste alınacak maddelerin ayırt etme, *a* özellikleri açısından birbirine benzer olmaları önemlidir. Oysa klâsik test kuramında ayırt etme önemli olmakla birlikte teste alınacak maddelere esas olarak maddenin toplam puanla yüksek derecede ilişkili olmasına bakılarak karar verilmekteydi. Klâsik test kuramında ayırt etme ön koşul değildir.[69]

Sonraki ölçümler için madde seçme. Madde-yanıt kuramında, bilim adamı kalibre edilmiş madde bankasından test sorusu seçerken sonuçta *test bilgi fonksiyonuna* en yüksek katkıyı yapacak maddeleri belirlemeye çalışır. Bu işlem daha çok madde-yanıt modelini test eden istatistikî analiz programları aracılığıyla yapılır. Analiz sonucunda maddelerin bilgi fonksiyonunu gösteren "kemer eğrileri" incelenerek belirlenen yetenek düzeyinde en yüksek bilgi fonksiyonuna sahip olan maddeler teste alınır. Birnbaum ve Lord (1980) madde seçim prosedürünü aşağıdaki gibi belirlemişlerdir (aktaran Hsu, 1993):[70]

1. Test bilgi fonksiyonu için arzu edilen eğrinin biçimini belirleyiniz. Arzu edilen eğri, *hedef bilgi fonksiyonu* olarak adlandırılır.
2. Hedef bilgi eğrisinin altındaki bölümü dolduran madde bilgi fonksiyonu eğrilerini saptayınız.
3. Madde bilgi eğrilerini kümülatif olarak toplayınız. Toplanan madde eğrilerinin hedef test bilgi fonksiyonuna benzer olmasına dikkat ediniz.
4. Hedef test bilgi eğrisinin altındaki bölüm doluncaya kadar madde seçmeye devam ediniz.

Lord tarafından önerilen bu yaklaşımda seçim işleminin el ile yapılmasından dolayı optimum çözüm bulunmasında bir takım güçlüklerle karşılaşılıyordu. Thenuissen (1985) ilk kez "hedef bilgi fonksiyonu" çerçevesinde çalışan matematiksel ikili veri programlama modeli geliştirmiştir. Thenuissen'in amacı, belirli bir yetenek aralığında önceden belirlenen bilgi düzeyinin üstündeki madde sayısını en aza düşürmekti (aktaran Hsu, 1993).[71]

Madde bankasından madde seçmekte farklı bir yaklaşım Van der Linden ve Boekkooi-Timminga (1989) tarafından önerilmiştir (aktaran Hsu, 1993). Önerdikleri "maksimin" adlı model ile hedef test bilgi fonksiyonunun şeklini nispî olarak belirlemişlerdir. Bu yaklaşımda test oluşturan bilim adamı, test bilgi fonksiyonunu kesin bir şekilde belirlemek zorunda olmuyordu.[72] Maksimin modeli test bilgi fonksiyonunu maksimize etme konusunda Thenuissen modelinden daha pratik bir yaklaşım getirmiştir. Aşağıdaki bölümde maddelerin iyileştirilmesi ve seçim süreciyle ilgili olarak Thenuissen'in "minimum test uzunluğu" yaklaşımı, Van der Linden ve Boekkooi-Timminga'nın "minimaks" modeli ile amaç programlama yaklaşımları tanıtım amaçlı olarak özet bilgiler verilerek ele alınmıştır.

Thenuissen'in minimum test uzunluğu yaklaşımı. Thenuissen kendisinden önce Lord tarafından da önerilen "doldurma" yaklaşımını benimsemiştir (aktaran Hsu, 1993).[73] Test tasarımında bilgi fonksiyonunu kullanarak algoritma yaklaşımından hareket etmiş ve test için madde seçiminde doğrusal programlama yöntemini kullanmıştır. Bu yaklaşımda kalibre edilerek madde bilgi fonksiyonu çıkarılmış n sayıdaki test maddesinden, hedeflenen test bilgi fonksiyonunun altındaki alanı tatmin edici bir biçimde dolduracak minimum sayıda maddenin nasıl seçileceği konusu üzerinde durulmuştur. Hedef bilgi fonksiyonu K yetenek düzeyi için belirlenebilmektedir.[74]

Maksimin modeli. Van der Linden, ve Boekkooi-Timminga yaptıkları incelemelerde Thenuissen tarafından geliştirilen yaklaşımın belirli koşullarda pratik sonuçlar vermediğini bulmuşlardır (aktaran Hsu, 1993).[75] Özellikle bilim adamının hedef bilgi fonksiyonunun ne olması gerektiği konusunda yeterince bilgiye sahip olmadığı durumlarda söz konusu yöntemin yetersiz kaldığını görmüşler ve kendi yaklaşımlarını geliştirmişlerdir.[76] Maksimin modelinde test uzunluğu sabitlenerek hedef bilgi fonksiyonu mutlak değerler yerine sabitlenen test uzunluğuna göre belirlenir. Böylece test oluşturucu öncelikle testteki madde sayısını belirlemekte ve daha sonra bu sayıya göre test bilgi fonksiyonunu şekillendirmektedir. Bu modeldeki ana fikir test bilgi fonksiyonunu maksimize edecek maddeleri

seçmektir. Doğrusal programlama modeli ile maddeler belirli bir sayı içinde test edilerek en yüksek bilgi fonksiyonunu verecek bileşim elde edilmeye çalışılır.

Amaç programlama yaklaşımı. Amaç programlama yaklaşımı doğrusal programlama yaklaşımının bir uzantısı niteliğindedir. Doğrusal programlama yaklaşımında belirli kısıtlar altında tek bir amacın maksimizasyonu veya minimizasyonu için çaba harcanırken amaç programlama yaklaşımının çok yönlü ve birbirileriyle çelişen amaçlarla ilgili kararları vermek için kullanılabileceği belirtilmiştir.[77]

ALINTI YAPILAN KAYNAKLAR

[1] S. Auchyn, "Characteristics of Measurement [Ölçümün Özellikleri]," <http://uregina.ca/~sauchyn/geog411/measurement_and_sampling.html> (25.08.2003).

[2] A. García-Pérez, "Fitting Logistic IRT Models: Small Wonder [MYK Uyuşum: Küçük Bir Merak]," 1999, <http://www.ucm.es/info/Psi/docs/journal/v2_n1_1999/art74.pdf> (25.08.2002).

[3] Winsteps, "Quality-control Misfit Selection Criteria [Uyuşumsuzları Seçim Kriterinde Kalite Kontrol]," <http://www.winsteps.com/winman/6_2.htm> (26.08.2003).

[4] L.S. Wang "Reliability."

[5] J.R. Emslie ve G.R. Emslie, "Using Statistical Criteria to Improve Classroom Multiple-Choice Tests: A Worked Example [Sınıf ortamında Yapılan Çoktan Seçmeli Testlerin İyileştirilmesi İçin İstatistiksel Kriterlerin Kullanılması]," (Toronto: Ryerson University, 2002), <http://www.ryerson.ca/lt/resources/trspaper.pdf> (26.08.2003).

[6] Michigan State University, "Item Analysis [Madde Analizi]," <http://www.msu.edu/dept/soweb/itanhand.html> (03.08.2003).

[7] Kent State University, "Interpreting the Reports [Raporların Yorumlanması]," <http://helpdesk.kent.edu/howto/testscoring/stat/> (03.08.2003).

[8] University of Minnesota, "Understanding the Item Analysis Report [Madde Analizi Raporunun Anlaşılması]," <http://www.ucs.umn.edu/oms/fceitemanal.htmlx.> (03.08.2003).

[9] D.M. Roberts "An Empirical Study on the Nature of Trick Test Questions [Hileli Test Sorularının Niteliği Üzerine Ampirik Bir Araştırma]," <www.roberts.ed.psu.edu/users/droberts/papers/TRICK.PDF> (12.09.2003).

[10] Roberts "An Empirical Study on."

[11] P. Schatz, "Bias and Error [Yanlılık ve Hata]," <http://schatz.sju.edu/methods/sampling/bias.html>(25.08.2003).

[12] J.R. Emslie ve G.R. Emslie, "Using Statistical."

[13] D. Burns, "Monday Question [Pazartesi Sorusu]," June 14, 2001, <http://www.feelinggood.com/Guru%20Questions/questions_of_the_week_3/monday_answer%202.htm> (08.08.2003).

[14] mecscore@www.utexas.edu, "MEC Item Analysis [Madde Analizi]," <http://www.utexas.edu/academic/mec/scan/scanitem.html> (03.08.2003).

[15] W. Martin, "Approaches to Evaluating Test Items [Test Maddelerinin Değerlendirilmesiyle İlgili Yaklaşımlar]," <http://www.ipmaac.org/mapac/meetings/1999/fall99/willmartin2.ppt> (13.09.2003).

[16] Data Analysis: Gathering Data – Questionnaires /Surveys [Veri Analizi: Veri toplama- Anketler/Ölçekler]," <www.preciousheart.net/chaplaincy/Auditor_Manual/7questid.pdf> (25.01.2004).

[17] D. Fesenmaier, "Response Rate [Yanıt Oranı]," ty, <http://www.tourism.uiuc.edu/itn/etools/eGuides_survey_responserate.htm> (01.12.2002).

[18] J.V. Brown, "Multivariate Data Analysis [Çok Değişkenli Veri Analizi]," <http://ocean.otr.usm.edu/~jwbrown/chapter_one.htm> (27.10.2002).

[19] A. Klein, "Some Aditional Remarks About Correlation [Korelasyon Hakkında İlave Bazı Görüşler]," <http://www.gseis.ucla.edu/courses/ed231a/lecture/Course4N.doc> (27.08.2003).

[20] M. R. Torabi ve K. Ding, "Selected Critical Measurement and Statistical Issues in Health Education Evaluation and Research [Sağlık Eğitiminin Değerlendirilmesi ve Sağlık Araştırmalarında Ölçüm ve İstatistik Sorunları]," <http://www.kittle.siu.edu/iejhe/paid/1998/number1/TORABI.HTM> (21.09.2002).

[21] Torabi ve Ding, "Selected Critical."

[22] D.G. Banet, "Sample Size Requirements For Testing and Estimating Coefficient Alpha [Testler ve Alfa Katsayısını Tahmin Etmek İçin Gerekli Örneklem Büyüklüğü]," *Journal of Educational and Behavioral Statistics*, Winter 27 (4), 335-341.

[23] Aynı.

[24] G.C. Thornton ve E.R. Oetting, *Exercises in Psychological Testing* [Psikolojik Test Egzersizleri], (London: Harper and Row,1982), 1.

[25] ERIC, "Basic Item Analysis for Multiple-Choice Tests [Çoktan Seçmeli Sorular İçin Temel Madde Analizleri]," <http://www.ericfacility.net/databases/ERIC_Digests/ed398237.html> (07.08.2003).

[26] C. Ho Yu, "Automation and Visualization of Distractor Analysis Using SAS/Graph [SAS/Grafik Araçları Kullanılarak Çeldirici Analizinin Otomasyonu ve Görselleştirilmesğ],"

[27] S. Stark ve O. Chernyshenko "Assessment of Model-Data Fit [Veri-Model Uyuşumu Değerlendirmesğ]," <http://io.psych.uiuc.edu/siop2001/Model_Data%20Fit.ppt> (09.08.2003).

[28] R.M. Smith, "Theory and Practice of Fit [Uyuşum Kuramı ve Uygulaması]," <http://www.rasch.org/rmt/rmt34b.htm.> (11.09.2003).

[29] R.K. Hambleton, H. Swaminathan ve H.J. Rogers, *Fundementals of Item Response Theory* [Madde Yanıt Kuramının Temelleri], (London: Sage), 1991, 55. Ayrıca *bk.*, <http://caacentre.lboro.ac.uk/dldocs/BP2final.pdf> (25.08.2003).

[30] Bu konuda daha fazla bilgi için *bk.*, S. Stark ve d., IRT Modeling Lab, "Assessment of Model-data Fit [Model-Veri Uyuşumunun Değerlendirilmesi]," <http://work.psych.uiuc.edu/irt/mdf_main.asp.> (24.08.2002).

[31] R.G. Downey, "Execise 7 [Egzersiz 7]," <http://www-personal.ksu.edu/~downey/ex7.htm.> (25.08.2002).

[32] S. Stark ve d. IRT Modeling Lab, "Assessment of Model."

[33] Aynı.

[34] R.M. Smith, "Common Oversights in Rasch Studies [Rasch Çalışmalarında Yapılan Genel Hatalar]," <http://www.rasch.org/rn9.htm> (11.09.2003).

[35] M. McAlpine, *A Summary of Method of Item Analysis* [Madde Analizi Yöntemlerinin Özeti], (Glaskow: CCA Centre, 2002), s, 26.

[36] C. Stage, "An Attemp to Fit IRT Models to the Subtest in te sweSAT [sweSAT'da MYK Modellerini Alt Testlere Uyuşturma Çalışması]," <http://www.umu.se/edmeas/publikationer/pdf/enr1996sec.pdf> (24.08.2002).

[37] F. Robin, "Carrying out Empirical Studies of Adapted Examinations [Uyarlanmış Sınavlarda Deneysel Çalışmalar]," <http://www.cesb.org/Carrying%20out%20Empirical.htm> (18.08.2003).

[38] N. Boiteau, Richard Bertrand, Eric Frenette ve Christina Saint, "Stability of IRT-based and non IRT-based DIF procedures [MYK Temelli ve MYK Temelli Olmayan DMF Prosedürlerinde İstikrarlılık]," <http://edtech.connect.msu.edu/Searchaera2002/viewproposaltext.asp?propID=3276> (19.08.2003).

[39] Y. Zhang, "DIF in a Large Scale Mathematics Assessment: The Interaction of Gender and Ethnicity [Büyük Ölçekli Matematik Değerlendirmelerde DMF: Cinsiyet ve Etnik Faktörün Etkileşimi]," <http://www.google.com.tr/search?q=cache:OFEgRI2g7hQJ:tigersystem.net/aera2002/viewproposaltext.asp%3FpropID%3D2591+Mantel-Haenszel+dif+&hl=tr&ie=UTF-8&inlang=tr> (19.08.2003).

[40] Terry L. Dickinson ve Glynn D. Coates, "Differential Item Functioning: Item Response Theory and Confirmatory Factor Analysis [Diferansiyel Madde Fonksiyonu: Madde Yanıt Kuramı ve Teyit Edici Faktör Analizi]," <http://courses.lib.odu.edu/psychology/tdickins/siop03.pdf> (25.01.2004).

[41] University of Washington, NACC, "Uniform DIF detection in DIFdetect [DIFdetect Programında Biçimli DMF Araştırılması]," <http://www.alz.washington.edu/DIFDETECT/uniform.html> (14.09.2003).

[42] Revista Electrónica de Metodología Aplicada, "Introduction [Giriş]," <http://www.uniovi.es/user_html/herrero/REMA/v5n1/a2/holanda1.html> (18.08.2003).

[43] University of Washington, "Non-uniform DIF detection in DIFdetect [DIFdetect'te Biçimsiz DFF Araştırılması]," <http://www.alz.washington.edu/DIFDETECT/nonuniform.html> (18.08.2003).

[44] Sahsa, "Detection of Differential Item/Test Functioning (DIF/DTF) Using IRT [MYK Çerçevesinde Diferansiyel Madde/Test Fonksiyonunun Teşhis Edilmesi]," <http://io.psych.uiuc.edu/siop2001/Detection%20of%20Differential%20Item2.ppt> (13.09.2003).

[45] L.H. Chiang ve P. Lam, "Differential Item Functioning... [Diferansiyel Madde Fonksiyonu]," <http://www.aare.edu.au/96pap/lamp96.613> (25.01.2004).

[46] IRT Modeling Lab, "Procedure for DIF Detection Using Lord's Chi-Square [Lord Kikare Yöntemiyle DMF Tedkiki]," <http://work.psych.uiuc.edu/irt/dif_iterlink.asp> (19.08.2003).

[47] IRT Modeling Lab, "Procedure for."

[48] Aynı.

[49] Revista Electrónica de Metodología Aplicada, "Introduction."

[50] Educational Testing Service, "Graduate Record Examinations [Mezuniyet Sınavları]," <http://www.etskorea.or.kr/download/994950.pdf> (20.08.2003).

[51] J.L Matthews, Kevin Larkin ve George A Johanson, "Using Differential Person Functioning To Detect Aberrant Response Patterns in A Standard-Setting Session For Teacher Licensure [Tipik Olmayan Yanıt Biçimlerini Saptamak İçin Diferansiyel Kişi Fonksiyonunun İncelenmesi]," <http://edtech.connect.msu.edu/Searchaera2002/viewproposaltext.asp?propID=5059> (20.08.2003).

[52] P. Habing, "R Templates [R Şablonları]," <http://www.stat.sc.edu/~habing/courses/778rS02.html> (20.08.2003).

[53] College Board, "Differential Item Functioning [Diferansiyel Madde Fonksiyonu]," <http://www.collegeboard.com/ap/techman/chap4/differential.htm> (19.08.2003).

[54] Ángel M. Fidalgo, "Effects of Amount of DIF... [DMF Miktarının Etkisi]," <http://www.uni-landau.de/~agmunde/mpr/issue11/art3/fidalgo.pdf> (25.01.2004).

[55] Ángel M. Fidalgo, Gideon J. Mellenbergh ve José Muñiz, "Effects of Amount of DIF, Test Length, and Purification [DMF Miktarının Etkisi, Test Uzunluğu ve Saflaştırma]," <http://www.mpr-online.de> (20.08.2003).

[56] S. Stark ve d., "Detection of DIF Using the SIBTEST Procedure [SIBTEST Prosedürü Kullanılarak DMF'nin İncelenmesi]," <http://work.psych.uiuc.edu/irt/dif_sibtest.asp> (25.01.2004).

[57] D. Bolt, "Differential Item Functioning [Diferansiyel Madde Fonksiyonu]," <http://www.stat.uiuc.edu/stoutlab/papers/Bolt961.ps> (17.08.2003).

[58] Y. Du, "When to Adjust for Differential Item Functioning [Diferansiyel Madde Fonksiyonu İçin Ne Zaman Gerekli Ayarlamalar Yapılır]," <http://www.rasch.org/rmt/rmt91e.htm> (11.09.2003).

[59] Aynı.

[60] Aynı.

[61] Aynı.

[62] William Stout Institute for Measurement, "DIFPACK - Dimensionality-Based DIF Analysis Package [Boyut Temelli DMF Analiz Yazılımları]," <http://www.assess.com/Software/DIFPACK.htm> (11.09.2003).

[63] Smith, "Theory and Practice.

[64] V.B. Lober , "The Identification and Interpretation of Item Bias [Madde Yanlılığının Tanımlanması ve Yorumlanması]," <http://www.google.com.tr/search?q=cache:hufv5w_YGVwJ:www.rasch.org/memo25.htm+Biased+Test+Items&hl=tr&ie=UTF-8&inlang=tr> (19.08.2003).

[65] R.J. Harvey , "Item Response Theory [Madde-yanıt Kuramı]," <http://harvey.psyc.vt.edu/Documents/TCP_IRT98.pdf> (21.08.2003).

[66] Stark ve Chernyshenko "Assessment of Model-Data."

[67] R. J. Harvey , "Item Response."

[68] Aynı.

[69] Institute for Objective Measurement, "Historic Misunderstandings of the Rasch Model [Rasch Modelinin Tarihsel Olarak Yanlış Anlaşılması]," <http://www.rasch.org/rmt/rmt142f.htm.> (25.09.2002).

[70] Y.C. Hsu, "The Goal Programming Approach for Test Construction [Test Oluşturmada Amaç Programlama Yaklaşımı], Master Thesis, Universitiy of Arizona, 1993.

[71] Aynı.

[72] Aynı.

[73] Aynı.

[74] Aynı.

[75] Aynı.

[76] Aynı.

[77] Aynı.

GÜVENİLİRLİK TESTLERİNİ İÇEREN ÖZEL YAZILIMLAR

Son yıllarda genel amaçlı istatistiksel analiz yazılımlarının dışında güvenilirlik analizlerini veya güvenilirlikle ilgili diğer testleri yapan özel programlar da geliştirilmeye başlanmıştır. Söz konusu programlar daha çok *yapısal eşitlik modeli*, *madde analizi*, *Rasch modeli* gibi belirli teknikleri uygulamaya yöneliktir. Bu bölümde, diğer testlerin yanı sıra güvenilirlik analizlerini de yapan veya değişik yollarla verilerin güvenilirliğini artıran bu yazılımların özellikleri üzerinde durulmuş ve yazılımlara ilgi duyacak okuyucuların karar vermelerini kolaylaştırmak amacıyla bazı bilgiler verilmiştir.

GENEL

Güvenilirlik analizlerini ölçüm modelleri çerçevesinde incelerken, özel yazılımları da söz konusu test kuramları çerçevesinde ele almakta yarar vardır. Geliştirilen yazılımların bir bölümü klasik test kuramına göre analiz yaparken, diğerleri modern yaklaşımlardan *madde-yanıt kuramını* temel alarak hazırlanmıştır. Özgün bir modeli test etmeye yönelik olarak hazırlanan bu yazılımlar değişik türdeki istatistiksel analizlerin yanı sıra aynı zamanda maddenin, testin veya modelin güvenilirliğini de analiz eder. Yazılımların bir bölümü ise sadece veri kalitesini iyileştirmeye yöneliktir. Bu tür yazılımlar güvenilirliği test etmekten çok eksik veri, uç veri, normal dağılım göstermeyen veri yapısı gibi durumlarda söz konusu verileri belirli işlemlere tâbi tutarak test veya ölçek verilerinin genel olarak güvenilirliğini artırmayı hedefler. Yapısal eşitlik modeli ve gizli yapıları test eden yazılımlar ise ayrı bir grubu oluşturur. Bu yazılımlarda test maddelerinin güvenilirliği yerine modelin güvenilirliği üzerinde odaklanılmıştır. Dar kapsamlı özel istatistikî yazılımları bir kaç başlık altında toplamak mümkündür:

1. Klasik madde analizini yapan yazılımlar.
2. Değişik türde korelasyon analizleri yapan yazılımlar.
3. Yapısal eşitlik modelini test eden yazılımlar.
4. Eksik veri analizi yapan yazılımlar.
5. Rasch modeli yazılımları.
6. Madde-yanıt kuramını temel alan yazılımlar.
7. Özel olarak sadece faktör analizi yapan yazılımlar.
8. Etki büyüklüğü ve meta analizi yazılımları.
9. Çok boyutlu gizli yapıları ortaya çıkaran yazılımlar.

İnternet kaynakları incelendiğinde bu yazılımların sürekli gelişme gösterdiği, isimlerinin değiştiği, yazılım kodlarının MS-DOS ortamından Windows ortamına aktarıldığı görülür. Bu kitapta büyük ölçüde İnternet kaynaklarından yararlanılarak Windows ve MS-DOS ortamında çalışan özel yazılımlara ilişkin bazı örnekler verilmiştir. Seçilen örneklerin küçük çaplı ve ücretsiz olarak dağıtılan yazılımlar yerine ticarî bir kuruluş tarafından pazarlanan daha ciddî çalışmalar olmasına dikkat edilmiştir. Öte yandan, bu yazılımlardan bazıları belirli bilim disiplinlerine özgü olarak geliştirilmiş olduğundan çok amaçlı kullanıma uygun olmayabilir. Örneğin, eğitim kuruluşlarında kullanılmak amacıyla geliştirilen madde analizi yazılımları daha çok okulların ölçme ve değerlendirme servisleri için uygundur. Yine sağlık kuruluşları ve psikiyatri kliniklerinde kullanılmak üzere geliştirilen yazılımlar hastaneler, psikiyatri klinikleri ve psikoloji laboratuarları için uygundur. Özel amaçlı yazılımlara ilgi duyan okuyucular seçtikleri programın diğer bilim alanlarında kullanılıp kullanılamayacağını araştırmalıdırlar.

MADDE ANALİZİNİ YAPAN YAZILIMLAR

Bu yazılımlar, klasik test kuramının varsayımları çerçevesinde çalışır. Güvenilirlik analizleri açısından, bu yazılımlarla sadece maddelerin ve testin güvenilirliği araştırılır. Testin maddeleriyle ilgili analiz, esas olarak testin bütününün güvenilirliğini sağlamaya yöneliktir.

ITEMAN

Iteman adlı yazılımın; doğru / yanlış şeklinde kodlanan soruların, çoktan seçmeli soruların ve alan araştırmalarında kullanılan Likert tipi soru ve ifadelerin madde analizlerini yaptığı bildirilmiştir. Yazılım, el ile girilen

verileri veya ASCII formatında tarayıcılarla alınan verileri analiz etme özelliğine sahiptir. Iteman'ın, 10 kadar farklı alt testi analiz edebildiği, her bir ölçek için özet istatistikleri verdiği, maddelerin aritmetik ortalamalarını, standart sapmalarını, basıklık ve çarpıklık değerlerini verdiği, KR-20 (alfa) güvenilirlik analizini yaptığı, her bir alt test için ölçümün standart hatasını hesapladığı ifade edilmiştir.[1]

Iteman, aynı zamanda örneklem grubunun üst ve alt %27'lik gruplarına ait istatistik hesaplamaları yapma özelliğine sahiptir. Ayırma indeks değeriyle, fonksiyonu zayıf olan maddeleri belirleyebilmekte ve demografik değişkenlere göre belirli alt gruplara ait istatistikî analizleri yapabilmektedir.

Program, birden fazla doğru cevabın bulunduğu çoktan seçmeli soruları da işleyerek daha iyi çalışan şıkkı belirleyebilme özelliğine sahiptir. Likert tipi ölçeklerde dokuz dereceye kadar işlem yapabilmektedir. MS-DOS ortamında çalışan yazılımın bir defasında 30.000 cevaplayıcıya ait 250 maddeyi, Windows ortamındaki yazılımın ise sınırsız sayıda cevaplayıcıya ait 750 maddeyi analiz edebildiği bildirilmiştir.[2]

LERTAP 5

Larry R. Nelson tarafından geliştirilen Lertap, Excell yazılımını kullanarak madde ve test analizini gerçekleştiren bir programdır. İstatistiksel analiz yazılımları SPSS ve SAS'ta bulunmayan farklı veri analizlerini yapan bu yazılımın ayrıca alan araştırmalarında kullanılan ölçekleri ve yetkinlik (mastery) testlerini analiz edebildiği bildirilmiştir. Yazılım, Windows 95 / 98 / ME / 2000 / XP / NT ve Macintosh OS 8.6 / 9.1 / 10 sistemlerinde çalışabilmektedir. Yazılımın veri dosyası Excell'in çalışma sayfasıdır ve programda normal olarak Excell'in fonksiyonları ve grafik özellikleri kullanılmaktadır. Güvenilirlik analizi bölümünde testten bir madde çıkarıldığında güvenilirlik katsayısı alfanın ne şekilde etkilendiğini görmek mümkündür.

Lertap'ın her bir madde için *güçlük*, *ayırma* ve *çeldirici* analizleri yaptığı, maddelere ait özet istatistikî analizleri verdiği, maddeler arasında çoklu korelasyon analizi yaptığı ve cevapların yüzde dağılımlarını verdiği bildirilmiştir.[3]

Lertap yazılımında, bilişsel (kognitif) testleri desteklemek üzere, kriter referanslı ve yetkinlik temelli test analizlerini yapacak formüller de kullanılmıştır. Ön tanımlı olarak belirlenen kesim/sınır puanı %70, araştırmacılar tarafından kendi belirledikleri başka bir değere göre de değiştirilebilmektedir. Yazılımın bu bölümünün "genellenebilirlik katsayısını",

Brennan'ın "genelleştirilmiş madde ayırt edicilik indeksini" ve Kohen kappa katsayısını hesaplayabildiği bildirilmiştir.[4]

Yazılım, alan araştırmalarında kullanılan sorularla ilgili olarak ki, bunlara Likert ve anlamsal farklılık ölçekleri de dahildir, her bir alt test için biri kapsamlı diğeri ise her bir maddenin işlevselliği ile ilgili olarak iki analiz yapabilmektedir.Yazılımın ayrıca aritmetik ortalama, medyan, standart sapma, varyans, iç ve dış kriterlerle ilgili korelasyon analizleri, puanların ters döndürülmesi, Cronbach alfa gibi güvenilirlik analizlerini yapabildiği ifade edilmiştir.

TESTFACT

R.Wood, D. Wilson, R. Gibbons, S. Schilling, E. Muraki ve D. Bock tarafından geliştirilen Testfact adlı yazılımın klasik madde analizi, test puanlaması ve iki dereceli puanlar için faktör analizi gibi istatistikî testleri yaptığı bildirilmiştir. Bunun yanında MYK'ye göre maddelere ait faktör analizi de yapabilmektedir. Yazılımın 1000 maddeyi işleyebildiği, alt testlere ve değişik gruplara ilişkin istatistiksel analizleri gruplandırmış olarak verdiği ifade edilmiştir. Yazılıma güvenilirlik katsayıları ve çeldirici analizi de eklenmiştir. Programın faktör analiziyle ilgili modülleri aşağıdaki gibidir:[5]

1. Tetrakorik korelasyon analizleri.
2. Temel bileşenler analizi.
3. Marjinal maksimum olasılık analizi.
4. Faktör analizleri (keşfedici faktör analizi, teyit edici faktör analizi, faktör ağırlıklarının hesaplanmasında varimax ve promax döndürme yöntemleri).
5. Tahmin yürütülerek işaretlenen maddelerde düzeltme yapılması.
6. Boş bırakılan maddelerin ele alınması.

Testfact yazılımının İnternet'te çevrimiçi çalışan ayrıntılı bir yardım dosyası bulunmaktadır. Bu dosyada açıklamalı analiz örnekleri ve program kod örnekleri verilmiştir. Yazılım, Windows 98, NT, ME, 2000, XP işletim sistemleri altında çalışabilmektedir.

SCRUTINY

Kelimenin sözlük anlamı *dikkatlice inceleme* anlamına gelen bu yazılım, test maddelerinde kopya çekilerek işaretleme yapılıp yapılmadığını belirler. Bu alanda yazılmış g2 (Frary ve diğerleri, 1977), ω (Wollack, 1997),

K-index (Holland, 1996) programıyla birlikte literatürde yer alan başlıca yazılımlardan birisidir (aktaran Wollack, 2001).[6] Testi alan kişilerin dürüst davranıp davranmadıklarını belirlemeye yardım eder.

Amerika Birleşik Devletlerinde *Assessment Systems Inc.* şirketi tarafından ticarî olarak pazarlanan Scrutiny adlı yazılımın istatistiksel güvenilirliği çok fazla araştırılmamıştır. Yazılımla Chason (1997) 1000 kişilik bir örneklem üzerinde 80 maddeyi test etmiş fakat sonuçların alfa hatası hakkında bilgi vermemiştir (aktaran Wollack, 2001).[7]

KORELASYON ANALİZİ YAPAN YAZILIMLAR

Korelasyon analizi yazılımları; test-yeniden test, paralel formlar güvenilirliği, gözlemciler arası değerlendirmenin güvenilirliği, madde-test güvenilirliği gibi konularda araştırmacılara yararlı olan araçlardır.

PRAM

K.A. Neuendorf tarafından geliştirilen ve Windows temelli olan bu program iki ve ikiden fazla kodlayıcının bulunduğu değerlendiriciler veya gözlemciler arası güvenilirlik hesaplamalarını yapmak üzere geliştirilmiştir. Programda veri dosyası Excell'e benzer biçimde düzenlenmiştir. Birinci sütuna değerlendiricinin kimlik kod numarası, ikinci sütuna ise vak'a numarası yazılır. Diğer sütunlar ise değişkenler için ayrılmıştır. PRAM isimli yazılımda aşağıdaki güvenilirlik analizlerinin yapıldığı bildirilmiştir:

1. Uyuşma yüzdesi (nominal veriler).
2. Scott Pi (nominal veriler).
3. Cohen Kappa (nominal veriler).
4. Spearman Rho (büyüklük sırasındaki veriler).
5. Pearson korelasyon analizi (eşit aralıklı – oranlı veriler).
6. Lin'in uyuşma korelasyon analizi (eşit aralıklı – oranlı veriler).

PRAM'ın, henüz geliştirilme aşamasında olması nedeniyle bazı hatalar içerebileceği bildirilmiştir. Yazılımın İnternet ortamında ücretsiz olarak dağıtılan deneme sürümü bulunmaktadır.

PRELIS

Karl Jöreskog ve Dag Sörbom tarafından geliştirilen PRELIS, Lisrel'le birlikte gelen çok değişkenli veri tarama ve özetleme yazılımıdır. Veri

taraması için, Lisrel programından önce kullanılır. Program, sürekli verileri, sınırlandırılmış verileri ve sıralı ölçek verilerini okuyabilmekte ve ayrıca korelasyon ve kovaryans matrisleriyle birlikte polikorik korelasyon analizi yapabilmektedir. Örneklem verilerine dayalı korelasyon ve koveryans matrislerinden hareket ederek semptomsuz (asymptotic) ana kütle korelasyon ve kovaryans matrisi tablolarının tahminine imkan vermektedir.[8] Sıralı ölçek verileri çoğunlukla normal dağılım özelliği göstermediğinden yapısal eşitlik modeli ve teyit edici faktör analizlerinde normal kovaryans matrisi değil, semptomsuz kovaryans matrisi (SKM) verileri kullanılır. Sıralı ölçek verileri, eğer eşit aralıklı değil de "sıralı" olarak kabul edilmişse bu değişkenler arasındaki ilişkilerin daha yansız bir resmini ortaya koymak için *polikorik korelasyon* analizi yöntemi uygulanır. Polikorik korelasyon katsayıları daha iyi tahmin göstergeleridir. Ayrıca Prelis eksik veriler için sıcak-deste yöntemiyle yeni değerler atama özelliğine sahiptir.

Rakamlar sürekli veri niteliğinde ise PRELIS'in kullanılmasına gerek yoktur. Ancak rakamlar sıralı ölçek verisi niteliğinde veya karma bir nitelikte ise LISREL'den önce PRELIS yazılımının kullanılması gerekir. Yeni sürüm, PRELIS-2'de bütün verilerin ön tanımlı olarak sıralı ölçek verisi niteliğinde olduğu kabul edilmiştir. Herhangi bir değişken PRELIS'te kolaylıkla normal dağılım verilerine dönüştürülebilmektedir. Yazılım küçük, orta ve büyük örneklem verilerindeki normal dağılım özelliği göstermeyen verileri normalleştirmesi açısından pratik bir kullanıma sahiptir. Yazılımın aynı zamanda polikorik modele ait G^2 testini de içerdiği bildirilmiştir. Prelis'in ikinci sürümünde analizi yapılan testin yaklaşık olarak uyuşma durumunu ortaya koymak için *yaklaşık hata ortalamasının karekökü değeri* de (root mean square error of approximation – RMSEA) hesaplanabilmektedir.[9]

POLYCORR

F. Drasgow (1988) ve U. Olsson (1979) tarafından geliştirilen Polycorr, ikili veya sıraya sokulmuş kategorik veriler üzerinde polikorik korelasyon analizini yapan bir programdır. Yazılım daha çok iki dereceli veya çok dereceli gözlemci değerlendirmeleri arasındaki uyuşmayı görmek için kullanılır. MS-DOS ortamında çalışan yazılımda model uyuşma istatistikleri olarak; G^2 testi (ki-kare olasılık oranı), Pearson ki-kare istatistiği ve onunla ilgili p istatistiğinin bulunduğu bildirilmiştir. Yazılım ayrıca polikorik korelasyon katsayısı rho değerini ve bu değerin standart hatasını hesaplamaktadır.[10] İnternet ortamından ücretsiz olarak indirilebilen yazılıma ilgi duyan okurlar ayrıntılı bilgileri İnternet ortamındaki kullanıcı rehberinden temin edebilirler (*bk.*, Tablo 14-1).

SHORTFORM

P. Levy (1967) tarafından önerilen *düzeltme formülünü* çalıştıran kullanımı kolay bir yazılımdır. Geliştirilen herhangi bir testin *uzun formuyla* kısa formu arasındaki paylaşılan hata varyansı nedeniyle şişkin gözüken korelasyon katsayıları arasındaki hatayı düzeltir. Bu yazılım, eğitim amacıyla veya uzun bir formdan kısa bir form üretilmek istendiği zaman kullanılır. Yazılımın çevrimiçi yardım dosyaları, problem çözüm örnekleri bulunmaktadır. İnternet ortamından ücretsiz olarak indirilebilir.

ATTEN2

Ölçümdeki hataları veya ölçüm ranjı kısıtlaması nedeniyle ortaya çıkan düşük korelasyon katsayılarındaki zayıflığı gideren bir yazılımdır. Programın ayrıca alfa katsayısındaki ranj kısıtlamasından kaynaklanan zayıflığı giderdiği bildirilmiştir. Program, *zayıflığı yenme formülünü* dört farklı şekilde hesaplar:[11]

1. Bilinçli olarak seçilen bir değişken için sadece ana kütle ve örneklem *SS* değeri biliniyorsa.
2. Tesadüfî olarak seçilen bir değişken için sadece ana kütle ve örneklem *SS* değeri biliniyorsa.
3. Üç değişken olması durumunda.
4. Değişkenlerden her ikisinde de veya yalnızca birisinde güvenilmezlik varsa.
5. Bir dizi test puanında ranj kısıtlaması varsa.

Yazılım, İnternet'ten ücretsiz olarak indirilebilir ve ayrıca İnternet ortamında çevrimiçi kullanım el kitabı bulunmaktadır.

DICHOT

Dichot, ikili veri yapılarında uyuşma indeks değerini hesaplayan bir yazılımdır. Daha çok öğrenciler için eğitime yardımcı olacak bir destek aracı olarak önerilmiştir. Ölçüm verilerinde ikili uyuşmanın "eğer olursa, ne olur" yaklaşımıyla hesaplanmasına imkan verir. Yazılımın 2x2 şeklinde düzenlenen verilerden hareket ederek aşağıdaki hesaplamaları yaptığı bildirilmiştir:[12]

1. Phi.
2. Phi/Phi Max.
3. Yule's Q (Gamma).
4. Jaccard.
5. G-Index.
6. Bennett B-index.
7. Kappa gözlemciler arası güvenilirlik.

Yazılım İnternet ortamından ücretsiz olarak indirilebilir. Program, yukarıda sayılan testlerin dışında tıbbî test değerlendirme istatistiklerini de verme özelliğine sahiptir.

ITRS

Gözlemciler arası değerlendirmenin güvenilirliğini ölçen bir yazılımdır. Bunun için değişik istatistikî tekniklerden yararlanıldığı bildirilmiştir. Söz konusu teknikler farklı kuramsal yaklaşımlara dayanmaktadır ve böylece ölçeğin veya testin farklı güvenilirlik rakamları elde edilmektedir. Yazılımda iki bölüm bulunmaktadır: gözlemciler arası güvenilirlik ve test güvenilirliği.

Gözlemciler arası güvenilirlik bölümünde gözlemciler için farklı değerlendirme kombinasyonları oluşturulmuştur ve araştırmacı bu kombinasyonlardan en yüksek tahmin değerini veren istatistiği seçebilmektedir. Analiz sonunda aşağıdaki istatistiksel sonuçların alındığı bildirilmiştir:[13]

1. Momentler çarpımı korelasyon katsayısı, gözlemci çiftleri arasında Kendall tau, korelasyonların ortalaması ve değişim aralığı.
2. Uyuşma yüzdesi.
3. Kullanıcı tarafından tanımlanan belirli bir noktadaki uyuşma yüzdesi.
4. Cronbach alfa.
5. Kappa katsayısı.
6. Küme içi korelasyon katsayısı.
7. Genellenebilirlik katsayısı.
8. Bireysel değerlendirmeyi grubun değerlendirmesiyle karşılaştıran William uyuşma indeksi.

Yazılımın madde analizi bölümünde ise iki tür analiz yapılmaktadır. Birincisi çoktan seçmeli maddelerin analizi ve ikincisi ise diğer tipteki testler

için madde analizinin yapılması şeklindedir. Bu bölümde aşağıdaki istatistiksel analizlerin yapıldığı bildirilmiştir:

1. Maddenin değişik şıklarına ait frekansların hesaplanması ve orantısız işaretlemelerin "bayrak" ile belirlenmesi.
2. Kullanıcı tarafından sağlanan anahtar şıkka ait puanlar.
3. Maddelere ve teste ait aritmetik ortalama ve standart sapma değerlerinin hesaplanması.
4. Madde–test ve madde–alt ölçek korelasyonlarının hesaplanması.
5. Maddeler arasındaki korelasyon katsayılarının dış kriterle birlikte değerlendirilerek hesaplanması.

Yazılım konusunda daha fazla bilgi edinmek isteyen okurların, *Advance Research and Data Analyses Center* isimli araştırma merkezinin İnternet sitesine başvurmaları önerilir.

Tablo 14-1. Tetrakorik ve Polikorik Korelasyon Analizi Yapan Yazılımların Karşılaştırılması

Türü	Veri yapısı	Yazılımlar	Örneklem büyüklüğü / değişken sayısı
Tetrakorik korelasyon	İkili (arka planda normal dağıldığı koşuluyla)	1. Prelis (kısıtlı olarak, matrisi tam olarak vermez). 2. EQS (Korelasyon matrisini tam olarak verir) 3. TASTFACT	$n > 100$, değişken sayısı, en az 4
Polikorik korelasyon	Sıralı (arka planda normal dağıldığı koşuluyla) İki, üç, dört, beş dereceli vb. sıralı veriler.	1. Prelis 2. POLYCORR 3. PROC FREQ (SAS)	$k(k + 1)/2$ k = Değişken sayısı

YAPISAL EŞİTLİK MODELİ YAZILIMLARI

Yapısal eşitlik modelini test eden yazılımlar doğrudan maddelerin veya ölçeklerin güvenilirliğiyle ilgili değildir, ancak gözlem verilerinin geliştirilen modele uygun düştüğünün veya geliştirilen modelle uyuştuğunun kanıtlanmasıyla birlikte, madde veya ölçeğin güvenilirliğinin de dolaylı olarak kanıtlandığı varsayılır. Ölçeklere ait maddelerin geçerlilik ve güvenilirliği önerilen veya test edilen kuramsal modelin uygunluğu kapsamında yapılır.

AMOS

İngilizce "Moment Yapıların Analizi" (Analysis of Moment Structures) anlamına gelen kelimelerin baş harflerinden oluşan bu yazılım SPSS yazılım şirketi tarafından Lisrel'in yerine geçmek üzere J. Arbuckle ve W. Wothke (1999) tarafından geliştirilmiştir. Yazılımın ABD'deki dağıtımı SPSS ve SmallWaters şirketleri tarafından yapılmakta ve İnternet ortamında öğrencilere yönelik olarak sınırlı kullanım imkanı sağlayan deneme sürümü bulunmaktadır. Yazılım ile *rota analizi, teyit edici faktör analizi, gizli değişkenlere ilişkin nedensellik analizi* ve *çoklu doğrusal regresyon analizleri* yapılabilmektedir.

Metin kodları veya grafik tanımlamasıyla çalışan AMOS yazılımının kullanılabilmesi için ölçekteki değişken başına en az 15 vakanın bulunması gerektiği bildirilmiştir. Veriler eğer mükemmel bir şekilde iyi davranıyorsa, diğer bir deyişle normal dağılım özelliği gösteriyorsa değişken başına düşen vak'a sayısı beşe kadar düşürülebilir.[14]

Yazılım bir ölçeğin veya test maddelerinin güvenilirlik analizi için değil, faktörler arasındaki ilişkileri açıklayan modelin güvenilirliği için kullanılır. Bilim adamının bu yazılımı kullanma amacı, gizli yapılar arasında veya faktörlerle gözlem değişkenleri arasında bulunduğunu varsaydığı ilişkileri matematiksel olarak kanıtlamaktır. Bunun için modele incelenen bütün değişkenler katılır, ancak modelde 20'den fazla kavramsal yapı veya gizli değişken varsa bu yapılar arasındaki ilişkileri yönetmek güçleşeceğinden bu tür kapsamlı analizlere başvurmak anlamlı değildir.[15]

Yapısal eşitlik modelinde göstergelerin[a] güvenilirliği ya modelin tahmin edilme süreci içinde deneysel olarak ortaya çıkarılır veya göstergelere ait güvenilirliğin "sabit" olduğu varsayılır. Tek bir maddenin veya ölçeğin

---

[a] Gösterge (indicator). Gizli kavramsal yapıya işaret eden, gizli kavramsal yapının tahmin edilmesini sağlayan veya gizli kavramsal yapıyı ortaya çıkarmaya hizmet eden maddelerden / değişkenlerden her biri.

güvenilirliği daha önceden yapılan güvenilirlik analizlerine dayandırılmışsa göstergelerin güvenilirliği sabittir. Daha önceden güvenilirlik analizleri yapılmamışsa, modelin güvenilir çıkmasıyla maddelerin veya testin de güvenilirliğinin sabit olduğu varsayılır.[16]

AMOS'ta esas olarak sürekli veriler kullanılır. Likert ölçeklerinde olduğu gibi sıralı ölçek verileri için teyit edici faktör analizini kullanmak doğru değildir. Sıralı ölçek verisine sahip araştırmacılar teyit edici faktör analizi için Lisrel yazılımını düşünebilirler.

Windows, OS/2 ortamlarında çalışan ve en son beşinci sürümü çıkmış olan AMOS'un kullanıcıları bilgilendirmeye yönelik olarak İnternet ortamında yardım dosyaları bulunmaktadır.

LISREL

K. Joreskog ve D. Sorbom (1993) tarafından geliştirilen Lisrel yazılımında, İngilizce LInear Structural RELations (Doğrusal Yapısal İlişkiler) kelimelerinden belirli harfler seçilerek yapılmış bir kısaltma, yazılımın adı olarak belirlenmiştir. Lisrel, ölçümlerdeki kovaryans yapıları analiz eden bir istatistikî yazılımdır. Faktörler arasındaki nedensel ilişkileri araştıran hipotezlerin test edilmesi için kullanılır. Son yıllarda bu yazılım regresyon analizinin yerine kullanılmaya başlanmıştır. Yazılım, regresyon analizinde olduğu gibi değişkenler arasındaki doğrusal ilişkileri araştırır. Lisrel aynı zamanda gizli değişkenlerin özelliklerini incelemek amacıyla da kullanılır.

Teorik veya soyut kavramlar dolaylı olarak, diğer bir deyişle gizli değişkenler ortaya çıkarılmak suretiyle ölçülür. Bu yöntem, ham puanların kullanılmasından daha elverişlidir. Böylece tüm test yerine testin alt boyutlarının güvenilirliği artmakta ve ölçüm hatası azalmaktadır. Araştırmacı geliştirdiği nedensel ilişkilere dayanan ölçüm modelinin gözlem verileriyle desteklenip desteklenmediğini bulmaya çalışır. Lisrel'de, ölçüm modelinin geçerlilik ve güvenilirliğini hesaplamak için *teyit edici faktör analizi* yöntemi uygulanır. Lisrel, genel amaçlı bir istatistikî analiz programı değil, bir dizi yapısal eşitlik katsayılarını tahmin eden bir programdır. Yazılımı kullanabilmek için ileri düzeyde istatistik ve matematik bilgisine ihtiyaç vardır. Lisrel programı ile aşağıdaki istatistiksel analizlerin yapıldığı bildirilmiştir:

1. Ölçüm modelleri ve teyit edici faktör analizi.
2. Gözlem değişkenleri için nedensel ilişki modelleri.
3. Gizli değişkenler için yapısal eşitlik modelleri.

4. Sıralı değişkenler ve normal dağılım özelliği göstermeyen diğer değişkenler için istatistikî analizler.
5. Çok örneklemli analizler.
6. Doğrusal ve doğrusal olmayan kısıtlayıcılar.
7. Uygunluk ölçüleri.
8. Rota diyagramları.
9. Sınıflandırıcı değişkenler için probit regresyon analizi.

Sıralı ve eşit aralıklı ölçek verileriyle çalışan bilim adamları Lisrel'i kullanarak bu verilerde polikorik ve poliseriyal korelasyon katsayılarıyla semptomsuz kovaryans matrisi elde edebilirler. Her ne kadar semptomsuz kovaryans matrisi konusunda bilimsel bazı tartışmalar devam etse de bu yöntem yanlılığı büyük ölçüde azaltmaktadır.[17] Ancak yazılımda en çok 25 değişkenin analiz edilebildiği ve örneklem sayısının ise, 2000 ilâ 5000 arasında olması gerektiği bildirilmiştir.

Matris veri tanımlamasına dayanan Lisrel yazılımının kullanıcı dostu olmadığı, mönü destekli değil, komut destekli olarak çalıştığı belirtilmiştir. Ancak son yıllarda skalar nitelikteki verileri tanımlamaya imkan veren SIMPLIS adlı ek bir modülün daha geliştirilmesiyle veri tanımlamasının kolaylaştığı ifade edilmektedir. Programı kullanmayı düşünen kişilerin doğrusal modeller, regresyon modelleri konusunda kendilerini yetiştirmiş olmaları gerekir. Lisrel yazılımının eğitim malzemeleri de yeterince zengin değildir. Bu açıdan bu yazılımı kullanmayı düşünen kişilere İnternet ortamındaki tartışma gruplarına katılmaları veya danışman öğretim üyelerinden destek almaları önerilmiştir. Kayıtlı kullanıcılar ise doğrudan satıcı firmadan destek alabilmektedirler. Halen sekizinci sürümü yayımlanmış olan LISREL 8 / PRELIS 2 yazılımının Windows, DOS ve Mac ortamlarında çalışabildiği belirtilmiştir.

SAS CALIS

Wolfgang Hartman tarafından geliştirilen yazılımın adı, "Covariance Analysis and LInear Structural Equations" kelimelerinin baş harflerinden oluşturulmuştur. Amerika Birleşik Devletlerinde yaygın kullanılan istatistiksel analiz programı SAS için geliştirilmiş *teyit edici faktör analizi* ve *yapısal eşitlik modellerini* analiz eden bir modüldür. *SAS Institute limitet* şirketi tarafından satılan programın içindeki PROC CORR komutuyla ölçeğin iç tutarlılık güvenilirliğinin hesaplandığı bildirilmiştir. Öte yandan PROC FACTOR komutuyla ölçeğin faktör yapısının araştırıldığı ve PROC CALIS komutuyla ise ölçeğin faktör yapısının doğrulamasının yapıldığı ifade edil-

miştir.[18] Windows, Mac, DOS, UNIX ve IBM 360 işletim sistemlerinde çalışan CALIS'ın en son 6.12 numaralı sürümü yayımlanmıştır ve modül SAS yazılımıyla birlikte ücretsiz olarak verilmektedir.[19]

EQS

P.M. Bentler (2000) tarafından geliştirilen Windows ve Mac sürümleri bulunan bu yazılımın, yapısal eşitlik modellerinden çoklu regresyon, teyit edici faktör analizi, rota analizi ve çoklu ana kütle karşılaştırmaları gibi testleri yaptığı bildirilmiştir. Lisrel programına göre kullanımının daha kolay olduğu ve matris cebrine ihtiyaç olmadığı ifade edilmiştir. EQS, çok değişkenli normalliği, Mardia çok değişkenli basıklık ve çarpıklık testlerini ve Mardia Kappa normallik tahmin değerini vermektedir. EQS, sıralı ölçek verilerinde AMOS'ta bulunmayan polikorik korelasyon analizini yapabilme özelliğine sahiptir, ancak kategorik değişkenlerin sayısının 20'yi aşmaması gerekir.[20] Bentler, sıralı ölçek verilerinde tatmin edici bir sonuç alınabilmesi için örneklem hacminin en az 2000 olması gerektiğini bildirmiş ve muhtemelen 5000 kişiden oluşan bir örnek kütlenin çok daha iyi sonuçlar vereceğini ileri sürmüştür. Kuşkusuz sosyal ve davranışsal bilimlerde sık uygulanan araştırmalarda bu rakamlara erişmek çok zordur. Bu nedenle, örneklem hacmi küçükse veya sıralı ölçek niteliğindeki verilerde eğer değişken sayısı 20'den fazlaysa ve bu ölçek örneğin bir Likert ölçeği ise, veriler eşit aralıklı ölçek verileri olarak kabul edilip böyle bir durumda semptomsuz dağılım fonksiyonunu görmek için *maksimum olasılık* (maximum likelihood) veya *genelleştirilmiş en küçük kareler* (generalized leased square) tahmin yöntemi kullanılır.[21]

Windows, Macintosh, DOS, UNIX, IBM 360 ve VAX sistemleri altında çalışan EQS'in[a] en son 5.7 sürümü yayımlanmıştır. Yazılımın deneme amaçlı sürümünü İnternet'ten ücretsiz olarak indirmek mümkündür. Deneme amaçlı sürümünde en fazla 18 parametre analiz edilebilmekte ve yazılım analiz sonuçlarının kayıt edilmesine, kopyalanmasına ve yazıcı çıktılarının alınmasına izin vermemektedir.

SEPATH

Psikoloji profesörü James Steiger tarafından, daha önce geliştirilen EzPATH isimli yazılım temel alınarak oluşturulmuştur. StatSoft limitet şirketinin yayımladığı Statistica isimli programın bir alt modülü olarak kullanılmaktadır. Yazılım, yapısal eşitlik ve rota analizlerini yapabilmek-

[a] Yazılım <İ>, <Q>, <S> şeklinde değil, X [iks] şeklinde okunur.

tedir. Yazılımın adı bu iki analiz türünün kısaltmasından oluşmuştur (Structural Equation Modeling / Path Analysis).

Sepath modülü sanatsal olarak düzenlenmiş ara yüzüyle, modelleme prosedürlerine ilişkin kapsamlı bir seçenekler listesi sunmakta ek program kodlarının yazımına ihtiyaç kalmadan karmaşık model tanımlamalarına dahi izin verebilmektedir. Programın içindeki *sihirbaz* ve *rota çizim araçları* kullanılarak yapılacak analizin basit fonksiyonel terimlerle tanımlanabildiği açıklanmıştır. Yazılım; korelasyon, kovaryans, moment matrisleri gibi analizleri yapabilmekte ayrıca bunun yanında kısıtlanmış optimizasyon teknikleri, uygun standart hata değerleri, standardizasyon modelleri, korelasyon matrislerine uyumlaştırılmış modelleri inceleyebilmektedir.

Sepath yazılımının literatürde yer alan çeşitli hesaplamalara ilişkin ayrıntılı açıklamalar yaptığı iddia edilmiştir. Bu hesaplamaların başlıcaları şunlardır: teyit edici faktör analizi, rota analizi, konjenerik testler için test modelleri, çoklu özellik-çoklu yöntem matrisleri, uzun dönemli araştırmalar için faktör analizi, birleşik simetri, yapılandırılmış ortalamalar analizi.

Yazılımda tek bir işlemde 300'e kadar madde işlenebilmekte ve ölçekteki tüm maddelerin güvenilirlikleri hesaplanabilmektedir. Ayrıca araştırmacı isterse yarıya bölme güvenilirliklerini ve alt ölçeklerin güvenilirliklerini hesaplama imkanına sahip olmaktadır.

Yazılımın bir diğer üstünlüğü grafik çizim imkanlarına sahip olmasıdır. "Eğer olsaydı" yaklaşımıyla çeşitli nitelikte grafikler (nokta dağılım, histogram ve çizgi grafikleri gibi) üretebilmekte ve ölçek geliştirilmesinde araştırmacının bu yardımcı araçları kullanmasına olanak sağlamaktadır.

MX

Virginia Psikiyatri ve Davranışsal Genetik Enstitüsü'nden Mike Neale ve Gary Xie tarafından geliştirilen bu yazılım Lisrel gibi matris yönelimlidir. Yapısal eşitlik modellerinde matris cebri yorumlayıcısı ve matematik en iyileştiricisi (optimizer) olarak tanımlanmıştır.[22] Bilim adamının, ölçüm verilerini yazılıma matris olarak tanıtması gerekir. Geliştirilen son sürümünde (1.44) programa Mx GUI adlı grafik ara yüzü de eklenerek araştırmacıların yapısal eşitlik modeline ilişkin diyagram çizmelerine olanak sağlanmıştır.[23] Bu ek modül sayesinde araştırmacıların doğrudan rota diyagramlarını kullanarak modelin uyuşma durumunu test edebilmeleri mümkün hale gelmiştir. Yazılım, LISREL, LISCOMP, EQS ve CALIS gibi programlarda bulunan değişik uyuşma fonksiyonlarını içerir. Ayrıca programın araştırmacıya kendi uyuşma fonksiyonlarını üretme imkanı sağ-

ladığı bildirilmiştir. DOS, VAX/VMS, UNIX, Win95 ve Win3.x ortamlarında çalışan yazılım İnternet'ten ücretsiz olarak indirilebilmektedir.

RAMONA

Michael Browne (1984) tarafından geliştirilen yapısal eşitlik modeli yazılımı, SYSTAT Windows adlı istatistiksel analiz programının ek modülü olarak dağıtılmaktadır. Windows ve MS-DOS ortamında çalışan yazılımın en son 10,2 numaralı sürümü yayımlanmıştır. Barrett (2003), RAMONA'yı sınırlı sayıda uyuşma istatistiği içermesi nedeniyle eleştirmiştir.[24] Bir diğer eleştiri, yapısal eşitlik modellerinde istatistikçiler için gerekli olan "topuk yükseltme"[a] modülünün RAMONA'da bulunmamasıdır. Bunun yanında komut dilinin basit ve kolay anlaşılır olması, tanımlanan ara yüzün çok iyi çalışması, eksik değer modülünün güçlü olması yazılımın önemli avantajları arasında sayılmıştır.[25]

TETRAD 3

P. Spirtes, R. Scheines, C. Meek, T. Richardson, C. Glaymour, H. Hoijtnik ve A. Boomsma tarafından 1996 yılında geliştirilen Tetrat 3 adlı yazılım, istatistiksel verilere dayalı olarak nedensel ilişki modelleri üretmek amacıyla yazılmış bir dizi modül ve araçtan oluşur. Yazılımın amacı başlıca iki tür istatistikî analizi yapmaktır: tekrarlamalı lineer yapısal eşitlik modellerini test etmek ve Bayesyen network şebekeleri oluşturmak.[26] Uygulamada, istatistik temelli nedensel model kurma çalışması, belirli aşamalarda gerçekleşir:[27]

1. Önce ölçüm değişkenleri saptanır.
2. Ölçümün yapılacağı örneklem belirlenir.
3. Örneklemden veriler toplanır.
4. Keşfedici istatistiksel analizler yapılarak veriler ayıklanır.
5. Nedensel ilişkileri tanımlayan bir model belirlenir.
6. Modelin parametreleri tahmin edilir.
7. Modelin uyuşma durumu test edilir.
8. Uyuşma yetersizse model yeniden tanımlanır.

---

[a] Bayağı bir dille ifade edilen bu terimin anlamı, *verilerin yapay olarak çoğaltılması* şeklinde açıklanabilir. İngilizce literatürdeki *bootstrapping* terimini "veri çoğaltması" şeklinde de çevirebilirdik, ancak literatürden kopuk yeni isimlendirmeler "köprülerin atılmasına" ve daha büyük anlam karmaşalarına neden olmaktadır. Bu yüzden "topuk yükseltme" terimini tercih ettik.

9. Son olarak veriler için, akla yatkın görünen başka açıklamalar yapılır.

Pek çok vakada, araştırmacılara model belirlerken çıkış noktası olarak teoriden hareket etmeleri önerilmiştir. Araştırılan konuda kuramsal bilgiler yeterli ise ve bilim adamı ilişkiler hakkında net bir fikre sahipse teorik yapının örnek alınması daha doğrudur. Eğer kuramsal bilgi birikimi model oluşturma konusunda yeterli bir temel oluşturamıyorsa Tetrad'ın bu konuda araştırmacılara yardımcı olacağı belirtilmiştir. Bu açıdan Tetrad belirli bir modeli önermekten çok uygun modelin ne olabileceğini araştıran bir yapılanmaya sahiptir.[28] Daha önce de belirtildiği gibi Tetrad bir dizi modülden oluşturulmuştur ve bu modüllerin adları aşağıdaki gibidir:

Kur.	MIM Kur.	STAT yaz.
Tahmin et.	Araştır.	Cequiv.
Güncelle.	Regres et.	Gibbs.
Dörtle.	Model yap.	
Saflaştır.	Monte et.	

Tetrad, belirlenen modele denk gelebilecek istatistiksel olarak eşit alternatifler üretirken, tahmin için gerekli olan regresörlerin sayısını azaltmakta, mevcut gizli değişkenlerin varlığını ortaya çıkarmakta, tek boyutlu ölçüm modellerini belirlemekte ve model belirleme araştırmasının güvenilirliğini tahmin etmekte yararlı bir yazılım olarak tanımlanmıştır. Programın İnternet ortamında kullanıcı el kitabı bulunmaktadır.

MPLUS

B. Muthen ve L. Muthen, (2001) tarafından geliştirilen bu yazılımın, keşfedici ve teyit edici faktör analizi, genel yapısal eşitlik modeli, çok düzeyli modelleme, gözlem değişkenleriyle rota analizi, Monte Karlo benzetim hesaplamaları, değişkenler için güvenilirlik analizi, gizli küme analizi, eksik veri ile modelleme yapma gibi hesaplamaları yapabildiği bildirilmiştir. Yazılımın kullanma kılavuzu, 30 bölümden ve teknik açıklamaları içeren 12 ekten oluşmaktadır.

PLS ve PLS-Graph

Herman Wold (1966) tarafından keşfedilen *kısmî en küçük kareler* (KEKK) yöntemi (Partial Least Squares – PLS) yapısal eşitlik modellerinden bir diğeridir. Yöntemde yapısal ilişkilerden çok tahmin algoritmasına daha fazla önem verilir. Kısmî en küçük kareler yaklaşımının güçlü bir

yöntem olduğu ifade edilmiştir. Bu uygulamada ölçeğin skalar yapısı, örneklem büyüklüğü ve artık değerlerin dağılımıyla ilgili olarak minimum düzeyde talepte bulunulmuştur. KEKK, her ne kadar kuramsal yapıların teyidi amacıyla kullanılsa da bunun yanında yöntemden değişkenler arasındaki ilişkileri araştırmak için de yararlanılabilir.[29]

Faktör temelli, kovaryans–gizli yapı uyuşum modelleriyle karşılaştırıldığında, bileşen temelli KEKK yaklaşımında iki önemli sorundan kaçınıldığı görülür. Bunlar; yetersiz çözümler ve belirsizliktir. Faktör temelli yapılarla bileşen temelli yapılar arasındaki temel farklılık birincisinin kuramsal yapıların teyidi ve ikincisinin ise, öngörme veya tahmin süreçleri üzerinde odaklanmasıdır. Kuramsal temel zayıfsa, gözlem değişkenleri veya ölçümleri dikkatli bir şekilde belirlenen ölçüm modeline tam olarak uymuyorsa bu gibi durumlarda KEKK yöntemi düşünülür. Wold, bu yaklaşımı "yumuşak modelleme" olarak isimlendirmiştir.[30]

KEKK, yapısal eşitlikle ilgili olarak faktör analizi yönteminden çok temel bileşenler ölçüm modelini esas alır. Bu modelde gizli yapılar, bileşik ölçüm değerleri / değişkenleri olarak görülür. Yöntemde uyuşmadan çok, tahminin maksimizasyonu ön plana çıkarılmıştır. Bu özelliğiyle YEM'in pratik alternatifi olarak görülmüştür.[31]

Araştırmacıların çoğu KEKK'i pratik bir yöntem olarak görürken bazıları da bu yöntemin kullanılmasına itiraz etmişlerdir. İtiraz eden bilim adamları modelde kullanılan istatistiksel analizlerin çok iyi anlaşılmadığını, araştırmacıların KEKK yönteminin ortaya koyduğu iddialara veya açıklamalara çok güçlü bir kanıt olmamasına rağmen çok fazla güvendiklerini ileri sürmüşlerdir.

EKSİK VERİ ANALİZİ YAPAN YAZILIMLAR

Ölçek veya testlerin güvenilirliğini düşüren bir diğer önemli faktör eksik verilerdir. Bilim adamı ölçüm verilerindeki eksikliği atama yöntemlerini kullanarak giderebilir. Literatürde, eksik veri analizi yaparak eksik verileri makul atama yöntemlerini kullanarak dolduran çeşitli yazılımlar vardır. Bu kitapta bunlardan birkaç tanesi üzerinde durulmuştur. Bu yazılımlar eksik veri problemini daha analizin başlangıç aşamasında çözerek verileri işlenebilecek hale getirirler. Ancak bu işlemin dikkatli bir şekilde yapılması gerekir. Acemi kişilerin ilkesiz ve nahiv bir şekilde yapacakları atama işlemi; tahmin ve standart hata değerlerini bozar, hipotez testlerini çarpıtır. Atama işlemi tek bir kez veya *çoklu atama* (ÇA) yöntemiyle yapılır. Çoklu atama yöntemi aslında bir Monte Carlo tekniğidir. Yaklaşımda eksik veriler $m>1$ benzetim tekniğine göre yapılır. Veri matrisinde boş olan her bir

hücreye *m* sayıdaki değerden biri atanır. Söz konusu *m* değerleri "tamamlanmış" veri setleridir. Tamamlanmış veri setleriyle gözlem veri setleri aynıdır ancak, eksik verilerin yerine farklı değerler atanarak eksik veri hakkındaki belirsizlik vurgulanır.

AMELIA

James Honaker, Anne Joseph, Gary King ve Kenneth Scheve tarafından geliştirilen ve kendi başına çoklu atama yapabilen bir yazılımdır. Ölçüm verilerindeki boş değerlerin yerine deterministik beklenen maksimizasyon[a] değerlerini otomatik olarak atar. Windows için Amelia, *girdi dosyası* formatını dosya uzantısına bakarak belirler. Bu açıdan çok geniş bir yelpazedeki dosya formatlarından bünyesine veri alabildiği bildirilmiştir. Veri transferinde sorunla karşılaşan araştırmacılara ASCII metin dosyalarını kullanmaları önerilmiştir. Yazılımın başlıca özellikleri aşağıdaki gibidir:[32]

1. Gözlemlenen kovaryans değerleri için koşullu model.
2. Hızı önemli ölçüde iyileştirilmiş dinamik bir kütüphane.
3. Güçlü model geliştirme için *t*-dağılımı algoritması.
4. Çeşitli olasılıkları değerlendirmek için örneklem oluşturma.

Amelia ücretsiz olarak dağıtılan bir yazılımdır. İlgi duyan okuyucular programı İnternet'ten indirebilirler. Ayrıca İnternet'te yazılımın ne şekilde kullanılacağına ilişkin açıklamalara da yer verilmiştir.

NORM

Windows işletim sistemi ortamında çalışan bu yazılımın çok değişkenli sürekli verilerdeki eksik verilere yönelik atama işlemleri için kullanıldığı belirtilmiştir.Yazılım, İnternet'ten ücretsiz olarak indirilebilmektedir. Programın ayrıca S-PLUS yazılımından veri alabildiği ve loglineer modellerdeki kategorik verilerde ortaya çıkan eksik verileri gidermek için de kullanılabileceği belirtilmiştir.

EMCOV

Amerika Birleşik Devletlerinde *Penn State Metodoloji Merkezi* tarafından geliştirilen yazılımın, MS-DOS ortamında çalıştığı, ücretsiz olduğu, *beklenti maksimizasyonu* yöntemini kullanarak kovaryans matrisi ve ortalama vektör hesaplamaları yapabildiği bildirilmiştir. Bunun yanında yazılım

[a] Deterministik beklenti maksimizasyonu (deterministic expectation maximization).

eksik verilerin yerine yeni değerlerin atanması işlemi için de kullanılabilmektedir.

SPSS Missing Value Analysis

Genel amaçlı istatistiksel analiz yazılımı SPSS'in ek modülü niteliğindeki bu yazılım aracılığıyla ölçüm verileri altı farklı rapor formatı kullanılarak analiz edilebilmektedir. Yazılımla özet analizleri yapılabilmekte ve istatistiksel algoritmalar kullanılarak eksik veriler için yeni değerler atanabilmektedir. SPSS'teki *Missing Value* adlı yazılımın aşağıdaki konularda araştırmacılara yararlı olduğu bildirilmiştir:[33]

1. Ciddi bir *eksik veri* sorunuyla karşı karşıya olunup olunmadığının belirlenmesi.
2. Eksik verilerin yerine tahmin değerlerinin atanması.
3. Eksik verilerin yerine çoklu atama değerlerinin veya regresyon analizi algoritma değerlerinin atanması.

Yazılımın kullanılmasıyla, araştırmacı analizlerini sadece maddeleri işaretlenmiş vakalarla sınırlamak yerine tüm anketleri kullanma imkanına sahip olur. Eksik verilerin yerine tahmin değerlerinin atanmasıyla istatistiksel olarak anlamlı sonuç elde etme olasılığı artır. Herhangi bir gruba ait kişilerin yanıt verme oranı düşük olsa bile bu yazılımla yapılan analizlerde tüm gruplar tam olarak temsil edilme özelliği kazanırlar.

MADDE-YANIT KURAMINI TEMEL ALAN YAZILIMLAR

Madde-yanıt kuramına göre test ve ölçeklerin oluşturulmasında, maddelerin güvenilirlik ve geçerliliklerinin saptanmasında klasik test kuramının ön kabullerinden bütünüyle farklı algoritmalardan hareket edilir. Maddelerin kalitesiyle insanların yetenek ve becerilerinin değerlendirilmesinde *olasılık modelleri* daha gerçekçi sonuçlar sağlar. Madde-yanıt kuramında ham puanlar değil, logaritmik dönüşüme uğratılmış değerler temel alınır. Bilim adamı test havuzu içinde biriktirdiği maddeleri kullanmadan önce kalibre etmelidir. Maddelerin kalibrasyonu ise, temel ölçüm parametrelerinin tahmin edilmesi suretiyle yapılır. Madde-yanıt kuramının temel alındığı yazılımlar hesaplama kapasiteleri, parametre sayıları ve hesaplama algoritmaları açısından büyük ölçüde farklılıklar gösterir. Son yıllarda mo-

dern ölçüm kuramına ilişkin olarak literatürde yer alan yazılımların en önemlileri aşağıdaki gibidir.

SCOREALL

Program, kağıt-kalem araçları kullanılarak yapılan testlerde madde-yanıt kuramındaki *maksimum olasılık* ve *bayesyen puanlarını* hesaplar. Scoreall yazılımı çıktılarından; güvenilirlik analizi sonuçları, doğru madde sayısı, Bayesyen ve maksimum olasılık yetenek puanları, her bir cevaplayıcı için Bayesyen yetenek tahmin değeri puanları elde edilir. Yazılımın MS-DOS ve Windows altında çalıştığı bildirilmiştir.

COSAN

İşletim sistemi olarak Windows ve MS-DOS altında çalışan bu yazılım Colin Fraser ve Roderick McDonald tarafından geliştirilmiştir. Yazılımın DOS sürümü İnternet ortamından ücretsiz olarak indirilebilir. COSAN'ı kullanmak isteyen bilim adamlarının öncelikle modelin kovaryans yapısına karar vermeleri gerektiği bildirilmiştir.[34] Belirlenen kovaryans yapısı daha sonra birinci düzey faktör analizi için yazılımın kendi içindeki modele dönüştürülür.

Girdi verileri, yazılıma korelasyon veya kovaryans matrisi olarak tanıtılır. Bunun için veriler simetrik matris şeklinde veya matrisin alt üçgenindeki veriler alınarak girilir. Yazılımda aşağıdaki tahmin parametrelerinin kullanıldığı bildirilmiştir:

1. Maksimum olasılık tahmini.
2. Genelleştirilmiş en küçük kareler yöntemi.
3. Ağırlıklandırılmamış en küçük kareler yöntemi.

Yazılımın Windows 95 ve Windows 98 altında çalışan sürümlerinde maksimum iş büyüklüğünün bilgisayarın Ram belleğine bağlı olduğu ifade edilmiştir.

EZDIF

Kullanıcı dostu olduğu bildirilen bu yazılım diferansiyel madde fonksiyonunun (DMF) analizi için kullanılır. Diferansiyel madde fonksiyonu Mantel-Haenszel testi ve lojistik regresyon analizi yöntemiyle belirlenir. Yazılımın özellikleri aşağıdaki gibidir:

1. Dinamik sıra düzenine sahip olması nedeniyle sanal olarak herhangi bir büyüklükteki verileri analiz edebilmesi.

2. Kapsamlı hata kontrol prosedürlerine sahip olması.
3. DMF etki büyüklüğü ve istatistiksel anlamlılık hesaplamalarını yapabilmesi.
4. Eşleştirme hesaplamalarını kontrol edebilmesi.
5. Madde etiketlerini okuyabilmesi.

Programın çıktılarında ise şu istatistik testler bulunmaktadır: (a) şans oranı, (b) Mantel-Haenszel ki-kare testi, (c) ki-kare anlamlılık düzeyi, (ç) Holland ve Thayer MH D-DMF testi, (d) MH D-DMF testinin standart hatası, (e) etki büyüklüğü, (f) ampirik madde özellikleri eğrisi, (g) lojistik regresyon analizi sonuçları.

Test Information

Testinfo, ikili olarak puanlanan maddelerde, madde-yanıt kuramına göre, lojistik madde analizi yapmak için kullanılan bir yazılımdır. Uyarlı test uygulamalarında test maddelerinin kısaltılmasının etkisini değerlendirmeye imkan sağlar. Bir dizi parametre değeri ile, testin ortalama ayırt edicilik puanı, zorluk puanı, tahmin etme parametresi, güvenilirlik katsayısı, beklenen test değeri, ortalama test değeri, dört farklı uzunluktaki testin bilgi verme fonksiyonu gibi konularda değerlendirmeler yapmaya imkan verir.[35] İnternet kaynaklarından sağlanan bilgilere göre yazılım MS-DOS ve Windows 95 altında çalışabilmektedir.

FACETS

Şikago üniversitesi, MESA psikometri laboratuarında görev yapan John Michael Linacre tarafından geliştirilen program, esas olarak Rasch madde-yanıt kuramına göre çalışır. Çevirisi, *yüzeyler* anlamına gelen adından da anlaşılacağı gibi, yazılım maddenin güçlük derecesi ve kişinin yetkinlik düzeyinin ötesinde ölçüme ilişkin başka yönleri / yüzeyleri de değerlendirmeye katar. Bu program birden fazla gözlemci / değerlendirici veya ölçülen nesnenin birden fazla yönü / yüzeyi veya özelliği varsa yararlıdır. Örneğin bir yabancı dil sınavında iki farklı öğretmen klasik sınav sorularını gramer bilgisi, kelime hazinesi, cümlelerin doğru kurulması gibi faktörlere göre değerlendirmek istiyorsa çok yüzeyli bir değerlendirme söz konusu olacaktır ve Facets bu amaçla geliştirilmiştir. Yazılım, değerlendirilen yüzeylerden her birisi için ayrı güvenilirlik katsayıları verir. Facets'in gerçek güvenilirliğin alt ve üst sınır değerlerini hesapladığı bildirilmiştir. Tam olarak bilinmesi mümkün olmayan *gerçek güvenilirlik* alt ve üst sınır değerlerinin arasında bir yerlerdedir. Verilerden pürüzlere neden olan çelişki

kaynakları çıkarıldıkça gerçek güvenilirlik *model güvenilirliğine* yaklaşır.[36]

MS-DOS ve Windows altında çalışan yazılımın DOS sürümü İnternet'ten ücretsiz olarak indirilebilmektedir. Facets'in Windows altında çalışan öğrenci sürümü MINIFAC ücretsiz olarak dağıtılmakta ve 2000 gözlem verisiyle sınırlı olarak orijinal sürümündeki aynı işlemleri yapmaktadır.Yazılımın İnternet ortamında aynı zamanda kullanıcı el kitabı bulunmaktadır.[37]

Facets yazılımının master aday öğrencilerinin seçimi, doktorların performansı, hastaların performansı, sporcuların performansı ve kamuoyu önünde konuşan bireylerin performanslarının ölçümü sonunda elde edilen verilerin analizi için kullanıldığı bildirilmiştir.[38]

XCALIBRE

Madde-yanıt kuramının temelini oluşturan parametrelerin tahminî değerlerini belirlemek için geliştirilmiş olan bir programdır. Yazılımda bunun için *maksimum olasılık analizi* kullanılarak belirli *tahmin* değerleri elde edilir.

Yazılımda MYK parametrelerinin maksimum olasılık ve Bayesyen tekniklerle doğru bir şekilde tahmin edilebilmesi için testin belirli bir uzunluğa ve belirli bir örneklem hacmine sahip olması gerektiği bildirilmiştir. Örneğin üç parametreli MYK modelinde maksimum olasılık teknikleriyle madde parametrelerini tahmin edebilmek için testte en az 50 madde ve örneklem hacminde ise 1000 cevaplayıcı bulunmalıdır. Programın, iki parametreli modellerde daha az sayıda ve daha küçük örneklemlerde *marjinal maksimum olasılık* tahmin değerini verdiği bildirilmiştir. Yazılımın MS-DOS ve Windows 95 / NT altında çalışan sürümleri bulunmaktadır.

SIMSTAT ve STATITEM

Simstat; ölçek ve test geliştirme, gözlemciler arası uyuşma, ölçeğin güvenilirliği ve duyarlılığı gibi konularda içerdiği alt modüller aracılığıyla kullanıcılara önemli ölçüde kolaylık sağlayan bir yazılımdır.

Programın Statitem adlı modülü ile çoktan seçmeli sorularda klasik madde analizlerini yapmak mümkündür. Yazılım her bir maddeye özgü istatistik tekniklerle maddenin ayırt edicilik özelliğini, ölçeğin iç tutarlılığına yaptığı katkıyı belirleyebilmektedir. Geliştirilen ölçeğin güvenilirlik ve geçerliliğini saptama; düşürülmesi, değiştirilmesi veya düzeltilmesi gereken maddeleri belirleme konusunda yazılımın araştırmacıya geniş olanaklar sağladığı bildirilmiştir.

Yazılım 2000 maddeye kadar madde analizi yapabilmektedir. Programda vak'alardaki eksik verilerin liste bazında çıkarıldığı veya bu yanıtların geçersiz sayıldığı bildirilmiştir.[39] Kaynaklardan edinilen bilgiye göre yazımla aşağıdaki istatistikî analizler yapılabilmektedir:

1. Madde istatistikleri (frekans dağılımları, madde-toplam iki serili korelasyon analizi, nokta-iki serili korelasyon analizi, Cronbach alfa, kullanıcı tanımlı ayırt edicilik indeksi).
2. Madde-toplam puan korelasyon matrisi ve madde özellikleri eğrisi (ham veya standardize edilmiş puanlara göre).
3. Eşit büyüklükteki çeşitli gruplar arasında madde-toplam puan karşılaştırmaları (2 ilâ 10 grup arasında).
4. Ölçek toplam puan istatistikleri (aritmetik ortalama, medyan, minimum ve maksimum değerler, standart sapma, standart hata, çarpıklık ve basıklık katsayıları, Cronbach alfa, Ferguson ayırt edicilik indeksi vb. gibi).
5. Doğru yanıt yüzdesi, iki serili korelasyon, nokta-iki serili korelasyon, toplam puanların frekans dağılımı gibi özet istatistikî analizleri.

MS-DOS ve Windows ortamlarında çalışan yazılımın yukarıda sayılan özelliklerinin dışında yüksek çözünürlüklü grafikleri çizebildiği, ASCII ve SPSS gibi çeşitli veri tabanlarından ve çalışma sayfalarından veri alabildiği bildirilmiştir.[40]

ETIRM

B. Hanson tarafından geliştirilmiş, iki şıklı maddelerde madde-yanıt kuramına göre geliştirilen, üç parametreli lojistik modeli ölçen bir yazılımdır. Yazılımın çok şıklı maddelerde ise "genelleştirilmiş kısmî kredi modelini" (generalized partial credit model) uyguladığı bildirilmiştir. ETIRM, parametre tahminlerinin yanında gruplar ve bireysel cevaplayıcılar için gizli değişkenin dağılımını tahmin etmek için de kullanılır. Ücretsiz olarak dağıtılan yazılım hakkında daha fazla bilgi İnternet ortamından sağlanabilir.[41]

TEK PARAMETRELİ RASCH YAZILIMLARI

Rasch yazılımları; çoktan seçmeli soru üretmek ve çok maddeli test geliştirmek için kullanılan istatistiksel analiz araçlarıdır. Rasch yazılımlarında

dört farklı ölçüm modeli test edilir: İkili veri yapılarının test edilmesi, kısmî kredili yanıtların test edilmesi, dereceleme ölçeklerinin test edilmesi ve çok yüzeyli verilerin test edilmesi. Rasch modelinde, klasik test kuramından farklı kavramlar kullanılır. Örneğin, tek boyutluluk, uyuşma, güçlük/yetenek tahmin değeri ve hata, maddelerin güçlük derecelerinin konumları, kişilerin yetenek düzeylerinin konumları gibi terimler bunlar arasındadır. Aşağıdaki bölümde önce bir parametreyi ve daha sonra iki ve üç parametreyi test eden Rasch yazılımlarına ilişkin özet bilgiler verilmiştir.

Tek parametreli yazılımlar, ikili veri yapısına sahip maddelerin analizinde, sadece maddenin güçlüğünü belirlemek için kullanılan programlardır. Bir parametreli lojistik modelde güçlüğün dışında, *maddenin ayırt edicilik özelliği* bilinen *sabit bir değer* olarak programa atanır. Tek parametreli modellerin kullanılabilmesi için ön test niteliğinde en az 100 kişi üzerinde ölçüm yapılmış olması gerektiği belirtilmiştir.[42] Bununla birlikte, pilot araştırma sırasında doğru olmayan uygulamalarla 200 kişilik bir örnek kütleye ulaşmak yerine doğru ve dikkatli bir şekilde yapılan 50 kişilik bir örneklemden elde edilen veriler daha sağlıklı sonuçlar verecektir.[43]

QUEST

Raymond J. Adams ve Siek-Toon Khoo tarafından geliştirilen yazılım iki şıklı, çoktan seçmeli, çok şıklı ve kısmî kredili maddeleri[a] analiz etme özelliğine sahiptir. Programın testteki maddelerin parametre tahminlerini, vak'aların tahmin değerlerini, uyuşma istatistiklerini ve alt gruplara ait istatistikî analizleri yaptığı bildirilmiştir. Bunların dışında değişik yüzde değerlerini, her bir maddenin nokta-iki serili korelasyon katsayılarını ve değişik güvenilirlik indeks değerlerini verdiği bilinmektedir.[44]

Quest'in veri işleme düzenlemesinde her defasında tek bir adımdan oluşan küme yaklaşımı, verilerin özelliklerini değişik açılardan aynı zamanda ve açılan pencereler içinde inceleme olanağı sağlayan etkileşim yaklaşımı veya ikisinin birleşiminden oluşan kombinasyon yaklaşımlarından biri seçilebilmektedir.

Quest'in diğer test analiz yazılımlardan farklı olan birkaç özelliğe sahip olduğu bildirilmiştir ve bunlar aşağıdaki gibidir:

Alt grup ve alt ölçek analizleri. Programın araştırmacı tarafından yapılacak alt grup ve alt ölçek tanımları içinde bu gruplara özgü istatistiksel analizleri yapması anlamına gelmektedir.

[a] Kısmî kredi, çoktan seçmeli ve iki doğru şıkkı bulunan sorularda sadece bir şıkkın işaretlenmesi halinde bu sorulara kısmî bir puan verilmesi.

Kullanıcı tanımlı değişkenlerin üretilmesi. Araştırmacının kendi tanımlayacağı değişkenlerle alt gruplarda, alt ölçeklerde ve testin tamamında diğer değişkenlerle birlikte değişik analizleri yapabilmesi.

Çipolu parametre tahmini. Bu özellik herhangi bir maddeye veya vak'aya ait tahmin değerinin daha önceki analizlerden elde edilmiş bilinen bir değere sabitlenmesi anlamına gelir. Bu özellik test ve madde bankasındaki soruların / maddelerin eşitlenmesini kolaylaştırır.

Eksik veri analizi. Quest'in veri matrisinde boşluk bulanan hücreleri değişik şekillerde ele aldığı, eksik verilerle ilgilenmek için değişik prosedürlere sahip olduğu, formlar arasında ilişki bulunması halinde birkaç test formunu birden aynı zaman diliminde kalibre edebildiği bildirilmiştir.

Dosya ihracı. Yazılımın bir diğer özelliği olarak, veri dosyalarını metin dosyası olarak değişik formatlarda (sekmeli, boşluklu, virgülle ayrılmış biçimde veya alan boşluğu bırakılmış olarak) diğer programlara gönderebilmesinden söz edilmiştir.

Diferansiyel madde fonksiyonu. Yazılımın Mantel-Haenszel ve Rasch diferansiyel madde fonksiyonunu analiz ettiği bildirilmiştir. Test maddelerinin farklı bir şekilde çalışıp çalışmadığını görmek için kullanılan bir yöntem olan diferansiyel madde fonksiyonu (DMF) ile iki veya daha fazla grup cinsiyet, ırk din, etnik köken ve kültürel özellikler gibi faktörler açısından karşılaştırılır. Bu karşılaştırma sonucunda test maddelerinin yetkinlik ve kabiliyet dışında bir grup lehine veya aleyhine yanlılık özelliği taşıyıp taşımadığına istatistiksel olarak karar verilir. Testte sadece tek bir maddenin DMF özelliğine sahip olması sorun olmamakla birlikte birden fazla maddenin DMF özelliğine sahip olması sonuçları büyük ölçüde çarpıtır.

RASCAL

İkili veri yapıları için Rash parametrelerini hesaplayan bir yazılımdır. Sabit tutulan madde parametrelerini temel alarak ölçeği yeniden oluşturur. Rascal, tarayıcılarla alınan veya el ile girilen ASCII formatındaki verileri analiz etme özelliğine sahiptir.[45] Yazılım, parametre tahminini maddenin güçlük ve kişilerin yetenek düzeylerine göre yapar. Rascal'ın MYK puanlarını otomatik bir biçimde ve doğrusal bir dağılım özelliği gösterecek şekilde yorumlanabilir hale getirdiği bildirilmiştir. Yazılım aynı zamanda

belirli madde parametrelerini değişik değerlere sabitleyerek ölçekteki diğer maddeleri otomatik olarak kalibre edebilmektedir. Bu özellik sayesinde ölçekteki birbiriyle bağlantılı olan maddelerin güvenilirliğini test etmek mümkündür.

Rascal, ölçekteki her bir madde için "maddenin güçlük derecesi parametresini" tahmin etmekte, Pearson ki-kare uyuşma istatistiği analizini yapmakta ve güçlük tahmininin standart hatasını hesaplayabilmektedir. Yazılım, ham puanları madde-yanıt puanlarına dönüştüren bir tablo ortaya koymaktadır. Bu tabloda yeteneğin en yüksek olasılık değerlerini, standart hatasını frekans ve yığışımlı frekans değerlerini, yüzdelik değerlerini ve dönüştürülmüş puanları görmek mümkün olmaktadır. Yazılım, ayrıca testi alan her bir kişi için maksimum olasılık (MYK) puanlarını dış bir metin dosyasına kayıt etmekte ve bu dosya gerektiğinde diğer istatistiksel yazılımlarla birlikte kullanılabilmektedir.

RSP

İngilizce *resource specialist program* kelimelerinin baş harflerini isim olarak alan bu yazılım, Rasch modeli çerçevesinde maddelere gelen ikili yanıtları analiz eder. Program; test tasarımı, madde bankası oluşturma, uyarlı test oluşturma, madde yanlılığını saptama, değişimin ölçümü ve eğitsel başarının değerlendirilmesi gibi amaçlarla kullanılabilmektedir.

Programın birbiriyle ilintili 32 kadar test formunu aynı zamanda analiz edebildiği bildirilmiştir. Her bir testte maksimum madde sayısı 96 olarak belirlenmiştir. Testin işleyebildiği maksimum madde sayısı ise, 600'dür. Kişi sayısı için bir sınırlama getirilmemekle birlikte yazılımda eksik verilerin kullanılmasına izin verilmemiştir. RSP, kişi parametrelerinin dağılımıyla ilgili herhangi bir varsayımda bulunulmadığı durumda madde parametrelerinin *şartlı maksimum olasılık tahminini* yapabilmektedir. Ayrıca programın normal dağılım varsayımı altında *marjinal olasılık tahmin değerlerini* hesapladığı bildirilmiştir.[46]

RUMMFOLD

D. Andrich tarafından geliştirilen Rummfold, tutum ölçeklerinde ve tercihleri belirlemeye yönelik olarak kullanılan diğer ölçeklerde *Rasch aşamalı açığa çıkaran ölçüm modelini* – RAAÇÖM (Rasch unfolding measurement model – RUMM) kullanan bir yazılımdır. Program bireyin hem özellik düzeyinin tahmin edilmesini temin etmekte ve hem de tek parametreli Rasch lojistik modelinde parametrelerin konumunu belirlemeye imkan sağlamaktadır. Bu modelde, maddelere ait madde-yanıt fonksiyonu için simetrik tek bir en yüksek nokta belirlenmiştir. Model, "bireyin özellik

seviyesi ve maddenin konumu her iki yöne doğru arttıkça, doğru yanıt verme olasılığının azaldığı" varsayımına göre çalışmaktadır.[47]

RUMM

D. Andrich, A. Lyne, B. Sheridan ve G. Luo (1998) tarafından geliştirilen yazılımın Rasch madde analizini yaptığı, Windows ortamında çalışan diğer yazılımlar arasında güçlü ve esnek bir program olduğu bildirilmiştir. Rumm'un (Rasch Unidimensional Measurement Models) ölçülmek istenen yapıyla ilgili olarak maddelere ilişkin tablolar, grafikler ve şemalar üreterek maddelerin yorumlanmasını kolaylaştırdığı ifade edilmiştir. Rumm çeşitli türdeki madde analizi yöntemlerini desteklemektedir ve bu analizler aşağıdaki gibidir:[48]

1. Rasch modellerini kullanarak dinamik ve etkileşimli olarak madde –yanıt analizlerini yapması.
2. İstatistiksel olarak tutarlı parametre tahminleri veren çift koşullu algoritmalar kullanarak Rasch ölçüm tahminlerini vermesi.
3. Çoktan seçmeli soru analizleri ve karmaşık çeldirici analizleri.
4. Eşit ve eşit olmayan kategorili çok dereceli madde analizleri.
5. Maddelerin güçlük derecesiyle kişilerin yeteneklerinin birlikte değerlendirilmesi.
6. Her bir soru için madde özellikleri eğrisinin çizilmesi ve kişi madde uyuşumunun belirlenmesi.
7. Farklı testlerden elde edilen ham puanların tek bir ölçek puanına dönüştürülmesi.

Rumm'un, güçlük ve yetenek konum parametrelerini verirken aynı zamanda belirli madde lokasyonlarına bağlı olarak çipolama yapma imkanı da sağladığı bildirilmiştir. Yazılımın içerdiği *madde-özellik testi*, madde parametrelerinin bütün katılımcılar için istikrarlı bir sonuç verip vermediğini belirler. Bu test aynı zamanda bütün maddeleri bir araya getirerek *genel test uyuşma indeks değerini* verir. Analiz sonucunda maddelerin ayırt etme katsayısı ,20'nin altında olan maddeler yetenekli kişilerle acemi kişileri iyi ayırt etmediği için ölçekten çıkarılır. Yazılıma ilgi duyan okurlar İnternet'ten deneme sürümünü bilgisayarlarına indirerek bu programı daha ayrıntılı bir şekilde inceleyebilirler.

WINMIRA

Kategorik değişkenlerle yapılan modelleri analiz etmek için geliştirilmiş olan bir yazılımdır. Nominal, sürekli veri veya her ikisinin kombinasyonu şeklindeki değişkenler için gizli değişkenlere dayalı modelleri test eder. Winmira, ikili ve çoklu verilerde Gizli Küme Analizi'ni (Latent Class Analysis – LCA)[a] Rasch modelini (RM), karma Rasch modelini (KRM) ve hibrid modelleri (HYBRID) test etmek için oluşturulmuştur. Yazılımın çoklu veri yapılarında dört farklı alt modeli test ettiği bildirilmiştir: kısmî kredi modeli, dereceleme ölçeği modeli, eşit uzaklık modeli ve yayılma modeli. Winmira ayrıca ölçek oluşturmak için de kullanılabilmektedir. Maddelerin seçilmesi, katılımcıların homojen kümeler şeklinde gruplandırılması, tipik olmayan yanıt biçimlerinin saptanması (kopya çekme veya sınav hakkında daha önceden bilgi sahibi olma) gibi konularda analizler yapabilen Winmira, genel amaçlı istatistik yazılımı olan SPSS'ten veri alabilecek ve ona veri verebilecek bir şekilde tasarlanmıştır.[49]

WINSTEPS

İstatistiksel analiz yazılımı SPSS, güvenilirlik analizlerini klasik ölçüm kuramındaki *gerçek puan modeline* göre yaparken Chicago Üniversitesi psikometri lâboratuvarı tarafından geliştirilen Winsteps adlı yazılımın modern yaklaşımlardan biri olan Rasch yöntemini kullandığı bildirilmiştir. Ancak Winsteps hesaplamaları tek parametre modeline göre yapmaktadır. Bu kısıtlayıcı özelliğinin yanında çıktı tablolarının oldukça işlevsel olduğu, testi alan kişinin yeteneği ile test maddelerinin güçlük derecesini yan yana verdiği ifade edilmiştir. Programın çıktılarından madde istatistiği, kişi istatistiği ve kalan (residual) istatistiği elde edilmektedir. Madde istatistiği, hangi maddelerin zayıf yazıldığına ilişkin bilgi verirken, kişi istatistiği her bir birey için *theta* yetkinlik değerini verir. "Kalan" istatistik değerleri ise, model ile gözlem verileri arasındaki uyuşma derecesini gösterir.[50] En son 2003 yılında 3.4 sürümü yayımlanan yazılımın 1 milyon kişiye uygulanabileceği, 30 bin maddeyi işleyebileceği ve her madde için 255 kategoriye kadar analiz yapabileceği iddia edilmiştir. Yazılımın öğrenci veya değerlendirme amaçlı sürümüne MINISTEP adı verilmiştir ve bu sürüm İnternet'ten ücretsiz olarak indirilebilmektedir.

---

[a] Gizli küme analizi, *kategorik gözlem değişkenlerinin* altında yatan gizli bir faktöre göre vak'aların veya araştırmaya katılan kişilerin gruplandırılmasıdır. Sıralı ve sınıflandırılmış ölçek verilerinde kullanılır.

İKİDEN FAZLA PARAMETRELİ RASCH YAZILIMLARI

Madde-yanıt kuramında maddenin özelliği olarak üç parametre vardır. Bunlar; (a) maddenin ayırt edicilik özelliği, (b) maddenin güçlük derecesi ve (c) şans faktörüdür. Maddeye verilecek yanıtların doğru çıkma oranı veya olasılığı bu üç parametreye göre belirlenir. Analize alınacak her madde bu üç parametre açısından incelemeye tâbi tutulur. Söz konusu üç parametre genellikle madde kişilere uygulanmadan önce hesaplanır ve bu hesaplamaya ön test adı verilir. İki parametreli yazılımlarda, bir *i* maddesine *j* yetkinliğine sahip bireyin doğru yanıt verme ihtimalinin hesaplanabilmesi için modele *güçlük derecesi* ve *maddenin ayırt edicilik özelliği* birlikte katılır. Üç veya dört parametreli yazılımlarda ise sayılan parametrelere şans faktörü de ilave edilir.

Bir yazılımda iki parametreli modelin kullanılabilmesi için ön test niteliğinde en az 200 kişi üzerinde ölçüm yapılmış olması gerektiği belirtilmiştir.[51] Üç parametreli analizlerde ise örneklemdeki kişi sayısı en az 1000 olmalıdır.

Aşağıdaki paragraflarda 2PL (iki parametreli lojistik) ve 3PL (3 parametreli lojistik) modelleri test eden yazılımlar hakkında özet bilgiler verilmiştir. Bu yazılımlar hakkında daha fazla güncel bilgi edinmek isteyen okurlara İnternet kaynaklarına başvurmaları önerilir.

BILOG-MG 3

Rasch yazılımları içinde üç parametreli modellerin analizi için en ideal yazılımın Biolog olduğu ileri sürülmüştür. Programın el kitabında farklı sayıda parametrenin model oluşturmada ne şekilde kullanılabileceğine ilişkin açıklamalar yapılmıştır.

Michele Zimowski, Eiji Muraki, Robert Mislevy ve Darrell Bock tarafından yazılan Biolog MG, MYK parametrelerini çoklu gruplar için analiz edebilmekte ve maddelerin farklı gruplarda nasıl bir davranış gösterdiğini test eden *diferensiyel madde forksiyonunun* (DMF) incelenmesine izin vermektedir.

Scientific Software International adlı şirket tarafından pazarlanan yazılımın İnternet'teki ağ kümesinde hesaplama özellikleri konusunda aşağıdaki bilgiler verilmiştir:[52]

1. Grafik kullanıcı ara yüzü.
2. İkili maddelerin etkili bir şekilde analiz edilmesi (çoktan seçmeli işaretleme, doğru yanlış işaretlemesi, boş bırakma, mevcut olmama gibi sorunları ele alabilmesi).
3. Büyük ölçek üretilmesini gerektiren analizleri yapabilmesi, aynı zamanda birden fazla grubun verilerini değerlendirebilmesi.
4. Testin tamamına veya alt testlere ilişkin madde analizleri yapabilmesi.
5. Eşit olmayan grupları eşitleyebilmesi.
6. Diferansiyel madde fonksiyonunu hesaplayabilmesi.
7. Madde uyuşma istatistiklerini, teorik ve ampirik güvenilirlik katsayılarını vermesi.
8. MYK grafikleri ve bunların MS-Wrod'e aktarılması.

Yazılımın İnternet ortamında çevrimiçi olarak ulaşılabilecek ayrıntılı bir YARDIM dosyasının bulunduğu bildirilmiştir. Programı kullanan araştırmacıların yardım dosyalarında ara yüzlerle ilgili gerekli açıklamaları, kod örneklerini ve diğer uygulama örneklerini bulabilecekleri bildirilmiştir.

LOGIMO

H. Kelderman ve R. Steen tarafından geliştirilen Logimo, sıralı ölçek verilerine dayalı loglineer modelleri ve loglineer MYK modellerini test ve tahmin etmek için kullanılan bir yazılımdır. Loglineer modellerde, kategorik değişkenlerin çeşitli kombinasyonları ve logaritmik olasılıkları doğrusal modellerle açıklanır. Logimo'da doğrusal model TYVA tipi parametrelerin ana etkisi ve etkileşim etkisiyle açıklanır.

Logimo'nun ayırt edici özelliğinin verileri çapraz tablolar içinde sunmak yerine, loglineer[a] model parametrelerine dayalı olarak maksimum olasılık tahminini yapması olduğu bildirilmiştir. Yazılım bunun dışında model parametrelerinin standart hatasını, gözlemlenen ve beklenen değerlerle ilgili istatistikleri, olasılık çekirdeğini, uyuşma istatistiğinin olasılık oranını ve Pearson uyuşma istatistiğini vermektedir.[53]

---

[a] Loglineer model analizinde, kategorik verilere dayanan çok değişkenli çapraz tablolarda her bir hücredeki frakans değerleri doğal logaritma değerlerine çevrilir ve daha sonra bu logaritmik değerler doğrusal veriler olarak ele alınır. Loglineer model analizinde, araştırmacı tabloya alınan değişkenlerden hangisinin belirli bir etkinin ortaya çıkmasında muhtemelen en fazla etkisi olduğunu bulmaya çalışır.

MSP

Yazılımın adı, İngilizce *Multidimensional Item Bank in the Polytomous Mokken IRT* kelimelerinin kısaltılmasından gelmektedir. Yazılım, ikili ve çok dereceli maddeler için Mokken ölçeği verilerini analiz etmek üzere I.W. Molenaar, P. Debets, K. Sijtsma ve B.T. Hemker (1994) tarafından geliştirilmiştir. Programda *Nonparametrik yığışımlı madde-yanıt kuramı* temel alınmıştır.

Yazılımın aynı zamanda Nonparametrik madde-yanıt kuramına göre ölçek veya test oluşturmak için de kullanılabileceği belirtilmiştir. Program cevaplayıcılar için sıralı ölçek ve tek tek maddeler için sıralı ölçek dereceleri oluşturmaktadır. Keşfedici araştırma kapsamında yapılan çalışmalarda yazılımın otomatik olarak çalışan madde seçim prosedürüne sahip olduğu bildirilmiştir. Yazılım kümeleme analizi ve ölçekleme analizi sonuçlarına göre maddeleri / göstergeleri tek boyutlu alt ölçeklere atayabilmektedir.[54] Verilen bir ölçeğin veya kavramsal yapının bir veya daha fazla boyut üzerinde ölçeklenip ölçeklenemeyeceğini belirleyen yazılım, diğer taraftan kişileri ölçek puanlarına göre sıralamaya tâbi tutmakta ve ölçümün kalitesinin değerlendirilmesine imkan sağlamaktadır. MSP yazılımının aşağıda belirlenen amaçlarla kullanılabileceği belirtilmiştir:[55]

1. Bir havuz içinde toplanan maddelerden yararlanarak aşmalı bir biçimde bir veya daha fazla boyutlu/faktörlü ölçek oluşturma.
2. Belirli bir ölçeğin model uyuşumunun değerlendirilmesi. Modelin uyuşumunun güvenilirlikle bağıntılandırılması ve modelle uyuşum göstermeyen maddelerin elenmesi.
3. Maddelere verilen cevap biçimlerinin ne ölçüde *yığışımlı Guttman skalogram* biçimine benzediğinin Loevinger *H* değeri ve ilgili anlamlılık testleriyle incelenmesi.

Programın 100 kadar çok dereceli maddeyi, maksimum 10 sıralı kategori içinde ve belirli bir ranj aralığında (900 madde adımı) ele alarak hesaplama yapabildiği bildirilmiştir. Yazılım 32.000 kişiye kadar genişleyen kapsamlı bir veri işleme kapasitesine sahiptir, ancak eksik veri tanımlamasına izin vermemektedir.

MULTILOG

Madde-yanıt kuramına göre çok dereceli maddelerin parametre tahminlerini yapan bir yazılımdır. Programın derecelendirilmiş ölçekler için *Samejima lojistik modelini,* çok kategorili sorular/maddeler için *multinomial logit modelini,* tahminen işaretleme yapmaya imkan veren çoktan seçmeli sorular/maddeler için ise *Bock-Samejima-Thissen* ve *Masters kısmî kredi* modelini temel alarak hesaplama yaptığı bildirilmiştir.[56]

Yazılımda aynı zaman diliminde birden fazla alt test analiz edilebilmektedir. Alt ölçek veya alt test sonuçları daha sonra birleştirilerek her bir katılımcı için testin tamamına ait ağırlıklı bir puan elde edilmektedir.

PARELLA

Tutumlar ve tercihler gibi gizli kişilik özelliklerini ölçmek için çoğunlukla bileşik test maddelerinden oluşan dereceleme ölçekleri kullanılır. Bu derecelemeölçeklerindeki maddeler/göstergeler öyle bir şekilde yapılandırılır ki, her bir birey tercih ettiği veya kabul ettiği maddeden küçük bir uzaklıkta konumlandırılır. Parella isimli yazılımın, iki şıklı tutum ve tercih ölçümlerinde madde-yanıt kuramını temel alarak kişiler ve maddelerin her ikisini de temsil eden tek boyutlu bir ölçek oluşturulmasına katkı sağladığı belirtilmiştir. Yazılımın hesaplama mantığında *yakınlık ilişkileri* bireyin vermiş olduğu cevapların niteliğini belirleyen temel öge olarak görülmüştür. Diğer bir deyişle, kişi ve madde arasındaki psikolojik mesafenin küçük olması nispetinde bireyden olumlu yanıt alma olasılığı da o ölçüde artmaktadır.

Program, maddelerin konumlarını tahmin etmek için *marjinal maksimum olasılık analizini* ve kişilerin konumlandığı yeri tahmin etmek için ise *nonparametrik yoğunluk fonksiyonu* analizini kullanmaktadır. Her bir bireyin yerini tahmin etmek için ise, *beklenen aposteriori tahmin değeri* ölçüsü kullanılır.

Parella'nın veri yükleme dosyasına en fazla 60 madde, 10 alt örneklem ve 300 kişiye ait verilerin yüklenebileceği belirtilmiştir. Yazılımın diğer özellikleri hakkında güncel bilgi ve fikir edinmek isteyen okurlara İnternet kaynaklarına başvurmalarını öneririz.

PARSCALE

Madde-yanıt kuramını test eden yazılımların içinde en kullanışlısı olduğu ifade edilmiştir. Program değerlendiricilerin katı bir biçimde işaretleme yapma özelliklerini de değerlendirebilmektedir. Analiz özellikleri açısından programın aşağıdaki modüllere sahip olduğu bildirilmiştir:

1. Bir, iki ve üç parametreli lojistik modeller.
2. Diferansiyel madde fonksiyonu.
3. Master kısmî kredi modeli.
4. Genelleştirilmiş kısmî kredi modeli.
5. Samejima derecelendirilmiş tepki modeli.
6. Değerlendirici etkisi analizi.
7. Çok dereceli maddelerde çoklu grup analizleri.

Yazılımın dereceleme ölçekleriyle çoktan seçmeli maddelerin tahminli veya tahminsiz olarak birlikte bulunmasına ve birlikte ele alınmasına izin verdiği bildirilmiştir. Yazılım Windows ortamında 98, NT, 2000, ME ve XP işletim sistemleriyle birlikte çalışabilmektedir. Programın kullanılmasıyla ilgili olarak İnternet ortamında kapsamlı yardım dosyaları bulunmaktadır.

ConQuest

Program madde-yanıt ve gizli regresyon modellerini değişik Rasch modellerine uygulama özelliğine sahiptir. İkili, çok dereceli, tek boyutlu veya çok boyutlu ölçekler programda analiz edilebilmektedir. ConQuest'in son psikometrik yöntemlerden çok yüzeyli madde-yanıt modelini, gizli regresyon modellerini işlediği ve makul sonuçlar verdiği bildirilmiştir. Yazılım, madde-yanıt kuramı ve regresyon analizinin bütünleşmesinden oluşmuş ve aşağıdaki analizleri yaptığı bildirilmiştir.

1. Rasch'ın basit lojistik modeli.
2. Dereceleme ölçeği modeli.
3. Kısmî kredi modeli.
4. Sıraya sokulmuş ayırma modeli.
5. Doğrusal lojistik test modeli.
6. Çok yüzeyli modeller.
7. Genelleştirilmiş tek boyutlu modeller.
8. Çok boyutlu madde yanıt modelleri.
9. Gizli regresyon modelleri.

ConQuest, belirlenen modellerin parametreleri için marjinal maksimum olasılık tahminlerini verir. Yazılım; grafik ara yüzlü, komut yönelimli,

veya konsol ara yüzlü olarak kullanılabilmektedir. Grafik ara yüzlü olan sürümün tüm Windows ortamlarında çalıştığı bildirilmiştir.

LPMC-WIN

Yazılımda *doğrusal lojistik test modeli* ve *çok boyutlu modeller* de dahil olmak üzere, Rasch modellerini tahmin etmek için doğrusal kısmî kredi modelleri kullanılmıştır.

FAKTÖR VE BİLEŞEN ANALİZİ YAZILIMLARI

Genel amaçlı istatistik yazılımlarının (SPSS, SAS, S-PLUS, Statistica, Sysas vb.) hepsinde faktör analizi yöntemi bulunur. Ancak bu yazılımlardaki faktör analizi teknikleri sürekli veriler için uygundur. Sosyal bilimciler ise daha çok *Doğru / Yanlış* şıklarından oluşan ikili verilerle veya *Kuvvetle Katılıyorum* şıkkından başlayıp *Şiddetle Reddediyorum* şıkkına kadar uzanan çok dereceli Likert tipi ölçeklerle çalışırlar. Sıralı ölçek niteliğindeki veriler için klasik faktör analizi yöntemi çok uygun değildir. Eğer klasik faktör analizi yöntemi uygulanmak isteniyorsa verilerin tetrakorik veya polikorik korelasyon katsayılarına dönüştürülmesi gerekir. Sıralı ölçek verilerinde faktör analizi yapabilmek için bu amaçla hazırlanmış bazı özel yazılımlardan da yararlanmak mümkündür.

MicroFACT

Niels G. Waller (1996) tarafından ikili ve çok dereceli sıralı ölçek verilerinde keşfedici faktör analizi yöntemini uygulamak için geliştirilmiştir. MicroFACT çok dereceli verileri, önce tetrakorik ve/veya polikorik korelasyon katsayılarına dönüştürmekte ve ondan sonra faktör analizi yöntemini uygulamaktadır. Böylece bu tür verilerin içerdiği boyut veya faktör sayısını tespit etmek mümkün olmaktadır. Sosyal bilimlerde, politik bilimlerde, davranış bilimlerinde, eğitim bilimleri ve psikolojide araştırmacılar daha çok eşit aralıklı ölçek verileriyle çalıştıklarından MicroFACT parametrik nitelikte olmayan bu verilerden faktör çıkarılmasını sağlar. Yazılımın özellikleri aşağıdaki gibi sıralanmıştır.

1. Dış dosyalardan veri alabilmesi.
2. Ağırlıklı varimaks, Harris-Kaiser döndürme yöntemi ve esnek döndürme yöntemleri gibi tekniklere sahip olması.

3. Lokal sorunlara yakalanmamak için tesadüfî oryantasyonlara göre döndürme yöntemleri.
4. Faktör modelini, yapısını ve faktörler arası korelasyonları önceki analizlerden ve yayımlanmış çalışmalardan alabilmesi.
5. Kalan değerler grafiği, uyuşma indeksi gibi yeni modellerin değerlendirmesini yapabilmesi.
6. Faktör yükleri grafiğini çizmesi.
7. Hataları ele alma prosedürleri.

Microfact yazılımının Windows 95/98/ME/2000/NT altında çalışabildiği ve bunun için bilgisayarın sabit diskinde en az 5 MB'lık bir yer olması gerektiği bildirilmiştir.

TASTFACT

Bu yazılım, madde-yanıt kategorilerine ait tüm istatistik testleri içerir. Yazılım; madde seçimi, çoklu alt test, katılımcıların çoklu gruplar altında incelenmesi, dış kriterle korelasyon analizi gibi teknikleri uygulayabilmektedir. Eksik verili veya eksik veri bulunmaksızın tetrakorik korelasyonlara dayalı olarak faktör analizlerini yapabilen programın sahip olduğu diğer özellikler aşağıdaki gibi sıralanmıştır:[57]

1. Marjinal maksimum olasılık (MMO) keşfedici faktör analizi.
2. İkili veriler için klasik madde analizi.
3. Tetrakorik korelasyonlar ve temel bileşenler analizi.
4. Ölçek maddeleri için klasik tanımlayıcı istatistikî analizler.
5. On faktöre kadar işlem kapasitesi.
6. On beş faktöre kadar Monte Karlo tekniklerinin uygulanması.
7. Faktör yükleri için dik açılı ve esnek döndürme yöntemleri.
8. Tahmin edilerek yapılan işaretlemeler ve ulaşılamayan maddeler için düzeltme fonksiyonu.

Yazılımın geniş kapsamlı çevrimiçi yardım dosyalarına sahip olduğu bildirilmiştir. Windows 95, 98, NT, ME, 2000, XP işletim sistemleri altında çalışabilen programın İnternet ortamında örnek çıktı dosyalarını görmek mümkündür.

VARCL

Maksimum olasılık yöntemiyle varyans bileşenleri analizini icra eden bir programdır. Çok düzeyli analiz için Fisher puanlama algoritmasını kullanır. Programın ana modülü üç düzeyli yuvalanmaya izin verir. Fakat paket programın kullanılmasıyla birlikte yuvalanma biçiminin dokuz düzeye kadar çıkabildiği bildirilmiştir. Program regresyon katsayılarını ve varyans parametrelerini geçici değerler şekline gelmesi için zorlar ve kovaryans parametrelerini sıfır şeklinde düzeltir. Yazılımda kullanıcılar, bağımlı değişken için normal veya normal olmayan dağılım biçimlerini tercih edebilirler.[58]

SCA

H.A.L. Kiers tarafından geliştirilen bu yazılım, iki veya daha fazla grupta karşılaştırmalı olarak temel bileşenler analizini yapar. Programda "bileşenin" anlamı bileşen ağırlıklarıyla tanımlanmıştır. Bileşenin anlamı bütün gruplar için aynıdır, çünkü tüm gruplarda aynı bileşen ağırlıkları kullanılmıştır. Program, birlikte vuku bulma analizini genelleştiren bir yönteme sahiptir. Bir gruptaki hangi bileşenlerin diğer gruptaki bileşenlerle benzerlik gösterdiğini belirlemeye yardım eder. Yazılımın 20 gruba kadar olan verileri analiz edebildiği bildirilmiştir. Değişken sayısı iki grup için 70 iken grup sayısı 20'ye çıktığında 30'a düşmektedir. Maksimum katılımcı sayısı ise 9999 olarak belirlenmiştir. Program MS-DOS altında çalışmaktadır.[59]

ETKİ BÜYÜKLÜĞÜ VE META ANALİZİ YAZILIMLARI

ES

Etki büyüklüğü anlamına gelen İngilizcedeki *effect size* kelimelerinin kısaltmasını ad olarak alan bu yazılım meta analizlerinde "standardize edilmiş ortalamadan farklılık istatistiğini" (Cohen *d*) hesaplar. Yazılımın 40 farklı türdeki veriyi analiz edebildiği bildirilmiştir. Yazılımın özellikleri aşağıdaki gibi belirlenmiştir:

1. İkili veya sürekli ham verileri işleyebilmesi.
2. Ortalamalar ve standart sapmaları göstermesi.
3. *t*-testi sonuçlarından hareket ederek etki büyüklüğünü hesaplayabilmesi.

4. *F*-testi sonuçlarından hareket ederek etki büyüklüğünü hesaplaması.
5. Olasılık düzeyleri ve korelasyon değerlerine bağlı olarak etki büyüklüğünü hesaplaması.
6. Tek yönlü kovaryans analizini yapması.
7. Yüzde değerlerine bağlı olarak etki büyüklüğünü hesaplaması.
8. Wilk lambda değerini hesaplaması.

Etki büyüklüğü hesaplamaları, meta analizleri için yararlı olmakla birlikte bunun yanında "güç analizleri" için de kullanılmaktadır. Raporlarında hipotez testi anlamlılık sonuçlarının yanında etki büyüklüğü değerlerini de vermek isteyen araştırmacılara bu yazılımın önemli ölçüde fayda sağlayacağı belirtilmiştir.

META

Meta, etki büyüklüğü ve meta analizlerini yapmak amacıyla geliştirilmiş olan bir programdır. Yazılım her bir çalışma için etki büyüklüğünü hesaplayarak etki büyüklüklerini birleştirmekte, homojenlik testini yapmakta ve ortalama etki büyüklüğünün sıfırdan farklı olup olmadığını belirlemektedir. Sonuçlar örneklem büyüklüğü, çalışmanın varyansı veya kullanıcının girdiği değerlere göre ağırlıklandırılabilmektedir.[60]

Kapsamlı Meta-Analizi

Kapsamlı meta analizi (comprehensive meta-analysis) adlı yazılımın bilimsel araştırmalarda ihtiyaç duyulan etki büyüklüğü ve meta analizi konularıyla ilgili değişik hesaplamaları yapabildiği ve değişik nitelikte grafikler ürettiği bildirilmiştir.

ÇOK BOYUTLU GİZLİ YAPILARI ORTAYA ÇIKARAN YAZILIMLAR

Bu yazılımlar maddelerin arka planındaki yapıların çok boyutlu olup olmadığını belirlemeye hizmet eder. Testi alan kişilerin yetenek düzeylerinin ortalama değerlerinden hareket ederek testin/ölçeğin boyutsallığını ortaya çıkarır.

CONCOV

Bu yazılım, ölçekteki her bir madde çifti için koşullu kovaryans eğrisini çizerek maddelerin arka planında yatan gizli boyutu ortaya çıkarmak için nonparametrik bir tahmin değeri verir. CONCOV maddelerin iki dereceli puanlama temeline dayalı olduğunu varsayar. Maddeler arka planda var olduğu düşünülen belirli sayıda psikolojik veya bilişsel yapılar altında toplanır.[61]

DETECT

Basit bir yapı altında, ikili puanlamaya dayanan test maddelerinin boyutsal tanımlamasını ayrıntılı bir şekilde veren bir yazılımdır. DETECT, kümeleme analizi prosedüründen hareket ederek test maddelerinin uygun bir şekilde gruplanmasını sağlar ve baskın boyutları ortaya çıkarır.[62]

DIMTEST

Dimtest, ölçüm aracında belirli boyutların bulunduğunu belirlemeye yönelik olarak geliştirilmiş bir tür hipotez test etme yazılımıdır. İki dereceli eğitsel ve psikolojik testlerde gizli tek boyutluluğun mevcudiyetini değerlendirir. Muhtemel boyutların birbirinden farklılığını ortaya çıkarmak için kullanıcının tanımladığı iki alt test arasında anlamlı bir farklılık olup olmadığını değerlendirir. DIMTEST ya teyit edici veya keşfedici modele göre çalışır. Teyit edici modelde kullanıcının öngördüğü boyutların ortaya çıkıp çıkmadığına bakılır. Keşfedici modelde ise, maddeler arasındaki geçerlilik çalışmaları yapılarak ortaya çıkan boyutun testin diğer maddelerinden farklı olduğu belirlenir. DIMTEST'in bütünüyle nonparametrik bir test olduğu bu nedenle analizlerde parametrik MYK modellemesine ve "madde yanıt fonksiyonu" tahminine ihtiyaç duyulmadığı bildirilmiştir.[63]

POLY-DIMTEST

POLY-DIMTEST, DIMTEST yazılımının değiştirilmiş başka bir sürümüdür. POLY-DIMTEST çok dereceli puanlarda gizli tek boyutluluğu belirlemeye yönelik olarak kullanılır. DIMTEST yazılımında olduğu gibi teyit edici veya keşfedici modele göre çalışır.

DIMTEST ve POLY-DIMTEST yazılımlarının 50 madde için 2000 katılımcıya kadar faktör analizi yaptığı, faktör analizinin kullanılmadığı durumda ise 100 maddeye kadar analiz yaptığı bildirilmiştir.

HCA/CCPROX

Hiyerarşik kümeleme analizini yapan bu yazılım, ikili ve çok dereceli maddelerle çalışabilmektedir. Nonprametrik çalışma prosedürüne sahip

olan yazılım, maddeleri aşamalı olarak homojen gruplar altında toplar. Kullanıcıya çeşitli toplam puanlar açısından maddelerin boyutsallığını değerlendirme imkanı sağlar. Daha çok basit yapıları ortaya çıkarmayı amaçlayan yazılım, veriler basit bir yapı ortaya koymuyorsa ayrı kümeler halinde diğer boyutları da ortaya çıkarır.[64]

YAZILIM SEÇİMİNDE DİKKAT EDİLMESİ GEREKEN ÖLÇÜTLER

Programın Analiz Modeline Uygunluğu

Araştırmacı verilerini analiz ederken klasik test kuramıyla modern test kuramlarından birini tercih edebilir veya verilerini her iki pencereden bakarak değerlendirmek isteyebilir. Araştırmacılar çoğunlukla bu modellerden birinin tercih ederek analizlerini yaparlar. Bilim adamı ölçüm sonuçlarını örneklem–evren bağlamından bağımsız olarak ele almak istiyorsa modern test kuramına göre hazırlanmış yazılımlara yönelecektir.

Yazılımın Hangi İşletim Sistemi Altında Çalıştığı

Yazılımlar işletim sistemlerine bağımlı olarak üretilirler. Unix, Windows NT Server, Windows, MS-DOS, MOS en yaygın kullanılan işletim sistemleridir. Yazılımların kullanılmasında sadece işletim sistemleri değil sürümleri de önemlidir. Bir programın Windows 3.11 sürümü altında çalışması son çıkan Windows sürümleri altında çalışmayabileceği anlamına gelir. Bilim adamı yazılım satın alırken şu soruları sormalıdır: Hangi işletim sistemi ve hangi sürüm altında çalışıyor? Yazılımın güncellenmesi sürekli olarak yapılıyor mu? Güncelleme için istenen ücretler makul mü? Güncelleme sürümleri İnternet ortamından sağlanabiliyor mu?

Bellek Kapasitesi

Yazımın çalışabilmesi için gereksinim duyduğu bellek kapasitesi vasat bir bilgisayar bellek kapasitesinin dörtte birinden daha fazla olmamalıdır. Aksi halde yazılım için bilgisayarda bellek artırımına gitmek gerekli olacaktır.

Mevcut Veri kapasitesi ve İhtiyaç Duyulan Kapasite

Yazılıma kaç adet vak'a yüklenebildiği araştırılması gereken bir diğer noktadır. Bazı yazılımlar veri yükleme kapasitesini sonsuz olarak belirlerken diğerleri 500, 1000 veya 32.000 gibi rakamlarla sınırlandırmışlardır. Araştırmacı, 1000 veya 2000 kişiden daha fazla cevaplayıcı üzerinde araştırma yapmayacaksa yüksek kapasiteli bir yazılımın peşinde koşmamalıdır.

Madde İşleme Kapasitesi

Yazılımın aynı zaman diliminde kaç maddeyi işleyebildiği ve kaç madde üzerinde analiz yapabildiğidir. Bazı istatistiksel analiz yazılımlarının madde işleme kapasitesi testten teste değişiklik gösterir. Örneğin SAS'ın alt modülü olan PROC FACTOR'ün madde sayısı 26'dan fazla olduğunda tetrakorik korelasyon analizini yapma konusunda güçlüklerle karşılaşıldığı bildirilmiştir.

Verileri Gruplandırma Özelliği

Analiz yazılımının kendi içinde verileri maniple ederek değişik kriterlere göre alt gruplar oluşturabilmesidir. Her tür analiz için gerekli olmamakla birlikte geçerlilik ve güvenilirliği klasik test kuramına göre analiz etmek isteyen kişiler bu özellikten büyük ölçüde yararlanacaklardır.

Veri Aktarması ve Veri Alması

Yazılımın dış ortamlardan ASCII formatında veya diğer formatlarda veri alabilmesi ve veri verebilmesidir. Dış ortamlar bir veri tabanı yazılımı, SPSS gibi bir istatistiksel analiz yazılımı, Excell gibi bir hesaplama yazılımı veya diğer sık kullanılan istatistik yazılımları olabilir. Bilim adamı değişik amaçlarla birden fazla yazılımla çalışıyor olacağından veri alma ve verme özelliği yoksa ham verileri her defasında bilgisayara yeniden girme durumunda kalabilir.

İstatistik Analizlerinin Zenginliği ve Amaca Uygunluğu

Yazılımlar amatör programcılar veya profesyonel kurumlar tarafından üretilir. Kendi heves ve merakını tatmin etmek için sadece belirli hesaplamaları yapmak üzere üretilmiş olan yazılımlar hem esnek ve hem de zengin hesaplama kapasitesine sahip değildir. Bu nedenle bilim adamı yazılım seçerken öncelikle amaca uygunluğu göz önünde bulundurmalı ve daha sonra hesaplama seçeneklerinin son kuramsal bilgi birikimine uygunluğu konusunda araştırma yapmalıdır. Artık terkedilmiş bir modele göre hesaplama yapan bir yazılımın teorik olarak ve pratikte bir yararı yoktur.

Güçlü Kurumsal Destek

Yazılımın bireysel uzmanların kişisel çabalarının ürünü olarak çıkmış olması güncelleme konusunda zayıf olacağının işaretidir. Yazılımı destekleyen, pazarlayan kuruluşların bir üniversite, araştırma kuruluşu, devlet kuruluşu olması arkasında güçlü bir destek olduğu anlamına gelir. Kurumsal destek daha çok araştırma kuruluşları, ticarî kuruluşlar ve üniversiteler tarafından sağlanır. Yazılımın arkasında kurumsal destek varsa bu durum

aynı zamanda programın kullanım el kitapları, bakım desteği, güncellenme, çevrim içi danışma olanağı veya tartışma gruplarına katılım fırsatı açısından araştırmacının çok yönlü kendisini geliştirme olanağına sahip olacağını gösterir.

Kullanıcı Dostu Olması

Yazımın etkileşimli pencerelerle çalışması, mönü destekli olması, ham verilerle çalışabilmesi o programın kullanıcı dostu olduğunu gösterir. Kullanıcı dostu olma ilkesi, aynı zamanda grafik özelliklerle, veri yükleme kolaylığı ile, örnek veri dosyaları, içerdiği yardım dosyaları ve analiz sonuçlarının nasıl yorumlanacağına ilişkin kılavuz bilgileriyle belli olur. Bu açıdan güçlü olan yazılımların İnternet ortamında deneme sürümleri bulunur. Bu sürümlerde, hayali veriler üzerinde hesaplamalar yapılarak gerekli bilgileri edinmek mümkündür.

Geliştirilme Tarihi, Sürümü ve Güncellenme Aralığı

Yazılımın geliştirilme tarihi önemli olan bir diğer konudur. Sürümü beş yıldan daha önceki bir tarihe sahip ve güncel uygulamaları zayıflamış yazılımlara kuşkuyla bakmak gerekir. Bilim ve yazılım dünyası hep birlikte hızlı bir değişim süreci içindedir. Bu nedenle satın alınması düşünülen yazılımın sürüm tarihi itibariyle ideal olarak üç yıldan daha fazla eski olmaması veya güncelliğini yitirmemiş olması tercih edilir. Geliştirilme tarihi kadar bir diğer önemli konu yazılımın güncelleştirilme aralığıdır. Arkasında güçlü kurumsal destek bulunan yazılımlar belirli periyotlarla güncellendiğinden güncellenme olgusunu kurumsal destekle birlikte değerlendirmek gerekir.

Ücretlendirilmesi

İstatistiksel analiz yazılımlarının ücretlendirilmesinde ticarî şirketler farklı fiyat politikası uygularlar. Özel kuruluşlara, akademik kuruluşlara, öğrencilere ve öğretim üyelerine yönelik fiyatlar farklı olabilir. Yazılımın ücretlendirilmesinde akademik nosyona sahip kişiler için ayrı bir fiyatın belirlenmiş olması yazılımın tercih nedeni olarak ortaya çıkabilir. Bazı yazılımlarda ticarî fiyatla akademik fiyat arasında %50'den fazla fark vardır. Değişik zamanlarda düzenlenen promosyon uygulamaları da yazılım ücretlerini önemli ölçüde düşürebilmektedir.

Grafik Özelliği

Yazılımın yüksek çözünürlüklü grafik özelliğine sahip olması, sonuçların aynı zamanda görsel açıdan temiz ve net bir şekilde alınmasını sağlar. Bir-

çok yazılımda grafik özelliği olmakla birlikte bu grafikler hem görsel kalite açısından hem de bilimsel açıklanabilirlik düzeyi açısından yetersizdir. Satın alınması düşünülen yazılımların grafik özellikleri deneme sürümleri üzerinde incelenebilir.

Kullanıcı Kılavuzu ve Çevrimiçi Destek Olanakları

Yazılım seçiminde dikkat edilmesi gereken bir diğer nokta programın kapsamlı bir kullanıcı kılavuzunun bulunup bulunmadığıdır. Kullanıcı kılavuzundaki örnekler araştırmacının işini kolaylaştırarak ona doğru hesaplamaları ve analizleri yapma imkanı sağlar. Kullanıcı kılavuzları ayrı bir kitapçık halinde veya İnternet ortamında çevrimiçi metinler halinde kullanıcıların yararlanımına sunulmuştur.

ALINTI YAPILAN KAYNAKLAR

[1] Assessment System Corporation, "ITEMAN: Classical Item Analysis [ITEMAN: Klasik Madde Analizi]," <http://www.assess.com/Software/iteman.htm> (18.05.2003).

[2] Assessment System Corporation, "ITEMAN: Classical."

[3] Assessment System Corporation, "Lertap 5 - Laboratory of Educational Research Test Analysis Package [Lertap 5: Eğitim Araştırmaları laboratuarı Test Analiz Paketi]," <http://www.assess.com/Software/Lertap.htm> (19.05.2003).

[4] Aynı.

[5] e-Academy, "TESTFACT version 4.0 for Windows [Testfact Windows Sürümü]," <http://www.e-academy.com/index.cfm?loc=estore/soft_browse/soft_display_product&ID_Product=236> (19.05.2003).

[6] James A Wollack, "Comparison of Answer Copying Indices with Real Data [Gerçek Veriler Üzerinde Kopya İndis Yanıtlarının Karşılaştırılması], <http://edtech.connect.msu.edu/Searchaera2002/viewproposaltext.asp?propID=598> (24.01.2004).

[7] J.A. Wollack, "Comparison of Answer Copying Indices with Real Data [Gerçek Veriler Üzerinde Kopya Cevaplarının Karşılaştırılması]," <http://tigersystem.net/aera2002/viewproposaltext.asp?propID=598> (21.05.2003).

[8] Tulane University, "LISREL and PRELIS in Unix [Unix İşletim Sistemi Altında LISREL ve PRELIS]," <http://tis.tulane.edu/How_To/Unix_System/LISREL_and_PRELIS.cfm> (06.07.2003).

[9] J. S. Uebersax, "The Tetrachoric and Polychoric Correlation Coefficients [Tetrakorik ve Polikorik Korelasyon Katsayıları]," <http://ourworld.compuserve.com/homepages/jsuebersax/tetra.htm#soft> (26.05.2003).

[10] Compuserve, "User Guide for POLYCORR 1.1 [POLYCORR 1.1 Kullanıcı Rehberi]," <http://ourworld.compuserve.com/homepages/jsuebersax/xpc.htm> (26.05.2003).

[11] P. Barrett, "Iddanet Program Library [Idanet Yazılım Kütüphanesi]," <http://www.liv.ac.uk/~pbarrett/programs.htm#SHORTFORM> (26.05.2003).

[12] Barrett, "Iddanet Program."

[13] Advance Research and Data Analyses Center, "Interrater/Test Reliability System (ITRS) [Gözlemciler Arası Değerlendirme ve Test Güvenilirlik Sistemi]," <http://www.uni-koeln.de/ themen/Statistik/software/itrs.txt> (28.05.2003).

[14] Academic Computing and Instructional Technology Services, "Introduction to Structurel Equation Modelling with AMOS [AMOS iye Yapısal Eşitlik Modeline Giriş]," [<http://www.google.com.tr/search?q=cache:snlOtMOkdFkJ:www.utexas.edu/cc/stat/tutorials/amos/ut-amos2.pdf+amos+reliability+path&hl=tr&ie=UTF-8> (17.05.2003).

[15] S.L. Sclove, "Notes on Path Analysis and Structural Equation Modeling (SEM) [Rota Analizi ve Yapısal Eşitlik Modeline İlişkin Notlar]," <http://www.uic.edu/classes/mba/mba503/981/503paths.htm> (17.05.2003).

[16] Sclove, "Notes on Path."

[17] Smallwaters, "Most Frequently Asked Questions about Amos [Amos Hakkında Sık Sorulan Sorular]," <http://www.google.com.tr/search?q=cache:UJzGEYnDrrMJ:www.smallwaters.com/amos/faq/faqa-mfaq.html+amos+ordinal&hl=tr&ie=UTF-8&inlang=tr> (06.07.2003).

[18] D. Suhr, "

[19] J. West, "Structural Equation Software [Yapısal Eşitlik Yazılımları]," <http://www.gsm.uci.edu/~joelwest/SEM/Software.html> (18.05.2003).

[20] Smallwaters, "Most Frequently ."

[21] Aynı.

[22] M.C. Neale, "About MX [Mx Hakkında]," <http://www.vcu.edu/mx/about-mx.html> (18.05.2003).

[23] Virginia Institute for Psychiatric and Behavioral Genetics, "Mx: Statistical Modeling [Mx: İstatistiksel Modelleme]," <http://www.google.com.tr/search?q=cache:PIHKvDMF4YsJ:www.vipbg.vcu.edu/~vipbg/Mx.html+mx+structural+equation+Matrix+algebra+&hl=tr&ie=UTF-8&inlang=tr>

[24] P. Barrett, "A Review of the SYSTAT v.10.2 Software Package [SYSTAT v.10.2 Yazılımının Gözden Geçirilmesi]," 2003, <http://www.google.com.tr/search?q=cache:KRr0-Qc4OwwJ:www.liv.ac.uk/~pbarrett/systat.pdf+structural+equation+ramona+systat&hl=tr&ie=UTF-8&inlang=tr> (07.07.2003).

[25] Aynı.

[26] P. Spirtes ve diğerleri, "Tetrad 3: Tools for Causal Modeling [Tetrad 3: Nedensel Modelleme Aracı]," <http://www.phil.cmu.edu/tetrad/tet3/master.htm.> (25.05.2003).

[27] P. Spirtes ve diğerleri, "Introduction [Giriş]," <http://www.phil.cmu.edu/tetrad/tet3/chp1.htm> (25.05.2003).

[28] Aynı.

[29] Wynne W. Chin, "Overview of the PLS Method [KEKK Yönteminin Genel Olarak Gözden Geçirilmesi]," 01 Mayıs 1998, <http://disc-nt.cba.uh.edu/chin/PLSINTRO.HTM> (08.07.2003).

[30] M.K. Teer, "Partial Least Squares [Kısmî En Küçük Kareler Yöntemi]," <http://www.gsu.edu/~mkteer/relmeth.html> (08.07.2003).

[31] Aynı.

[32] J. Honaker, A. Joseph, G. King ve K. Scheve, "Amelia: A Program for Missing Data [Amelia: Eksik Veri Analizi Programı]," <http://gking.harvard.edu/amelia/ whatsnew> (25.05.2003).

[33] SPSS Inc., "Create Higher-Value Data And Build Better Models When You Estimate Missing Data [Yüksek Değerli Veri Yaratma ve Eksik Verilerle Çalışırken Daha İyi Model Oluşturma]," <http://www.spss.com/spssbi/missing_value/> (25.05.2003).

[34] C. Fraser ve R.P. McDonald, "Cosan User Guide [Cosan: Kullanıcı Rehberi]," <http://www.unt.edu/rss/class/rich/5840/mcdonald/cosan/COSAN.htm.> (31.05.2003).

[35] Assesment Systems Corporation, "Test Information Program [Test Bilgi Yazılımı]," <http://www.assess.com/Software/TESTINFO.htm.> (31.05.2003).

[36] Winsteps, "Reliability and Chi-square Statistics [Güvenilirlik ve Ki-kare İstatistiği]," <http://www.winsteps.com/facetman/table7summarystatistics.htm.> (06.06.2003).

[37] Winsteps, "Facets: DOS Version," <http://www.winsteps.com/facdos.htm> (06.06.2003).

[38] Winsteps, "Facets," <http://209.130.54.113/facets.htm> (06.06.2003).

[39] Simstat, "Classical Item Analysis Module for Simstat [Simstat için Klasik Madde Analizi Modüllü]," <http://www.simstat.com/statitem.htm> (04.06.2003).

[40] Kovach Computing Services "Simstat" <http://www.kovcomp.co.uk/simstat/index.html> (04.06.2003).

[41] B.A. Hanson"Estimation Toolkit for Item Response Models 8Madde Yanıt Modelleri İçin Tahmin Araç Kiti]," <http://www.b-a-h.com/software/cpp/etirm.html.> (31.05.2003).

[42] Polyglot, "Notes [Notlar]," <http://polyglot.cal.msu.edu/llt/vol2num2/article4/notes.htm.> (06.06.2003).

[43] B.D, Wright ve A. Tennant, "Sample Size Again [Tekrar Örneklem Büyüklüğü]," <http://www.rasch.org/rmt/rmt94h.htm.> (06.06.2003).

[44] Assesment System Corporation "QUEST: Interactive Analysis System [QUEST: Etkileşimli Analiz Sistemi], <http://www.assess.com/Software/quest.htm.> (31.05.2003).

[45] Assesment System Corporation, "Rascal-Rasch Analysis Program [Rascal-Rasch Analiz Yazılımı]," <http://www.assess.com/Software/rascal.htm> (24.05.2003).

[46] Rasch.org, "RSP: A Program for Rasch Scaling [RSP: Rasp Ölçekleme Yazılımı]," <http://www.rasch.org/rmt/rmt81s.htm.> (04.06.2003).

[47] Assesment System Corporation "Rasch Unidimensional Models for Measurement for Unfolding Response Models [Aşamalı Oarak Açığa Çıkarma Modellerinde Rasch Tek Boyutlu Modeli]," <http://www.assess.com/Software/rummfold.htm.> (04.06.2003).

[48] Rasch org, "Rasch Software [Rasch Yazılımları]," <http://www.rasch.org/rmt/rmt114d.htm> (25.07.2003).

[49] mvondavier@hotmail.com, "Winmira 32 Pro." <http://planet.ipn.uni-kiel.de/_planet_alt/winmira/32pro/> (04.06.2003).

[50] C. Ho Yu, "Developing Data Systems to Support the Analysis and Development of Large-Scale, On-line Assessment [Çevrimiçi Değerlendirmeye İmkan Verecek Büyük Ölçekli Veri Sistemlerinin Geliştirilmesi]," <http://seamonkey.ed.asu.edu/~alex/pub/aera2001.pdf.> (06.06.2003).

[51] Polyglot, "Notes [Notlar]," <http://polyglot.cal.msu.edu/llt/vol2num2/article4/notes.htm.> (06.06.2003).

[52] Scientific Software International, "SSI Products [SSCI Ürünleri]," <http://www.ssicentral.com/product.htm#la3> (28.05.2003).

[53] Assesment Systems Corporation, "Loglinear and Loglinear IRT Model Analysis [Loglineer ve Loglineer MYK Modelleri Analizi]," <http://www.assess.com/Software/logimo.htm> (03.06.2003).

[54] K. Sijtsma, "Item Analysis and Test/Questionnaire Construction Using Nonparametric Item Response Theory [Nonparametrik Madde-Yanıt Kuramı Çerçevesinde Test/Anket Oluşturma ve Madde Analizi]," <http://kubnw5.kub.nl/web/fsw/Mto/SMABS/BACKUP/W1.PDF> (04.06.2003).

[55] Assesment Systems Corporation, "Nonparametric IRT Scaling [Nonparametrik MYK Ölçeklemesi]," < <http://www.assess.com/Software/msp.htm.> (04.06.2003).

[56] Rasch.org, "Some Rasch-capable Computer Software [Rasch Kabiliyetine Sahip Bazı Bilgisayar Yazılımları]," <http://www.rasch.org/rmt/rmt133l.htm.> (04.06.2003).

[57] Assesment Systems Corporation, "TESTFACT 4 - Classical Item and Item Factor Analysis TESTFACT 4 – Klasik Madde Analizi ve Faktör Analizi]," <http://www.assess.com/Software/testfact.htm> (26.07.2003).

[58] Assesment Systems Corporation, "VARCL - Variance Component Analysis by Maximum Likelihood [VARCL – Maksimum Olasılık Yöntemlerine Göre Varyans Bileşenleri Analizi]," <http://www.assess.com/Software/SCA.htm> (26.07.2003).

[59] Assesment Systems Corporation, "SCA - Simultaneous Principal Components Analysis [SCA – Eş Zamanlı Temel Bileşenler Analizi]," <http://www.assess.com/Software/SCA.htm> (27.07.2003).

[60] "META Program Information [META Yazılım Bilgisi]," <http://users.rcn.com/dakenny/metain.htm> (26.07.2003).

[61] Assesment Systems Corporation, "CONCOV - Covariance Curve Estimation [CONCOV – Kovaryans Eğrisi Tahmini]," <http://www.assess.com/Software/CONCOV.htm> (26.07.2003).

[62] Assesment Systems Corporation, "DETECT - Dimensionality Analysis [Boyutsallık Analizi]," <http://www.assess.com/Software/DETECT.htm> (26.07.2003).

[63] Assesment Systems Corporation, "DIMTEST and POLY-DIMTEST - Latent Unidimensionality Assessment [Gizli Çok Boyutluluk Değerlendirmesi]," <http://www.assess.com/Software/DIMTEST.htm#POLY> (26.07.2003).

[64] Assesment Systems Corporation, "HCA/CCPROX - Hierarchical Cluster Analysis [Hiyerarşik Kümeleme Analizi]," <http://www.assess.com/Software/HCA.htm> (26.07.2003).

GEÇERLİLİK

Geçerlilik, kullanılan ölçüm aracının ölçülmek istenen özelliğe uygun olması, verilerin ölçülmek istenen özelliğin niteliğini tam olarak yansıtması ve aynı zamanda verilerin amaca yönelik olarak yararlı olmasıdır. Bu nedenle kısaca, "test puanlarının sonuç çıkarmak için uygun, anlamlı ve yararlı olması"[1] biçiminde tanımlanmıştır. Verilerin ölçüm amacı hakkında doğru bilgi verme derecesi *yararlılığı* ve *kullanışlılığı* ifade eder. Doğru bilgi verme derecesi düşükse söz konusu bilgiler kullanılamaz. Herhangi bir araştırmada, ölçüm verilerine bakarak genelleme yapmadan önce bilim adamı araştırma süreci ve toplanan verilerin geçerliliği hakkında bilgi vermelidir. Araştırmada uygulanan istatistikî analizlerin ve elde edilen bulguların değeri, "geçerliliğe" bağlıdır. Sürecin ve toplanan verilerin geçerlilik sorunu varsa, yapılan istatistikî analizler ne kadar iyi yapılmış olursa olsun araştırmanın veya ölçümün bilimsel değeri sınırlıdır. Geçerlilik analizi yöntemleri, ölçüm aracının türüne göre değişir. Öğrenmeyi ölçen bilgi testlerinde, bilişsel testlerde (genel yetenek, sözel-sayısal yetenek gibi testlerde) ve duygusal testlerde (kişilik envanterlerinde, tutum ölçeklerinde) geçerlilik analizleri farklı biçimlerde uygulanır. Geçerlilik analizlerinin psikolojide, tıp bilimlerinde, antropolojide, işletme biliminde, eğitim bilimlerinde ve sosyolojideki uygulamalarında da bazı farklılıklar vardır.

Ölçüm verilerinin doğrulamasını yapmak için geçerlilik analizinden önce güvenilirlik analizleri yapılır.[2] Fakat, güvenilirlik analizleri tek başına yeterli değildir, aynı zamanda geçerlilik analizinin de yapılması gerekir.[a]

---

[a] Önce güvenilirlik analizinin mi yoksa geçerlilik analizinin yapılacağı literatürde tartışmalı bir konudur. Bazı yazarlara göre önce geçerlilik analizleri yapılır, güvenilirlik analizi ikinci sırada gelir. Çünkü güvenilirlik analizi, ölçeğin kaç boyutlu olduğu konusu ile madde-yapı ilişkisi hakkında bilgi vermez. Geçerlilik analizi daha geniştir ve parça parça kanıtların toplanmasıyla oluşur. Ancak, uygulamaya baktığımızda ölçüm sürecinde geçerlilik-güvenilirlik analizlerinin ardışık değil, karmaşık bir süreç içinde gerçekleştirildiği görülür. Yüzey ve içerik geçerliliği ile başlayan süreç, hem güvenilirlik, hem de geçerlilik için kullanılabilecek faktör analizi yöntemiyle devam eder. Yapısal geçerlilik analizlerinin önemli bir bölümü güvenilirlik analizlerinden sonra gerçekleştirilir.

Yüksek geçerlilik aynı zamanda yüksek güvenilirlik anlamına gelebilir, ama tersi doğru değildir. Diğer bir deyişle yüksek güvenilirlik geçerlilik hakkında hiçbir bilgi vermez. Yeni geliştirilen bir test veya ölçek, güvenilirlik ve geçerlilik analizi yapılmadan ve bu analiz sonuçları hakkında bilgi verilmeden yayımlanamaz. Eğer "kullanılan ölçüm verileri *geçerli* değilse, yapılan yorumlar bütünüyle anlamsızdır."[3] Araştırmacı açısından güvenilirlik analizlerini yapmak nispeten daha kolay iken geçerlilik analizleri için ayrıntılı ve kapsamlı bir çalışma yapmak gerekir. Güvenilirlik analizlerinde teknik hesaplamalar ön plana çıkarken, geçerlilik analizinde yargısal değerlendirmeler ile teknik hesaplamaların birlikte kullanılması gerekir. Bilim adamı, araştırma sürecinin daha başlangıç aşamasında hangi tür geçerlilik analizlerini yapacağını belirlemelidir. Bunun için çalışmanın deneysel, temel veya uygulamalı araştırma olmasına göre uygun bir *geçerlilik analizi planı* hazırlanır. Bu planda, araştırma tasarımı göz önünde bulundurularak yapılması uygun görülen geçerlilik analizi türleri ve bu analizlerin nasıl yapılacağı saptanır. Örneğin, yüzey geçerliliğinin nasıl belirleneceği, içerik geçerliliğinde hangi yöntemin uygulanacağı, kriter geçerliliği yöntemi kullanılacaksa kriter puanların ne olacağı, hangi benzer ölçeklerin kullanılacağı, yapısal geçerlilik analizi için hangi istatistiksel yöntemlerin uygulanacağı belirlenir. Geçerlilik analizleri, dikkatli bir şekilde belirlenmiş bir plan çerçevesinde gerçekleştirilir.

TANIMI

Geçerlilik, bilim disiplinlerinin yaklaşım biçimlerine göre literatürde farklı biçimlerde tanımlanmıştır. Geçerlilik kavramının tek bir tanımı veya değişik disiplinlerden bilim adamlarının üzerinde anlaştıkları ortak bir tanımı yoktur.[4] Geçerlilikle ilgili tanımlarda genel olarak üç tema vurgulanır. Birincisi, kullanılan ölçüm aracının ölçülmek istenen özelliğe uygun olmasıdır. İkinci tema, ölçümün kurallara uygun olarak doğru yapılıp yapılmadığıdır. Üçüncü tema ise, ölçüm verilerinin gerçekten ölçülmek istenen özelliği yansıtıp yansıtmadığıdır. Geçerliliğin ilk tanımı 1937 yılında Garrett tarafından yapılmıştır. Ona göre geçerlilik "ölçülmek istenen özelliğin amaca uygun olarak ölçülme derecesidir."[5] Son yıllarda en fazla referans gösterilen tanım ise, Hammersley (1987) tarafından yapılmıştır. Ona göre geçerlilik, "belli bir olguya ait ölçüm rakamları olguyu doğru bir şekilde yansıtıyor, doğru bir şekilde tanımlıyor veya doğru bir şekilde kuramsal açıklamalar getiriyorsa geçerlidir" (aktaran Winter).[6] Bir başka tanımda geçerlilik, bir ölçeğe veya teste ait gözlemlenen rakamlar arasındaki farklı-

lıkların gerçek hayattaki farklılıkları tam olarak yansıtması olarak belirlenmiştir. Buna göre geçerlilik, "realiteye" yaklaşma derecesidir. Geçerliliğin yaygın kabul görmüş bir başka tanımı Amerikan Psikoloji Derneği (APA) tarafından yapılmıştır. Bu derneğin yayımladığı, *Standards for Educational and Psychological Testing* isimli eserinin 2000 yılı baskısında geçerlilik kavramı aşağıdaki gibi tanımlanmıştır.

> "Geçerlilik, kuramsal bilgilerin ve gözlenebilir kanıtların geliştirilen test veya ölçek puanlarını doğrulamasıdır. Geçerlilik süreci, yapılan test/ölçek yorumlarının geçerli ve güçlü bir bilimsel temele sahip olduğu konusunda kanıtlar toplamaktır. Yorumlar testin/ölçeğin kendisine göre değil, elde edilen puanlara göre yapılır. Test puanları birden fazla şekilde yorumlanıyorsa amaçlanan her bir yorum için geçerlilik analizi ayrıca yapılmalıdır. Geçerlilik kanıtları, testin/ölçeğin içeriğine, cevaplama sürecine, iç yapısına, diğer değişkenlerle olan ilişkisine ve test sonuçlarına dayalı olarak yapılabilir, ancak sadece sayılan bu yöntemlerle de sınırlı değildir."[7]

Bu tanımda; hayatta yaşanan olayların, davranışların veya gözlemlerin test puanlarını doğrulması olgusuna önem verilmiştir. Bu açıdan araştırmacıya düşen görev, değişik yöntemleri uygulayarak sonuçların doğru olduğuna ilişkin *kanıt* toplamaktır. Kanıtların fazlalığı, az kanıt bulunmasına göre daha iyidir. Fakat bazen öyle durumlar olur ki, tek bir kanıt zayıf bir çok kanıttan daha ikna edici sonuçlar ortaya koyar. Bu nedenle kanıtların sayısı kadar, kalitesi de önemlidir. Araştırmacı birden fazla ve aynı zamanda güçlü kanıtlar bulmaya çalışmalıdır.[8]

Geçerlilik kavramıyla ilgili olarak 1990'lı yıllardan itibaren daha geniş içerikli tanımlamalar yapılmaya başlanmış ve literatürde "klasik geçerlilik tanımları" ve modern geçerlilik tanımları" sınıflandırması yaygınlık kazanmaya başlamıştır.

Klasik Geçerlilik Tanımları

Klasik geçerlilik tanımlarında geçerlilik kanıtlarının başlıca üç yöntemle toplanması üzerinde odaklanılır: içerik geçerliliği, kriter geçerliliği ve yapısal geçerlilik.[b] Buna göre geçerlilik, test maddelerinin alan örneklemesi-

---

[b] Amerikan Psikoloji Derneği'nin geçerlilik için uyguladığı bu üçlü yaklaşım *üçleme doktrini* (trinitarian doctrine) olarak isimlendirilmiştir. Bu doktrin psikoloji ve istatistik ki-

ne uygun olarak temsil edicilik özelliği çerçevesinde belirlenmesi, bir dış kriterle doğrulanması ve hipotetik yapıyı tam olarak temsil etmesidir. Günümüze kadar geliştirilen tüm test ve ölçeklerde en fazla klasik geçerlilik tanımları etkili olmuştur.

Modern Geçerlilik Tanımları

Son yıllarda Messick (1996) tarafından yapılan araştırmalarla geçerlilik kavramına farklı bir yaklaşım getirilmiş ve bu yaklaşım çerçevesinde yapılan açıklamalar "modern geçerlilik tanımı" olarak isimlendirilmiştir. Messick geçerliliği değişik özellikleri kapsayan *üniter bir kavram* olarak incelemiştir. Üniter olması, sadece test sonuçlarına bakmayı değil, bunun yanında testin nasıl kullanıldığına da önem verilmesini gerektirir. Bu yaklaşımda üniter bir yapıya sahip olan geçerlilik kavramının altı farklı yönünün olduğu vurgulanmıştır. Söz konusu altı yön (veçhe) birbirinden bağımsız değil, birbiriyle ilişkili ve bütünleşiktir.

İçerik. Geçerliliğin, *içerik* yönünü tanımlar. Ölçüm aracının içeriği ölçülmek istenen yapıyla ilgili olmalı ve o yapıyı temsil etmelidir. Son yıllarda başarıyla ilgili kavramsal yapıların giderek daha da karmaşıklaşmasıyla içerik zenginleşmiş, karmaşıklaşmış ve ilgililik çok daha fazla ön plana çıkmaya başlamıştır.

Gerçeklik. İkincisi, geçerliliğin *gerçeklik*[c] yönüdür. Gerçeklik, veçhesinde test sonuçlarının görevlerin değerlendirilmesinde somut bir temele veya ilişkililiğe sahip olup olmadığı konusu ele alınır. Gerçeklik, kuram temel alınarak veya proses modellemesi yöntemiyle belirlenir.[9]

Yapı. Üçüncüsü, geçerliliğin *yapısal yönüdür*. Bir değerlendirme veya ölçme aracının yapısı, kavramsal alanın o güne kadar bilinen iç yapılarıyla uyumlu olmalıdır.

Genellenebilirlik. Dördüncüsü, *genellenebilirliktir*. Değerlendirme veya ölçüm aracı incelenen kavramsal yapıya ait alanı ve süreçleri tam olarak

---

taplarında bir standart olarak ele alınmış ve Anastasi, *Psychological Testing* isimli kitabında 1961 yılından itibaren sürekli olarak hep aynı planı uygulamıştır.

[c] Substantive (gerçeklik). Hayalî değil, fiilî veya gerçek. Ölçümün özünü oluşturan esasa ait. Verileri deney ve gözleme dayalı olarak etüt etme yöntemini öneren akım *bk.*,<www.dictionary.com>; Oxford Talking Dictionary.

temsil etmeli ve o yapıya genellenebilmelidir.

Dış faktörler. Beşincisi, *dış faktörlerdir*. Dış faktörler, ölçüm aracının puanlarıyla diğer ölçümler veya değerlendirilmeyen davranışlar arasındaki ilişkilerin ne ölçüde ayırt edici olduğudur.

Müteakip sonuçlar. Altıncısı, geçerliliğin *müteakip sonuçlarıdır*. "Müteakip sonuçlar" incelemesinde, elde edilen kanıtların gerçek hayatta ne derece doğru çıktığı ve ne derece istenen sonuçları verdiği konusu üzerinde durulur. Test ve ölçüm işlemi neticesinde arzulanan sonuçların yüksek, istenmeyen yan etkilerin ise minimum olması hedeflenir. Messick tarafından belirlenen bu altı geçerlilik yaklaşımı, psikolojik nitelikteki ve eğitimle ilgili tüm ölçüm olaylarına uygulanır.[10]

TARİHSEL GELİŞİMİ

Geçerlilik kavramının tarihsel gelişimini 1930'lu yıllardan itibaren izleyebiliriz. Bu yıllarda geçerlilik, basit bir istatistiksel korelasyon analizi olarak görülmüştür. Ölçüm verilerinin objektif bir kriterle yüksek derecede ilişkili olması, ölçümün geçerli olduğu anlamına gelmekteydi. 1950'li yıllara gelindiğinde geçerlilik değerlendirmesinde "çoklu korelasyon analizi" ve "geçerlilik türleri" konusu gündeme gelmiştir. Ellili yılların ilk dönemlerinde faktöriyel, içsel ve mantıksal geçerlilik yaklaşımlarından söz edilir. Bu yıllarda geçerlilik analizlerine katkı sağlayan bilim adamları Gullickson (1950), Guilford (1946), Jenkins (1946) ve Rulon'dur (1946).

Geçerlilik konusunda ilk önemli çalışmaları yapan L. Cronbach (1949) olmuş ve Cronbach geçerliliği *deneysel* (ampirik) ve *mantıksal* olmak üzere iki grupta değerlendirmiştir. Mantıksal geçerlilikte gevşek bir şekilde organize olmuş bilgiler topluluğu vardır. Mantıksal geçerlilikte içerik analizi, test alma süreci, uygulama sorunları gibi çok sayıda konu üzerinde durulur. Mantıksal geçerlilik yaklaşımı, bu gün *içerik geçerliliği* olarak isimlendirdiğimiz alanın gelişmesine katkıda bulunmuştur.[11] Deneysel geçerlilik ise, faktör analizi yöntemine dayanır. Bu yıllarda ölçme konusunda araştırmalar yapan Anastasi (1950) olguya farklı bir bakış açısı getirmiş, geçerlilik için *test puanlarıyla kriter puanları* arasındaki korelasyon katsayılarına önem vermiştir.

Amerikan Psikoloji Derneği (American Psychological Association – APA) 1954 yılında dört tür geçerlilik standardı belirlemiştir. Bunlar; (a) içerik geçerliliği, (b) tahmin geçerliliği, (c) birlikte vuku bulma –eş zaman-

lılık– geçerliliği ve (d) yapısal geçerliliktir. Ancak APA 1966 yılında tahmin geçerliliği ve birlikte vuku bulma geçerliliğini *kriter geçerliliği* adı ile tek bir başlık altında toplamış ve geçerlilik analizlerinin sayısını üçe indirmiştir.[d] Amerikan Psikoloji Derneği tarafından geliştirilen bu yaklaşım aynı zamanda Amerikan Eğitim Araştırmaları Derneği (American Educational Research Association – AERA) tarafından da kabul görmüştür. Daha sonra 1996 yılında, Thompson ve Daniel, test ve ölçeklerin geçerliliğini test etmeye yönelik olarak *birleşme* ve *ayrılma* analizlerinin yapılmasını önermişlerdir.

"Ölçme ve değerlendirme konusunda saygı duyulan bir bilim adamı olan Cronbach 1971 yılında geçerliliğin anlamının daraltılmasının, geçerliliğin sadece belirli bir tahminin doğrulanması için test ve kriter puanlarından hareket ederek sonuç çıkarılması olgusu olarak görülmesinin ileriki yıllarda bir takım tartışmalara neden olacağını belirtmiştir."[12] Bu nedenle olsa gerek, bazı yazarlar geçerliliğe daha geniş bir anlam verme eğilimi içinde olmuşlardır. Bu yazarlar, geçerliliği yapılan *tahminin gerçek çıkması* olgusunun ötesine taşımışlar ve geçerliliği daha geniş bir çerçeve içinde tanımlamışlardır. Bu kişilerden biri olan Messick'e göre geçerlilik (1989), değişik alanlardan elde edilen ve kanıtlara dayalı olarak geliştirilen yargıların bir araya getirilmesi ve birleştirilmesiyle elde edilir. Geçerlilikte tecrübî kanıtlar ve teorik mantık birlikte kullanılır. Bilim adamı, ampirik araştırma sonuçlarına ve mantığına dayanarak ölçüm puanlarının makul olduğu, bu puanlardan sonuç çıkarılabileceği ve bu puanlara göre hareket edebileceği sonucuna ulaşır.[13]

Amerikan Psikoloji Derneği ile Amerikan Eğitim Araştırmaları Derneği'nin 1974'te birlikte geliştirdikleri test standartlarında geçerlilikle birlikte test veya ölçeğin toplumsal sonuçları üzerinde de durulmuş, bu tür testlerin yaratabileceği olumsuz yan etkilere dikkat çekilmiştir. Bu açıdan test uygulayıcısının, kullandığı testin amacı ve doğurabileceği muhtemel sonuçlar hakkında bilgili olması gerektiği belirtilmiştir. Yapısal geçerliliğe sahip olmayan çeşitli testlerle yapılan ölçüm ve değerlendirmelerin mahkemelerde dava konusu olabildiği görülmüştür.[14]

Son yıllarda geçerlilik kavramı bütünleşik bir kavram olarak ele alınmaya başlanmıştır. Geçerlilik; test/ölçek puanlarının uygunluğu, anlamlılığı ve kullanışlılığı anlamına gelir. Günümüzde yapılan çok sayıda araştırma ile her gün yeni katkılar ortaya çıkmakta ve bilimsel bilgi birikimi genişlemektedir. Fakat tüm bilgiler, yaklaşımlar farklı geçerlilik türlerini değil birbiriyle ilgili tek bir olguyu açıklar.

Test ve ölçüm konusunda önde gelen bilim adamlarından biri olan Sam Messick 1990'lu yıllarda geçerlilik analizlerinin *içerik*, *kriter* ve *yapısal geçerlilik* şeklinde sınıflandırılmasına karşı çıkmıştır. Ona göre bu sınıflandırma olguyu parçalamakta ve geçerliliği açıklamakta yetersiz kalmaktadır. Messick kendi geçerlilik anlayışını *yön (veçhe) yaklaşımı* olarak isimlendirmiştir. Bu yaklaşımda test sonuçlarıyla toplumsal hareket tarzları arasındaki ilişkiye bakılmaktadır. Ona göre geçerlilik, basit korelasyon katsayılarının ötesinde bilimsel ve politik rolü de olan bir tür toplumsal sıçrama değeridir.[15] Messick geçerlilikte yön yaklaşımını başlıca iki boyut üzerinde temellendirmiştir: Amaç, sonuçlar (*bk.*, Tablo 15-1.)

Tablo 15-1. Sam Messick'in Geçerlilik Analizinde Yön Yaklaşımı

Doğruluğu Kanıtlar / Sonuçlar	Amaç: Yorum getirme	Amaç: Kullanma
	Yapısal geçerlilik	Tahmin geçerliliği
	Değerlerle ilgili sonuçlar	Toplumsal sonuçlar

Messick'in yaklaşımı, *sonuçlar* faktörüne gereğinden fazla önem vermesi nedeniyle eleştirilmiştir. Bazı bilim adamları *sonuçlar* yerine *yapısal geçerliliğe* daha fazla önem verilmesi gerektiğini belirtmişlerdir.[16]

KAPSAMI

Geçerlilik bir araştırmanın tamamı için mi yoksa sadece belirli bir aşaması için mi düşünülmelidir? Geçerlilik; saran ve kuşatan bütüncül bir kavramdır. Ancak bu kavram, bir araştırmayı bütün olarak temize çıkarma anlamında ele alınmamalıdır. Geçerlilik kavramını evrensel olarak açıklamaya yetecek tek bir tanım veya form yoktur. Herhangi bir araştırma projesinde geçerlilik, araştırma sürecinin belirli bir kategorisine, evresine veya evrelerine aittir. Geçerlilik, tüm bir araştırma sürecine uygulanacak bir tür *onaylama testi* değildir. Geçerlilik ölçümleri ihtiyaca göre farklı şekillerde ya-

pılabilir. Tüm bir araştırma uygulamasının geçerliliğinden çok, veri toplama biçiminin, ölçüm sonuçlarının, gözlem sonuçlarının, deney sonuçlarının, anket sonuçlarının, karşılaştırma sonuçlarının geçerliliği söz konusudur. Bu yaklaşıma göre geçerlilik, bir araştırma uygulamasının belirli bir safhasına veya safhalarına aittir. Öte yandan geçerlilik bir sonuç değil, bir süreçtir. Araştırmacı bu süreçte verilerini sürekli gözden geçirmeli ölçeğinde/testinde gerekli ince ayarları, değişiklikleri ve düzenlemeleri yaparak geçerliliği artıracak önlemleri almalıdır.

Maxwell (1992) bir araştırma sürecinin değişik aşamalarında uygulanabilecek beş farklı geçerlilik türünden söz etmiştir. Bunlar; tasarım geçerliliği, yorumlama geçerliliği, kuramsal geçerlilik, genellenebilirlik ve değerlendirme geçerliliğidir (aktaran Winter).[17]

Tasarım geçerliliği. araştırmanın başlangıç aşamasındaki veri toplama süreciyle ilgilidir. Bu aşamada örneklemin yetersiz olması, konuyla doğrudan ilgili olmayan verilerin toplanması veya yanlış kişilere başvurulması geçerliliği büyük ölçüde tehdit eder. Araştırmacının dürüst davranmayarak bu konuda doğru bilgiler vermemesi de onun saygınlığına gölge düşürür.

Yorumlama geçerliliği. Özellikle niteliksel araştırmalarda önem kazanan bir konudur. Açık uçlu sorulara gelen yanıtlarla ilgili yorum yapmak kaçınılmaz bir durumdur. Niceliksel araştırmalarda bu tür yorumlar oldukça sınırlı olsa da niteliksel araştırmalarda bu tür yorumlara sık baş vurulur. Araştırmacı ne kadar *makul* bir yorum yapmış olursa olsun bu yorumların geçersiz olma tehlikesi vardır. Geçerli bir yorum, ilgili tüm aktörlerin görüşlerine saygı duymayı gerektirir.

Kuramsal geçerlilik. Geçerlilik, kullanılan terimlerin ve kavramsal yapıların anlamı üzerinde bilimsel dünyada konsensüs olmasına bağlıdır. Kuramsal geçerlilik soyut bir yaklaşımdır. Araştırmacının çalışmanın başlangıç aşamasında literatürü tarayarak belirlediği kuramsal çatılar veya mülahazalar veri toplama amacıyla uyumlu olmalıdır. Kişinin belirli bir olguyu kendi düşüncelerine göre sınıflandırması, kavramlamlaştırması onun sübjektif teoriysen olmasına yol açar.

Genellenebilirlik. Araştırma sonuçlarının üzerinde çalışma yapılan örneklemin dışında daha geniş bir ana kütleye, zamana, duruma veya düz-

leme genellenebilmesidir. Ana kütleye genellenemeyen sonuçların geçerliliği düşüktür. Genellenebilirlik, ölçüm aracının kendisiyle ilgili değildir, daha çok örneklemin yeterliliğini, örnekleme sürecini ve bu çerçevede elde edilmiş sonuçları dikkate alır. Bazı yazarlar genellenebilirliği *ekolojik geçerlilik* olarak isimlendirmişlerdir.[18] Niceliksel araştırmalarda genellenebilirlik özelliği yüksek iken niteliksel araştırmalarda genellenebilirlik özelliği düşüktür. Niteliksel araştırmalarda daha çok iç geçerlilik üzeride odaklanılırken, niceliksel araştırmalarda iç ve dış geçerliliğin her ikisine de önem verilir.

Değerlendirme geçerliliği. Değerlendirme geçerliliği, araştırma sonucunda varılan yargıları ifade etmek ve araştırmanın *gerçeğini* ortaya koymak değildir. Tersine bir değerlendirme çerçevesi, planı veya yaklaşımı geliştirmek ve bu plan çerçevesinde araştırılan *iddiayı* analiz etmektir. Değerlendirmede yorum yapmak kaçınılmazdır; ancak değerlendirme ve tartışmalar *geçerlilik* verilerine dayalı olarak yapılmalıdır. Maxwell'in yaklaşımında, *ölçüm aracı* ile birlikte araştırma sürecinin *belirli aşamalarının* da geçerli olması tezi ele alınmıştır. Ölçüm aracına ait verilerin tek başına geçerli olması, genellenebilirlik ve değerlendirme geçerliliği olmadığı sürece bilimsel açıdan fazla bir anlam ifade etmez.

AİDİYETİ

Geçerliliğin test formuna mı yoksa toplanan verilere mi ait olduğu konusuyla ilgilidir. Meslekten olmayan kişiler, çoğunlukla kullanılan "testin veya ölçeğin geçerli olup olmadığını" sorarlar. Örneğin, "ölçeğin geçerlilik analizi yapıldı mı?" şeklindeki bir soruyla karşılaşırız. Geçerlilik, sadece sürece ve ölçüm aracıyla toplanan verilere aittir.[19] Araştırma yapılan örneklemin büyüklüğü, örnekleme yöntemi, örneklemin homojen veya heterojen olması ve ölçek maddelerinin[e] anlaşılırlığı kullanılan test verilerinin geçerliliğini etkiler. Ayrıca, bir testin veya ölçeğin geçerliliği tek bir araştırmaya bağlı olarak belirlenemez. Geçerlilik devam eden ve sonu olmayan bir süreçtir. Yurt dışında geçerlilik analizi yapılmış bir testin/ölçeğin bazen Türkçeye çevrilerek kullanıldığı ve araştırma sonuçlarının tekrar geçerlilik analizleri yapılmadan raporlandığı görülür. Buradaki

[e] İngilizcedeki *item* kelimesinin karşılığı olarak kullanılmıştır. Başarı testlerinde soru cümlesi, tutum ölçeklerinde görüş cümlesi, yetenek testlerinde ise bir problem cümlesi veya grafik ünitesi olabilir.

yanlış anlayış, testin geçerli olduğu felsefesine dayanır. Aynı şey ülke içinde geçerlilik analizleri yapılmış diğer testlerin ikinci kez kullanılması sırasında da ortaya çıkar. Kullanılan her bir test veya ölçeğin önceki uygulamalarında geçerlilik analizleri yapılmış olsa bile daha sonraki kullanımlarında bu analizlerin yeniden yapılması gerekir.

Araştırmacı incelediği yapıyla ilgili olarak ölçek/test maddelerini kendisi geliştirebilir veya daha önceden geçerlilik ve güvenilirlik analizi yapılmış bir başka ölçeği kullanabilir. Bu konuda araştırmacıların mümkün olduğunca standart test/ölçek maddelerini kullanmaları uygun olur. Standart maddelerin kullanılması halinde araştırmacı kendi sonuçlarıyla önceki araştırmaların sonuçlarını karşılaştırma imkanı elde eder.[20] Fakat bu tür bir ölçek kullanılmış olsa bile geçerlilik ve güvenilirlik analizleri yeniden yapılmalıdır.

AŞAMALARI

Araştırma sürecine ilişkin geçerlilik çalışmaları, araştırmanın her aşamasında ayrı ayrı ele alınır. Bu aşamalardan en önemlisi ölçüm aracının geliştirilmesi, uygulanması ve verilerin kullanılmasıdır. Ölçüm aracıyla toplanan verilere ilişkin geçerlilik analizleri üç aşamada gerçekleştirilir: pilot araştırma sonuçlarına dayalı olarak, esas araştırmanın devam ettiği sırada ve esas araştırmanın bitiminden sonra. Birinci aşamada pilot araştırması sonucunda elde edilen veriler kullanılır. Hazırlanan ölçeğin pilot araştırma ile sınanması tercihe bağlı bir olay değildir, tersine iş işten geçmeden bazı yetersizliklerin saptanması için gerekli olan bir süreçtir. Bu aşamada uygun olmayan maddelerin çıkarılması, bazı maddelerin yeniden ifade edilmesi, ölçüm boyutlarının yeniden belirlenmesi gibi çalışmalar yapılır. Değiştirilen maddeler için araştırmacı bunları esas araştırmada değiştirilmiş şekliyle uygulayabilir veya yeniden başka bir pilot araştırma daha yapabilir. Pilot araştırmaya alınan kişilerle gerçek örnek kütledeki kişiler aynı ana kütleden seçilmeli, fakat farklı olmalıdır. Nihaî araştırma raporuna; pilot araştırma bulguları, süreç içinde yapılan araştırma bulguları ve esas araştırma sonucunda elde edilen bulgular birlikte alınır. Bir test veya ölçeğe ait verilerin geçerliliğine esas araştırma bulgularına bakılarak karar verilir. Pilot araştırma bulguları, testi/ölçeği rafine hale getirmek veya gelişmeleri izlemek için kullanılır. Varılan yargıların doğruluğu, gerçek örnek kütledeki verilerin geçerliliğine bağlıdır.

VERİLERİN NİTELİĞİ VE GEÇERLİLİK

Herhangi bir ölçek veya testle toplanan veri türleri nominal, sıralı, eşit aralıklı veya oranlı olabilir. Veri türüne göre geçerlilik değerlendirmesi için kullanılacak istatistikî analiz türleri de değişir.

Nominal Verilerde

Nominal nitelikteki değişkenlerin (demografik değişkenler ve diğer ölçüm değişkenleri olabilir) geçerliliği için, gerek duyulmuşsa değerlendiriciler arası uyuşma indeksi olan Cohen kappa formülü uygulanır. Burada değerlendiriciler demografik değişkenin kökü ile bu köke ait şıkların yeterli olup olmadığını değerlendirirler. Örneğin, sosyoekonomik sınıflandırmanın hangi ölçütler temel alınarak yapılacağı gerçek anlamda bir sorundur. Sosyal sınıf gruplandırmaları, eğitimden sağlığa, meslekten kültüre kadar pek çok değişkenden etkilenebilmektedir. Bu sınıflandırma doğru yapılmamışsa verilerin geçerliliği tehlikeye girer. Örneğin, bazı anketlerde meslek sorusu altında "ev hanımı" şıkkının yazıldığı görülmektedir. Ancak bu şıkkı bir meslek olarak nitelendirmek doğru değildir. Hakemlerin sorunun şıklarının yeterli ve uygun olduğunu belirlemeye yönelik olarak yaptıkları değerlendirmeler yüzey geçerliliğiyle ilgilidir. Literatürde bu uygulamaya aynı zamanda "mutabakat geçerliliği" adı verilmiştir.

Nominal olarak adlandırılan her tür sınıflandırılmış ölçek verisi için istatistiksel geçerlilik analizi yapmak anlamlı değildir. Eğitim değişkeni için; *okul bitirmemiş, ilk öğretim, orta öğretim* ve *yüksek öğretim* şıkları araştırma amacına uygunsa, geçerlilik analizi yapmaya gerek yoktur. Nominal ölçek niteliğindeki soru listeleri için genelde istatistiksel geçerlilik analizleri yapılmaz.

Sıralı Ölçek Verilerinde

Ölçüm verileri büyükten küçüğe veya küçükten büyüğe doğru belli bir sıralamaya sahipse "sıralı ölçek verisi" olarak isimlendirilir. Sıralı veri; çoktan seçmeli sorunun şıklarına ait, Likert ölçeğindeki her bir maddenin derecelerine ait, Bogardus-Guttman ölçeğinin *Evet* yanıtlarına ait, çiftli karşılaştırma ölçeğine ait, tercih sıralı ölçeklere ait (rank-order scaling), Q sıralama ölçeğine ait veya *sabit toplam değerli ölçek* (constant sum scaling) verilerine ait olabilir. Sıralı ölçek verileri iki dereceli veya çok derecelidir. Sıralı verilerin hakemler tarafından yapılacak geçerlilik değerlendirmesi istatistiksel olarak *kappa katsayısı* ile hesaplanır.[21] Veriler çok dereceli bir

özelliğe sahipse geçerlilik analizi için yüzey, içerik ve duruma göre kriter geçerliliği analizleri uygulanır. Maddeleştirilmiş dereceleme ölçekleri için kavramsal bir yapı ölçülmediğinden yapısal geçerlilik analizleri yapılmaz.

Eşit Aralıklı Ölçek Verilerinde

Eşit aralıklı ölçekler, şıklara ait puanların veya etiket derecelerinin şıklar/dereceler arasındaki farklılığın büyüklüğü hakkında okuyucuya belli bir fikir verebildiği ölçüm araçlarıdır. Fakat bu tür ölçeklerde sabit bir sıfır noktası bulunmaz. Eşit aralıklı ölçeklerin sık kullanılanları şunlardır: grafik dereceleme ölçeği (aynı zamanda "sürekli dereceleme ölçeği" olarak da isimlendirilmiştir), Likert ölçeği, anlamsal farklılık ölçeği, stapel ölçeği, Thurstone ölçeği, davranışa çıpalı dereceleme ölçeği (behaviorally anchored rating scale), yansıtıcı ve oluşturucu ölçekler. Veriler eşit aralıklı bir ölçeğe aitse söz konusu ölçek arka plandaki gizli bir yapıyı ölçüp ölçmeme durumuna göre oluşturucu ölçek ve yansıtıcı ölçek olmak üzere iki genel başlık altında incelenir ve geçerlilik analizleri buna göre yapılır.

Yansıtıcı ölçeklerde geçerlilik. Yansıtıcı ölçeklerde yapısal geçerliliği belirlemeye yönelik olarak; (a) yüzey, (b) içerik, (c) nomolojik ağ, (ç) birlikte vuku bulma, (d) tahmin geçerliliği –bilişsel testlerde–, (e) yapısal geçerlilik çerçevesinde ise faktör analizi, birleşme-ayrılma geçerliliği ve yine duruma göre çoklu özellik - çoklu yöntem geçerliliği hesaplamaları yapılır.

Oluşturucu ölçeklerde geçerlilik. Oluşturucu ölçekleri klasik anlamda "ölçek" olarak isimlendirmek doğru değildir. Oluşturucu ölçekler aslında indeks değişkenlerine sahiptir.[22] Fakat oluşturucu ölçek tanımlamasındaki "indeks" kavramıyla Babbie'nin tanımladığı "indeks" kavramları birbiriyle örtüşmez. Babbie'nin indeks tanımlamasında maddeler arasındaki korelasyonlar önemli iken oluşturucu ölçeklerde bunun herhangi bir anlamı yoktur. Oluşturucu ölçeklerde yansıtıcı ölçeklerden farklı bir yol izlenir. Oluşturucu ölçeklerde; (a) yüzey, (b) içerik, (c) nomolojik ağ, (ç) birlikte vuku bulma geçerliliğinin yanında (d) yapısal geçerliliği saptamak için Kısmi En Küçük Kareler (Partial Least Squares – PLS) analizi ve (e) çok boyutlu oluşturucu ölçek vak'asında ise ayrılma geçerliliğini saptamaya yönelik olarak teyit edici faktör analizi yöntemlerinden yararlanılır. Oluşturucu ölçeklerde rasyonel geçerlilik, ampirik geçerlilikten daha önemlidir ve rasyonel geçerlilik de büyük ölçüde içerik geçerliliğine dayanır. Yapısal geçerliliği belirlemeye yönelik olarak uygulanan Kısmi En Küçük Kareler yönteminde geliştirilen göstergelerin amaçlanan yapıyla ilişkili olduğunu

belirlemek için *gösterge ağırlık değerleri* kullanılır. Oluşturucu ölçeklerde maddeler "yapıyı" yansıtmayıp oluşturması nedeniyle geçerlilik çalışması gerekleri yansıtıcı ölçekler kadar katı değildir.[23] Oluşturucu ölçeklerde göstergelerin birbirleriyle ilişkili oldukları varsayılmadığından, belirli bir yapıyı temsil edip etmedikleri, iç tutarlılık derecesi, maddelerin tek boyutluluğu veya çok boyutlu indekslerde boyutların literatüre uygunluğu soruşturulmaz. Bu nedenle, ölçeğe ait madde-toplam puan korelasyonlarında ilişki katsayısı düşük çıkmış olsa bile bir maddenin ölçekten düşürülmesi düşünülmez. Maddeler arasındaki çoklu korelasyon katsayıları ve madde çiftlerine ait kovaryans değerleri önemli olmadığından düşük katsayılı maddeler de ölçekte bırakılır. Çünkü tek bir madde dahi ölçek için büyük ölçüde anlamlı olabilir. Oluşturucu ölçeklerde, yansıtıcı ölçeklerin tersine maddelerin çoklu doğrusallık özelliğine (koşutluk) sahip olması istenmez. Oysa yansıtıcı ölçeklerde, maddeler arasındaki koşutluk özelliği istenen ve aranılan bir niteliktir. Oluşturucu ölçeklerde iki madde birbiriyle yüksek derecede ilişkili ise bu maddelerden birinin düşürülmesi yoluna başvurulur. Oluşturucu ölçeklerde yapısal geçerlilik için *madde ağırlık değerleri* ve madde ağırlık değerlerinin istatistiksel anlamlılığı önem kazanır. Chwelo ve Benbasat'a (2004) göre, geçerli bir oluşturucu ölçekte göstergeler arasındaki korelasyon katsayıları pozitif işaretli, negatif işaretli veya sıfır değerli olabilir. Oluşturucu ölçeklerde faktör-madde arasındaki ilişkiyi gösteren *faktör yükü değerleri* değil, PLS analizi sonucunda elde edilen ve R^2 simgesiyle gösterilen *gösterge ağırlıkları* araştırılır. Gösterge/madde ağırlıkları, standart regresyon denklemindeki beta katsayıları gibi yorumlanır (1,0'e ne ölçüde yakın olduğu). Ölçekte yer alan göstergelerin ağırlıkları karşılaştırılarak oluşturucu yapıyı hangi maddenin daha çok temsil ettiğine bakılır. Bunun yanında bilim adamı, ağırlık değerlerinin istatistiksel anlamlılığını gösteren t değerlerini inceleyebilir. Normal koşullarda oluşturucu göstergelerin *ağırlık değerleri* yansıtıcı ölçeklerin *faktör yüklerinden* daha düşüktür. Bu nedenle oluşturucu bir ölçek düşük *ağırlık değerlerine* sahip göstergelerden meydana gelmişse zayıf bir ölçüm aracı veya modeli olarak değerlendirilmemelidir.[24] Bir oluşturucu ölçek tek boyutlu veya çok boyutlu olabilir. Eğer çok boyutlu ise, birinci düzeyde "boyutlar ve göstergelerden" ve ikinci düzeyde ise "kavramsal yapı ve boyutlardan" söz edilir. Herhangi bir ölçekte birinci ve ikinci düzey boyutların her ikisi de yansıtıcı, birinci düzey yansıtıcı, ikinci düzey oluşturucu veya tersi, ayrıca her iki düzey oluşturucu nitelikte olabilir.[25] Bu şekildeki karma bir niteliğe sahip ölçeklerde ölçüm modeline uygun olarak yansıcı ve oluşturucu ölçekler için uygun olan geçerlilik analizleri hep birlikte yapılır.

Diamantopoulos ve Winklhofer (2001) oluşturucu bir ölçek oluştururken başlıca dört kritik sorun belirlemişlerdir. Bunlardan birincisi, ölçülmek istenen kavramsal alana ait içerik belirlemesidir. İkincisi, kavramsal alana ait göstergelerin saptanmasıdır. Üçüncüsü, göstergeler arasındaki koşutluk (çoklu doğrusallık) özelliğinin araştırılmasıdır. Oluşturucu ölçeklerde göstergeler arasında koşutluk özelliği varsa, bu durum göstergelerin ağırlık değerlerini etkiler. Bu nedenle maddelerin koşutluk özelliğine sahip olması istenmez, koşutluğu sağlayan maddelerden biri düşürülür. Dördüncüsü ise, göstergelerin dış geçerliliğinin belirlenmesidir (aktaran, Reinartz, Krafft, ve Hoyer, 2004).[26] Dış geçerlilik, bir yapıyı temsil eden tüm göstergelerin ölçeğe alınmış olduğunun kanıtlanmasıdır. Yansıtıcı ölçeklerde olduğu gibi literatüre uygun olarak tam bir gruplandırma veya sınıflandırma şeklinde olmasa bile araştırmacı göstergelerin gerçek hayattaki, örgütlerdeki, toplumdaki uygulamalara uygun olduğu ve yelpazenin bütün alanlarını kapsadığı konusunda makul ve tutarlı açıklamalar yapmalıdır. Bu aşamaların içinde en önemlisi ikinci sırada gelen göstergelerin belirlenmesi aşamasıdır. Yansıtıcı ölçeklerden farklı olarak oluşturucu ölçeklerde kavramsal alana ait göstergelerin alanı kapsayacak kadar geniş tutulması gerekir. Yansıtıcı ölçeklerde güvenilirliği ,70'in üzerine çıkaracak optimum sayıda madde sayısı alan örneklem büyüklüğü için yeterli sayılırken oluşturucu ölçeklerde içerik geçerliliğinin sağlanması için madde sayısının kavramsal alanın bütün yönlerini (veçhelerini) kapsayacak kadar geniş olması gerekir. Yapısal geçerlilik, gözlem puanlarıyla varsayılan teori arasındaki ilişkilere dayandığından ve oluşturucu ölçeklerde de alan örneklemesi yapılmadığından geliştirilen oluşturucu ölçekte gözlem-kuram örtüşmesinin büyük ölçüde içerik geçerliliği ile sağlanmış olduğuna dikkat edilmelidir.

ÖLÇEK DERECELERİ VE GEÇERLİLİK

Araştırmada kullanılan ölçek derecelerinin geçerliliği konusunda güvenilirliğe göre nispeten daha az araştırma yapılmıştır. Ancak bu araştırmalarda da güvenilirlikte olduğu gibi geçerlilik katsayılarının yedi dereceye kadar arttığı bulunmuştur.[27] Değerlendiricilerin veya hakemlerin yaptıkları değerlendirmede iki üç dereceli, basit dereceleme şıkları kullanıldığı zaman *değerlendirici yanlılığının* ortaya çıktığı ve derece sayısının dörtten fazla olduğu zaman ise bu yanlılığın azaldığı bildirilmiştir.[28] Araştırmacı ölçeğindeki derece sayısını belirlerken 5, 7 ve 9 gibi tek rakamlı dereceleri düşünmelidir. Tek rakamlı derece sayıları verilerin normal dağılım özelliği açısından daha uygundur. Ölçeklerin derecelendirilmesi yapılırken tam

ortada nötr ifadelere yer verilmeli, fakat bunun yanında cevaplayıcılarla ilgili olmayan ifadeler için, diğer ölçek dereceleriyle karışmayacak bir şekilde ayrı bir sütunda *Bilmiyorum/Görüşüm yok* şıkkı belirlenmelidir. Yapılan araştırmalar, çift numaralı ölçek derecelerini kullanıldığı durumlarda nötr noktanın bulunmaması nedeniyle insanların olumlu yanıtları seçme eğilimi içine girdiğini ve bu nedenle de sonuçların pozitif çarpıklığa sahip olduğunu göstermiştir.[29] Derece sayısı kadar önemli olan bir diğer nokta derecelerin etiketlenme sorunudur. Cevaplayıcılar derecelerin her birine belirli etiketler verildiği zaman daha anlamlı bir şekilde işaretleme yapabilmektedirler. Bu nedenle ölçek derecelerinin her birine etiket verilmiş olmalıdır.

TEST TÜRLERİ VE GEÇERLİLİK

Bilişsel Yetenek Testleri ve Geçerlilik

Bilişsel yetenek testleri, kişilerin zihinsel kabiliyetlerini saptamak ve değerlendirmek için kullanılan ölçüm araçlarıdır. Zeka testleri, genel yetenek testleri, sözel kavrama, sözel akıcılık, sayısal muhakeme, görsel takip, dikkat, üç boyutlu uzay ilişkilerini kavrama, şekil algısı ve sayısal yetenek testleri bilişsel yetenek testlerinin tipik örnekleridir. Bilişsel yetenek testlerinde aşağıdaki geçerlilik analizleri yapılır.

1. Yüzey geçerliliği. Test maddelerinin türdeş, anlaşılır ve iyi düzenlenmiş olması.
2. İçerik geçerliliği (problem cümlesi içeren maddelerde). Bilgiyi ölçen testlerin kavramsal alanın bütün yönlerini içermesi. Örneğin, sayısal muhakemede dört işlem, kesirli işlemler, köklü işlemler vb. temel matematiksel işlemlerin test içine alınmış olması.
3. Tahmin geçerliliği. Gelecekteki bir başarıyı doğru bir şekilde tahmin etme gücü.
4. Birlikte vuku bulma geçerliliği. Testin altın standardıyla karşılaştırılması veya zayıf ve iyi performans gösteren gruplarda sınanması.
5. Yapısal geçerlilik (birleşme ve ayrılma geçerliliği). Testin aynı kavramsal yapıyı ölçen diğer psikometrik testlerle yüksek, farklı kavramsal yapıları ölçen testlerle ise düşük korelasyona sahip olması. Belirli bir yeteneği ölçmek için oluşturulan test bataryasına alınan testlerin g faktör yükünün ,70'nin üzerinde olması. Faktör

analizi uygulandığında testin değil, test bataryasının geçerlilik analizi yapılmış olur.

6. Artışsal geçerlilik (Incremental validity). İlave diğer testlerin katılmasıyla elde edilen toplam puan ile performans puanları arasındaki korelasyon analizi sonucunda geçerlilik katsayısındaki artış. Örneğin, PM-38 testinin tahmin geçerliliği katsayısı ,45 çıkmış olsun. PM-38+B53 testinin artışsal geçerlilik katsayısı ,55 olabilir.

Bilişsel yetenek testlerinde arka plandaki gizli kavramsal yapılar ölçülüyor olması nedeniyle bütün geçerlilik analizi yöntemleri uygulanabilir.

Tutum Ölçekleri ve Geçerlilik

Tutum ölçekleri duygusal nitelikleri belirlemeyi amaçlayan ölçüm araçlarıdır. Tutum ölçüm araçları; ölçek, indeks veya maddeleştirilmiş sorular şeklinde olabilir.

Literatürde indeks ve ölçek kavramları konusunda bilim adamları belirli bir fikir birliği içinde değildirler. Babbie, "ölçek" kavramında maddeler arasındaki hiyerarşik sıralanmaya ve Likert ölçeğinde olduğu gibi katılım yoğunluğunu gösteren çok dereceli yanıtlama modellerine; "indeks" kavramında ise hiyerarşik bir sıralamaya sahip olmayan maddelerin sadece seçilmesi ve bu maddelere atanan puanların toplanması veya aritmetik ortalamasının alınması olgusuna önem vermiştir. İndekslerde bireysel vasıflar önemlidir ve bireysel vasıfların bulunma durumu puanlar toplanarak belirlenir. Basit bir indekste, her maddeye işaretlenmesi halinde 1 puanı verilir. Fakat araştırmacı isterse indeks göstergelerini Likert ölçeklerinde olduğu gibi çok dereceli olarak da puanlandırabilir. Babbie'nin yaklaşımında indeks ölçüm aracı, basit bir şekilde yapılanmasıyla "ölçeklerden" teknik anlamda ayrılır.

Son yıllarda ise indeks kavramı aynı zamanda *oluşturucu ölçek* anlamında kullanılmıştır. Fakat, oluşturucu indekslerde maddeler arasındaki ilişkinin güçlü olmasına izin verilmez. Oluşturucu indekslerde maddeler birbirinden bağımsızdır.

İndekslerin geçerlilik analizleri. İndekslerin tipik özelliği tek boyutlu olmasıdır. Babbie (1973), indeksleri tek bir boyut veya kavramın ölçüldüğü araçlar olarak görmek istemiştir.[30] Fakat daha sonraki yıllarda indeksler de ölçeklerde olduğu gibi çok boyutlu olarak geliştirilmeye başlanmıştır. Bilim adamlarının bir bölümü, indeksin tek boyutlu olmasının onun yapısal geçerliliğini azaltacağını ileri sürmüşlerdir. Ancak bir indeks çok boyutlu olsa

bile, boyutlar arasındaki korelasyon düşükse her bir alt boyut, bir alt indeks gibi değerlendirilip maddeler ona göre seçilmelidir.

Bir indeks Babbie'nin tanımladığı biçimde tek boyutlu veya sonraki gelişmelere bağlı olarak çok boyutlu bir şekilde oluşturulmuşsa geçerlilik analizleri için; (a) yüzey, (b) içerik, (c) iç tutarlılık, (ç) maddelerin varyans değerleri incelemesi, (d) yapısal geçerlilik ve (e) dış geçerlilik incelemesi yapılır. Bu anlamda güvenilirlik analizleri ile geçerlilik analizi belirli bir ara kesite sahiptir. Maddelerin iç tutarlılığı çerçevesinde maddeler arası korelasyon, madde-toplam puan korelasyonu analizleri yöntemine başvurulur. Maddeler arası korelasyon katsayıları ,20'nin üzerinde olan maddeler indekse alınır. Varyans değerleri incelemesi sonucunda indekste farklı varyans değerlerine sahip maddeler bulunmalıdır. Ayrıca madde derecelerinin yüzde dağılım tabloları incelenerek kişileri yaklaşık olarak iki yarıya bölen maddeler seçilir. Yine bu aşamada maddeler arası korelasyon analizi yapılarak koşutluk özelliğine sahip olanlar ölçekten düşürülür.[31] İndeksin dış geçerliliği için ise geliştirilen anket formu aynı kavramsal yapıyı ölçen başka bir indeks veya ölçekle karşılaştırılır. Görüldüğü gibi Babbie'nin indeks tanımlaması ölçek tanımlamasına benzemekle birlikte hassas bir alan örneklemesi yöntemine başvurulmaması, toplam/ortalama puanların katılımın yoğunluğu yerine sadece bir indeks değeri olarak gözükmesi, literatür taraması yapmaya gerek kalmadan basit bir şekilde masa çalışması yapılarak hemen oluşturulabilmesi gibi özelliklere sahiptir. Bu nedenle de indekslerde ayrıntılı bir şekilde faktör analizi, birleşme-ayrılma geçerliliği analizleri yapılmayabilir.

İndeks, oluşturucu ölçek niteliğinde ise bu kez oluşturucu ölçekler için uygun olan geçerlilik analizleri yapılır. (Bu konuda "Eşit Aralıklı Ölçek Verilerinde Geçerlilik" başlığına bakınız.).

Ölçeklerin geçerlilik analizleri. Ölçekler, katılım yoğunluğunun; (a) maddelerinin hiyerarşik bir sıralanmaya sahip olduğu düzenleme ile veya (b) maddelerin şıklarının derecelendirilmesi yoluyla belirlendiği hassas ölçüm araçlarıdır. Ölçeklerde, indekslerin tersine maddeler alan örneklemesi yöntemine göre belirlenir ve maddeler dikkatli bir kalibrasyon çalışmasının sonucunda ölçek haline gelir. Ölçeklerin ayırt edici en önemli niteliği, titiz bir güvenilirlik ve geçerlilik çalışmasının sonunda katılımın yoğunluğunu belirlemeye yönelik olarak oluşturulmasıdır. İndekslerde olduğu gibi ölçeklerin de başlangıçta tek boyutlu olduğu konusu vurgulanmasına karşılık pek çok bilim adamı ölçeklerin de çoğunlukla tek değil, çok boyutlu olduğunu göstermişlerdir. Tutum ölçekleri, yansıtıcı veya oluşturucu nitelikte olabilir ve geçerlilik analizleri de buna göre değişir. (Bu konuda "Eşit Aralıklı Ölçek Verilerinde Geçerlilik" başlığına bakınız.).

Maddeleştirilmiş soruların geçerlilikleri. Likert tipi bağımsız maddeler veya diğer çoktan seçmeli bağımsız soruların geçerliliği birleşik ölçüm olmaması nedeniyle nispeten daha kolay bir şekilde yapılır. Bu tür maddeler için; yüzey, içerik ve duruma göre yapısal eşitlik modellerinde nomolojik ağ geçerlilik çalışması yapılır.

Bilgi Testleri ve Geçerlilik

Bilgi testleri daha çok okullara giriş sınavlarında, eğitim süreci içinde yapılan sınavlarda, mezuniyet sınavlarında, sertifikasyon sınavlarında, lisans alma sınavlarında ve işletmelerde personel seçimi sırasında meslekî bilgiyi belirlemeye yönelik olarak yapılır. Bilgi sınavlarında aşağıdaki geçerlilik analizi yöntemleri uygulanır.

1. Yüzey geçerliliği.
2. İçerik geçerliliği.
3. Birlikte vuku bulma geçerliliği.
4. Yapısal geçerlilik (faktör analizi, birleşme ve ayrılma geçerliliği).
5. Çoklu özellik - çoklu yöntem analizi.

Bir test uygulamasında sayılan geçerlilik analizlerinin hepsinin yapılması gerekmeyebilir. Geçerlilik analizlerinin genişliğini araştırma veya ölçüm tasarımı belirler.

Kişilik Envanterleri ve Geçerlilik

Kişilik envanterlerinde gizli olan "özellikler" ölçülmeye çalışılır. Özellikler, teorik kavramsal yapılardır. Bu nedenle tutum ölçekleri için uygulanan geçerlilik analizleri kişilik ölçekleri için de uygundur. Bu çerçevede kişilik envanterlerinde yüzey, içerik, tahmin, yapısal, birleşme ve ayrılma geçerlilik analizleri yapılır. Tahmin geçerliliğini yapmak için, kişilik ölçeği içerdiği boyutlara göre belirli özellikleri gösteren özel gruplarla genel gruplara ayrı uygulanır ve özel grupların incelenen boyutta yüksek puan alıp almadıklarına bakılır. Kişilik envanterlerinin uygulama geçerliliğini sağlamak için iki yaklaşımdan yararlanılır. Bunlardan birincisi ortalama yöntemi ve ikincisi ise karşılaştırma yöntemidir. Ortalama yönteminde belirli kişilik özellikleri için norm değerlerin saptanmasına yönelik bir çalışma yapılır. Norm değerlerini oluşturmak için çok sayıda kişide (10.000 gibi) anket uygulanır ve bu anketlere bağlı olarak norm değerleri saptanır. Ortalama yönteminde, her bir maddeye gelen cevapların artı-eksi (±) beklenen norm-

lar düzeyinde olmasına dikkat edilir. Bu kriteri tutturamayan maddeler yeniden gözden geçirilir. Karşılaştırma yönteminde ise kişilik ölçeğinin puanları aynı kavramsal yapıları ölçen tanınmış başka bir testin puanlarıyla karşılaştırılır. Geçerlilik açısından iki ölçek sonuçları arasında en azından %75 oranında bir benzerlik olması gerekir.

Fiziksel Yetenek Testleri ve Geçerlilik

Fiziksel yetenek testleri literatürde değişik şekillerde ele alınıp sınıflandırılmıştır. Bu sınıflandırmalardan biri; statik güç (sıkma gücü, kolu kaldırma gücü), dinamik güç (kolun dayanıklılığı), vücut kuvveti (oturma, kalkma), anaerobic güç (bacakların gücü) ve esneklik gücü şeklindedir.[32] Burada sınıflandırma mantığı üzerinde durmadan bu tür testlerde yapılabilecek geçerlilik analizi yöntemlerine işaret edilmek istenmiştir. Fiziksel yetenek testlerinde aşağıdaki geçerlilik analizi yöntemleri uygulanır.

1. Yüzey geçerliliği. Fiziksel yetenek testinin işle ilgili olma derecesi.
2. İçerik geçerliliği. Fiziksel yetenek testi içindeki vücut hareketlerinin kapsamı, yeterliliği ve işe uygunluğunun belirlenmesi.
3. Tahmin geçerliliği. Fiziksel yetenek testinin iş başarısıyla ilişkisinin araştırılması.
4. Birlikte vuku bulma geçerliliği. Fiziksel yetenek testinin zayıf ve güçlü birey gruplarında sınanması. Ayrıca fiziksel yetenek testlerinin altın standardı diğer test sonuçlarıyla karşılaştırılması.
5. Yapısal geçerlilik. Fiziksel yetenek test bataryasının toplam varyansı açıklama yüzdesinin araştırılması.

Fiziksel yetenek testlerinin geçerliliği büyük ölçüde iş başında elde edilen performans sonuçlarıyla karşılaştırma yapılarak sağlanır.

Psikomotor Testler ve Geçerlilik

Psikomotor yetenek testleri ilk olarak 1920'li yıllarda bilim dünyasına tanıtılmıştır. Fleishman (1972) ölçülebilecek 11 adet Psikomotor yetenek belirlemiştir ve bunlar aşağıdaki gibidir.[33]

1. Nişan alma.
2. El-kol kımıldamazlığı.
3. Kontrol hassasiyeti.

4. Parmak becerisi.
5. El becerisi.
6. Çoklu kas koordinasyonu (hareket halinde olmaksızın; otururken, yatarken veya dururken iki kolu, iki bacağı veya bir kolu ve bir bacağı aynı zamanda birlikte hareket ettirme).
7. Hız kontrolü.
8. Reaksiyon zamanı.
9. Tepki yönelimi.
10. Kol hareketinin hızı.
11. Bilek-parmak hızı.

Hava trafik kontrolörlüğü gibi mesleklerde önemli olan psikomotor yetenek testlerinin geçerliliğini analiz etmek için daha çok içerik, birlikte vuku bulma ve tahmin geçerliliği yöntemleri uygulanır. Araştırmalar tahmin geçerliliği rakamlarının pratik yapma etkisi nedeniyle genelde düşük çıktığını göstermiştir. Psikomotor yetenek testlerinin içerik geçerliliği, kullanılan test ile performans, iş, görev veya problemler arasında bir ilişki olmasını gerektirir. Geliştirilen psikomotor yetenek testinin faaliyetle olan ilgisi önemlidir. Psikomotor yetenek testi, doğrudan ilgili faaliyeti minyatürize edilmiş bir ortamda ölçme niteliğine sahip olmalıdır. Ayrıca testin geçerliliği, psikomotor niteliğin işteki ağırlığıyla da orantılıdır. Görevde psikomotor faaliyetin ağırlığı %10 gibi bir düzeyde ise psikomotor yetenek testi - performans puanları korelasyon katsayısı düşük çıkar.

GÖSTERGELERE İLİŞKİN GEÇERLİLİK ANALİZİ YÖNTEMLERİ

Ölçüm aracıyla toplanan göstergelere ait verilerin geçerlilik analizi değişik şekillerde yapılabilir. Uzun yıllardan beri bu konuda öncü bir rol üstlenen APA ve AERA'nın yaklaşımlarıyla söz konusu uygulamalar bir ölçüde standart hale gelmiştir. APA'ya göre üç tür geçerlilik söz konusudur: içerik, kriter ve yapısal geçerlilik. Ancak yapılan araştırmanın türü, akademik düzeyi ve ölçümün niteliği bu yöntemlerin kullanılmasını genişletebilir veya daraltabilir. Bu bölümde ölçüm araçlarıyla elde edilen verilerin geçerlilik analizleri herhangi bir sınıflandırma sistemine bağlı kalınmadan genel bir yaklaşım içinde, dört başlıkta ele alınmıştır: yüzey geçerliliği,

içerik geçerliliği, kriter geçerliliği ve yapısal geçerlilik. Ölçüm verilerinin geçerliliğinin saptanmasında bu yöntemlerin her üçünün veya dördünün birlikte kullanılması gerekmez. Hangi analizlerin uygun olacağına ölçümün niteliğine bağlı olarak araştırmacının kendisi karar verir.

YÜZEY GEÇERLİLİĞİ

Yüzey geçerliliği, bir testin/ölçeğin araştırılan yapıyı ölçüp ölçmediğine ilişkin olarak araştırmacının (a) kendisinin, (b) yakın çevresindeki arkadaşlarının, (c) araştırılan konu hakkında uzman olmayan diğer kişilerin ve (ç) pilot araştırmaya katılan cevaplayıcıların kanaat ve görüşlerinin toplanmasıyla belirlenir. Bu yaklaşım, aynı zamanda *mantıksal geçerlilik* olarak da isimlendirilmiştir.

Yüzey geçerliliği ilk aşamada araştırmacının kendisi tarafından yapılır. Araştırmacı, yüzey geçerliliğinde başlıca iki noktaya dikkat eder. Birincisi ifadelerin ölçüm amacına uygun olmasıdır. İkincisi ise, ifadelerin aynı zamanda hedef kitledeki kişilerin eğitim, kültür ve bilgi seviyelerini zorlamamasıdır.[f] Testin veya ölçeğin maddeleri ile araştırılan kavramsal yapı arasında bir şekilde anlamlı bir bağ kurulabiliyorsa test yüzeysel geçerliliğe sahiptir denilir. Yüzey geçerliliğinde ölçek maddesinin düzgün ve anlamlı bir şekilde ifade edilmesi, doğru terimlerin kullanılması, uygun kelimelerin seçilmesi, anlamın açık ve net olması; belirsiz, birden fazla anlama gelecek kelimelerden kaçınılması önemlidir. Yine yüzey geçerliliğinde ölçeğin okunurluk analizi,[g] terimlerin anlaşılırlık analizi ve cümlelerin uzunluk analizleri yapılır.[34] Kullanılan ölçek/test cevaplayıcıların eğitim düzeylerine, sahip oldukları bilgi donanımlarına, kültürel yapılarına ve yetenek düzeylerine uygun olmalıdır. Eline hiç iskambil kağıdı değmemiş bir kişiye iskambil kağıtlarından oluşan bir test verilmesi yüzey geçerliliği açısından uygun değildir. Araştırmacılar bazen bir testin yüksek yüzey geçerliliğine sahip olmasını istemeyebilirler. Yüksek yüzey geçerliliği, bazı durumlarda ölçüm hatası riskini arttırabilir.

---

[f] Bazı araştırmacılar yüzey geçerliliğini sadece testi/ölçeği alan kişilerin algılarına dayandırmışlardır.

[g] Okunurluk ve anlaşılırlık analizi için literatürde değişik formüller vardır. Bunlardan sık kullanılanlar Flesh-Kincaid ve Fog indeksidir. Ancak bu formüller daha çok İngilizce metinler için hazırlandığından Türkçe cümlelerin değerlendirilmesinde gerekli uyarlama çalışmaları yapıldıktan sonra kullanılmalıdır. Fog indeksi için *bk.*,<http://www.shared learning.org. uk/fog _ index. htm> (30.08.2002); H. Şencan, *Bilimsel Yazım*, İstanbul:İÜ İşletme Fakültesi Yayını, 2002.

Yüzey geçerliliği için test ikinci aşamada araştırmacının arkadaşları tarafından kontrol edilir. Araştırmacı arkadaşlarına anlaşılmayan ifadeleri, terimleri belirlemelerini ve bunları anlaşılacak hale getirmelerini söyler. Bunun dışında her bir maddeyle ilgili olarak bir sorun olup olmadığını, ifadelerin yeterince açık olup olmadığını belirlemelerini ister. Kendilerinden gelecek uyarılara uygun olarak gerekli değişiklikleri yapar.

Üçüncü aşamada ölçek veya test yakın arkadaş çevresinin dışında daha objektif olma şansına sahip üçüncü kişilere değerletilir. Araştırmacı bu kişilere de arkadaşlarına söylediği talimatları verir.

Dördüncü aşamada pilot uygulama sırasında katılımcılardan gelen görüşler değerlendirilir. Uygun büyüklükteki bir örnek kütlede sadece yüzey geçerliliğini belirlemeye yönelik olarak yapılacak bir pilot araştırma ile yüzey geçerliliği tekrar sınanır. Pilot araştırma yapma şansı kısıtlı olan kişiler tek başına yüzey geçerliliği için değil, genel amaçlı pilot araştırma verilerinden yararlanırlar. Bu aşama ile aynı zamanda içerik geçerliliği başlamış olur. Pilot araştırma sırasında kavramsal evrenden seçilen örneklem verileri de soruşturuluyorsa yüzey geçerliliği ile içerik geçerliliği birlikte paralel olarak sürdürülür. Araştırmacı yüzey geçerliliğini belirlemek üzere ana kütleyi temsil eden en az 40-50 kadar katılımcıya veya örnek kütlenin %20'lik bir dilimine pilot araştırma uygular. Pilot araştırma yapılacak kişilerin büyüklüğü konusunda kesin bir kural yoktur, ancak büyük örnek kütleler küçük örnek kütleye göre her zaman daha güvenilir bulunur.[h] Pilot araştırmada katılımcılara ölçekteki ifadelerin anlaşılırlık du-

[h] Pilot araştırma sırasında yüzey geçerliliği için nispeten küçük bir örnek kütle büyüklüğü yeterli olurken yapısal geçerlilik için faktör analizinin uygulanmak istenmesi halinde daha büyük örneklem hacmine ihtiyaç duyulur. Bu konuda standart bir ölçü verilmemiştir. Literatürde değişik yaklaşımlar bulunmaktadır. Örneklem büyüklüğü ayrıca faktör analizinin türüne göre de değişir. Sudman'a (1983) göre bir pilot araştırmada 20 ilâ 50 arasındaki vak'a sayısı anket formundaki yetersizlikleri tespit etmek için yeterlidir. Sheatsley (1983) ise yüzey geçerliliği için çok daha küçük örneklem hacminin yeterli olabileceğini belirtmiş ve 12–25 vak'a sayısından söz etmiştir (*bk.*, Zukerberg ve diğerleri). Garson (2002) geniş bir literatür taraması sonucunda bu konudaki yaklaşımları beş başlık altında toplamıştır. (a) On kuralı: Boyal, Stankov ve Cattell (1995) gibi araştırmacılar her bir ölçek maddesi için en az 10 cevaplayıcının olmasını önermişlerdir (*bk.*, D.T. Grifee, Questionnaire Construction and Classroom Research, <http://langue.hyper. chubu.ac.jp/jalt/pub/tlt/99/jan/griffee.html> (24.08.2002). (b) Kişi-madde Oranı (KMO): Bryant and Yarnold (1995), KMO'nun 5'ten az olmaması gerektiğini belirtmişlerdir. (c) 100 Kuralı: Hatcher'e (1994) göre kişi sayısı değişken sayısının 5 katından fazla veya en az 100 olmalıdır.(ç) 150 kuralı: Hutcheson ve Sofroniou'ya göre (1999) en az 150 kişi olmalıdır. (d) 200 kuralı: Gorsuch'a (1983) göre en az 200 kişi olmalıdır (bk., Garson, "Factor Analysis"<http://www2.chass.ncsu.edu/ garson/ pa 765/ factor.htm> (24.08.2002). Pilot araştırmaya katılacak kişilerin sayısı, pilot araştırtmanın amacına ve uygulanacak test türüne bağlıdır. Bir ölçekte hangi maddenin kalması ve

rumu, ölçeğin uzunluğu, kolay okunma ve doldurulma durumu, yazıların punto büyüklüğü, satırların sıkışık veya rahat görünmesi, açık uçlu sorular hakkındaki düşünceleri ve ölçeği doldururken sıkılıp sıkılmadıkları gibi sorular yöneltilir. Katılımcılardan alınan yanıtların analizi yapılır ve buna göre anket formunda gerekli değişikliklere gidilir.

Cevaplayıcıların kendilerini mağdur olmuş göstermek istememeleri, mutlu ve huzurlu olduklarını yansıtmaya çalışmaları, araştırmacının beklediği cevapları vermeleri veya test sonuçlarını sabote etmeye çalışmaları beklenilen cevapların alınması açısından olumlu olmakla birlikte yüksek düzeyde ölçüm hatası içerir.

Belirsiz ve öznel olması nedeniyle psikometrisyenlerin önemli bir bölümü yüzey geçerliliğini kullanmayı uzun bir zaman önce terk etmişlerdir. DeVellis (1991) ve Kerlinger (1973) gibi yazarlar yüzey geçerliliğini içerik geçerliliğinden farklı olduğun bildirirlerken Carmines ve Zeller (1979), Nunnally (1967) gibi yazarlar ise yüzey ve içerik geçerliliğini bir paranın iki yüzüne benzetmişlerdir (aktaran, Vickery).[35] Son zamanlarda Lacity ve Jansen'in (1994) "Kuram ve Geçerlilik" alanında yaptığı çalışmayla yüzey geçerliliği yeniden önem kazanmıştır. Bu yazarlara göre yüzey geçerliliği *sağ duyu* ve *ikna etme* özelliğidir (aktaran, Yu).[36] Hipotez test eden araştırmalarda bilim adamı sadece yüzey geçerliliği ile yetinemez. Görünüşteki geçerlilik yeterli değildir ve bilimsel gerçekler çoğunlukla insanların sağ duyularıyla kavrayabildiklerinin ötesindedir.

İÇERİK GEÇERLİLİĞİ

İçerik geçerliliği, örneklem olarak belirlenen test veya ölçek maddelerinin belirli bir amaca yönelik olarak kavramsal ana kütleyi temsil etme derecesidir. Bu nedenle literatürde bazen "örneklem geçerliliği" olarak isimlendirilmiştir. Seçilen örneklem maddeleri kavramsal ana kütleyi temsil ettiği oranda içerik geçerliliğine sahiptir. İçerik geçerliliğinde ölçüm aracının ölçmek istediği yapıyı ölçüp ölçmediği ölçeği geliştiren kişilerin kendileri-

---

hangi maddelerin çıkarılması gerektiğine karar vermek için yapılacak bir pilot araştırma en az 100 kişi üzerinde uygulanmalıdır (*bk.*, Andrew L. Zukerberg, Dawn R. Von Thurn and Jeffrey C. Moore, "Practical Considerations in Sample Size Selection For Behavior Coding Pretests [Davranış Ölçeklerinin Pilot Araştırma Aşamasında Seçilecek Örnek Büyüklüğü İçin Pratik Düşünceler]," <http://www.census.gov/srd/papers/pdf/az9501.pdf> (24.08.2002). Öte yandan, keşfedici faktör analizi için 100 civarındaki örneklem büyüklüğünün yeterli olabileceği bildirilirken teyit edici faktör analizi için örneklem büyüklüğünün en az 200 olması gerektiği vurgulanmıştır (*bk.*, Trochim a.g.k).

ne değil, uzman[i] kararlarına bırakılmıştır. İçerik geçerliliğinin arka planında yatan felsefe, uzmanların meslekten olmayan kişilere göre araştırılan yapıya/kavrama ilişkin nüansları ve ayrıntıları daha iyi bilecekleridir. Bu nedenle içerik geçerliliğinde "konu içeriği uzmanlarından" yararlanılır. Bazı yazarlar içerik geçerliliği yapılmadan yapısal geçerlilik ve güvenilirlik analizlerinin yapılmasının bir anlam ifade etmeyeceğini, çünkü öncelikle içeriğin belirli bir kavramsal yapı modeline[j] uygun olduğunun kanıtlanması gerektiğini belirtmişlerdir.[37] Anastası'ye (1988) göre içerik geçerliliği aynı zamanda yapısal geçerliliğe ilişkin kanıt sağlar (aktaran, Haynes).[38] İçerik geçerliliği, bilgi ve başarı testlerinde nispeten daha kolay yapılırken, soyut anlamlardan oluşan psikolojik kavramsal yapılarda daha zordur. Çünkü psikolojik kavramsal yapının sınırlarını çizmek üst düzeyde uzmanlık bilgisi gerektirir.

İçerik Geçerliliğinin Aşamaları

İçerik geçerliliğinde başlıca beş aşama vardır: (a) kavramsal yapı veya test evrenin tanımlanması, (b) kavramsal yapıya ait boyutların ortaya çıkarılması, (c) ölçek veya test maddelerinin oluşturulması, (ç) ölçeğin hakemlere değerletilmesi. (d) matematiksel analizlerin yapılması. Bu beş aşama birbirini ardışık bir biçimde izlemez. Araştırılan kavramsal yapının niteliğine göre bazen önce faktörler bazen de, kavramsal yapıyı oluşturan faktörler bilinmediğinden ölçek maddelerinin geliştirilmesi yoluna başvurulabilir.

Kavramsal yapı ve test evrenin tanımlanması. İçerik geçerliliği kavramsal yapının tanımlanmasıyla başlar. Kavramsal yapı soyut bir nitelendirme olduğundan ilgili pek çok diğer kavramsal yapılarla ilişki içindedir. Kavramsal yapı tanımlanırken nüans farkları ortaya konur, kavramın ortaya çıkış tarihi incelenir ve söz konusu kavramsal yapıya ilişkin yapılan önceki araştırmalardan söz edilir. Kavramsal yapı aynı zamanda test evrenini belirler. Okullarda uygulanan bilgiyi ölçmeye yönelik testlerde kavramsal yapı, belirli bir dersle ilgili olarak yaş düzeyi, sınıf ve müfredat konularına ilişkin kapsamın belirlenmesidir.

---

[i] Literatürde bazı yazarlar *uzman*, başka yazarlar da *hakem* sözcüklerini kullanırlar. Bu kişilerin en önemli özelliği konuyu bilmeleri ve daha objektif bir değerlendirme yapabilmeleridir.

[j] Kavramsal yapıların *içerik analizi* yöntemi kullanılarak çıkarılmasıyla *istatistikî yöntem* kullanılarak çıkarılması literatürde bilim adamları arasında tartışmalı olan bir konudur. Deney yönelimli bilim adamları istatistikî analize dayanmayan içerik analizi yöntemlerinin bilimsel bir yaklaşım olmadığı görüşündedirler.

Konunun kavramsal boyutlarının ortaya çıkarılması. Araştırmacı içerik geçerliliği için kapsamlı bir literatür taraması yapmalı ve konunun önceki yıllarda yapılmış araştırmaların istatistiksel analizleriyle ortaya çıkmış temel boyutlarını belirlemelidir. Söz konusu kavramsal boyutlar (faktörler) kuramsal bir temele sahiptir. Araştırmacı veya bilim adamı yaptığı literatür taraması sonucunda incelediği kavramsal yapıyla ilgili olarak araştırma temelli bir bulguya rastlayamamışsa, elde ettiği bulgular kendi araştırdığı kavramsal yapıyla dolaylı olarak ilgiliyse literatürdeki bilgilere, gözlemlerine, odak grubu araştırmalarına, panel toplantılarına, deneyimlerine ve yaptığı görüşmelere bağlı olarak geçici nitelikte belirli sayıda faktör belirler.

Kavramsal boyutlar önceki araştırma sonuçlarına dayandırılmış ve bu çerçevede örneğin, belirli bir kavramsal yapıyla ilgili olarak beş boyut belirlenmiş olabilir. Araştırmacı bu aşamada daha önceden belirlenen söz konusu beş boyuta uygun olarak yeteri kadar ölçek maddesi geliştirmeye çalışır. Bu uygulamaya *çıkarsama*[k] *yöntemi* denir (Hinkin 1995, aktaran Young).[39] Maddelerin geliştirilmesinde genelden özele doğru gidilir. Ancak bu boyutların yeterli olup olmadığı konusunda araştırmacı kesin bir şekilde emin olamaz.

İncelenen kavramsal yapıyla ilgili olarak çok az araştırma yapılmış ve bu alandaki kuramsal bilgiler yetersizse ölçek maddelerinin geliştirilmesinde gözlem verilerinden yararlanarak tümleme (tüme varım) yaklaşımı benimsenir.[40] Değişik maddeler bir araya getirilerek boyutlar oluşturulur. Araştırılan kavramsal yapının temel boyutlarını (faktörleri) belirleme işini araştırmacı kendi üstlenir. Bilim adamı, daha çok niteliksel araştırmalardan hareket ederek kavramsal yapının iskeletini oluşturmaya çalışır. Bunun için panel çalışması yapar, ilgili kişilerle görüşerek onların düşüncelerini derler ve kendi kişisel gözlemlerini dikkate alır.

Kavramsal yapının temel boyutlarını tümleme yöntemiyle ortaya çıkarmak ve belirli bir mutabakat oluşturmak için *kavram haritası* (concept mapping) tekniğinden yararlanılabilir.[41] Kavram haritası, Trochim (1989) tarafından geliştirilmiş, planlı bir şekilde kavramlaştırma ve istatistiksel model oluşturma sürecidir. Bu yöntemde uzmanlardan oluşan bir panel toplantısı yapılır ve katılımcılara kavramsal yapı ile ilgili çok sayıda gö-

[k] Inductive (çıkarsama – tümden gelim): Genelden özele gitme; bilinenlerden hareket ederek bilinmeyenleri ortaya çıkarma.

rüş/ifade geliştirmeleri ve daha sonra bu görüşleri kavramsal yapıyı ne ölçüde temsil ettiklerini belirlemek üzere beş dereceli bir ölçek üzerinde derecelendirmeleri istenir. Süreçte katılımcılar beyin fırtınası tekniğini uygularlar. Kavram haritası bu puanlara göre değerlendirilir. Analiz için hiyerarşik *kümeleme analizi* ile birlikte *iki boyutlu çok yönlü ölçekleme tekniği* kullanılır. Analiz sonucunda ortaya çıkan haritada ifadelerin her biri iki boyutlu bir uzayda konumlanmış olur. Birbirine yakın ifadeler bir araya gelerek bir grup oluşturduğundan panelistlerle yapılacak ikinci bir toplantıda bu kümeleri yorumlamaları ve anlamlı bir isim bulmaları istenir. Panele katılacak kişi sayısı konusunda kesin bir rakam vermek güçtür, fakat bu sayı 5 ilâ 20 arasında değişebilir. İkinci aşamada boyutların isimlendirilmesine katılacak panelistler konunun uzmanları arasından seçilir.

Ölçek maddelerinin geliştirilmesinde çıkarsama yöntemi uygulanmışsa eldeki ifadeler hakemler grubuna değerletilirken literatürdeki faktör sayısından bir fazla grupta değerlendirmeleri istenir. Böylece sonuncu guruba kavramsal yapıyla ilgili olmayan ifadeler alınır. Bu konuda başvurulabilecek bir diğer yöntem, hakemlere literatürdeki boyut sayısı bildirilmeden bu sayıdan bağımsız olarak değerlendirme veya gruplama yaptırmaktır. Hakemler ifadeleri gruplandırdıklarında grup sayısı (faktör) literatürdeki boyut sayısıyla bir paralellik gösteriyorsa ölçek maddeleri geçerlidir denilir. Literatürdeki boyutlarla uyuşmayan maddeler ise ölçekten çıkarılır.[42] Ölçek maddelerinin geliştirilmesinde tümleme yöntemi uygulanmışsa, yine çıkarsama yöntemindeki aşamalardan geçilir.

Türkçe, matematik, sosyal bilgiler gibi bilgi testlerinde ise geliştirilen testin kavramsal boyutları tam olarak temsil etmesi için kaç üniteyi ve her bir ünite içinde kaç konuyu kapsadığı incelenir. Araştırmacı bunun için "test özellikleri" ve "test planı" tablolarını oluşturur. Test özellikleri tablosunda işlenen üniteler, ünitelerin ağırlığı, sorulacak soruların niteliği (Bloom veya diğer bilim adamlarının sınıflandırma sistemi temel alınabilir) ve ağırlıkları saptanır (*bk.*, Tablo 15-2).

Tablo 15-2. Test Özellikleri Tablosu

Üniteler	Bloom'un Taksonomisi			
	Bilgi ve kavrama	Uygulama	Analiz, sentez ve değerlendirme	Toplam
A Ünitesi	%10	%20	%15	%45
B Ünitesi	%15	%20	%20	%55
Toplam	%25	%40	%35	%100

Test özellikleri tablosu, kişilere ne öğretildiği veya neyi bilmeleri gerektiği konusunu temel alır. Test planı tablosu ise test özellikleri tablosuna bağlı olarak sorulacak soru sayısını belirler (*bk.*, Tablo 15-3).

Tablo 15-3. Test Planı Tablosu

Üniteler	Bloom'un Taksonomisi			
	Bilgi ve kavrama	Uygulama	Analiz, sentez ve değerlendirme	Toplam
A ünitesi	1, 2	6, 7, 8, 9	14, 15, 16	9 (%45)
B Ünitesi	3, 4, 5	10, 11, 12, 13	17, 18, 19, 20	11 (%55)
Toplam	5 (%25)	8 (%40)	7 (%35)	20 (%100)

Testin içerik geçerliliğine sahip olması için, kapsadığı tüm üniteleri ve konulara ait soruları yeterli ölçüde temsil etmesi gerekir. Örneğin, sayılar ünitesinden çarpma konusuyla hiç soru sorulmaması testin geçerliliğine gölge düşürür. Bilgi testlerinde konunun değil, bir ünitenin iyi öğrenilip öğrenilmediğine karar verebilmek için testin o üniteyle ilgili olarak en az altı soru içermesi gerekir. Ayrıca test sorularının zorluk derecesi sınıf düzeyine uygun olmalıdır. Bir üst sınıf öğrencilerinin çözebileceği zorluktaki bir sorunun alt sınıf öğrencilerine sorulması test maddeleri açısından uygun olmayan örneklem konusunu gündeme getirir.

Personel seçim testlerinde herhangi bir özelliği ölçecek test için yeteri kadar iş örneği (uyaran) ve davranış (tepki) biçimi belirlenir. İşin içerdiği özelliklerle testin veya test bataryasının içerdiği özellikler benzer olmalıdır.

Başarı değerlendirme formlarında ise maksimum ve normal başarı standartları belirlenir. Başarı değerlendirme formu başarı faktörlerini veya değişkenlerini yeterli ölçüde kapsamalıdır. Ölçek sadece normal başarı de-

ğişkenlerinden oluşturulmuşsa üst düzeyde başarı gösteren kişiler söz konusu ölçekle doğru bir şekilde ölçülemez.

Maddelerin belirlenen boyutlara göre oluşturulması. İçerik geçerliliği için ölçekteki/testteki ifadeler veya birimler kavramsal yapıyı (veya faktörü) temsil edecek sayıda ve yeterlilikte olmalıdır. Gereğinden fazla maddenin okuyucuları sıkarak yorgunluğa neden olduğu bildirilmiştir. Maddelerin uzun ve çok sayıda olması halinde *yanıt verme yanlılığı*[1] ortaya çıkacağından sonuçta testin/ölçeğin yüzeysel ve yapısal geçerliliği yara alır. Teste bir boyutla ilgili, muhtemel tüm ifadeler değil, o boyutu temsil edecek sayıda ifade alınmalı, boyutla birinci derecede ilgili olmayan ifadeler dışarıda bırakılmalıdır. Geliştirme aşamasında, nihaî ölçekte yer alması düşünülen madde sayısının üç katı kadar ifade geliştirilir.[43] Nihaî ölçekte 20 kadar madde olması öngörülüyorsa başlangıç aşamasında 60 civarında madde tespit edilir. Madde sayısı belirlenirken faktör sayısı da göz önünde bulundurulur. "Teyit edici faktör analizinde, bir faktör altındaki değişken sayısı için herhangi bir sınır değer söz konusu değildir. Fakat keşfedici faktör analizinde Thurstone, her bir faktör için en az üç değişken/madde bulunması gerektiğini belirtmiştir."[44] Bu görüş aynı zamanda Cook, Hepworth, Wall ve Warr (1981) tarafından da desteklenmiştir (aktaran Young).[45] Bir ölçekte az sayıda madde varsa bu ölçeklerin içerik ve yapısal geçerlilikleri düşüktür (Kenny, 1979; Nunnally, 1976, aktaran Young).[46] Kavramsal bir yapının sadece tek bir madde ile ölçülmeye çalışılması geçerlilik konusunu gündeme getirir. Öte yandan bir faktör altındaki madde sayısının dört veya beş olması dahi yeterli olmayabilir. Geçerlilik açısından yeterli görülse bile, bu maddelerin Cronbach alfa güvenilirlik katsayılarının ,70'in üzerinde olup olmadığına bakmak gerekir. Yapılan araştırmalar, normal olarak bir faktöre ait sekiz ve üzerindeki madde sayısıyla güvenilirlikte ,70 kriterinin rahatlıkla sağlanabildiğini göstermiştir.

Bilgi ve başarı testlerinde ise belirli bir boyut veya faktöre genelleme yapılacaksa ilgili madde sayısı en az altı olmalıdır. Burada ölçümün amacı önemlidir. Ölçümün amacı belirli bir konunun öğrenilme durumunu saptamak ise, maddeler o konunun içerdiği ünitelerden ünitelerin ağırlık ve önemine göre serbestçe seçilir. Bazı ünitelerden bir madde seçilirken önemli olarak değerlendirilen ünitelerden üç veya dört madde alınabilir. Bu

[1] Response pattern bias (yanıt yanlılığı): Cevaplayıcıların değişik faktörler nedeniyle doğru yanıt vermemeleri ve sonuçta verilerin çarpık olmasıdır.

anlamda evrenden madde seçmek *sübjektif değerlendirmelere* dayanır. Hangi ünitenin önemli olduğu ve o üniteden kaç soru seçilmesi gerektiği bilim adamının kişisel değerlendirmelerine ve konunun müfredattaki önemine bakılarak belirlenir.

Tutum ölçeklerindeki ifade sayıları da *apriori* olarak boyutların önemine orantılı bir biçimde belirlenir. Gerçek madde sayısı daha sonra ampirik olarak belirlenecektir. Kavramsal yapıyı temsil etme konusunda daha önemli ve çok işlenmiş boyutlar daha fazla madde ile ve varyansı düşük olan boyutlar ise daha az madde ile temsil edilebilir. Seçilen ifadelerin (örneklem) ölçülmek istenen yapıyı temsil edip etmediği büyük ölçüde tartışmalı olan bir konudur. Bu ifadelerin temsil kabiliyetini artırmak için araştırma literatürü dışında; kişilerin yazılarından, kişilerle yapılan röportajlardan, mülakatlardan ve onlara yöneltilen açık uçlu sorulardan yararlanılır.

Araştırmacı, raporunda içerik geçerliliği ile ilgili olarak bilgi verirken her bir boyutu hangi aşamalardan geçerek belirlediğini, bu boyutları literatürde hangi araştırmacıların bulduğunu veya önerdiğini, her bir boyutla ilgili kaç tane ifade geliştirdiğini ve her bir ifadede o boyuta ait hangi farklı özelliği vurguladığını açıklamalıdır. Ayrıca aynı veya benzer bir konuyu ölçen diğer ölçekler hakkında da bilgi vermelidir.

Geliştirilen ölçeğin uzmanlara değerletilmesi. İçerik geçerliliğinin üçüncü ögesi, test maddelerinin uzmanlara değerletilmesidir. Uzman değerlendirmesi, kavramsal yapıya ilişkin temel faktörleri ortaya çıkarmaya veya geliştirilen maddelerin belirli bir kavramsal veya faktöriyel yapıya uygun olup olmadığını belirlemeye yöneliktir.

Uzman değerlendirmesinde değişik yöntemlerden yararlanılır. Bunların başlıcaları aşağıdaki gibidir:

1. Odak grubu çalışması yapılması.
2. Kavram haritası yöntemi.
3. Hambleton yöntemi.

Odak grubu araştırmasında konunun uzmanları bir araya getirilerek kendilerine ölçülmek istenen kavramsal yapının temel boyutlarını belirlemeye yönelik sorular sorulur. Kavram haritası yönteminde ilgili taraflar bir araya getirilerek kendilerine önceden hazırlanan ve "kavram haritası" adı verilen bir şema sunulur. Bu şemayı değerlendirmeleri ve kavramları önem

derecesine göre sıralamaları istenir. Hambleton ve arkadaşları tarafından geliştirilen, bu nedenle de kısaca *Hambleton yöntemi* olarak adlandırılan yaklaşımda ise dört aşama vardır. Birinci aşamada ölçek veya testin içeriğine vâkıf veya konuyu iyi bilen uzmanlar belirlenir. İkinci aşamada bu uzmanlara, araştırma alanının ve incelenen kavramsal yapının tanımları bir mektupla anlatılır ve bu çerçevede geliştirilen anket formu yollanır. Üçüncü aşamada her bir uzman birbirinden bağımsız olarak geliştirilen anket formunu üç veya dört dereceli bir ölçek üzerinde değerlendirir. Dördüncü aşamada içerik geçerliliği için uzmanların verdikleri puanların ortalama değerleri temel alınır.[47]

İçerik değerlendirmesini yapacak olan hakemler veya uzmanlar dikkatli bir şekilde belirlenmelidir. Hakemler sadece aynı ana bilim dalındaki öğretim üyelerinden veya konuya sıcak bakmayan kişilerden, konuya uzak olan kişilerden veya araştırma görevlileri gibi deneyimi yetersiz kişilerden oluşmamalıdır. Çünkü hakemlerin yanlılığı, kültürel özellikleri ve uzmanlık düzeyleri değerlendirmeyi etkileyebilir. İçeriği değerlendirecek kişiler ölçeğe/teste değişik açılardan katkı yapma imkanına sahip olmalıdırlar. Araştırmacı, raporunda içerik geçerliliğinin bu aşamasında kaç hakemden yararlandığını, hakemleri nasıl belirlediğini, hakemlerin özelliklerini, kaç yıllık bir deneyime sahip olduklarını, uzmanlık alanlarını ve değerlendirmeyi nasıl yaptığını özet olarak açıklamalıdır. Değerlendirmenin kaç dakika/saat sürdüğü, hakemlerin ne gibi öneriler getirdikleri hakkında bilgi verilmelidir. Ayrıca ölçeğin genel olarak düzenleme biçimi, soru sayısının yeterliliği, ölçek dereceleri, ifadelerin anlaşılırlığı ve örnek kütleye uygunluğu hakkında da hakemlerin görüşleri alınır. Hakem değerlendirme çalışmaları pilot araştırması öncesinde yapılır. Hakem sayısı konusunda kesin bir rakam verilemez, ancak bu sayının en az üç olması gerektiği belirtilmiş beş kişiden oluşan bir hakem grubunun ise ideal olacağı söylenmiştir.[48] Değerlendirmede yararlanılacak hakem sayısı ölçeğin kullanım amacına, insanlar üzerinde yaratacağı etkiye ve bilimsel çalışmanın hassasiyet derecesine göre değişir.

İstatistiksel ve matematiksel analizlerin yapılması. Hakemlerin yaptıkları değerlendirmelere ait sonuçların kantitatif analizinde hangi yöntemin uyuşmayı daha iyi ölçtüğü konusunda bilim adamları arasında tam bir mutabakat yoktur. Fakat literatürde çoğunlukla iki yöntemden biri uygulanır; birincisi Lawshe'nin İçerik Geçerliliği Oranı ve ikincisi ise Cohen kappa formülüdür.

Lawshe'nin İçerik Geçerliliği Oranı. Lawshe'nin İçerik Geçerliliği Oranı formülünde hakemlerin her bir ifadeyi nasıl değerledikleri dikkate alınır. Lawshe katsayısının yüksekliği veya düşüklüğü, hakemlerin her bir ifadeye verdikleri *uygun* cevabının sayısına göre belli olur. İçerik Geçerliliği Oranı formülü Eşitlik 15-1'deki gibidir:

$$İGO_i = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}. \qquad (15\text{-}1)$$

$İGO_i$ = Ölçeğin *i*'ninci maddesinin İçerik Geçerliliği Oranı.
n_e = İfadenin *Uygun* olduğunu belirten hakemlerin sayısı.
N = Hakemlerin toplam sayısı.

İçerik Geçerliliği Oranı formülünün uygulanmasıyla birlikte her bir ölçek maddesi için bir yüzde değeri elde edilir ve bu katsayı –1 ilâ +1 arasında değişir. Panele katılan hakemlerin yarısından daha azı bir madde için "uygun" işaretlemesi yapmışsa sonuç eksi çıkar. Eksi işaretli maddeler ölçekten çıkarılır. Hakemlerin tamamı bir ifadenin gerekli olduğunu bildirmişlerse sonuç 1,00 çıkacaktır. Öte yandan panele katılan uzmanların %50'si bir madde için "uygun" işaretlemesi yapmışlarsa İçerik Geçerliliği Oranı sıfır çıkar. Bir maddenin içerik geçerliliği için hakemlerin yüzde 50'sinden fazlasının *uygun* veya *gerekli* şıkkını işaretlemeleri gerekir. Hakemlerin yarısından fazlası bir maddenin o ölçek için "gerekli" olduğunu işaretlemeleri halinde söz konusu madde bir ölçüde içerik geçerliliğine sahiptir. Lawshe (1975) farklı panel büyüklükleri için $p = ,05$ (tek yönlü) güven aralığında asgarî içerik geçerliliği oranlarını bir tablo halinde belirlemiştir. Örneğin, panelde 25 uzman varsa ve bazı ifadelerin *İGO* değeri ,37'den küçükse bu maddeler ölçekten çıkarılır (*bk.*, Tablo 15-4) (aktaran Vickery).[49] İçerik geçerliliği hesaplamasında hakemler arasındaki uyuşmanın tesadüfe bağlı olup olmadığı belli olmaz. Murphy ve Davidshofer'e göre *Lawshe'nin İçerik Geçerliliği Oranı*, değerlendiriciler arasındaki uyuşmayı gösterir. Bu yöntem tam bir içerik geçerliliği analizi olarak değerlendirilemez (aktaran Miles).[50]

Tablo 15-4. Lawshe Minimum İçerik Geçerliliği Oranları

Panelist sayısı	Minimum Değer
5	,99
6	,99
7	,99
8	,78
9	,75
10	,62
11	,59
12	,56
13	,54
14	,51
15	,49
20	,42
25	,37
30	,33
35	,31
40	,29

■ Örnek uygulama.

Bir araştırmada 10 maddeden oluşan bir ölçek geliştirilmiş ve bu ölçek 12 uzmana ifadelerin belirlenen kavramsal yapıya uygunluğu açısından değerletilmiştir. Her bir maddeye *uygun*, *kalabilir* ve *uygun değil* şeklinde cevap veren uzmanların işaretleme sıklıkları Tablo 15-5'te çizelge haline getirilmiştir.

Tablo 15-5'teki değerlerden hareket edilerek daha sonra ikinci bir tablo hazırlanır. Bu ikinci tablo ile *Lawshe'nin İçerik Geçerliliği Oranı*'nı kolay bir şekilde hesaplamak mümkündür. Bu tabloda n_e bir ifadenin *uygun* ve *kalabilir* olduğunu belirten hakemlerin sayısını ve *N*/2 ise toplam hakem sayısının yarısını gösterir (*bk.*, Tablo 15-6). Yapılan hesaplama sonucunda elde edilen katsayı içerik geçerliliği tablosu ile karşılaştırılır. Hesaplanan değer bu tablodaki asgari değerlerden yüksekse uzmanlar arasında uyuşma olduğuna karar verilir. Hesaplanan değer eksi çıkmışsa veya içerik geçerliği tablosundaki değerden düşükse uyuşma olmadığı şeklinde yorumlanır ve bu ifadeler ölçekten çıkarılır.

Tablo 15-5. Uzmanların Değerlendirme Sonuçları

İfadeler	Uygun	Kalabilir	Uygun değil	Toplam uzman sayısı
Madde 1	10	2	0	12
Madde 2	2	10	0	12
Madde 3	0	2	10	12
Madde 4	8	2	2	12
Madde 5	8	3	1	12
Madde 6	7	4	1	12
Madde 7	8	2	2	12
Madde 8	8	2	2	12
Madde 9	6	3	3	12
Madde 10	4	4	4	12

Tablo 15-6. İçerik Geçerliliği Oranı Hesaplama Tablosu

İfadeler	n_e	$N/2$	IGO $\frac{n_e - \frac{N}{2}}{\frac{N}{2}}$	$İGO$	Karar
Madde 1	12	6	6/6	+ 1,00	Kabul
Madde 2	12	6	6/6	+ 1,00	Kabul
Madde 3	2	6	–4/6	– ,66	Ret
Madde 4	10	6	4/6	+ ,66	Kabul
Madde 5	11	6	5/6	+ ,83	Kabul
Madde 6	11	6	5/6	+ ,83	Kabul
Madde 7	10	6	4/6	+ ,66	Kabul
Madde 8	10	6	4/6	+ ,66	Kabul
Madde 9	9	6	3/6	+ ,50	Ret
Madde 10	8	6	2/6	+ ,33	Ret

Araştırmacı uyuşma olmadığına karar verdiği maddelerle ilgili olarak bir değerlendirme yapmalı bu maddelerin ifadelendirmesinde ne gibi sorunlar olduğunu tespit etmeye çalışmalıdır.

Cohen Kappa. Kohen kappa formülü, hakemlerin tesadüfen benzer kararlar verebilecekleri olgusunu da dikkate alarak yapılan ve yüksek katsayılarda bu etkiyi ortadan kaldıran bir hesaplama yöntemidir. Kohen kappa formülü gözlemci içi ve gözlemciler arası değerlendirme yöntemlerinin her ikisinde de kullanılabilir. Bu hesaplama yönteminde, frekans kategorilerinin birlikte gerçekleşme olasılığı hesaplanır. Kappa hesaplaması doğrudan testin geçerliliği ile ilgili değildir, hakemlerin test maddelerinin uygun olduğu konusundaki görüşlerin ne ölçüde tutarlı olduğunu belirler. Kohen kappa formülünün uygulanabilmesi için hakemler birbirlerinden bağımsız hareket etmelidirler. Hakemlerin bir araya gelerek birlikte değerlendirme yapmaları doğru değildir. Hakemler arasındaki uyuşmayı gösteren Cohen kappa formülü iki şekilde hesaplanabilir (*bk.*, Tablo 15-7):

1. İki hakem olması halinde.
2. İkiden fazla hakem olması halinde.

İki hakem olması halinde, bu hakemler Örneğin, 200 maddeden oluşan ölçeğe ait ifadeleri ölçülmek istenen kavramsal yapıyı dikkate alarak *Uygundur, Kalabilir* ve *Uygun Değil* şıklarına göre ayrı ayrı işaretlerler. İşaretlemede araştırmacı isterse iki dereceli bir ölçekten de yararlanabilir. Böyle bir durumda ölçek dereceleri, *Uygun* ve *Uygun Değil* etiketlerine göre belirlenir. Kappa formülü "gözlem değerleri" veya "gözlem oranları" dikkate alınarak iki farklı şekilde hesaplanır. Gözlem değerlerinin dikkate alınması halinde Eşitlik 15-2'deki formül kullanılır.[51]

$$K = \frac{gözlemlenen\ uyusma\ sayisi - beklenen\ uyusma\ sayisi}{toplam\ gözlem\ sayisi - beklenen\ uyusma\ sayisi} . \qquad (15\text{-}2)$$

Beklenen uyuşma sayısı, [Σ (sütun toplamı x satır toplamı / *N*)] formülüyle bulunur. Gözlem oranlarının hesaplanması halinde ise, kappa formülü Eşitlik 15-3'teki gibi uygulanır:[52]

$$K = \frac{gözlemlenen\ uyusma\ orani - beklenen\ uyusma\ orani}{1 - beklenen\ uyusma\ orani} . \qquad (15\text{-}3)$$

İkinci formüldeki "beklenen uyuşma oranı", her bir sütun toplamıyla her bir satır toplamının çarpılması ve genel değerlendirme sayısına bölüne-

rek çıkan rakamların toplanması ve bu genel toplamın tekrar genel değerlendirme sayısına bölünmesi suretiyle bulunur (*bk.*, Tablo 15-7). "Gözlemlenen uyuşma oranı" ise, birinci hakemle ikinci hakemin uyuştukları rakamları gösteren ve tablonun köşegeninde yer alan rakamların toplanması ve genel değerlendirme sayısına bölünmesiyle elde edilir.

Tablo 15-7. Gözlemlenen ve Beklenen Uyuşma Oranları Tablosu

İkinci hakem		Birinci Hakem			
		Uygun	*Kalabilir*	*Uygun değil*	*Toplam*
	Uygun	11	2	4	17
	Kalabilir	1	14	2	17
	Uygun değil	4	2	10	16
	Toplam	16	18	16	50

Gözlemlenen uyuşma oranı = Köşegendeki değerler/*N* = (11+14+10)/50 = ,70.

Beklenen uyuşma oranı (BUO) = Σ (satır toplamı x sütun toplamı / *N*) / *N*.

BUO = (16*17/50 + 18*17/50 + 16*16/50)/50 = (5,44 + 6,12 + 5,12)/50 = ,33.

Kappa = (Gözlemlenen uyuşma oranı – Beklenen uyuşma oranı) / (1 – Tesadüf eseri ortaya çıkabilecek beklenen uyuşma oranı) = (,70 – ,33) / (1 – ,33) = ,55.

Kohen Kappa formülünün uygulanmasıyla elde edilen katsayı –1 ilâ +1 arasında değişir. Değerin +1.00 çıkması hakemler arasında tam bir mutabakat olduğunu gösterir. Değerin sıfır çıkması hakemler arasındaki mutabakatın olmadığını, –1 çıkması değerlendiricilerin birbirlerinin tam tersine bir işaretleme yaptıklarını gösterir. Crocker ve Algina'ya göre (1986) $\kappa = 0$ çıkması hakemlerin *uygun, uygun değil* kararlarının çok tutarsız olduğunu göstermez, tersine şans faktörünün tam olarak giderilemediğini, gözlenen uyuşma değerinin şans faktörünün üzerinde olmadığı şeklinde yorumlanır. Öte yandan negatif kappa değeri, uyuşma değerinin şans faktörünün altında olduğu anlamına gelir. "Diğer bir deyişle bir hakem *Evet* cevabını vermişken diğer hakemin *Hayır* cevabını vermiştir. Kappa katsayısı ,40'tan

küçükse hakemler arasındaki uyuşma zayıf; ,40 ilâ ,70 arasında ise orta derecede bir uyuşma ve ,70'ten büyükse güçlü bir uyuşma ver demektir.[m] Bilim adamları Kappa katsayısını yorumlarken belirli bir kesim noktasını temel alma eğilimindedirler ve bu kesim noktası çoğunlukla ,60 olarak belirlenmiştir. Şans faktöründen arındırılmış ,60'ın üzerindeki hakemler arasındaki uyuşma iyi bir değerdir.[n]

Tablo 15-8. Değerlendiriciler Arasındaki Uyuşma

	İki hakem	İkiden fazla hakem
Nominal veriler	• Uyuşma yüzdesi. • Kappa (Cohen, 1960).	• Uyuşma yüzdesi. • İkiden fazla hakem için Kappa.
Sıralı ölçek verileri	• Spearman sıra korelasyonu (*rho*). • Kendall uyuşma katsayısı.	• Kendall *W*. • Uyuşma katsayısı.
Eşit aralıklı ölçek verileri	• Pearson korelasyon katsayısı.	• Sınıflar arası korelasyon (Pearson *r*).

Değerlendiriciler arası mutabakatın belirlenmesi için Kappa istatistiğinin kullanılması yöntemine literatürde eleştiriler geliştirilmiştir. Bazı bilim adamlarına göre, (a) Kappa istatistiği uyuşmayı doğru bir şekilde ölçen tartışmasız bir teknik olarak görülmemelidir. (b) Çok fazla tartışma kaynağı olan böyle bir istatistiğin kullanılması konusunda ihtiyatlı olunmalıdır. (c) Bilinçli bir karar vermek için aynı zamanda diğer tekniklerin de kullanılması düşünülmelidir.

Kappa gerçekte şans faktörünü tam olarak ortadan kaldırmaz. Çünkü değerlendiriciler bağımsız değildirler, onlar aynı ölçeği değerlendirmektedirler. Kappa herkesin aklına ilk gelen genel amaçlı bir uyuşma indeksidir. Hakemlerin uyuşmazlık kaynakları hakkında önemli ölçüde bilgi verme-

[m] SPSS'te kappa hesaplaması, iki değerlendiriciye ait uyuşma ve uyuşmama verileri için yapılabilir. (Analyze, Descriptive Statistics, Crosstabs, Statistics, Kappa). S. Siegel and J. N. Castellan'ın *Nonparametric Statistics* isimli kitabında belirttikleri ikiden fazla değerlendirici vakasında SPSS programı için yazılmış özel makrolar da kullanılabilir.

[n] Bazı bilim adamları ise kesim noktasını ,70 olarak belirlemişlerdir. Buna göre ,60 veya ,70 değerinin kabul edilmesi araştırmanın doğuracağı sonuçlar dikkate alınarak tercih edilmelidir.

mektedir. Kappa, hakemlerin aynı ölçek derecelerini kullanmalarını gerektirir. Bir değerlendirici 1-3 dereceli bir ölçek kullanırken diğeri 1-5 dereceli bir ölçek kullanamaz. Kappa yöntemi, iki dereceli (evet, hayır) nominal ölçeklerde ve sıralanmış ölçeklerde kullanılabilir. Kappa eşit aralıklı ölçekler için uygun değildir.

Kappa ve SPSS. İstatistik paket programı SPSS, kappa formülünü sadece iki hakem için hesaplayabilmektedir. İkiden fazla hakem için hesaplama yapmak üzere İnternet ortamında SPSS'te kullanılmak üzere yazılmış makrolar bulunmaktadır fakat bu makroların programa tanıtılması oldukça zordur. Bunun yerine hesaplamaların ikili olarak yapılıp sonuçların ortalaması alınabilir. İkili hesaplama ortalamalarının çoklu hesaplama yöntemiyle aynı sonucu verdiği bildirilmiştir.[53] İki hakemin değerlendirme sonuçları arasındaki uyuşmayı görmek için İstatistiksel analiz programı SPSS'teki hesaplama iki şekilde yapılabilir. Birinci yöntemde veriler ağırlıklandırma yapılmadan hesaplatılır. İkinci yöntemde ise ağırlıklandırma yöntemine baş vurulur. Ağırlıklandırma yapılmadığı durumda ölçeğin madde numaraları vak'a olarak ve hakemlerin 1 = *Uygun*, 2 = *Kalabilir* ve 3 = *Uygun Değil* şeklinde verdikleri puanlar ise programa değişken olarak tanıtılır. Daha sonra SPSS'te Analysis mönüsü altında Descriptive düğmesinden Crosstabs bölümüne girilir ve buradaki kappa kutusu seçili hale getirilerek hesaplama yaptırılır. Cells mönüsü altında da Gözlenen (Observed) ve Beklenen (Expected) kutuları ile Sıra Temelli Yüzde dağılımı (row percentage) kutusu seçili hale getirilir.

Ağırlıklandırma yönteminde ise *uygun, kalabilir* ve *uygun değil* şıklarının ikili olarak toplam frekansları önceden tespit edilir. Kappa değerini hesaplamak için bilgisayara sadece Tablo 15-9'daki değerler girilir. Hesaplama yapmadan önce ağırlıklandırma yapmak gerekir. Bunun için Data mönüsü altında Weight Cases düğmesi tıklanarak *Toplam Frekans* ağırlıklandırma değişkeni olarak tanıtılır. Daha sonra birinci yöntemde açıklanan prosedürler uygulanır.

Tablo 15-9. Kappa değerinin Ağırlıklandırma Yöntemiyle Hesaplanması

Birinci hakem	İkinci hakem	Toplam frekans
1 (uygun)[a]	1 (uygun)	88
1 (uygun)	2 (kalabilir)	10
1 (uygun)	3 (uygun değil)	2
2 (kalabilir)	1 (uygun)	14
2 (kalabilir)	2 (kalabilir)	40
2 (kalabilir)	3 (uygun değil)	6
3 (uygun değil)	1 (uygun)	18
3 (uygun değil)	2 (kalabilir)	10
3 (uygun değil)	3 (uygun değil)	12
Toplam		200 ifade

Not. Tabloda parantez içindeki ifadeler gösterilmez.

Hesaplama sonucu şu şekilde yorumlanır. "Şans faktörünü içermeksizin hakemler arasındaki uyuşma %48 oranındadır (N = 200; Kappa = ,481; $p < ,005$).

İçerik Geçerliliğinin Güçlü ve Zayıf Yönleri

İçerik geçerliliği, ölçülmek istenen kavramsal yapının temel boyutlarının ortaya konulması ve ölçüm alanının makul bir oranda kapsanması açısından önemlidir. Ancak içerik geçerliliği kavramsal yapının iskeletini ampirik olarak ortaya koymaz. Gözlemci ve hakem değerlendirmelerine dayalı bir ön değerlendirme niteliğindedir. Bir ölçüm aracının geçerliliğini sadece içerik geçerliliğine dayandırmak zayıf bir yapılanma ortaya koyar.

İçerik geçerliliğinin zayıf yönü, kavramsal alanın sınırlarının net ve kesin bir şekilde belirlenmesi konusunda yaşanan zorluktur. Araştırmacı hiçbir zaman kavramsal alanının gerçek sınırlarına ulaştığını tam olarak bilemez, sadece tahmin eder. Literatürdeki veriler de bazen yeterince bilgi vermeyebilir. Bir ikinci zorluk, geliştirilen maddelerin kavramsal alanın veya boyutların içeriğine ne ölçüde uygun olduğudur. Ampirik olarak deneme yapmadan bu maddelerin kavramsal yapıyı iyi temsil ettiğini söyleyemeyiz. Bu nedenle içerik geçerliliğinin yapısal geçerlilikle desteklenmesine ihtiyaç vardır.

KRİTER GEÇERLİLİĞİ

Kriter geçerliliği, geliştirilen test veya ölçek ile elde edilen sonuçların standart olarak tespit edilen bir ölçüm kriterine ait puanlarla karşılaştırılması ve bu karşılaştırma sonucunda elde edilen korelasyon katsayısının yüksek çıkmasıdır. Kriter geçerliliği, geliştirilen test sonuçlarını yorumlama amacıyla değil, ileriye yönelik tahmin yapma amacıyla kullanılır. Kriter geçerliliği, içerik geçerliliğinden daha duyarlı ve daha somut sonuçlar verir. Kriter geçerliliğinde, geliştirilen ölçek veya test sonuçlarıyla karşılaştırma yapmak için pratik hayattan daha önceden geçerlilik ve güvenilirlik analizi yapılmış standart bir ölçek veya birden fazla ölçüt temel alınır. Geliştirilen ölçeğin uygulama sonuçları kriter olarak kabul edilen ölçekten elde edilen değerlerle karşılaştırılır. Kriter geçerliliğinde, ölçüm yapılan her bir birey için iki puan vardır: tahmin puanları ve kriter ölçekten elde edilen puanlar. Bir kişi için, her iki puandan biri eksikse bu kişi analizden çıkarılmalıdır. Üç tür kriter geçerliliği vardır: (a) tahmin geçerliliği, (b) birlikte vuku bulma geçerliliği (eş zamanlı geçerlilik) ve (c) geriye dönük geçerlilik.[54]

Kriter geçerliliğinde karşılaştırma standardı olarak kullanılacak ölçüm değerlerinin (puanların) belirli özelliklere sahip olması gerekir ve bunlar aşağıdaki gibidir:

1. Karşılaştırma kriterinin kendisi ve bu kriterden elde edilen puanlar güvenilir olmalıdır.
2. Kriter puanlar, başarıyla gerçekten ilgili olmalıdır.
3. Örnek alınan kriter puanları, araştırmacının veya bilim adamının kullanım amacına uygun olmalıdır.
4. Kriter puanları dış etkenler nedeniyle kirlenmiş olmamalıdır.

Kriter puanların *güvenilir* olması, herkes tarafından kabul edilebilecek nesnel kanıtlara dayanmasıdır. Sübjektif değerlendirmelere dayanan kriter puanlarının güvenilirliği düşüktür. Ayrıca kriter puanları temsil edici bir örneklemden elde edilmiş olmalıdır.

Kriter puanlarının *ilgili* olması, başarı sonuçlarını yansıtmasıdır. Başarı sonuçları örgütle ilgili, grup veya takımla ilgili veya bireysel sonuçlarla ilgili olabilir. İlgililik; üretim miktarı, istenen davranışların gösterilmesi,

eğitim programı sonrasında yapılan sınavlarda başarı gösterme gibi doğrudan çalışılan işe ilişkin veya ölçme amacına ilişkin olan sonuçlardır. İlgili bir kriter bulunamıyor veya geliştirilemiyorsa "kriter geçerliliği" tekniğini uygulamaktan vazgeçilmelidir.

Amaca uygunluk, kriter puanlarının karşılaştırma yapılabilir nitelikte olmasıdır. Kriter puanları oldukça makul bir büyüklüğe sahip örneklemden elde edilmiş olmalı ve aynı zamanda işi ve adayları temsil edebilmelidir. Kriter geçerliliği çalışmasında araştırmacı örneklemle ilgili olarak istatistiksel güç hakkında bilgi verelidir. Kendilerine test uygulanan kişilerle kriter puanların temin edildiği örneklem grubu arasındaki önemli farklılıklar kriter puanları geçersiz hale getirir.

Dış etkenlerden etkilenmeme ise, kriter puanlarının tüm kişiler için aynı anlama gelmesi ve bünyesinde yabancı etkilerden dolayı bir bozulmanın bulunmamasıdır.

Tahmin Geçerliliği

Geliştirilen bir ölçeğin/testin tahmin geçerliliğine sahip olması, söz konusu ölçekten veya testten yüksek puan alan bir kişinin elde ettiği bu sonuçların aynı zamanda o kişinin daha sonraki davranışlarında veya başarılarında da görülmesidir. Tahmin geçerliliği için test/ölçek ilgili kişilere uygulandıktan sonra bir süre beklenir. Bekleme süresi değişkendir. En az altı aydan başlayıp beş yıla kadar sürebilir. Tahmin geçerliliği genellikle iki yıllık sürenin sonunda sorgulanır. Bu bekleme süresi içinde kişilerin tutum ve davranışları, başarıları gözlenerek fiili hayata ilişkin veriler toplanır ve bu verilerin ortalaması alınarak söz konusu değer "kriter puan" olarak adlandırılır. Daha sonra kişilerin kriter puanlarıyla test puanlarının korelasyonu hesaplatılarak ölçeğin geçerliliği saptanır. Bu nedenle tahmin geçerliliği hemen yapılamaz. Tahmin geçerliliğinin yapılabilmesi için aradan en az altı ay gibi bir sürenin geçmesi ve hangi davranışların kriter olarak alınacağının belirlenmesi gerekir. Tahmin geçerliliği; tutum ölçeklerinden çok başarı testleri ve bilişsel testler için uygulanır. Örneğin, LES'ten yüksek puan alan öğrenciler düşük puan alan öğrencilere göre yüksek lisans eğitimi sırasında daha başarılı olmuşlarsa LES sınav sorularına ait puanların geçerli olduğu söylenir. Bir kişinin gelecekteki davranışları veya başarısı tahmin edilmek istendiğinde tahmin geçerliliği yöntemi uygulanır. Bir test ikna kabiliyetini ölçüyorsa, iş müracaatı sırasında pazarlamacı olarak istihdam edilecek adaylara bu test uygulanır. Adaylar test sonucuna göre değil, mülakat ve özgeçmiş sonuçlarına göre istihdam edilirler. Aradan altı ay

gibi bir sürenin geçmesinden sonra işe alınan kişiler iş ilişkileri ve davranışları, fiili satış rakamları, kaçırdıkları müşteri sayısı gibi somut göstergelere göre değerlendirilirler. Bu değerlendirmeden yüksek puan alan kişiler başlangıçta uygulanan ikna kabiliyeti testinden de yüksek puan ve fiili değerlendirmeden düşük puan alan kişiler başlangıçta uygulanan ikna kabiliyeti testinden de düşük puan almışlarsa İkna Kabiliyeti Testi'nin tahmin geçerliliğine sahip olduğu söylenir. Psikometrik test bataryasına göre alınan kişilerin işlerinde yüksek performans göstermeleri bataryayı oluşturan testlerin kriter geçerliliğine sahip olduğunun kanıtıdır. Ancak test puanlarıyla fiili davranışlar arasında tam tersi bir durum ortaya çıkmışsa testin tahmin geçerliliği düşüktür.

Tahmin geçerliliğinde korelasyon analizi katsayılarından yararlanılır. Korelasyon katsayısının en az ,30 olması gerekir. Bu değer *birlikte vuku bulma* korelasyon kat sayısından daha düşüktür çünkü zaman içinde davranışın gerçekleşme olasılığını tespit etmek büyük ölçüde güç bir iştir. Aslında test puanlarıyla davranışın tahmin edilmesi için aradan geçen süre uzadıkça korelasyon katsayıda da azalır.[55] Tahmin geçerliliği için korelasyon analizi yerine birden fazla kriter temel alınarak regresyon analizi yöntemi de kullanılabilir Örneğin, *sürücü simülasyon testi* tahmin değişkeni olarak alınmış olsun. Kriter değişkeni olarak fiili yol sürüş testi, fiili görme pratiği, işitme pratiği kriter değişkenleri olarak alınır ve regresyon analizi bu verilere dayalı olarak yapılır.

Tahmin geçerliliğinde duyarlılık ve ayırt edicilik. "Duyarlılık" ve "ayırt edicilik" tahmin geçerliliğinde sık kullanılan iki önemli kavramdır. *Duyarlılık*, sosyal bilimlerde bir testin veya ölçeğin içerdiği hata payı anlamında ele alınırken tıp bilimlerinde farklı bir anlamda kullanılır. Tıpta teşhis amaçlı olarak kullanılan bir testin araştırılan hastalığa ilişkin gerçek pozitif sonucu verme oranıdır. Teşhis amaçlı testten yüksek puan alan kişilerin yüzde kaçının gerçekten hastalandığına ve yüzde kaçının ise hastalanmadığına bakılır. Hastalanma oranının yüksekliği testin *duyarlılığını* gösterir. *Ayırt edicilik* özelliği ise, testin/ölçeğin sağlıklı, normal ve istenen özelliklere sahip kişileri ortaya koyma veya açığa çıkarma başarısı ile ilgilidir. Burada testten düşük puan alan kişilerin aradan belli bir süre geçtikten sonra yüzde kaçının hastalanmadığına bakılır. Testin ayırt edicilik özelliği güçlü ise büyük bir kesimin hastalanmaması gerekir. Test puanı yüksek olup hasta olanlar *doğru pozitif* ve hasta olmayanlar ise *yanlış pozitif* olarak isimlendirilir. Bu iki kriter, bir testin etkililik derecesini ortaya koyar. Normal bir ana kütlede, *ayırt edicilik özelliğinin* ortaya çıkardığı

kişilerin %97,5 ve *duyarlılık* özelliğinin ortaya çıkardığı kişilerin ise %2,5 seviyesinde olması arzulanır. Teşhis amaçlı test ve ölçeklerle ilgili olarak literatürde güvenilir bilgi bulmak zordur, çünkü bu testlerin "referans ana kütleleri" sürekli değişkenlik gösterir. Elde edilen sonuçlar genelde belirli bir araştırma evrenine aittir.[56]

Tahmin geçerliliğinde kriter kirliliği. Kriter puanların kirlenme olgusu kısaca *kriter kirliliği* olarak isimlendirilir. Kriter kirliliği, haricî etkilerin belirlenen kriterler üzerinde etkili olması ve bu nedenle ölçüm verilerinde sistematik varyansın ortaya çıkması halini ifade eder. Örneğin, satış rakamları kriter puan olarak belirlenmişse kişilerin satış yaptıkları bölgelerin özeliklerinin göz önünde bulundurulması gerekir. Bazı bölgelerde satış yapma şansı yüksek iken diğer bölgelerde satış yapma şansı düşük ise satış rakamları kriter kirliliğine sahiptir. Kriter puan olarak kişilerin amirlerinin verdikleri performans değerlendirme puanları temel alınıyorsa; bazı amirlerin personeline yüksek puanlar verme ve bazı amirlerin de düşük puanlar verme eğilimi içinde olacakları gözden uzak tutulmamalıdır. Performans puanları belli ölçüde kriter kirliliğine sahiptir. Dış etkenlerden hangilerinin kriter puanlarını etkilediğini ve hangi ölçüde etkilediklerini tam olarak saptamak her zaman mümkün olmayabilir. Ancak araştırmacı bunu belirlemeye ve söz konusu bulaşma etkisini sınırlandırmaya yönelik bir çaba içinde olmalıdır.[57]

Ek faktör desteği alarak sağlanan geçerlilik. Buna kısaca "ek faktör geçerliliği" veya "artışsal geçerlilik" (incremental validity) adını verebiliriz Ek faktör geçerliliği, kriter puanlarla tahmin puanları arasındaki korelasyon katsayısı düşük çıktığında ilave tahmin puanlarından yararlanma anlamına gelir. Örneğin yüksek lisans öğrencilerinin başarısını sadece LES puanlarıyla açıklamak yeterli olmayabilir. Bunun için LES puanlarının dışında başka tahmin puanlarının veya göstergelerinin de modele alınması gerekir. Lisans puanları ortalaması, yabancı dil puanı, mülakat puanları hep birlikte regresyon eşitliğine alınırsa sonuçta korelasyon/regresyon katsayılarının yükselmesine ek bir artış sağlayacaktır. Bu şekilde elde edilen geçerliliğe "artışsal geçerlilik" veya *ek faktör geçerliliği* adı verilir. Ek faktör geçerliliğini hesaplamak için regresyon analizinden yararlanılır. Regresyon analizi sonucunda elde edilen R^2 değeri, birden fazla değişken olarak belirlenen tahmin puanlarının hep birlikte kriter puanlardaki varyansın yüzde kaçını açıkladığını gösterir. Açıklanan yüzde düşükse ölçüm modeline daha fazla tahmin puanı alınır. Daha fazla tahmin puanı a-

lınması demek, ölçümün tek bir test yerine bir test bataryası ile yapılması anlamına gelir.

Beklenti Tabloları. Beklenti tabloları, bir testin kriter geçerliliğinin değerlendirilmesinde yararlanılan bilgi sağlama araçlarıdır. Beklenti tabloları test sonuçları ile kriter değişken arasındaki korelasyon katsayılarını yansıtır. Bu tabloya bakarak, örneğin 40 gibi belirli bir tahmin puanına sahip bir kişinin o testten başarılı veya başarısız olma ihtimalinin yüzde kaç olduğunu söyleyebiliriz. Ancak üretilen bu tabloların geçerli olabilmesi için çalışılan örneklem büyüklüğünün oldukça büyük olması ve tahmin geçerliliğinin de ,70'in üzerinde çıkması gerekir. Aksi halde beklenti tablolarının sonuçları doğru olmayabilir. Beklenti tabloları, örneklem büyüklüğünün yanında "baz oranı"ndan da etkilenir. Baz oranı, herhangi bir özelliğin toplumda, ana kütlede bulunma oranıdır. Örneğin, toplumda şizoreflerin bulunma olasılığının ,01 olduğu söyleniyorsa bu rakam bir baz oranıdır. İş hayatında işler için de baz oranları tespit edilebilir. Bazı işlerin baz oranları yüksek bazılarının ise düşüktür. Bir iş için herhangi bir test veya seçim yöntemi uygulanmasa bile, başvuran kişilerin ,50'si işe alınacaksa söz konusu işin baz oranı ,50'dir. Baz oranı %50 gibi bir rakamda sabit tutulup tahmin geçerliliği katsayısı düştüğünde iş için başvuran kişilerden yüksek puan alanlar daha az oranda kabul edilebilir olarak değerlendirilir. Geçerlilik katsayısı sabit tutulup bu kez baz oranı değiştirildiğinde, bu kez iş için başvuran kişilerden yüksek puan alanlar çok daha az oranda işi kazanmış olacaklardır.[58] Beklenti tablolarının geçerliliği; örneklem büyüklüğü, baz oranı ve geçerlilik katsayısı dikkate alınarak değerlendirilmelidir. Baz oranı yüksekse, test puanları (tahmin puanları) araştırmacının tahmin etme çalışmasına çok fazla bir katkı sağlamaz.

Tahmin Geçerliliğinde korelasyon katsayılarının yorumlanması. Tahmin geçerliliği sonucunda elde edilen korelasyon katsayıları iki değişken arasındaki ilişkinin gücünü gösterir. Kriter "başarı" faktörü ise tahmin puanlarının kriteri açıklama yüzdesi korelasyon katsayısının karesi alınarak bulunur. Örneğin, tahmin geçerliliği katsayısı ,60 çıkmışsa tahmin puanlarının kriter puanlarını tahmin etme gücü ,36'dır. Tahmin puanlarının çoklu test bataryası kullanılmadığı sürece kriter puanları açıklama olasılığı düşük kalır.

Birlikte Vuku Bulma Geçerliliği

Birlikte vuku bulma geçerliliği, geliştirilen ölçek veya test puanlarının aynı

zamanda yapılan başka nitelikteki ölçüm puanlarıyla karşılaştırılması ve bu karşılaştırmaya dayalı olarak korelasyon katsayısının yüksek çıkmasıdır. Örneğin, öğrencilerin eğitimi değerlendirmelerine yönelik olarak uygulanan *Öğretim Üyesini Değerlendirme Ölçeği*'nden bir öğretim üyesinin yüksek puan alması halinde bu puanın aynı zamanda o kişinin başarısına ilişkin elde mevcut olan diğer başarı göstergeleriyle doğrulanıp doğrulanmadığına bakılır. Söz konusu diğer göstergelere ait puanlar da yüksek ise ölçeğin geçerli olduğu söylenir. Diğer göstergelere ait puanlar düşük ise, ölçek geçersizdir.

Birlikte vuku bulma geçerliliği değişik şekillerde analiz edilebilir. Bu yöntemler aşağıdaki gibidir.

1. Test puanlarının kriter olarak önceki yıllarda, dönemlerde yapılmış olan diğer testlerden elde edilmiş puanlarla karşılaştırılması (bu uygulamaya bazı kaynaklarda *geriye dönük geçerlilik* adı verilmiştir).
2. Test puanlarının paralel form (veya duruma göre standart form) puanlarıyla karşılaştırılması.
3. Test puanlarının kriter olarak önceki yıllarda veya dönemlerde yapılmış fiili performans puanlarıyla karşılaştırılması (*geriye dönük geçerlilik* grubunda da değerlendirilebilir).
4. Test puanlarının kriter olarak aynı gün içinde veya günlerde yapılan fiili performans puanlarıyla karşılaştırılması.
5. Testin, özellikleri birbirinden önemli ölçüde farklı olduğu bilinen iki ayrı grupta uygulanması.

Birlikte vuku bulma geçerliliğinin türleri. Aşağıdaki bölümde, birlikte vuku bulma geçerliliğine ilişkin analiz yöntemleri üzerinde durulmuştur.

Kriter puan olarak önceki yıllarda yapılmış benzer test puanlarının temel alınması. Birlikte vuku bulma, elde mevcut olan bir puan dizisiyle ölçüm puanlarının karşılaştırılması esasına dayanır. Elde mevcut olan puanlar önceki yıllarda yapılmış olan test değerleri olabilir. Yapılan çalışma bilişsel yetenek testi ise önceki yıllarda yapılmış aynı özelliği ölçen bir başka teste ait sonuçlar birlikte vuku bulma geçerliliği olarak kullanılabilir. Çalışma bir tutum ölçeği niteliğinde ise aynı kişilere veya aynı performans düzeyindeki farklı kişilere ait önceki yıllarda uygulanmış benzeri veya aynı kavramsal yapıyı ölçen tutum ölçeği puanları kriter puan olarak kullanılabilir.

Test puanlarının kriter olarak paralel form puanlarıyla karşılaştırılması. Bu uygulamada önceki dönemlerden elde edilmiş test puanları yoktur. Tam tersine kişilere aynı zaman diliminde aynı kavramsal yapıyı ölçen iki farklı form verilir. Uygulamada bilişsel testlerle duygusal testler arasında fark yoktur. Kriter olarak uygulanabilecek test sayısı bir, iki veya üç tane olabilir. Bazı araştırmacılar tek bir testin veya ölçeğin belirli bir olguyu ortaya çıkarmak için yetersiz kalabileceğini belirtmişlerdir. Bu nedenle bu araştırmacılara göre, belirli bir olgu hakkında daha sağlıklı karar verebilmek için aynı şeyi ölçen değişik sayıda testten/ölçekten oluşan bir batarya kullanılmalıdır. Bu uygulama, belirli bir olguyu yakalamaya yönelik olarak kişi/kişiler üzerinde değişik araçlarla bir tür *tarama* çalışması yapmaktır. Kullanılan ölçeklerin puanları arasındaki korelasyonun yüksek olması bulguların geçerliliğinin de yüksek olduğu anlamına gelir. Ancak çok sayıda test uygulanması halinde (örneğin, 20 test uygulanmışsa) bu testlerden en az birinin istenmeyen sonuçlar vermesi olasıdır ve böyle bir durumda araştırmacı gereksiz yere daha dikkatli ve duyarlı olmaya yönelecektir. Bu nedenle *çok sayıda* sözcüğünü *belirli sayıda* şeklinde anlamakta yarar vardır. Kişilere özellikle psikiyatrik veya psikolojik tanı koymaya yönelik uygulamalarda, tedavi amaçlı ölçümlerde birlikte vuku bulma geçerliliği için belirli sayıda testten yararlanmak gerekir. Kullanılan ölçekler/testler birbirlerinin yerine tam olarak ikame edilebilir nitelikte olmalıdır.

Birlikte vuku bulma geçerliliğinde en önemli problem geçerliliğin test edilmesinde kullanılacak alternatif formun/testin gerçekten aynı kavramsal yapıyı ölçüp ölçmediğine ilişkin somut kanıtların bulunması ve kullanılacak alternatif formun kendisinin kaliteli bir ölçüm aracı olmasıdır. Ölçüm aracının kaliteli olması o ölçüm aracıyla toplanan verilerin geçerlilik ve güvenilirliğinin yüksek olması anlamına gelir. O halde paralel formlar uygulamasında karşılaştırma yapmak amacıyla kullanılacak form daha önceden geçerlilik ve güvenilirlik çalışması yapılmış ve yüksek değerler elde edilmiş bir form olmalıdır. Literatürde karşılaştırma yapmak amacıyla kullanılan ölçüm değerine veya forma "altın standardı" adı verilmiştir. Ancak kıyaslama amacıyla kullanılacak hiçbir ölçüm aracı mükemmel değildir ve 24 ayar olarak nitelendirilemez. Her ölçüm aracının geçerlilik ve güvenilirlik açısından bazı yetersizlikleri söz konusu olabilir. Mükemmel bir kıyaslama ölçüsü veya ölçüm aracı bulunamayacağından kullanılan kıyas aracıyla elde edilen verilerin bir şekilde kendi geliştirdiğimiz ölçüm aracından daha güçlü veya sağlıklı olduğu "varsayımından" hareket ederiz. Bu şekilde kaliteli bir ölçüm aracı bulunmuşsa bu ölçüm aracına dayalı olarak ya-

pılacak analizlerde birlikte vuku bulma geçerliliği korelasyon katsayısı en az ,70 olmalıdır. İstatistik duyarlılığı daha yüksek olan bilim adamları ise söz konusu korelasyon katsayısını ,80 olarak belirlemişlerdir.

Paralel formlar uygulaması, testlerin güvenilirliğine benzerlik gösterir. Farklılık, burada testlerin kavramsal yapıları arasındaki benzerliğin üzerinde durulmasıdır. Bu açıdan uygulanan testler boyutlar arasındaki benzerliği ortaya koymayı amaçlıyorsa bu yönüyle yapısal geçerliliği destekleyen bir uygulama haline gelir. Birlikte vuku bulma geçerliliği sonucunda paralel nitelikteki test formlarının faktörlerine göre aritmetik ortalama puanları pilot araştırma ve esas araştırma sonuçları dikkate alınarak sütun grafiği ile raporlanır. Sütun grafiğinde aritmetik ortalama değerleri arasında önemli bir farklık yoksa geliştirilen testin/ölçeğin geçerli olduğu sonucuna varılır.

Test puanlarının kriter olarak önceki yıllarda veya dönemlerde yapılmış fiili performans puanlarıyla karşılaştırılması. Bir diğer uygulama; test sonuçlarının değil, fiili başarı sonuçlarının kriter olarak kabul edilmesidir. Bunun için aynı kişilere ait önceki en yakın dönemden elde edilmiş performans puanları temel alınır. Bu uygulama daha çok personel seçim testleri için geçerlidir. Personel seçim testlerinde birlikte vuku bulma geçerliliği şu şekilde test edilir. Belirli bir endüstride, belirli bir işletmede ve belirli bir işte çalışan kişilerin iş başarıları değerlendirilir ve üç grup altında sınıflandırılır: ortalamanın bir standart sapma diliminin üstünde yer alanlar A dilimi, ortalamanın bir standart sapma dilimi içinde yer alanlar B dilimi ve ortalamanın bir standart sapma diliminin altında yer alanlar C dilimi. Bu grupların puanları kriter olarak kabul edilir. Söz konusu kişiler geliştirilen personel seçim bataryasından da aynı veya benzer puanları almışlarsa personel seçim testinin geçerli olduğuna karar verilir.

Test puanlarının kriter olarak aynı günlerde yapılan fiili performans puanlarıyla karşılaştırılması. Her zaman önceden belirlenmiş puanlara ulaşmak mümkün olmayabilir. Böyle bir durumda kriter puanları test uygulamasının hemen arkasından yapılan fiili performans sonuçlarıyla karşılaştırılır. Fiili performans sonuçları; iş örneklemesi yöntemi, bir satış işlemini gerçekleştirme, günlük üretim miktarı ve ölçülebilir diğer başarı kriterleridir. Her tür iş için sağlıklı bir şekilde gerçek performansı gösterecek fiili gösterge bulmak kolay olmayabilir. Bu yöntem sadece fiili göstergelerin somut ve hemen uygulanabileceği ortamlar için söz konusudur.

Testin, özellikleri birbirinden önemli ölçüde farklı olduğu bilinen iki ayrı grupta uygulanması. Eş zamanlı geçerliliğin değişik bir uygulaması, yüksek ve düşük performansa sahip grupların açık bir şekilde bilindiği durumlarda uygulananıdır. İki farklı grup vardır ve bu grupların performansları, özellikleri ve tutumları arasında önemli ölçüde farklı olduğu herhangi bir ölçüm yapmaya gerek kalmadan zaten bilinmektedir. Böyle bir durumda her iki grupta aynı zamanda ölçüm yapılarak iki grubun puanları arasındaki korelasyona bakılır. Korelasyon katsayılarının düşük olması uygulanan test veya ölçeğin geçerliliğinin yüksek olduğunu gösterir.

Birlikte vuku bulma geçerliliği istatistikî analiz yöntemleri. Birlikte vuku bulma (eş zamanlı) geçerliliğini test etmek için korelasyon analizinden yararlanılır. Aynı kavramsal yapıyı ölçen testlerde korelasyon katsayılarının yüksek çıkması geçerliliğin de yüksek olduğunu gösterir. Ancak farklı özelliklere sahip gruplara uygulandığında korelasyon katsayılarının düşük olması geçerliliğin yüksek olduğu anlamına gelir. Paralel formlar uygulamasında araştırmacı birinci test sonuçları ile ikinci test sonuçları arasındaki farkın anlamlı olup olmadığını görmek istiyorsa böyle bir durumda p değerlerine bakar. Ortalama puanlar arasındaki farklılık anlamlı değilse testin birlikte vuku bulma geçerliliğine sahip olduğu söylenir. Birlikte vuku bulma geçerliliği, araştırma veya ölçümün niteliğinde göre cinsiyet, yaş ve meslek grupları veya grubun geneli açısından da test edilebilir. İki paralel form arasındaki ilişkiler korelasyon analizi dışında regresyon analiziyle de sınanabilir. Regresyon analizi bir değişkende bir birimlik bir artışın diğer değişkende ne kadar bir artış ortaya çıkacağını tahmin eder. Hesaplanan R^2 değeri eğer yüksekse uygulanan testin veya ölçeğin tahmin geçerliliğinin de yüksek olduğu belirtilir. Buradaki tahmin sonraki zamanda elde edilen verilere dayandırılmamış, varsayımsal olarak hesaplanmıştır.

Geriye Dönük Geçerlilik

Geriye dönük geçerlilik, geliştirilen test veya ölçekten elde edilen puanlarının kriter olarak, önceki yıllarda veya dönemlerde yapılmış fiili performans puanlarıyla karşılaştırılması sonucunda korelasyon katsayısının yüksek çıkmasıdır. İki dizi test puanının aynı zaman diliminde karşılaştırılıyor olması nedeniyle bu uygulamaya aynı zamanda birlikte vuku bulma geçerliliği adı da verilir. Geriye dönük geçerlilik uygulamasında, kriter puanları önceki zamanlara aittir.

Kriter Geçerliliği Analizinde Dikkat Edilmesi Gereken Hususlar

Bilim adamı kriter geçerliliği analizini yaparken ve analiz sonuçlarını raporlarken belirli noktalara dikkat etmelidir. Bunlar aşağıdaki gibidir:

1. Tahmin ve kriter puanları doğrusal nitelikte olmalıdır. Verilerin dağılımı eğrisel bir niteliğe sahipse sonuçlar yanıltıcı olur.
2. Ölçüm yapılan grup homojen veya heterojen olabilir. Bilim adamı kriter geçerliliği rakamlarını buna göre uygun bir biçimde yorumlamalı ve okuyucularını bu konuda bilgilendirmelidir.
3. Kriter geçerliliği rakamlarıyla birlikte ölçek veya testin güvenilirlik katsayısı da göz önünde bulundurulmalıdır.
4. Verilerin türdeşsellik (sabit varyans) özelliği gösterip göstermediği incelenmelidir.

İki seri veri arasındaki ilişkiler yüksek türdeşsellik özelliğine sahipse korelasyon rakamlarının yüksek olması geçerliliğin de yüksek olduğu anlamına gelmez.

Kriter Geçerliliği Analiz Yöntemleri

Test ölçüm sonuçlarını gösteren tahmin puanlarıyla kriter puanları arasındaki ilişkiler değişik istatistikî analiz yöntemleriyle test edilir. Aşağıdaki bölümde özet olarak bu teknikler üzerinde durulmuştur.

Korelasyon analizi. Sürekli verilere ait kriter geçerliliğinin belirlenmesinde en çok Pearson korelasyon analizi yöntemi kullanılır. Bu şekilde elde edilen korelasyon katsayısına, "geçerlilik katsayısı" adı verilir. Tahmin puanları, kriter referanslı test kullanılarak *geçti-kaldı* şeklinde belirlenmiş ve kriter puanları da buna uygun olarak *başarılı-başarısız* şeklinde saptanmışsa geçerliliği hesaplamak için *phi katsayısından* yararlanılır. Araştırmacı tahmin puanlarına dayalı olarak başarıyı kestirmek istiyorsa korelasyon katsayısının karesini alarak (belirlilik katsayısı) ortak varyans değerini belirler. Örneğin, geçerlilik katsayısı ,50 çıkmışsa belirlilik katsayısı ,25'tir. Buna göre, test puanlarındaki geriye kalan değişkenliğin ,75'i başlıca iki faktörle açıklanır. Birincisi bu değişkenlik belirli ölçüde ölçüm hatalarından kaynaklanmış olabilir. İkincisi ise, değişkinliğin ölçüm hatalarının yanında tesadüfî olaylardan da etkilenebileceğidir. Diğer bir deyişle Y puanlarındaki değişkenlik, sadece X tahmin değişkeniyle açıklanamaz.

Denkleme, X tahmin değişkenin dışında başka değişkenlerin daha alınması gerekir.

Tahmin puanlarıyla kriter puanları arasındaki geçerlilik katsayıları literatürde genellikle hep düşük çıkmıştır. Geçerlilik katsayıları, güvenilirlik katsayıları kadar yüksek değildir. Uygulamada r_{xy} değerlerinin ,30 ilâ ,50 arasında değiştiği görülür. Geçerlilik katsayısı nadiren ,60'ın üzerine çıkar. Eğer ,30 ilâ ,50 arasında bir değer elde edilmişse testin geçerli olduğu sonucuna varılır.[59] Bazı bilim adamları, genelde yüksek değerler elde edilmemesi nedeniyle geçerlilik katsayısını ,20'ye kadar düşürmüşlerdir.

Tahminin standart hatası. Geçerlilik katsayısı r_{xy}, test ve kriter arasındaki ilişkiyi gösterirken tahminin standart hatası (TSH), yapılan tahmindeki hata oranını gösteren indeks değeridir. Bir anlamda *ölçümün standart hatası* değerine benzer. Ölçümün standart hatası güvenilirlik katsayıları için kullanılırken tahminin standart hatası geçerlilik katsayıları için kullanılır. Tahminin standart hatası, testin geçerlilik katsayısı etrafında güven aralığı değerlerinin belirlenmesine imkan sağlar. Böylece karar vericiler, vermiş oldukları kararlarında isabet derecesini artırırlar. Tahminin standart hatasını hesaplamak için regresyon eşitliği kullanılır ve daha sonra TSH değeri hesaplatılır. Örneğin, lise son sınıftaki öğrencilerin üç yıllık lise başarı ortalamalarının üniversite giriş sınavındaki puanları tahmin etme amacıyla kullanılması söz konusu olsun. Önceki yıllarda üç yıllık lise başarı ortalamalarıyla üniversite giriş puanları sonuçları Tablo 15-10'daki gibi belirlenmiş olsun. Fahri'nin üniversite sınavlarından 190 puan alacağı tahmin edilmektedir. Bu tahmin puanı ne ölçüde doğrudur? Tahmin puanının doğruluğu iki test puanları arasındaki geçerlilik katsayılarına dayalı olarak hesaplanır (*bk.*, Eşitlik 15-4; 15-5; 15-6).[60]

Tablo 15-10. Tahminin Standart Hatası İçin Veri Tablosu

	Öğrencilerin üç yıllık lise başarı puanı ortalamaları (x)	ÖSYM Sonuçları (y)
Aritmetik ortalama	4,58	184,5
Standart sapma	,68	25,3

r_{xy} = ,40 Geçerlilik katsayısı.

$$TSH = \sigma_y \sqrt{(1-r^2)}, \qquad (15\text{-}4)$$

$$TSH = 25{,}3\sqrt{(1-(,40)^2)}, \quad (15\text{-}5)$$
$$TSH = 23{,}2. \quad (15\text{-}6)$$

Buna göre, %68 güven aralığında (%68 ihtimalle) Fahri'nin puanı 166,8 ilâ 213,2 puan arasında değişecektir. Tahminin standart hatası değeri, bir kişinin alması düşünülen puanının kesim puanını içirip içermediğini belirlemek açısından önemlidir. Kesim puanı bir önceki yıl belirlenen fakülteye giriş puanı olabilir.

Doğrusal regresyon eşitliği. Tahmin değişkeni (X) ile kriter değişkeni (Y) arasındaki ilişkiler oldukça iyi çıkmışsa, test sonuçları başarıyı tahmin etmek için kullanılabilir bir özelliğe sahip olur. Bunun için regresyon eşitliğinden yararlanılır ($Y' = a + bX$). Regresyon eşitliği, test sonuçlarında bir puanlık bir artışın başarı üzerinde kaç birimlik bir etki doğuracağını gösterir.

YAPISAL GEÇERLİLİK

Yapısal geçerlilik, araştırmacı somut bir kriter veya standart yerine belirli bir *davranış alanına, kavramsal yapıya* veya belirli bir *faktöre* ilişkin sonuçlar elde etmek istediği zaman uygulanır. Yapısal geçerliliğin en basit bir şekilde anlamı, test veya ölçek maddelerinin ölçülmek istenen hipotetik faktörle (veya faktörlerle) yüksek derecede ilişkili olması ve faktörler arasındaki ilişkilerin de kurama uygun düşmesidir. Değişkenlerin bir faktör üzerindeki *faktör ağırlıkları* yüksekse söz konusu değişkenlerin yapısal geçerliliğe sahip olduğu söylenir. Fakat bu yeterli değildir, faktör sayısının ve faktörler arasındaki ilişkilerin de kuramla bir şekilde mutabakat içinde olması gerekir. Uygulamada yapısal geçerlilik, önce literatür çalışmasına veya gözlemelere dayalı olarak değişkenler ve faktörler arasında belirli ilişkilerin kurulmasıyla başlar. Daha sonra bu ilişkilerin ampirik test sonuçlarıyla doğrulanması halinde "yapısal geçerlilik" koşulu sağlanmış olur. Yapısal geçerlilik, sadece matematiksel modeller veya hesaplamalarla kanıtlanmaz. Yapısal geçerlilik birbirini teyit eden ve bir kısmı yargısal nitelikte pek çok kaynaktan toplanan kanıtların toplam sonucuna dayanır.

Yapısal geçerlilik kavramını tanımak için öncelikle "yapı" kavramı hakkında bilgi sahibi olmak gerekir. *Eğitimsel ve Psikolojik Test Standartları* kitabına göre, "yapı" insan davranışlarının niteliği hakkında yapılan bir soyutlama veya teorik bir açıklamadır.[61] Kavramsal yapılar; sınırları belirli, herkesin aynı şeyi anladığı terimler değildir. Tam tersine farklı kav-

ramsal yapılarla bir ölçüde iç içe geçmiş olması nedeniyle bulanık bir içeriğe sahip soyut tanımlamalardır. Literatürde *yapı* kavramı aynı zamanda özellik, yetenek, beceri, yetkinlik terimleriyle de anlatılmaya çalışılmıştır.[62] Örneğin; *iç denetimlilik, kendini gerçekleştirme, karizma, gerilim* sözcükleri kavramsal yapıları tanımlar. Kavramsal yapılar bazen tek boyutlu, bazen çok boyutlu bazen de diğer kavramsal yapıların toplamından veya çarpımından oluşur. Örneğin güç; kütle x hız değişkenlerinin bileşimini yansıtır. Sosyoekonomik statü; gelir, eğitim, yaşanan semt, servet faktörlerinin bileşiminden oluşur.

Yapısal geçerlilik, kavramsal yapının tam olarak açığa çıkarılma konusuyla ilgilidir ve diğer geçerlilik analizlerine göre daha geniş kapsamlıdır. Yapısal geçerliliği analiz etmek için kuramsal bir çerçeveye ihtiyaç vardır. Araştırmacının tek bir yapıyı diğer yapılardan soyutlanmış bir şekilde ele alıp incelemesi doğru değildir. Kuram, yapıyla göstergeler arasındaki ilişkilerin ve ayrıca farklı yapılar veya değişkenler arasındaki ilişkilerin niteliğini ortaya koyar ve değerlendirme bu çerçevede yapılır.[63] Yapısal geçerlilikte sadece geliştirilen ölçek veya testin kendisi değil, aynı zamanda kurulan hipoteze dayalı ilişkiler de test edilir. Bu nedenle araştırmacı, kavramsal yapının başka değişkenle / değişkenlerle olan ilişkilerini gösteren hipotezlerini de yapısal geçerlilik çerçevesinde test eder. Örneğin, yüksek stres altında çalışan personelin daha fazla devamsızlık yapacağını varsayabilir. Analiz sonucunda bu varsayım doğrulanmamışsa ya kullanılan stres ölçeği geçerli değildir veya hipotez yanlıştır. Araştırmacı hipotezinin kesin doğru olduğuna inanıyorsa, böyle bir durumda ampirik araştırma sonuçlarının ölçek veya test sonuçlarının geçerliliğini soruşturulabilir haline getirdiği sonucuna varılır.[64]

Yapısal geçerlilik daha önce sözü edilen geçerlilik türlerinin hepsini kapsar. Bu anlamda yüzey geçerliliği, içerik geçerliliği, kriter geçerliliği, birleşme ve ayrılma geçerliliği ile iç tutarlılık aynı zamanda yapısal geçerliliğin göstergeleri olarak değerlendirilir.[65]

Araştırmacı yapısal geçerliliği, geliştirmiş olduğu ifadelerin arka planda açıkça görülemeyen belirli bir kavramsal yapıyla (veya duruma göre yapılarla) ilgili olduğunu kanıtlamak için uygular. Geliştirilen ölçek tek bir yapıyı ölçüyorsa genelde faktör analizi sonucunda tek bir faktör çıkar. Birden fazla faktör çıkmışsa bunun iki anlamı vardır: Birincisi, ölçekte karmaşık kavramsal yapı nedeniyle gerçekten birden fazla faktör vardır İkincisi, ölçekteki ifadelerin madde evrenini temsil etme konusundaki yetersizlikleri nedeniyle sanki birden fazla faktör varmış gibi bir durum ortaya çıkmıştır.[66] Birden fazla faktör çıkmasında ayrıca cevaplayıcıların eğitim durumu,

dikkatsiz cevaplama, maddelerin anlamının tam olarak anlaşılamaması gibi etkenler de rol oynayabilir.[o]

Yapısal geçerliliği test etmek için literatürde çok sayıda yöntem, teknik ve hesaplama biçimi önerilmiştir. Bu yöntem ve teknikleri kullanılan ölçüm aracı çerçevesinde değerlendirmek gerekir. Araştırmacı, çok fazla zaman gerektirmeyen ve oldukça düşük maliyetli teknikleri tercih etmelidir.

Bazı bilim adamları, yapısal geçerliliğin korelasyon analizi, faktör analizi, regresyon analizi gibi tekniklerin kullanıldığı "niceliksel bir analiz" olduğunu savunurlarken; Angoff (1988), Cronbach ve Quirk (1976) gibi diğer bilim adamları yapısal geçerliliğin "tek bir katsayı" şeklinde ifade edilemeyeceğini, geçerlilik kanıtlarının "niteliksel içeriğe sahip" olması gerektiğini bildirmişlerdir (aktaran Yu).[67] Bilim adamı kullandığı test veya ölçeğin yapısal geçerliliğine ilişkin bilgi verirken tek bir yönteme dayanmamalı bir çok yöntem ve tekniği birlikte kullanarak çok sayıda kanıt içeren bir dosya oluşturmalıdır. Ölçek veya testin yapısal geçerliliği daha sonraki uygulamalarla da gelişeceğinden ilk bulgular okuyuculara bir ön analiz olarak sunulmalıdır. Çok sayıda teknik ve yöntemden kaç tanesi uygulanırsa yapısal geçerlilik için yeterli analiz yapılmış olur sorusunun cevabı, araştırmacının doygunluk hissine bağlıdır. Araştırmacı birkaç yöntemde benzer sonuçları alıyor ve kavramsal yapının net bir şekilde ortaya çıktığını düşünüyorsa daha fazla geçerlilik analizi yapmaktan vazgeçebilir. Önemli olan, yeteri kadar ikna edici kanıt toplamaktır. Aşağıdaki başlıklarda yapısal geçerliliğe ilişkin kanıt toplamak için kullanılabilecek yöntemler üzerinde durulmuştur.

İçerik Analizi İle Yapısal Geçerliliğin Test Edilmesi

İçerik analizi, kullanılan ölçek veya teste ilişkin kavramsal boyutların araştırılan konuyla ne ölçüde ilgili olduğudur. Bilim adamı öncelikle teorik yapı ile ölçüm aracı arasında bir bağ kurmak zorundadır. Bunun için kavramsal yapıyı irdelemeli, kavramsal yapıyı mümkün olduğunca anlaşılır bir şekilde tanımlamalı, kavramsal yapının faktörlerini veya alt boyutlarını net bir şekilde ortaya koymalıdır. İçerik analizinde kuramsal yapının "tanım-

[o] Bu konuda daha fazla bilgi için *bk.*, (measurement artifacts theory). Bazı araştırmalarda, *olumsuz (negatif) cümle yapılarına* sahip maddelerin ayrı bir faktör altında toplandığı bunmuştur. Eğitim düzeyi düşük olan kişiler ters döndürülmüş bu maddeleri ayırt etmekte güçlük çekmektedirler. Bu nedenle bu maddeler sanki ayrı bir faktörmüş gibi ortaya çıkmaktadır. Aynı şekilde katılımcıların %10 kadar küçük bir kısmı bile anketi dikkatsiz bir şekilde cevaplandırsa yine bağımsız ayrı faktör yapıları ortaya çıkmaktadır.

lanması" ve ilineklerinin ortaya çıkarılması üzerinde odaklanılır. İçerik analizi yöntemi, nomolojik ağ yöntemiyle birlikte gerçekleştirilir. İçerik analizi ile konunun kavramsal haritası ortaya çıkarılırken nomolojik ağ yöntemiyle kavramlar arasındaki ilişkilerin rotası belirlenir. İçerik analizi niteliksel bir inceleme yöntemidir. Okuyucular bu konu hakkında kitabın içerik analizi bölümünde ayrıntılı bilgi bulabilirler. Araştırmacı içerik analizini kendi başına yapmamalı bu konuda kavramsal yapıyı iyi bilen konunun uzmanlarına, hakemlere, konu içeriği uzmanlarına danışmalıdır.

İç Tutarlılık Analizi İle Yapısal Geçerliliğin Test Edilmesi

Ölçüm işlemi bir dereceleme ölçeğine dayanıyorsa (daha çok yansıtıcı ölçek veya bir indeks olabilir) söz konusu ölçeğin yapısal geçerliliğini test etmek için ilk başvurulacak yöntem iç tutarlılık analizidir. Ölçek veya testin iç tutarlılığı Chronbach alfa değeri ile belirlenir. Aynı zamanda güvenilirlik analizinde de kullanılan alfa katsayısı ile, ölçeğin maddeleri arasındaki birlikte değişim (kovaryans) değerleri dikkate alınarak iç tutarlılık hesaplaması yapılır.

Cronbach alfa değerleri, söz konusu kavramsal alana ilişkin geçerlilik katsayılardır. Bazı bilim adamları bu nedenle güvenilirlik kavramıyla geçerlilik kavramı arasında ayrım gözetmezler.

Dış Testler İle Yapısal Geçerliliğin Test Edilmesi

Bu uygulamada araştırmacı geliştirdiği test veya ölçeğin geçerliliğini kanıtlamak için bir taraftan aynı veya benzer yapıları ölçen başka testler bulur ve diğer taraftan da farklı kavramsal yapıları ölçen testleri de aynı kişilere uygulayarak elde ettiği sonuçları karşılaştırma yoluna başvurur. Bu testler *dış kriter* olarak adandırılır.

Aynı kavramsal yapıyı ölçen dış kriter testlerle yapılan karşılaştırmalar kriter geçerliliğine benzer. Bu anlamda birlikte vuku bulma geçerliliği aynı zamanda "yapısal geçerlilik" olarak değerlendirilir. Bu uygulamanın kriter geçerliliği veya yapısal geçerlilik başlığı altında sunulması, anlatılması çok fazla önemli değildir. Önemli olan ölçüm olayının dış kriter ölçümlerle destekleniyor olması ve ölçülmek istenen yapının net bir şekilde ortaya konmasıdır. Karşılaştırmalarda "aynı" veya "ilgili" olan yapılar arasındaki korelasyona bakılır. Aynı kavramsal yapıyı ölçen testlerde korelasyon katsayılarının ,70-,80 gibi güçlü değerler olması gerekir. Karşılaştırma yapmak için aynı nitelikte bir dış kriter test bulunamamışsa böyle bir durumda ölçülmek istenen kavramsal yapıya benzer "ilgili" testlerden yararlanılır. İlgili testlerle yapılan karşılaştırmalarda orta derecede ilişkiyi gösteren ,50-

,70 büyüklüğündeki korelasyon katsayıları yine yapısal geçerlilik kanıtı olarak değerlendirilir.[68] Örneğin, geliştirilen *Girişimcilik İndeksi*'nin "ilgili ölçek" dış kriter *Bağımsız Çalışma Envantari*'yle ilişkisi ,55 çıkmışsa *Girişimcilik İndeksi*'nin geçerli olduğuna karar verilir. Dış kriter testle yapılan karşılaştırmalarda korelasyon katsayısı ,30 gibi düşük bir değer olarak ortaya çıkmışsa hatanın A testinden mi yoksa B testinden mi kaynaklandığını ortaya koymak çok zordur. Bu gibi durumlarda faktör analizi tekniğine başvurulur.

Grup Farklılıklarıyla Yapısal Geçerliliğin Analiz Edilmesi

Geliştirilen bir testten farklı iki grubun farklı puanlar almasını bekliyorsak bu beklentinin doğru çıkmasıyla yapısal geçerlilik de sağlanmış olur. Örneğin, dindarlığı ölçen bir ölçeğin geçerliliği, namaz kılma alışkanlığı olan kişilerle namaz kılma alışkanlığı olmayan kişilere uygulandığında farklı sonuçlar veriyorsa ölçek yapısal geçerliliğe sahiptir. Benzer sonuçlar çıkmışsa ölçeğin geçerliliği kuşkuludur. Aynı şekilde yetişkinlerin genel yeteneğini ölçen bir test maddesi ilk okul öğrencilerine uygulandığında aralarındaki korelasyon katsayısı ,95 çıkmışsa testin veya test maddesinin geçerliliğinin düşük olduğu sonucuna varılır.[69]

Faktör Analizi Yöntemiyle Yapısal Geçerliliğin Test Edilmesi

Faktör analizi, çok sayıda değişkenin arka planında yatan temel yapıyı ortaya çıkarmak için yapılır. Faktör analizi yapmak için araştırmacının öncelikle araştırmanın/ölçümün *kavramsal alanını* belirlemesi gerekir. Araştırma alanı; kişilik, tutumlar, yetenekler veya beceriler olabilir. Kavramsal araştırma alanı kuramsal temelin dayandığı sınırlardır. İkinci aşamada, sonuçların genelleneceği ana kütle belirlenmelidir. Ana kütle genellikle araştırma yapılamayacak kadar büyük bir kesimi kapsar.

Faktör analizi yapmak için bilim adamı, kavramsal alandan belirli bir ölçüm konusuyla ilgili olarak maddeleri/testleri ve ana kütleden ise örneklem birimlerini seçer. Kavramsal alandan seçilen maddeler *yüzey değişkenleri* olarak isimlendirilir. Yüzey değişkenlerinin arka planında, "faktör" olarak isimlendirdiğimiz gözle görülmeyen gruplandırma değişkenleri vardır ki bunlara *iç değişkenler* adını veririz. Faktörler, bir anlamda yüzey değişkenlerinin "nedeni" olarak değerlendirilir. İç değişkenler veya faktörler de kendi içinde iki alt grupta incelenir: *genel faktörler* ve *spesifik faktörler*. Genel faktörler, birden fazla yüzey değişkeni üzerinde etkili olan faktörlerdir. Bir ölçekte/test bataryasında yüzey değişkenlerinin önemli bir bölümü hakkında tanımlama yapma imkanı sağlayan az sayıda faktör var-

dır. Spesifik faktörler ise bir veya daha az yüzey değişkeni üzerinde etkili olan faktörlerdir. Bir ölçekte spesifik faktörlerin sayısı çok daha fazladır. Bu nedenle birden fazla spesifik faktörün etkisi birleştirilerek ölçekte sanki tek bir spesifik faktör varmış gibi değerlendirilir. Ölçekteki maddelerde değişiklik yapıldığında spesifik faktörler genel, genel faktörler de spesifik faktör olabilir. Yüzey faktörleri üzerinde üçüncü bir etken daha vardır ki bu da *ölçüm hatası*dır. Ölçüm hatası, iç değişken anlamında bir faktör olmamakla birlikte mantık yürütme açısından bir faktör olarak değerlendirilir. Eğer güvenilir bir ölçüm yapılmışsa daha az hata; özellik ve yöntem hatası çok olan bir ölçüm yapılmışsa daha fazla hata faktörü söz konusu olur. Bir araştırmada düşük güvenilirlikli bir ölçüm yapılmışsa hata faktörü yüksek olacak, fakat genel faktör ve spesifik faktör bu durumdan pek fazla etkilenmeyecektir. Herhangi bir ölçekte veya test bataryasında *hata faktörü* ile *spesifik faktör* birleşik tek bir faktör gibi düşünülebilir. Bu iki faktörün birleşimine *tekil faktör* adını verelim. Böyle olunca herhangi bir yüzey değişkeni, spesifik faktörle hata faktörünü içeren *tekil faktöre* denk gelir. Ölçek maddelerinde yapılacak herhangi bir değişiklik hem spesifik faktörü ve hem de hata faktörünü etkiler.

Bilim adamı bir ölçek veya test bataryası ile ilgili olarak karar verirken genel faktörleri dikkate alır. Genel faktörler söz konusu ölçeği veya test bataryasını büyük ölçüde açıklama özelliğine sahip maddelerden oluşur. Ancak burada her bir yüzey değişkenin sadece tek bir faktöre ait olması gibi bir durumdan söz edilmez. Bir yüzey değişkeni birden fazla genel faktörün etkisi altında olabilir. Diğer bir deyişle bir yüzey değişkeni hem Faktör 1'in altında ve hem de Faktör 2'nin altında yer alabilir.[70] Öte yandan genel faktörler aynı zamanda birbirleriyle yüksek derecede ilişkili veya tam tersine bağımsız olabilir. Ölçek veya testin yapısal geçerliliğini ortaya koymak isteyen bir araştırmacı öncelikle faktör analizi sonuçlarının ne anlama geldiğini bilmesi ve hangi koşullarda faktörlerin kavramsal yapıya işaret ettiği konusunda fikir sahibi olması gerekir.

Faktör analizi sonuçlarını yapısal geçerlilik kanıtı olarak sunma biçimleri. Araştırmacı geliştirdiği ölçeğin tek boyutlu veya çok boyutlu olduğunu kanıtlamak istiyor olabilir. Hangi kanıtın peşinde koşuyorsa buna uygun analizleri yapmalı ve sonuçları buna uygun olarak raporlamalıdır. Faktör analizinin yapısal geçerlilik için sunumunda aşağıdaki seçenekler kullanılabilir.

1. *Tek boyutluluğu kanıtlama.* Araştırmacı ölçeğin tek boyutlu olduğunu göstermek ve kanıtlamak istiyorsa elde edilen birinci faktör toplam varyansın en az %40'ını açıklamalı ve diğer faktörlerin ağırlığı ise giderek azalan bir seyre sahip olmalıdır. Tek boyutluluğu kanıtlamak isteyen araştırmacılar kavramsal yapının her bir boyutunu veya faktörünü bir alt ölçek imiş gibi değerlendirip faktör analizini her bir boyut için faktör bazında ayrı ayrı yapmalıdırlar.
2. *Çok boyutluluğu kanıtlama.* Araştırmacı ölçeğin/testin çok boyutlu olduğunu kanıtlamak istiyorsa böyle bir durumda elde ettiği birden fazla faktörün "açıklanan toplam varyans yüzdesi", "özdeğer" veya "yamaç-birikinti" grafiğine göre faktör sayısını belirlemeli ve literatürdeki araştırma sonuçlarıyla karşılaştırmalıdır. Literatürdeki araştırma sonuçları, kendi ölçeğini oluştururken kullandığı boyutlar ile faktör analizi sonucunda ortaya çıkan boyutlar arasında bir ilişki varsa veya paralellik söz konusu ise ölçeğin yapısal geçerliliğe sahip olduğu söylenir. Ölçüm aracında birden fazla bağımsız faktör varsa ve bu faktörler birbirinden bağımsız ise bu kez "ikinci düzey faktör analizi" yöntemi uygulanarak her bir faktörün kendi içinde basit yapılı tek bir faktörü ortaya çıkarıp çıkarmadığına bakılır. Bu uygulamada bağımsız faktörlerin her biri "yapı" gibi bir işleve sahiptir.
3. *Boyutlar arasındaki ilişki.* Faktör analizi sonucunda iki, üç veya dört faktör elde edilmişse ve bu faktörlerin kendi aralarındaki korelasyon katsayıları yüksekse (,60 ve daha yukarı bir değer olabilir) boyutların bağımlı olduğundan ve hepsinin birlikte tek bir kavramsal yapıyı ölçtüğünden söz edilir. Böyle bir durumda faktörler veya boyutlar birer alt ölçek gibi değerlendirilmez. Boyutlara ait ifadelerin tek bir kavramsal boyutu ölçtüğü varsayılır.71
4. *Faktöriyel yapıyı teyit etme.* Bu uygulamada araştırmacı "teyit edici faktör analizi" yönteminden yararlanır. Literatürde kavramsal yapıya ilişkin faktör sayısı belli ise veya araştırmacı kendi gözlemlerine göre belirli sayıda faktör çıkmasını istiyorsa böyle bir durumda teyit edici faktör analizi yöntemini uygular. Teyit edici faktör analizi yönteminde alt kavramsal yapıların bir araya gelip bir küme oluşturup oluşturmadıkları incelenir.

Araştırmacı, faktör analizi yöntemini uygulamışsa hesaplama yöntemi olarak (temel bileşenler analizi, ortak faktör analizi, maksimum olasılık analizi vb.) hangi yöntemi seçtiğini, niçin bu yöntemi uygulamaya karar

verdiğini gerekçeleriyle birlikte açıklamalıdır. Hesaplamada rotasyon yöntemi (ör, orthogonal varimax, oblique promax) kullanılmışsa hangi yöntemin kullanıldığı ve söz konusu yöntemin niçin tercih edildiği ayrıca belirtilmelidir.

Faktör analizi sonucunda maddelerin faktör ağırlıkları değerlendirmeye alınır. Bir maddenin faktör ağırlığı ,40'ın altındaysa ya ölçekten çıkarılır veya yeniden ifade edilerek kalibrasyona tâbi tutulur. Keşfedici faktör analizi yöntemini uygulayarak geliştirilen ifadelerden faktöriyel yapıyı ortaya çıkarmayı amaçlayan araştırmacılar birden fazla örneklemde pilot araştırması yaparak her bir örneklemde aynı yapının ortaya çıkıp çıkmadığını gözlemelidirler. Eğer aynı faktöriyel yapılar ortaya çıkıyorsa söz konusu faktörler bilimsel bir temelle sahiptir ve ölçek geçerlidir. Farklı örneklemlerden farklı sonuçlar alınmışsa, diğer bir deyişle farklı faktöriyel yapılar ortaya çıkmışsa anket formu üzerinde çalışma yaparak aynı yapıyı ortaya çıkarmaya yönelik değişiklikler yapılmalıdır. Araştırmacı ikinci kez uygulayacağı faktör analizinde teyit edici faktör analizi veya keşfedici faktör analizi yöntemlerinden birini kullanabilmekle birlikte teyit edici faktör analizi yöntemini tercih etmelidir. Teyit edici faktör analizi maddelerin aynı faktörle yüklü olup olmadığı hakkında bilgi vermesine karşılık, söz konusu faktörün amaçlanan yapıyı ölçüp ölçmediği hakkında fikir vermez.[72]

Sheperd (1993) yapısal geçerliliğin, kriter ve içerik geçerliliğinin her ikisini de kapsadığını iddia etmiş ve böylece geçerlilikte, ampirik ve mantıksal çalışmanın her ikisinin de yapılmış olacağını belirtmiştir. Anastasi de (1986) yapısal geçerliliğin içerik geçerliliği ve kriter geçerliliği koşullarının her ikisini de karşıladığını kabul etmiştir (aktaran Stapleton).[73]

Birleşme ve Ayrılma Analizi

Yapısal geçerliliği belirlemek için kullanılabilecek bir diğer analiz birleşme ve ayrılma geçerliliğidir. *Birleşme geçerliliği*, geliştirilen ölçek veya testin tek boyutlu ise tüm göstergeleri arasındaki ilişkinin yüksek çıkmasıdır. Test veya ölçek çok boyutlu ise böyle bir durumda boyutlar veya faktörler arasındaki ilişkiye bakılır. Boyutlar birbirinden bağımsız ise her bir faktör veya boyut altındaki göstergeler arası ilişkinin yüksek çıkmasıdır. Faktörler birbiriyle ilişkili ise bu kez hangi faktör veya boyut altında bulunduğuna bakılmaksızın tüm göstergeler arasındaki ilişkinin yüksek çıkması anlamına gelir. *Ayrılma geçerliliği* ise geliştirilen ölçek veya testin sonuçlarıyla "farklı, fakat ilgili" başka bir kavramsal yapıyı ölçen test sonuçlarının düşük korelasyon rakamları vermesidir. Birleşme ve ayrılma geçerliliği ikisi birlikte

uygulanır. Araştırmacının sadece birleşme veya sadece ayrılma geçerliliği analizini yapması yapısal geçerliliği test etmek için yeterli değildir.[74]

Birleşme geçerliliği. Birleşme geçerliliği, aynı kavramsal yapıyı ölçen "test" veya "göstergelerin" kendi aralarında en azından orta derecede ilişkili olması anlamına gelir. Birleşme geçerliliği iki şekilde yapılır.

1. Aynı kavramsal yapıyı ölçen iki teste ait sonuçlar arasındaki korelasyon katsayısı ile.
2. Belirli bir kavramsal yapıyı ölçen teste ait maddelerin/göstergelerin kendi aralarındaki korelasyon katsayıları ile.

Bilim adamı birleşme geçerliliğini test etmek için daha başlangıç aşamasında aynı kavramsal yapıyı ölçen paralel bir form daha bulmalıdır. Bu şekildeki paralel formlar geçerliliği aynı zamanda birlikte vuku bulma geçerliliğine benzer. Bir faktöre ait göstergelerin veya maddelerin yüksek derecede ilişkili olması ise verilerin bir noktada birleştiğini gösterir. Göstergeler arasındaki korelasyon katsayıları düşükse bunun anlamı modelden daha fazla faktör çıkarılabileceğidir.[75] Böyle bir durumda faktör çıkarma durumu gözden geçirilmeli ve birleşme geçerliliği yeni faktörlere ait göstergeler üzerinde yeniden sınanmalıdır.

Ayrılma geçerliliği. Ayrılma geçerliliğinde ilgili fakat farklı olan kavramsal yapılar arasındaki ilişkinin düşük olacağı varsayımından hareket edilir. Burada en önemli sorun farklı kavramsal yapıyı ölçen testin veya ölçeğin nasıl belirleneceğidir. Diyelim ki bir araştırmacı örgütsel stresi ölçmek istemektedir. Birleşme geçerliliği için örgütsel stresi ölçen başka bir ölçek kullanılabilir, fakat ayrılma geçerliliği için motivasyon veya liderlik gibi hiç ilgisi olmayan bir ölçek kullanmak anlamsızdır. Onun yerine aynı aileden gelen (ilgili) fakat kavramsal alanı farklı olan bir ölçek kullanılır. Bu çerçevede kaygı ölçeği, endişe ölçeği, bireysel stres ölçeği kullanılabilir. Bu konuya araştırmacının kendisi karar verecektir. Bu nedenle daha araştırmaya başlamadan ayrılma geçerliliğini saptamaya yönelik olarak "ilgili, fakat farklı bir yapıya" ait bir ölçek daha bulunmalı ve katılımcılara iki ölçek birlikte verilmelidir. Ayrılma geçerliliğinde, geliştirilen ölçeğin maddeleriyle, ilgili ve fakat farklı kavramsal yapıyı ölçen testtin/ölçeğin maddeleri çoklu korelasyon analizine tabi tutulur. Analiz sonucunda iki farklı ölçeğin maddeler arasındaki korelasyon katsayıları düşükse ayrılma geçerliliğinin sağlanmış olduğuna karar verilir.

Geçerlilik katsayıları. Birleşme ve ayrılma geçerliliğinde korelasyon katsayılarının ne kadar yüksek ve ne kadar düşük olması gerektiği konusunda bilim adamları kesin bir ölçü vermemişlerdir. Birleşme geçerliliğinde mümkün olduğunca yüksek bir değer ve ayrılma geçerliliğinde de mümkün olduğunca düşük bir değer elde edilmesi gerektiği dile getirilmiştir. Bu konuda birleşme geçerlilik katsayılarının her zaman ayrılma geçerlilik katsayılarından daha yüksek olduğu ifade edilmiştir.[76]

Kullanıldığı yerler. Birleşme ve ayrılma geçerliliği analizleri daha çok kişilik analizlerinde kullanılmıştır, fakat aynı zamanda bilişsel ve duygusal içerikli testlerin geçerliliğini test etmek için de bu yöntemden yaralanılabilir.

Birleşme ve ayrılma geçerliliği analiz yöntemleri. Birleşme ve ayrılma geçerliliğini analiz etmek için literatürde en çok Campbell ve Fiske'in çoklu özellik - çoklu yöntem matrisi (ÇÖÇYM) kullanılmıştır. Bu yöntemde aynı ve farklı kavramsal yapıları ölçen test sonuçları tek bir matris tablosu üzerinde gösterilir. ÇÖÇYM farklı ölçeklerin göstergeleri arasındaki korelasyon katsayılarına dayanır.

Birleşme ve ayrılma geçerliliğini analiz etmenin ikinci bir yöntemi LISREL isimli istatistiksel analiz programındaki teyit edici faktör analizi tekniğinden yararlanmaktır. Bu yöntem bir ölçüde daha karmaşık işlemlerin uygulanmasını gerektirir.

Nomolojik Ağ Grafiği İle Geçerlilik Analizinin Yapılması

Nom Lâtince hukuk veya kural anlamına gelir. Nomolojik ağ ise, hukuka uygun veya *kuralara uygun ağ* anlamındadır. Bir kavramsal yapıda, kavrama bağlı faktörler ve faktörlere bağlı olan göstergeler arasındaki ilişkiler belirli kurallara göre çalışır. Kavramsal yapılar arasındaki ilişkiler, "gerçek hayattaki ilişkiler hukukuna" uygun olmalıdır. Fakat ilişkiler gerçekte ya rasgeledir veya istatistikseldir. Nomolojik ağ kavramı 1955 yılında Lee Cronbach ve Paul Meehl tarafından geliştirilmiştir. Araştırmacı, yaptığı ölçümlerin yapısal geçerliliğe sahip olduğunu göstermek için nomolojik ağdan yararlanabilir. Bu ağ, araştırmacının kuramsal olarak belirlediği; (a) göstergelerle boyutlar arasındaki ilişkileri, (b) boyutların kendi aralarındaki ilişkileri ve (c) dış değişkenlerle kavramsal yapılar arasındaki ilişkileri temsil eder. Sonuncusu, belirlenen hipotez çerçevesinde ölçek dışında kalan ölçüm maddelerini de kapsar. Araştırmacı yapılar arasındaki ilişkileri gözlemlerine veya literatür çalışmasına dayalı olarak belirler. Ancak her zaman önceki araştırma bulgularına veya kurama uygun olarak hareket e-

dilmez. Araştırmacı bazı gözlemlerine dayanarak ilişkiler ağını "nomolojik geçerlilikten yoksun" bir biçimde de oluşturabilir. Bunun anlamı, araştırmacının yeni bir yapının varlığından şüpheleniyor olmasıdır. Önceki ölçüm sonuçları veya kurulu ilişkiler bütünüyle yanlış da olabilir. Bu nedenle bilim adamı, önerdiği ilişkiler ağının geçerli olduğunu kanıtlamak zorundadır. Yaptığı ölçümlerin yapısal geçerliliğine sahip olduğunu göstermek için nomolojik ağdan yararlanmak isteyen araştırmacılar belirli ilkelerden hareket ederler.[77]

1. Bilimsel olarak bir yapının var ve diğer yapılarla ilişkili olduğunun açık bir biçimde literatürden de destek alınarak açıklanması.
2. Kavramlar, faktörler ve göstergeler arasındaki ilişkilerin doğal bir şekilde birbirilerine kenetlenmiş ilişkiler yumağı halinde görüntülenmesi.
3. Yapılar arasındaki gözlenebilir ilişkilerin açık bir şekilde ortaya konulması.
4. Teorik kavramsal yapılarla gözlenebilir göstergelerin birbirinden ayırt edilmesi ve net olarak ortaya çıkarılması.
5. İlişkiler ağındaki hukuk veya kurallardan (kavramsal yapılardan) en azından birinin gözlenebilir göstergeleri içermesinin sağlanması.
6. Kavramsal yapı, faktörler ve göstergeler arasındaki ilişkilerin karmaşık olmayacak ve net bir şekilde bilgi verecek şekilde ortaya konulması.
7. Kurama yeni bir faktör ilave edilmesi veya çıkarılması işleminin bazı gözlemlerle teyit edilmesi.

Bu yaklaşımda Cronbach ve Meehl teorik ve kavramsal yapılarla gözlenebilir göstergeleri birleştirmeye ve aralarındaki ilişkilerin haritasını çıkarmaya çalışmışlardır. Nomolojik ağ yapısal geçerliliğin felsefi temelini oluşturmasına karşın somut ve pratik bir geçerlilik ölçüm değeri veya geçerlilik katsayısı vermemektedir.[78] Sadece yapısal ilişkileri gösteren bir çıkış veya başlangıç haritası olarak kullanılabilir. Nomolojik ağın bu yetersizliği daha sonra Campbell ve Fiske (1959) tarafından bulunan çoklu özellik - çoklu yöntem matrisi ile giderilmiştir. Bilim adamı ampirik araştırma sonuçlarına dayalı olarak kavramsal yapıya ilişkin nomolojik ağı yeniden çizerek önceki ağ ile sonraki ağ arasındaki benzerlik ve farklılıkları etüt eder. Tahminle sonuçlar uyum içindeyse testin yapıyı ölçtüğü iddiası teyit

edilmiş olur. Uyum içinde değilse nomolojik ağ yeniden oluşturulur. Nomolojik ağ yapının "doğru" olduğunu göstermek için kullanılmaz.[79] Cronbach ve Meehl'e göre biz önce teorinin doğru olduğunu kabul edip daha sonra testin geçerliliğini kanıtlamaya çalışıyor değilizdir. Tam tersine, testin geçerliliğine göre teoriyi yeniden oluşturma peşinde de koşmayız. Nomolojik ağ yapısının çizilmesindeki temel felsefe gözlem verileri ve teori arasındaki ilişkilerin bir bütün olarak ele alınması ve incelenmesidir.

Cronbach ve Meehl'e (1955) göre nomolojik ağ, bir ölçeğin geliştirilme aşamasında ciddiye alınmalıdır. Ampirik araştırma sonucunda ölçeğin nomolojik ağ geçerliliğine sahip olduğunu göstermek isteyen araştırmacılar, göstergeler ve diğer kavramsal yapılar arasındaki ilişkileri korelasyon katsayılarını vererek gösterebilirler. Nomolojik geçerlilik farklı yapılara ait göstergeler arasında ampirik ilişkiler kurulmasını gerektirir. Nomolojik geçerliliğin arkasındaki temel güç, ampirik sonuçların nomolojik ilişkiler ağını destekleyip desteklemediğini görme imkanı sağlamasıdır.[80]

Çoklu Özellik - Çoklu Yöntem Matrisi Geçerlilik Analizi

Herhangi bir özelliğin/yeteneğin ölçümünde sadece tek bir yöntemin kullanılması halinde söz konusu karakteristiğin resmini tam olarak elde etmek mümkün olmayabilir. Bunun için, birden fazla test ve birden fazla yöntemden birlikte yararlanılarak sonuçların birbiriyle ilişkili olup olmadığına bakılır. Birden fazla test, birden fazla ölçüm aracı kullanma anlamına gelir. Birden fazla yöntem ise; kişisel öz değerlendirmelerin yaptırılması, araştırmacının gözlemde bulunması, kişiye test uygulanması gibi ölçümlerdir. Böylece, hipotezi birden fazla ölçüm aracı ve birden fazla yöntemle sınama imkanı ortaya çıkar. Çoklu özellik - çoklu yöntem matrisinin oluşturulabilmesi için en az iki ölçüm aracı (kavramsal yapı veya boyut) ve yine en az iki yöntem kullanılmış olmalıdır. Bir kişinin *iki farklı özelliğini iki farklı ölçme yöntemiyle* belirlemeye kalkışmışsak çoklu özellik -çoklu yöntem matrisinden (ÇÖÇYM) yararlanırız.

Çoklu özellik - çoklu yöntem matrisinde örneğin; bir öğrencinin "dürüstlüğünü", "uyumluluğunu" ve "zekasını" üç farklı yöntemle ölçmek isteyebiliriz. Bu çerçevede "gözlemlerimizden" yararlanabilir, "test" uygulayabilir veya "öğretmeninin değerlendirmelerini" kullanabiliriz. Böyle bir durumda 9x9 boyutlu bir korelasyon matrisi elde ederiz (*bk.*, Tablo 15-11).

Toplanan *verilerin güvenilirliği* matrisin sol köşesinden başlayıp sağ alt köşesine doğru uzanan köşegen üzerindeki hücrelerde yer alan verilerle değerlendirilir. Bu köşegene "tek özellik-tek yöntem köşegeni" adı verilir.

Ölçüm güvenilirliğinin sağlanabilmesi için bu köşegendeki bütün değerlerin sıfırdan farklı ve yüksek korelasyon katsayısına sahip olması gerekir. Sıfırdan önemli ölçüde farklı değilse ek araştırmalar yapmak gerekir.

Grafikte noktalı çizgilerle gösterilen üçgenler farklı yöntemlerle farklı özellikler arasındaki ilişkilere ait korelasyon değerlerini gösterir. Bu korelasyon katsayılarının bir anlamı yoktur.

Tablo 15-11. Çoklu Özellik - Çoklu Yöntem Matrisi

		Yöntem 1			Yöntem 2			Yöntem 3		
	Özellik	A	B	C	A	B	C	A	B	C
Yöntem 1	X	(G)								
	Y	ayrılma geçerli	(G)							
	Z	ayrılma geçerli	ayrılma geçerli	(G)						
Yöntem 2	X	BİR-LEŞME			(G)					
	Y		BİR-LEŞME		ayrılma geçerli	(N)				
	Z			BİR-LEŞME	ayrılma geçerli	ayrılma geçerli	(N)			
Yöntem 3	X	BİR-LEŞME			BİR-LEŞME			(L)		
	Y		BİR-LEŞME			BİR-LEŞME		ayrılma geçerli	(I)	
	Z			BİR-LEŞME			BİR-LEŞME	ayrılma geçerli	ayrılma geçerli	(R)

Güvenilirlik katsayılarının hemen altında yer alan üçgenlerdeki değerler ise aynı yöntem çerçevesinde farklı özellikler veya yapılar arasındaki korelasyon katsayılarını gösterir. Bu rakamlar ayrılma geçerliliğini ifade eder.

Çoklu özellik - çoklu yöntem matrisi, kullanılan ölçeklerin ve yöntemlerin yapısal geçerlilik derecesini tek bir katsayı halinde vermez. Birden fazla birleşme katsayıları ve birden fazla ayrılma katsayıları vardır. Sadece bu katsayıların ne ölçüde yüksek olduğuna ve ne ölçüde düşük olduğuna bakılır. Korelasyon katsayıları açısından ± ,10 ilâ ± ,39 değerleri düşük; ± ,40 ilâ ± ,69 arasındaki değerler orta ve ± >,70'ten yüksek olan değerler güçlü ilişki olduğunu gösterir.[p] Birleşme geçerliliği için kendi satırındaki ve kendi sütunundaki katsayılar diğer rakamlardan daha yüksek olmalıdır.

[p] Artı işaretleri değişkenlerin aynı yönde artış veya azalış gösterdiklerini, eksi işareti ise değişkenlerin ters yönlerde artış veya azalış gösterdiklerini belirtir.

SPSS'te ÇYÇÖ matrisini hesaplamak için iki benzer ölçeğin verileri ile iki farklı yöntemle toplanan veriler aynı veri matrisine girilir. Daha sonra Statistics – Correlate – Bivariate komutlarıyla çoklu korelasyon analizi yapılır.

Özellik-Yöntem Matrisinin avantaj ve dezavantajları. Bu yöntem birleşme ve ayrılma geçerliliğini tek bir matris üzerinde görme imkanı sağlar. Bu avantajına rağmen literatürde az kullanılmıştır. Bunun nedeni birden fazla özelliği birden fazla yöntemle ölçmek için dikkatli düşünülerek oluşturulacak ve tam olarak birine karşılık gelecek bir matris oluşturma zorluğudur. Campell ve Fiske matrisin eşit sayıda yöntem ve özellik içerme gibi bir zorunluluğa sahip olmadığını belirtmelerine karşılık, özellikle belirli sayıda uygun yöntem bulma konusunda güçlüklerle karşılaşılmıştır.[81]

Model Denkleştirme Yöntemi İle Geçerlilik Analizi Yapılması

Yapısal geçerliliği test etmenin bir başka yöntemi, "model denkleştirme yöntemi"dir (pattern-matching). Bu uygulamada öncelikle analiz sonucunun ne olacağına ilişkin belirli tahminlerden, sezgilerden, kuramdaki bilgilerden veya bunların kombinasyonundan hareket edilerek bir "model" oluşturulur. Model, kavramsal yapılar arasındaki ilişkiler ağıdır.

Daha sonra gerçek uygulama sonucunda elde edilen verilere dayalı olarak nasıl bir modelin ortaya çıktığına bakılır. Gerçek uygulama sonuçları gözlemlere, izlenimlere, alan notlarına veya anket sonuçlarına dayandırılmış olabilir. Öngörülen model ile ortaya çıkan model arasındaki ilişkiler karşılaştırılarak ne ölçüde örtüşme sağlandığı incelenir. Örtüşme sağlandığı ölçüde teorik model bazı pratik sonuçların tahmin edilmesinde kullanılabilir.[82] Model denkleştirme uygulaması korelasyon analizi, *t*-testi ve TYVA gibi anlamlılığı test eden istatistikî analiz yöntemleriyle test edilir.

Model denkleştirme ilkesi, nedensel ilişkiler hakkında karmaşık tahminler yapılmasını gerektirir. Karmaşık modeller, denkleştirme sağlanabilirse daha geçerli sonuçlar sağlar. Model denkleştirme yöntemi, klasik anlamdaki *hipotez test etme* ve *model kurma* yaklaşımından önemli ölçüde farklı değildir. Teorik model, verilerden ne beklediğimize ilişkin olarak belirlediğimiz hipotezlerdir. Gözlemlenen model ise toplanan verilerdir. Hipotezle model denkleştirme yöntemi arasındaki farklılık, model denkleştirme yönteminin araştırmacıyı daha karmaşık hipotezler kurmaya sevk etmesi, gözlem verilerini tek boyutlu bakış açısı yerine çok boyutlu bakış açısıyla değerlendirmeye almasıdır.[83] Model denkleştirme yaklaşımı, araş-

tırmacıya sistematik bir yöntem önermez; karmaşık araştırma modelleri geliştirerek araştırmacıyı bu modellerin veriler tarafından doğrulanıp doğrulanmadığı sorunuyla baş başa bırakır.

Yapısal Geçerlilik Analizinin Aşamaları

Araştırmacı geliştirmiş olduğu test veya ölçek için yapısal geçerlilik analizi yapmak istiyorsa bunu belirli aşamalarda gerçekleştirir. Aşağıdaki bölümde örnek bir uygulamaya ait yapısal geçerlilik çalışmasının aşamaları üzerinde durulmuştur.

1. Öncelikle incelemek istediğiniz kuramsal alanı tanımlayınız. Kuramsal alanda hangi kavramsal yapılar rol üstlenmiştir? İnceleme odağı içindeki kavramsal yapılar ne anlama gelmektedir, ilgili ve benzer kavramlar nelerdir. Kavramsal yapılar üzerinde kimler hangi tür çalışmaları ve araştırmaları yapmıştır. İncelenen kavramsal yapılar daha önce hangi test veya ölçeklerle ölçülmüştür?
2. Ölçmek istediğiniz kuramsal alandaki yapılara ilişkin nomolojik ağı oluşturunuz. Nomolojik ağda sadece ölçek ve testin göstergelerini, boyutlarını değil, kanıtlamak istediğiniz hipoteze ilişkin diğer ölçüm değişkenlerini de ağa alınız.
3. Keşfedici faktör analizi yöntemiyle ölçekle ilgili kavramsal alanın tek boyutlu olduğunu kanıtlayınız. Aynı kavramsal yapıyı ölçen göstergeler birden fazla faktör olduğunu gösteriyorsa yapısal geçerliliğin sağlandığı iddia edilemez. Ayrıca bu gibi durumlarda Cronbach alfa değeri de anlamsızdır.84
4. Kavramsal alan alt ölçek niteliğinde birbirinden bağımsız boyutlardan oluşuyorsa temel boyutlar/faktörler hakkında bilgi veriniz. Kimler hangi boyutları ortaya çıkarmıştır. Literatürde bu konuda mutabakat var mıdır?
5. Bağımsız kavramsal boyutların her biri kaç göstergeyle ölçülmüştür. Her bir boyut altındaki gösterge sayısı yeterli midir?
6. Kavramsal alanın boyutları arasındaki ilişkiler hakkında bilgi veriniz. İlişkilerin oluşturucu veya yansıtıcı olma durumunu belirleyiniz.
7. Kavramsal yapının geçerliliğini belirlemek için kullanacağınız dış kriterler veya testler hakkında bilgi veriniz. Birleşme geçerliliği için hangi testleri ve ayrılma geçerliliği için hangi testleri veya kavramsal yapıları kullandınız?

8. Kavramsal alanın içerik geçerliliği hakkında bilgi veriniz.
9. Ölçek veya testin iç tutarlılığı hakkında bilgi veriniz.
10. Faktör analizi sonuçları hakkında bilgi veriniz.
11. Başlangıçta öngördüğünüz nomolojik ağ ile ampirik araştırma sonucunda ortaya çıkan nomolojik ağı karşılaştırınız ve bilgi veriniz.
12. Eğer geliştirdiğiniz test teşhis amaçlı ise zıt gruplardan elde edeceğiniz sonuçlar hakkında bilgi veriniz.
13. Geliştirdiğiniz testi birden fazla yönteme ait veriler ve birden fazla özellikle desteklemişseniz çoklu özellik - çoklu yöntem matrisini oluşturunuz ve matris sonuçları hakkında bilgi veriniz.
14. Son olarak uyguladığınız değişik yöntem ve tekniklerin ne ölçüde birbirine benzer ve ne ölçüde farklı sonuçlar verdiğini yorumlayınız ve nihaî hükmünüzü veriniz.

Yapısal geçerlilik hiçbir zaman kesin bir şekilde saptanamayacağından elde edilen tüm bulgular geçici niteliktedir. Bu bulguların daha sağlıklı olması yapılacak sonraki araştırmalarla teyit edilmesine bağlıdır. Bu nedenle araştırmacı kesin yargılardan uzak olarak değerlendirme yapmalıdır.

Yapısal Geçerliliği Tehdit Eden Faktörler

Bilim adamı daha araştırmasının başlangıç aşamasında veya ölçüm işlemine girişmeden önce yapısal geçerliliği belirlemeye yönelik bir tasarım yapmamışsa araştırmanın ileriki bölümlerinde zorluklarla karşılaşır. Bir ölçümün yapısal geçerliliği çeşitli faktörlerden etkilenir ve bunlar aşağıdaki gibidir.

1. Kuramsal çerçevenin iyi tanımlanmamış olması.
2. Kavramsal yapıların iyi tanımlanmamış olması.
3. Verilerin sadece tek bir yöntemle toplanmış olması.
4. Sadece kağıt-kalem testlerinin kullanılmış olması.
5. Benzer ve farklı yapıları ölçecek alternatif ölçüm araçlarının kullanılmamış olması.
6. Kişilerin doğal bir şekilde davranmayıp araştırmacının istediği yönde rol oynamaları.
7. Kişilerin iyi, olumlu veya zeki görünmek istemeleri.

8. Kişilerin araştırma sonuçlarını sabote etmek istemeleri.
9. Araştırmacının faktör sayısını literatüre uygun bir şekilde konfigüre etmesi.
10. Kavramsal yapılara ait korelasyon katsayılarının net bir farklılığı ortaya koymaması.
11. Kavramsal yapının kültüre bağımlı olması veya içinde yaşanılan kültürden etkilenmesi.
12. Araştırmacının yapısal geçerliliği sadece tek bir yöntemle kanıtlamaya çalışması.

Yapısal geçerliliği tehdit eden faktörler, aynı zamanda iyileştirme yapmak için önlem alınması gereken etkenleri gösterir. Yapısal geçerlilik, bütün her şey bittikten sonra sorgulanacak bir özellik değildir. Yapısal geçerlilik, araştırma veya ölçüm süreci boyunca ayarlama, düzenleme ve kontrolü gerektiren bir süreçtir.

Yapısal Geçerliliğe İlişkin Olumsuz Kanıtlarla Karşılaşılması

Bilim adamı yapısal geçerlilik konusunda beklediğinin tam tersine olumsuz sonuçlarla karşılaşabilir. Araştırmacının öngörüleriyle ampirik araştırma sonuçları birbiriyle uyuşmuyorsa bu gibi durumlarda sonuçlar üç şekilde yorumlanır.[85]

1. Test veya ölçeğin kavramsal yapıyı ölçmediği kararına varılır.
2. Hipotezin dayandırıldığı teorik ilişki ağının doğru olmadığı kararına varılır.
3. Deneysel tasarımın hipotezin doğru bir şekilde test edilmesi için yetersiz kaldığı sonucuna varılır.

Yukarıda sayılan şıkların üçüncüsünden şüpheleniliyorsa araştırtmacı, sonuçlarını yorumlamadan araştırma tasarımını iyileştirmeye yönelik önlemleri alır. Diğer iki alternatif vakasında ise testinin veya ölçeğin yapısal geçerlilikten uzak olduğu belirtilir. Böyle bir durumda araştırmacı yeni bir test geliştirme veya mevcut testi kalibre etme çabası içine girer.[86]

ARAŞTIRMANIN BİR BÜTÜN OLARAK GEÇERLİLİĞİ

Araştırmanın bir bütün olarak geçerliliği; *ölçüm geçerliliği, iç- dış geçerlilik ve istatistiksel sonuç geçerliliği* alt başlıkları altında incelenir. Sonuç çıkarıcı araştırmalara uygulanabilecek iç ve dış geçerlilik sınıflandırması Cook ve Campbell (1979) tarafından gözlem değişkenleri arasındaki nedensel ilişkileri belirlemek ve örnek kütleden elde edilen verilerin daha büyük ana kütlelere ne ölçüde genellenebileceğini ortaya koymak için geliştirilmiştir.[87]

ÖLÇÜM GEÇERLİLİĞİ

Ölçüm geçerliliği, araştırmada kullanılan değişkenlerin veya göstergelerin incelemeyi amaçladığımız davranışları doğru bir şekilde temsil etme derecesidir. Bir araştırmacı ölçmek istediği kavramsal alanla çok fazla ilgili olmayan bazı değişkenleri de ölçüm aracının içine katmışsa söz konusu anketin ölçüm geçerliliğine sahip olmadığı söylenir. Örneğin, bir araştırmacı kişilerin İnternet'te sörf yapma alışkanlıklarını ölçmek istiyor olsun. Geliştirdiği İnternet Sörf Ölçeği'nde aynı zamanda intranet ile ilgili bazı sorular da varsa anket veya araştırma ölçüm geçerliliğine sahip değildir. Çünkü İntranet İnternet değildir.

İÇ GEÇERLİLİK

İç geçerlilik kavramı; literatürde birkaç farklı şekilde tanımlanmıştır. Bunlardan birincisi "uygun araştırma tasarımının kullanılmasıdır."[88] Bilim adamı incelediği konuya uygun olarak deneysel tasarım, yarı deneysel tasarım, alan araştırması tasarımı, niteliksel araştırma tasarımı veya tarihsel araştırma tasarımı gibi yöntemlerden birini doğru olarak seçmiş ve bu yöntemin gereklerine tam olarak uymuşsa çalışması iç geçerliliğe sahiptir.

İkinci tanıma göre iç geçerlilik araştırılan veya ölçümü yapılan olgunun değişkenleri arasındaki nedensel ilişkilerin gerçeği yansıtma derecesidir. Bu tanıma göre iç geçerlilik daha çok bağımlı ve bağımsız değişkenler arasındaki nedensel ilişkilerin araştırıldığı deneysel bazlı araştırmalar için uygundur. Araya giren başka faktörlerin etkisi olmaksızın sadece ileri sürdüğümüz nedenler dolayısıyla tespit ettiğimiz sonuçlar ortaya çıkıyorsa araştırma iç geçerliliğe sahiptir.

Üçüncü tanımda ise araştırma uygulaması yerine ölçüm aracı üzerinde durulmuştur. Ölçüm aracının madde puanları ile ölçeğe ait bileşik puan arasında belirli ölçüde bir ilişki varsa ölçek iç geçerliliğe sahiptir.[89] Bu son tanımda iç tutarlılık aynı zamanda iç geçerlilik olarak görülür.

Bir araştırmada iç geçerlilik dış geçerlilikle bütünleşmelidir. İç geçerlilik sağlandığı halde sonuçlar ana kütleye genellenemeyebilir. Mikro nitelikte, tümevarımsal (endüktif)[q] ve niteliksel araştırmaların iç geçerliliği daha yüksektir. Çünkü bu araştırmalarda nüanslara önem verilir ve olgu daha ayrıntılı bir şekilde ele alınır.[90] İç geçerliliği, ölçüm koşulları etkiler. Alan araştırmalarının dış geçerliliği yüksek iken, bu araştırmalarda iç geçerliliğin sağlanması bir ölçüde risklidir. Tıp bilimlerinde iç geçerlilik, uygulanan tedavi yöntemiyle elde edilen sonuçlar arasında bir ilişki olduğunun kanıtlanmasıdır. Bu ilişki üzerinde, "ölçülmeyen veya kontrol altına alınamayan diğer faktörlerin etkisinin olmaması" ve ilişkinin sadece "tedavi yöntemiyle ilgili olduğunun ortaya konması" (nedensellik) iç geçerliliğin yüksek olduğunu gösterir.[91] Campell (1967); Calero, Piattini ve Genero'nun (2004) yaptığı sınıflandırmalar ile diğer kaynaklardan yararlanarak iç geçerliliği etkileyecek faktörleri aşağıdaki gibi belirleyebiliriz.[92]

1. Ara değişkenlerin etkisi. Elde edilecek sonuçları etkileyeceğinden tedavi / müdahale değişkenlerinin dışında kalan diğer ara değişkenlerin etkisi kontrol altında tutulmalıdır.
2. Tarihsel etki. Bağımlı değişkenle bağımsız değişkenler arasındaki ilişkiler farklı iki zaman diliminde araştırıldığında, ikinci zaman diliminde yapılan ölçüm sonuçlarının aradan geçen süre içinde ortaya çıkan bir çok olaydan etkilenmediği veya etkilenmemesi garanti altına alınmalıdır.
3. Kişiler arasındaki farklılıklar. Ölçüme alınan tüm kişiler benzer deneyimlere ve özelliklere sahip olmalıdırlar.

---

[q] Tümevarımsal incelemelerde araştırmacı ayrı ayrı gözlem sonuçlarının onu genel bir yargıya veya kurama götüreceği düşüncesinden hareket eder. Niteliksel bir araştırmada, belki sadece tek bir işletmede veya birkaç işletmede inceleme yapılmıştır. Bu işletmelerden elde edilen sonuçların çok daha büyük sayıdaki işletmelere genellenebilmesi, daha sonraki yıllarda aynı konuda yapılacak başka araştırmaların sonuçlarıyla doğrulanmasını gerektirir. Bu nedenle *kesit araştırmalarında* çok sayıda örneklem üzerinde ölçüm yapılmadığı sürece iç geçerlilik yine de düşük kalır.

4. Seçim etkisi. Deney ve kontrol grupları içinde ölçüm işlemine tâbi tutulacak kişilerin bu gruplara tesadüfî olarak atanmamalarıdır. Ölçümde iradî belirleme, seçim etkisi yaratır.
5. Olgunlaşma etkisi. Bağımlı değişkendeki gelişmeler deney ve kontrol gruplarına ait sonuçlar karşılaştırıldığında normal gelişme sürecinin bir sonucu olarak ortaya çıkmış olmalıdır.
6. Ölçüm şemaları arasındaki farklılıklar. Araştırmacı farklı şemalarla çalışarak tedavi / müdahale yöntemiyle sonuçlar arasındaki ilişkileri araştırabilir. Kullanılan şema farklılıkları bir şekilde ölçüm sonuçlarını etkileyebilir.
7. Zamanla ilgili kayıtların tutulmasındaki hassasiyet. Test uygulamasında zamanla ilgili kayıtların araştırmacı yerine ölçümü yapılan kişiler tarafından tutulmasının daha etkili olduğu bilinmektedir. Ancak bu uygulamada kişilerin hepsi zaman çizelgelerinin doğru bir şekilde tutulmasına aynı hassasiyeti göstermeyebilir.
8. Öğrenme etkisi. Testlerin farklı kişilere farklı sıralarda uygulanarak öğrenme etkisinin önüne geçmek gerekir.
9. Yorgunluk etkisi. Kişilere uygulanan testler belirli bir süreyi aştığında yorgunluk etkisi yaratır.
10. İzlenim etkisi. Kişiler bir dizi test veya deneyime tabi tutulduklarında önceki dönemlerde almış oldukları diğer test ve deneyimlerden etkilenmemiş olmalıdırlar. Yakın zamanda alınmış benzer bir test veya uygulama izlenim etkisi yaratır. Literatürde buna aynı zamanda "test etkisi" adı verilmiştir.
11. Araç etkisi. Ölçüm uygulaması sırasında ölçüm yöntemi değiştirilmemeli, hep aynı yöntem uygulanmalıdır. Ayrıca ölçüm aracında ortaya çıkan gelişmelere göre kalibrasyon çalışması yapılmamalıdır.
12. Bulaşma etkisi. Ölçüm uygulaması deney ve kontrol grubu şeklinde yapılıyorsa kontrol grubundaki kişilerin araştırma hakkında hiçbir şey bilmemeleri, anlamamaları ve hissetmemeleri gerekir. Her iki gruptan herhangi birinin araştırmanın veya ölçümün başarılı olması veya başarısız olması gibi bir gayretleri varsa bulaşma etkisi ortaya çıkar.
13. Kişi motivasyonu. Kişilerin sonuçların ne şekilde kullanılacağına ilişkin düşünceleridir. Sonuçların yöneticileri tarafından kullanılacağını bilmeleri, isimlerinin belli olması sonuçları olumsuz etkiler.

14. Etkileşim etkisi. Ölçüm yapılan kişilerin kendi aralarındaki etkileşimleri, birbirlerinden ölçüme ilişkin fikir almaları ve birbirlerini yönlendirmeleri ölçüm sonuçlarını etkiler. Buna aynı zamanda bulaşma etkisi adı verilir.
15. İstatistiksel regresyon. Ortalamaya çekilme davranışını ifade eder. Bir ölçüm uygulamasında belirli bir gruptaki kişiler testten ekstrem puanlar almışlarsa bu ekstremiteyi gidermek için söz konusu kişilere yeniden test uygulaması yaparak ekstrem puanları ortalamaya doğru çekmekle daha doğru bir ölçüm yapılmış olacağına inanılır. İstatistiksel regresyon, bir değişkenin genel aritmetik ortalama değeriyle bir veya daha fazla bağımsız değişkene göre aritmetik ortalama değerleri arasında fark görülmesi ve bu farklılığı azaltma çabasıdır. Ancak burada araştırmacı "ekstrem puanların gerçeği yansıtmadığı" düşüncesiyle hareket ederken bu iddiasından yüzde yüz emin olmaz. Ekstrem puanlar değil, belki de genel ortalamanın düşüklüğü / yüksekliği soruşturulabilir bir niteliğe sahiptir. Genel toplam puanların düşüklüğü veya yüksekliği örneklemdeki bireylerin daha az olgun tutum ve davranışlara yönelmelerinin sonucu olabilir.93 Araştırmacı ölçüm değişkenine ait genel aritmetik ortalama değerini daha yüksek veya daha düşük bir değere çekme çabalarından vazgeçmelidir.
16. Ölüm etkisi. Uzun süren araştırmalarda üzerinde ölçüm yapılan kişilerden önemli sayıda kişinin ölmesi ve sonuçların bu nedenle gerçeği tam olarak yansıtmamasıdır.

Bilim adamının araştırmasının iç geçerliliğini artırmak için başvuracağı birkaç yol vardır. Bunlardan birincisi iç geçerliliği tehdit eden faktörlerin etkisini mümkün olduğunca azaltmak veya bu tür bir etkinin doğmasına hiç meydan vermemektir. İkincisi meslektaş veya danışman öğretim üyesi değerlendirmesinden yararlanmaktır. Üçüncüsü, hedef ana kütleden seçilecek bir örneklem üzerinde pilot araştırma yapmaktır. Pilot araştırma sonuçları hem ölçüm aracındaki hem de ölçüm uygulamasındaki aksaklıkları görme imkanı sağlar.

DIŞ GEÇERLİLİK

Dış geçerlilik, "neden-sonuç ilişkilerinin kurulduğu" ölçüm sonuçlarının diğer kişilere, yerlere, düzlemlere ve zamanlara genellenebilmesidir. Genelleme spesifik bir ana kütleye yönelik olabilir veya çok daha geniş ola-

rak elde edilen sonuçlar bütün ana kütlelere genellenebilir.[94] Bu nedenle, dış geçerlilik kısaca genellenebilirlik olarak tanımlanır. Sonuçlar, sadece örnekleme katılan kişiler için bir anlam ifade ediyorsa dış geçerlilik koşulu sağlanamamıştır Dış geçerliliği sağlamak için aşağıdaki faktörlere dikkat edilir.

1. Coğrafî alan sınırları. Ölçüm sonuçları dar bir coğrafi alanın sınırlarına dayanıyorsa daha büyük ana kütlelerle genellenemez.
2. Hedef ana kütle. Ölçüm sonuçlarını genellemek istediğimiz ana kütlenin kimlerden veya hangi kesimlerden oluştuğu net bir şekilde tanımlanmalıdır.
3. Örneklem çerçevesi. Hedef ana kütledeki kişilere ulaşmamızı sağlayacak ve örneklem planı oluşturmak için kullanabileceğimiz, listesi çıkarılabilecek kişilerin kapsamı veya sınırları tam olarak belirlenmelidir.
4. Örneklem büyüklüğü. Genelleme yapabilmek için örneklem hacmi istatistiksel olarak belirli bir büyüklüğe sahip olmalıdır. Örneklem büyüklüğü hata payı ve alfa güvenilirlik düzeyinde belirlenir.
5. Örnekleme yöntemi. Elde edilen sonuçları genelleyebilmek için tesadüfi örnekleme yöntemi uygulanmalıdır.
6. Örneklemin temsil edicilik özelliği. Ölçüm yapılan kişiler genelleme yapılacak ana kütleyi temsil etmelidir. Sadece belirli kişilerin örnekleme alınması "seçim yanlılığı" olgusunu ortaya çıkarır.
7. Ana kütleden seçilen örneklem sayısı. Elde edilen sonuçları genelleyebilmek için mümkün olduğu kadar birden fazla örnek kütle üzerinde çalışma yapılır.
8. Örneklemde yapılan ölçüm sayısı. Araştırma birden fazla zamanda tekrarlanarak aynı sonuçların elde edilip edilmediğine bakılır.
9. Veri örneklemi. Seçilen örneklem büyüklüğünden toplanan verilerin ne kadarının kullanılabilecek türde olduğunun incelemesi yapılır. Veri örneklemi yetersizse ikâme yöntemi araştırılır.
10. Deneysel düzenlemelerin reaktif etkisi. Sonuçlar, kişilerin gerçek durumunu yansıtmak yerine deney seriminin kişiler üzerinde yarattığı etkiden kaynaklanmış olabilir ve bu nedenle sonuçlar daha bü-

yük örnek kütleye genellenemez. Hawthorne ve John Henry etkisir tipik reaktif etki örneklerdir.

11. Veri analizi için seçilen istatistiksel testlerin uygunluğu. Araştırmada ölçüm verileri arasındaki ilişkileri belirlemek için uygun olmayan istatistiksel yöntemler uygulanmışsa sonuçlar yanlıştır ve bu nedenle ana kütleye / ana kütlelerle genellenemez.
12. Kullanılan malzemeler, araçlar veya görevler. Ölçüm amacıyla kullanılacak malzeme, araç ve görevler gerçek vak'aları temsil etme özelliğine sahip olmalıdır.
13. Zaman dilimi. Özellikle bilişsel testlerin geçerliliği çok daha uzun zaman diliminde çok sayıda örneklem üzerinde yapılan araştırmalarla belirlenir. Zamana genelleme farklı zamanlarda yapılacak ölçümler sonucunda ortaya çıkar.
14. Ön test-tedavi etkileşimi. Deneysel araştırmalardaki ön test uygulaması üzerlerinde tedavi veya müdahale uygulaması yapılacak kişileri aşırı biçimde duyarlı hale getirir ve daha sonraki ölçüm sonuçlarını etkiler.95
15. Seçim-tedavi / müdahale etkileşimi. Deneysel araştırmalarda ölçüme katılacak kişilerin rasgele seçilmemeleri ölçüm sonuçlarını etkiler ve elde edilen sonuçlar ana kütlelerle genellenemez.96
16. Çoklu tedavi/müdahale uygulamaları arasındaki etkileşim. Katılıcılara birden fazla müdahale yapılmışsa veya birden fazla tedavi yöntemi uygulanmışsa önceki tedavi uygulamalarının sonuçları sonraki tedavi / müdahale uygulamalarının sonuçlarını etkiler ve bu nedenle nihaî sonuçların genellenebilirlik özelliği zayıflar.97

Bilim adamları dış geçerliliği, kendi içinde ayrıca iki alt bölüm halinde ele almışlardır: Ekolojik geçerlilik ve ana kütle geçerliliği.

Ekolojik geçerlilik. Ekolojik geçerlilik (ecological validity), "laboratuar ortamında" veya "araştırma yapılan düzlemde" elde edilen davranışsal sonuçların doğal ortamda, gerçek dünyada, işletmelerde veya diğer düzlemlerde ne ölçüde anlamlı olduğu konusuyla ilgilidir. Laboratuar ortamında elde edilen sonuçlar gerçek dünyada da elde edilebilecek türden ise,

---

[r] Kontrol grubundaki kişilerin deney grubundaki kişilerden daha yüksek performans göstermeleri.

gerçek dünyaya genellenebilecekse araştırma sonuçları ekolojik geçerliliğe sahiptir. Laboratuar ortamında iki değişken arasındaki ilişkiler incelenirken ilişkinin niteliğini bozan bir çok faktör kontrol altında tutulur. Oysa gerçek dünyada bu faktörlerin etkisini kontrol altına almak çok zordur. Ekolojik geçerlilik sonuçların pratik hayatta kullanılabilmesini ve yararlı olmasını ifade eder. Araştırmacı ekolojik geçerliliği belirlemek için şu soruları sorar:[98]

1. Ölçüm sonuçları, doğal ortam için olağan dışı bir niteliğe sahip mi?
2. Elde ettiğimiz sonuçlar yapay (sun'î) mı?
3. Elde ettiğimiz sonuçların görünür gerçekle ilgisi ne?
4. Elde ettiğimiz sonuçlar başka ortamlarda kullanılabilir mi?
5. Elde ettiğimiz sonuçlar gerçek hayattaki kişilerin, grupların davranışlarına ne ölçüde uygun?

Ekolojik geçerliliği düşük olan çalışmalardan elde edilen sonuçlar içinde bulunulan ortamın dışına genellenemez. Ölçüm yapılan ortam; bir laboratuar, bir işletme, işyeri veya atölye olabilir. Geniş alan araştırmalarının (survey) ekolojik geçerliliği genellikle yüksektir. Araştırmanın ekolojik geçerliliği değerlendirilirken deneysel çalışma çok boyutlu olarak ele alınır. Deney yapılan düzlem, inceleme altında tutulan uyaran ve gözlemcinin tepkisi birlikte ele alınıp değerlendirilmelidir. Ekolojik geçerliliğin standart bir uygulaması yoktur.

Ana kütle geçerliliği. Ana kütle geçerliliği, bulguların diğer ana kütlelere ne ölçüde genellenebileceği konusuyla ilgilidir. Ana kütle-örneklem seçim prosedürlerine tam olarak uyulmuşsa araştırma bulguları ana kütle geçerliliğine sahiptir. Araştırmaya ait örneklem verileri, hedef ana kütle yerine "ulaşılabilir ana kütleden" seçmişse ulaşılabilir ana kütle sonuçlarını hedef ana kütleye genellemek riskli olur.[99] Araştırma sonuçları belirli bir ana kütleden seçilen örneklem verilerine dayanıyorsa ana kütle geçerliliğini artırmak için aynı ana kütleden başka örneklemler seçilerek ölçüm işlemi bu örneklemler üzerinde de tekrarlanır. Gerektiğinde başka ana kütlelerden seçimler yapılarak benzeri sonuçların elde edilip edilmediğine bakılır.

İSTATİSTİKSEL SONUÇ GEÇERLİLİĞİ

İstatistiksel sonuç geçerliliğinde, değişkenler arasında kurulan ilişkilerin uygun istatistiksel analiz uygulamalarının sonuçlarına dayanıp dayanmadığı araştırılır. İlişkiler uygun istatistik testlerle kanıtlanmış ve sonuçlar doğru yorumlanmışsa bulgular *istatistiksel sonuç geçerliliğine* sahiptir. İstatistiksel sonuç geçerliliği istatistikçiler, meslektaşların gözden geçirme çalışmaları ve araştırma yöntem bilimcileri tarafından değerlendirilir. İstatistiksel sonuç geçerliliğinde iki değerlendirme yapılır. Birincisi kullanılan testin, ikincisi yapılan yorumun doğru olup olmadığıdır. Kurulan neden sonuç ilişkilerinde değişkenler gerçekten birlikte değişiyorsa, istatistik teknik yeterince güçlü ise sonuçların doğru olduğu kararına varılır. İstatistiksel sonuç geçerliliğini tehdit eden faktörler aşağıdaki gibidir:[100]

1. İstatistiksel gücün düşük olması. (İstatistiksel güç daha araştırmaya başlanmadan önce planlanmalıdır. İstatistiksel gücü azaltan en önemli etken, örneklem hacminin küçük olmasıdır. İstatistiksel gücü belirleyen diğer faktörler iki değişken arasındaki ilişkinin gücü, belirlenen alfa anlamlılık düzeyi –,05'e göre ,01 çalışmanın istatistiksel gücünü azaltır– ve etki büyüklüğüdür.)
2. Ölçek veya test sonuçlarının güvenilir olmaması. (Ölçek veya test sonuçları farklı zamanlarda hep aynı sonuçları veriyor olmalıdır.)
3. Örneklem verilerindeki ranj kısıtlaması. (Örneklem verileri normal dağılım özelliğine sahip olmalıdır.)
4. Cevaplayıcıların heterojen olması (bağımlı değişken üzerindeki etkiler araştırıldığında cevaplayıcıların homojen nitelikte olması gerekir).
5. Müdahale uygulamalarının güvenilir olmaması. (Müdahale ve tedavi uygulamaları istikrarlı ve standart bir şekilde yapılmamışsa farklılıkları ortaya çıkarma ihtimali azalır. Uygulamadaki değişkenlikler ölçüm hatasını artırır ve gerçek farklılıkları yakalama şansını azaltır.)
6. Yanlış bir testin seçilmiş olması. (Araştırmacı verilerin niteliğine uygun bir istatistik test yöntemini seçmiş olmalıdır.)
7. Test istatistiklerindeki varsayımların veya ön kabul şartlarının yerine getirilmemiş olması. (Test varsayımları ihlal edilmişse veya karşılanamamışsa elde edilen bulgular doğru değildir.)

8. Test sonuçlarının yanlış yorumlanması. (Test sonuçlarının istatistiksel olarak anlamlı olması, gerçek hayatta da iki değişken arasında farklılık olacağı anlamına gelmez.)
9. Aynı veriler üzerinde küçük değişikliklerle birden fazla istatistiksel test yapıldığında Tip I hatasının ortaya çıkmasıs (TipI hatası var olmayan ilişkinin var imiş gibi gösterilmesidir. Bu hatayı düzeltmek için Benferroni düzeltmesi yoluna başvurulur).
10. Birden fazla test uygulamalarında, belirlenen alfa anlamlılık düzeyinde düzeltme yapılmamış olması. (Araştırmacı alfa düzeyini ,05 olarak belirlemiş ve bağımlı değişkeni 10 farklı örneklemde analiz etmişse alfa değerini düzelterek ,05/10 = ,005 şeklinde belirlemelidir.)

İstatistiksel sonuç geçerliliğinde araştırmacı "sinyal" ile (ilişkilerin gerçek gücü) "gürültü" (ilişkiyi bozucu etkenler) arasındaki ilişkileri netleştirmeye çalışır. Sinyalin gücüne oranla gürültünün etkisi çok daha az olmalıdır. Gürültünün etkisi fazlaysa araştırma sonuçlarının *istatistiksel gücü* yoktur.[101] İstatistiksel güç, tedavi veya müdahale etkisinin gerçek anlamda ortaya çıkma oranıdır. İstatistiksel güç; örneklem büyüklüğü, etki büyüklüğü ve alfa anlamlılık düzeyinin fonksiyonu olarak tanımlanmıştır.

GEÇERLİLİK ANALİZİ SORUNLARI

Geçerlilik analizlerinin yargısal ve istatistiksel alanın her ikisini de kapsaması nedeniyle içeriği güvenilirlik analizlerine göre hem daha geniş hem de daha sorunludur. Bilim adamı geçerlilik analizi sonuçlarını raporlarken bir dizi faktörü göz önünde bulundurmalıdır. Aşağıdaki bölümde bu faktörler üzerinde durulmuştur.

---

[s] Literatürde bu uygulamaya "balık tutma" adı verilmiştir. Araştırmacı topladığı verileri biraz farklı ön kabuller ve koşullarda birkaç kez teste tabi tutarak istediği sonucu alabilir. Gerçekte iki değişken arasında bir ilişki olmadığı halde bazı uyarlamalar yapılarak test sayısının artırılmasıyla istatistiksel olarak anlamlı bir ilişki varmış gibi bir sonuç elde edilir. Sosyal bilimlerde araştırmalar ,05 güven aralığında yapılır. Bunun anlamı her 100 araştırmadan elde edilecek sonuçların beş tanesinde aradığımız ilişkinin tesadüfen ortaya çıkmasıdır. Literatürde bu olgu "balık tutma ve hata oranı problemi" (fishing and the error rate problem) olarak isimlendirilmiştir. Araştırmacı bazı düzeltmelerle birden fazla test yaparak istediği "balığı" tutabilir. Oysa bu gibi durumlarda araştırmacının hata oranından kaçınmak için test veya örneklem sayısını dikkate alarak düzeltme formülünü uygulaması gerekirdi.

KURAMLA UYUŞMAMA SORUNU

Geçerlilik analizlerinden yapısal geçerlilik, başlangıçta öngörülen boyutlarla gerçek hayattaki ampirik araştırma sonuçları arasında paralellik olmasını sağlamaya çalışır. Fakat her zaman ampirik araştırma sonuçları başlangıçta öngörülen ilişkileri doğrulamayabilir. Bu gibi durumlarda genellikle ampirik ölçüm verilerinin arka planda yatan kuramsal çatıyı desteklemediği ve verilerin yapısal geçerliliğinin olmadığı yorumu yapılır. Ülkemizde bu konuda karşılaştığımız önemli sorunlardan biri öncül kuramsal boyutların yabancı ülkelerde yapılan araştırma sonuçlarına dayandırılıyor olmasıdır ve bu boyutlar bir çok kez ülkemizde yapılan ampirik sonuçlarla örtüşmez. Kavramsal yapıların kültüre bağımlı olabileceği gerçeği dikkate alınmadığında kuram-ampirik sonuç uyuşmazlığı iddiası tutarlı değildir.

Ölçümü yapılan kavramsal yapı kültürden bağımsız ise ve yine buna rağmen kuram-ampirik sonuç uyuşmazlığı ile karşılaşılmışsa araştırmacının bu kez arka plandaki kuramsal çerçeveden, araştırma veya ölçüm uygulaması yönteminden veya belirlediği ölçüm değişkenlerinden şüphelenmesi gerekir. Araştırmacılara iyileştirmeyi sağlayacak somut bir çözüm önerisi getirmek zordur. Bu gibi durumlarda araştırmacı çok yönlü çalışmalarla her bir modeli, kuramsal yapı çerçevesini, uyguladığı yöntemi, ölçüm aracına aldığı değişkenleri, temel boyutları tek tek gözden geçirmeli duyarlı bir çalışma yaparak elde ettiği sonuçları okuyucusuyla paylaşmalıdır.

ZAMAN İÇİNDE FARKLI DEĞERLER ELDE ETME SORUNU

Ölçüm aracının belirli bir zaman boyutunda geçerli olmasıyla uzun zaman boyunca geçerli olması iki farklı olgudur. Psikometrik testlerin on veya yüz yıllarla ifade edilen bir sürede hep geçerli olması istenir. Tutum ölçeklerinin ise genellikle çok daha kısa zaman dilimindeki geçerliliği ön plana çıkar. Ancak tutum ölçeklerinde bile makul uzunlukta bir süre içinde geçerli sonuçlar alınması o ölçüm aracının güçlü olduğunu ortaya koyar. Bir ölçüm aracının zaman boyutunda geçerli olması insanlar arasında dönemsel karşılaştırmalar yapmaya imkan verir. Dönemsel geçerlilik aynı zamanda kavramsal yapının değişkenlik özelliğiyle de bağıntılıdır. Kavramsal yapının içeriği sık değişiyorsa, aynı ölçek daha sonraki dönemlerde kullanılamaz. Örneğin, günümüzde zaman zaman dile getirilen "Türk tipi demokrasi" kavramına dayalı olarak geliştirilen bir ölçeğin geçerlilik değerlendirmeleri 10 yıl sonra bütünüyle iptal edilmiş olabilir.

ÜLKESEL VE BÖLGESEL FARKLILIK SORUNU

Farklı ülke ve bölge verileri üzerinde sınandığında yapısal geçerlilik analizlerinden farklı sonuçlar elde edilebilir. Bu farklılığın istatistiksel olarak anlamlı olması, değişik ülke ve bölgelerdeki insanların kafalarındaki kavramsal şemaların da farklı olduğunu yansıtır. Buna göre geçerlilik analizi sonuçları ülkeler ve bölgeler arasında değil, içinde bulunulan ülkedeki ve/veya bölgedeki ana kütleler için geçerlidir. Geniş topraklara sahip ülkelerdeki kültürel, iklimsel, idarî coğrafya alanları bölge tanımlaması yapmak için kullanılabilir. İtaat, özgürlük, zeka, takım çalışması, değerler, liderlik gibi kavramsal yapılar farklı kültürlerde farklı boyutlar altında incelenir ve bu nedenle bir kültür için belirlenmiş olan boyutlar bir başka kültüre tam olarak uygun düşmeyebilir.

ÖZEL AMAÇLI ÖLÇÜMLERDE GEÇERLİLİK ANALİZLERİ

Bu bölümde kısaca özel amaçlı ölçümlerde yapılabilecek geçerlilik analizlerine değilmiştir. Özel amaçlı ölçümler, işletmelerde yapılan durum değerlendirme araştırmaları, personel memnuniyeti araştırmaları, müşteri memnuniyeti araştırmaları, personel seçimi amacıyla yapılan ölçümler, performans değerlendirme ölçümleri, değer ve kişilik ölçümleridir. Bu tür ölçüm ve araştırmalarda bilim adamı ön inceleme yapmalı ve geçerlilik açısından hangi yöntem ve yaklaşımı uygulayacağına dikkatli bir literatür taramasından sonra karar vermelidir. Kitabın daha sonraki baskılarında genişletmeyi düşündüğümüz bu bölümdeki bilgiler okuyucu tatmin etmeyebilir. Bu nedenle bu konularda ek araştırma yapılması önerilir.

ÖRGÜTLERDE YAPILAN ALAN ARAŞTIRMALARINDA

Örgütsel alan araştırmaları, örgütün değişik faktörler açısından değerlendirilmesi ve söz konusu faktörleri iyileştirilmeye yönelik belirli önlemlerin alınması amacıyla yapılır. Örgüt değerlendirme çalışmalarında kavramsal yapılardan çok, ölçüm sürecinin geçerliliği önem kazanır. Kendilerine anket uygulanan çalışma gruplarına veya tüm personele anketlerin uygun bir zamanda verilmesi, uygun süre tanınması, uygulamanın belirli bir disiplin içinde yapılması, sorunların doğru bir şekilde tespit edilmesi önemlidir. Geçerlilik çalışması için kullanılan anket formlarının soruları üzerinde değil, daha çok süreç üzerinde odaklanılır.[102] Bu tür araştırmaların yapılma

amacı; sorunların, yetersizliklerin ve aksaklıkların tespit edilmesi ve iyileştirme önerilerinin geliştirilmesidir. Örgütsel alan araştırmalarında oluşturucu ölçekler veya maddeleştirilmiş Likert tipi ölçekler kullanılır.

PERSONEL SEÇİMİ PROSEDÜRLERİNDE

Personel seçimi prosedürleri değişik uygulamaları içerir. Personeli işe alma, nakletme, ilerletme, terfi ettirme, tenzil ettirme, tavsiye etme, işte alıkoyma, işten çıkarma ve eğitim uygulamasına gönderme bir şekilde seçim işlemini gerektirir. Personel seçimi prosedüründe iş gerekleri, öz geçmiş, mülakat, performans testleri, kağıt-kalem testleri ve değerlendirme merkezi gibi yaklaşımlar kullanılır. İstatistikler, yetenek ve mizaçlarına göre doğru bir şekilde seçilmiş kişilerin işlerinden büyük ölçüde memnun olduklarını ve daha üretken davranışlar sergilediklerini ortaya koymuştur. Bu açıdan personel seçim sürecinde, geçerlilik analizi yapılmış ve sonuçların geçerli olduğu saptanmış testlerin kullanılması gerekir. İş için müracaat eden adayların hazırlamış olduğu CV'ler gerçeği tam olarak yansıtmıyor ve yapılan mülakatlarda da değerleyiciler sübjektif yargıların etkisinde kalıyor olabilirler. Personel seçiminde başlıca dört tür geçerlilik uygulanır ve bunlar aşağıdaki gibidir.

İçerik geçerliliği. Personel seçimi amacıyla kullanılan testlerin işin kapsamıyla ilgili olmasıdır. Bu amaçla önce işin içeriğindeki kritik hareketler, yetenekler ve beceriler belirlenir. Bu hareketler testlerin ögelerinin belirlenmesinde temel alınır. Kullanılacak testlerdeki ögelerin veya maddelerin bu davranışları, bilgiyi, becerileri ve yetenekleri gerçek anlamda *temsil etmesi* gerekir. Temsil olgusu "yaklaşık olarak öyledir" anlayışıyla değil pratik uygulamalar ve araştırmalar yaparak belirlenmelidir. İçerik geçerliliğinin belirlenebilmesi için öncelikle iş analizinin yapılması gerekir. İş analizi yapıldıktan sonra görev ve sorumluluklar, gerekli olan davranışlar, bilgi, yetenek ve beceriler belirlenir (BYB). İş analizinde davranışlar iki grupta incelenir. Gözlenen davranışlar ve gözlenmeyen davranışlar. Bilgi, yetenek ve beceriler davranış değildir. Bilgi, yetenek ve beceriler davranışları etkilemekle birlikte kendileri somut davranış olarak değerlendirilmez. Bilgi başarılı bir iş ortaya koymak performans göstermek için zorunlu değildir. Bilgi testlerinin test bataryasında yer alması için şu sorunun sorulması gerekir. "Personel bu bilgileri iş yaşamında gün be gün kullanacak mıdır?" Günlük iş yaşamında başarılı olması için bu bilgileri sürekli olarak hatırlaması gerekecek midir?" Personelin davranışlarını ve hareketlerini

genel olarak yöneten davranışlarını olgunlaştıran bilgileri test bataryasına almaya ve içerik geçerliliği yapmaya gerek yoktur. Eğer bilgi testi yapılacaksa test maddeleri bilgi alanını aynen içermelidir. Davranışlar hareketlerdir. Fakat belirli koşullarda bilgi yetenek ve becerilerin de içerik geçerliliği kapsamında düşünülebileceği belirtilmiştir. İkinci aşamada iş analizi ile belirlenen yeteneklere uygun testler belirlenir. Böylece her bir testin başlığı iş gruplarıyla uyumlu ve tutarlı hale getirilir. Ancak literatürde bu yaklaşım "semantik geçerlilik" adı verilerek eleştirilmiştir.[103] Bu yaklaşım ne işin içeriğine ilişkin yeterli ölçüde bilgi vermekte ve ne de yüzey geçerliğinin yeterli koşullarına sahiptir. İsim vermekle belki en fazla yüzey geçerliliğinden söz edilebilir.

İçerik geçerliliği, bir testin ölçüm amacı veya fonksiyonudur. Daktilo etme, bir makineyi kullanabilme, araç kullanabilme, fiziksel performans göstergeleridir. Psikoteknik testlerin içerik geçerliliği testin türüne göre değişir. Fiziksel performans testlerinde hareketler ve güç ön plana çıkarken, zihinsel ve sözel testlerde içerik geçerliliğini yapmaya gerek yoktur. Ayrıca personelin iş başında öğreneceği bilgi, beceri ve yetenekler için de içerik geçerliliğini uygulamaya gerek yoktur.[104] Ölçek geliştirilirken özellik veya kavramsal yapılar içerik analizi ile belirlenmeye çalışılırken gözle görülmeyen yapılar için içerik geçerliliği analizleri yapılmaz. Yapı olarak nitelendirilen ölçüler aslında gözlemlenebilir davranışlardır. Öte yandan bir eğitim programı sonunda personelin öğrenme durumunu belirlemeye yönelik olarak düzenlenen testlerde içerik geçerliliği analizleri uygulanır. Çünkü testin bütün konuları yeterince kapsaması gerekir.

Yapısal geçerlilik, aynı özelliği/yapıyı ölçen farklı test sonuçlarının birbirleriyle yüksek korelasyona sahip olmasıdır. Yapısal geçerliliğin bir türü özellikle kağıt-kalem testlerinde uygulanabilir. Örneğin, bireyin kişiliğiyle ilgili olarak test maddelerinin arka planında onun dürüstlüğünü, dışa dönüklüğünü, girişkenliğini ölçen faktörlerin bulunduğu iddia ediliyorsa bunu kanıtlamaya yönelik istatistikî analizlerin yapılmış olması gerekir. İçerik ve yapısal geçerlilikten önce test uygulanacak işler için tam bir iş analizinin yapılması gerekir. İş analizi gözlenebilir davranışlar temel alınarak yapılır. İş analizinde gerekli olan davranışlar, üretim oranı, hata oranı, tepki sürati gibi sonuçlara ilişkin göstergeler belirlenir. Uygun işlerde iş analizi yapılmadan *personel performans puanları* da kullanılabilir, ancak performans değerlendirmesinin dikkatli bir şekilde yapılmış, standardize edilmiş ve puanlarının dikkatli bir şekilde hesaplanmış olması gerekir.

Birlikte vuku bulma geçerliliği, belirli bir endüstride, belirli bir firmada ve belirli bir işte çalışan başarılı bir kişiye test uygulandığında elde edilen puanların, o endüstride, benzer işletmede ve benzer işlerdeki ilk %25'lik dilimde yer alan kişilerin puan ortalamalarıyla benzerlik gösterip göstermediğidir.[1] Personelin işe yerleştirilmesinde, eğitim fırsatlarından yararlandırılmasında ve terfi ettirilmesinde çoğunlukla ilk %25'lik başarı dilimi kullanılır. İkinci bir uygulama biçimi ise kişiye belirli bir zaman diliminde test uygulanır ve aynı zamanda bu kişilerin üstlerine performans değerlendirmesi yaptırılır. İkisi arasındaki korelasyon yüksekse birlikte vuku bulma geçerliği de yüksek denir. Birlikte vuku bulma geçerliliği aday işgörenlere değil, çalışanlara uygulanır.

Birlikte vuku bulma geçerliliğinin dezavantajı aday işgörenlerin durumu hakkında tam bir bilgi vermemesidir. Çünkü aday işgörenler iş bulma kaygısı içinde olacağından testi aldıkları puanları daha farklı olabilir. Bu yöntemin bir diğer sakıncası ise, işletmede çalışanların benzer özelliklere ve tutumlara sahip olmasıdır. Oysa iş başvurusunda bulunan kişiler çok farklı özelliklere sahiptirler. [105]

Tahmin geçerliliği. Tahmin geçerliliğinde başlıca iki kriter kullanılır. Bunlardan birincisi *genel performans puanları* ve *ikincisi ise üretimle ilgili rakamlardır.* Üretimle ilgili rakamlar; kaç birim üretildiği, ne kadar hata yapıldığı, ne kadar ıskarta olduğu gibi konular veya ürünün kalitesiyle ilgilidir. Örneğin, birinci, ikinci ve üçüncü kalitede ne kadar üretim yapıldığına ilişkin rakamlar üretimle ilgili performans puanları olarak kullanılabilir.

Kriter geçerliliğini uygulama prosedüründe şu adımlar uygulanır. İşveren personeli her zamanki seçim prosedürüne göre (öz geçmiş incelemesi, mülakat ve referans kontrolü gibi) belirler. Fakat aynı zamanda seçilen adaylara test uygulanır, ancak işe alım kararında bu testlerin puanları dikkate alınmaz. Aradan altı ay veya bir yıl gibi bir süre geçtikten sonra söz konusu elemanların *performans puanları* incelemeye alınır. Performans puanları, kriter olarak kabul edilir. Daha sonra performans puanları ile test puanlarının karşılaştırması yapılır. Aradaki korelasyon yüksekse testin tahmin geçerliliğine sahip olduğu söylenir.

Tahmin geçerliliğinin dezavantajı sağlıklı bir geçerlilik analizi yapılabilmesi için yeteri sayıda adayın olmasına ve yeteri sayıda da personelin

[1] Puan ortalamaları çok daha farklı şekillerde hesaplanabilir. Bu konuda daha ayrıntılı bilgiler için psikoteknik kitaplarına bakılmalıdır.

istihdam edilmiş olmasına bağlıdır. Sağlıklı bir analiz için en az 75 personelin bulunması gerektiği belirtilmiştir.[106] Eğer sayılar yetersizse geçerlilik analizleri güvenli bir şekilde yapılamaz. Bir diğer sakıncası performans ölçümlerinin altı ay veya daha uzun bir süre sonunda yapılmasıdır. Çünkü bu süre içinde işin içeriğinde değişiklikler olabilir ve performansla test aynı şeyi ölçmüyor olabilir.

Ancak kriter puanlar işletmeden işletmeye ve işten işe farklılık gösterebilir. Bu nedenle başka bir işletmede ve başka bir iş için yapılan geçerlilik çalışmaları testi satın alan bir işletme için geçerli olmayabilir. Geçerlilik çalışmalarının işletmeye özgü olarak güncelleştirilmesi gerekir. Personel seçiminde uygulanan testlerde içerik ve yapısal geçerlilikten çok kriter geçerliliğine daha fazla önem verilir. Bunun nedeni karşılaştırmalarda somut bir karşılaştırma rakamına sahip olunması ve bu rakamın sınama yapmaya elverişli olmasıdır.

Bu arada halen işletmede çalışan ve yüksek performans gösteren kişilere de aynı testler uygulanır. Bu uygulamanın iki amacı vardır. Birincisi bu kişilerin söz konusu testlerden yüksek puan alıp alamayacaklarını tespit etmek ve ikincisi yüksek puanlar almışlarsa bu puanlara bakarak psikometrik test bataryasının norm değerlerini belirlemek.

Bir araştırmacı personel seçim prosedüründe kullandığı psikometrik testler için üç farklı normdan yararlanabilir. Birincisi o işletmede çalışan en başarılı İşgörenlerin test değerlerini kullanabilir. İkincisi, İşle ilgili sorulara ve gereklere cevap veren bir anket uygulayabilir ve üçüncüsü ise, ülke çapında aynı sektörde ve aynı işte çalışanların başarı puanlarını bir kriter olarak ele alıp değerlendirebilir.

Personel seçiminde bir kişinin iş için seçim kararının verilmesinde psikometrik testlerin ağırlığı %30'dan daha fazla tutulmaz. Psikoteknik testlerin yanında görüşme, mülakat, deneyim, bilgi, kişilik, iş yaşamıyla ilgili öz geçmiş, yabancı dil gibi faktörler ayrıca değerlendirmeye alınır.

Personel seçiminde uygulanacak testlerin geçerlilik analizlerinin uygun bir örnek kütle büyüklüğü üzerinde test edilmesi gerekir. Örnek kütle hacmi yeterince büyük değilse, test sonuçları başvuran kişileri olumsuz yönde etkiler. Bunun için olumsuz etki istatistikî analizlerinin yapılması gerekir.

ABD gibi gelişmiş ülkelerde personel seçiminde uygulanan testlerin geçerli olmasına büyük ölçüde önem verilmiştir. Testler için geçerlilik analizinin yapılması zorunlu tutulmamakla birlikte, uygulanan test puanları nedeniyle haksızlığa uğradığını ve olumsuz etkilendiğini iddia eden bir kişinin şikayeti üzerine söz konusu testlerin geçerli olup olmadığı mahkemelerde dava konusu edilebilmektedir. Adayların haksızlığa uğradıklarını

iddia etmeleri üzerine yapılan puan sıralamaları dikkate alınmamakta, uygulanan testlerin geçerliliği araştırılmaktadır. Olumsuz etkilenme veya haksızlığa uğrama iddiası testlerin objektif olmadığı, kişilerin kendilerine cinsiyet, yaş (40 veya daha yaşlı), ırk, fiziksel engelli olma, din veya etnik köken nedeniyle ayrım yapıldığı iddiasıyla birlikte gelebilmektedir. Bu gibi durumlarda mahkemeler testleri uygulayan kurumdan veya işyerinden testlerin uygulanma amacıyla ilgili geçerlilik analizi verilerini veya istatistikî analiz sonuçlarını isteyebilmektedir. Bu gibi durumlarda sorumluluk testi üreten değil testi kullanan kişi ve kurumların sorumluluğunda görülmüştür.[107] Testi üreten veya testi kullanan kişilerin söz konusu testlerin geçerli olduğuna ilişkin yazılı veya sözlü beyanlarına itibar edilmemekte geçerliliğine ilişkin somut kanıtlar gösterilmesi istenmektedir. Herhangi bir test bir ırk, cinsiyet, etnik gurup veya dinî grup üzerinde haksız ve olumsuz bir etki yaratıyorsa bu tür testler için geçerlilik analizlerinin yapılması gerekir. Yurt dışında ayrımcılık yaratan testlerin personel seçimi amacıyla iş hayatında kullanılması yasaklanmıştır.

KÜLTÜRLER ARASI ÇALIŞMALARDA

Bir ölçek veya testle elde edilen veriler esas olarak kendi geliştirildiği ülkesi için geçerliliğe sahiptir. Karşıt kültürel çevrelerde geçerliliğe sahip bir ölçek veya test oluşturmak hem maliyetli ve hem de büyük ölçüde zordur. Testin/ölçeğin yapısal içeriği, test uygulama biçimi, testin düzenleme biçimi, cevaplama süresi, cevaplama biçimi, insanların tahminde bulunma eğilimi ülkeler arasında farklılık gösterir ve bu nedenle bir ülkede geçerli olan ölçek verileri, başka bir ülkede geçerli olmayabilir.

Önceki yıllarda araştırmacılar, sadece şekillerden oluşan ve *kültürden bağımsız* olarak adlandırılan testleri farklı kültürlerde uygulayabileceklerini düşünüyorlardı (Cattell, 1940). Fakat uzun yıllar boyunca yapılan araştırmalar kültürden bağımsız test veya görev diye bir olgunun olmadığını göstermiştir. Tam tersine testler kültüre büyük ölçüde bağımlı veya kültüre daha az bağımlı bir boyut üzerinde yer alırlar.[108]

Bir testin/ölçeğin kültürel olarak geçerli olması ölçüm yapıldığı toplumdaki değerlere, inançlara, deneyimlere, haberleşme biçimlerine ve bilgi temeline uygunluğuna bağlıdır.[109] Felsefi görüş açısından gelişmiş ülkelerin kültürleri, bir testin/ölçeğin içeriğini, uygulama biçimini, kavramsal yapısını büyük ölçüde belirlemektedir. Yöntem bilim açısından ise bir testin/ölçeğin uyarlanmış veya tercüme edilmiş biçimi soruşturulabilir bir niteliktedir. Uyarlama prosedürleri çok farklıdır ve bu nedenle uyarlanmış

veya tercüme edilmiş testlerin orijinallerine göre oldukça zayıf olduğu iddia edilmiştir.[110]

Solano-Flores ve Nelson-Barber'e göre, kültürel geçerliliğin birinci aşamasında geliştirilen testin/ölçeğin maddelerinin her biri cevaplayıcılar tarafından dikkatli bir şekilde okunarak yorumlama farklılıklarının bulunup bulunmadığına bakılır. Farklılıklar hem niteliksel olarak hem de istatistiksel olarak analiz edilir. İstatistiksel analizde birden fazla grup ortalamaları karşılaştırılarak algılama farklılığı olup olmadığına bakılır. Kültürel geçerlilik bir testin/ölçeğin son haline gelmiş maddeleri üzerinde olmaktan çok geliştirilme aşamasında uygulanır.[111]

EYLEM ARAŞTIRMALARINDA

Eylem araştırmalarının tarihi 1930'larda Kurt Levin'in yaptığı çalışmalara dayandırılır. Kurt Levin eylem araştırmalarıyla esas olarak kurumsal değişimi gerçekleştirmeye yönelik çalışmaları kastetmiştir.[112] Eylem araştırmaları tek bir işletmede, tek bir kurumda, tek bir kişi üzerinde veya toplum çapında yapılan vak'a araştırmalarıdır. Bilimsel bir araştırma yöntemi olmaktan çok bir tür sorgulama yöntemidir. Bu yöntemde ana kütle ve örneklem seçimi gibi prosedürler yoktur. Eylem araştırmalarında her hangi bir şey ispat edilmez. Araştırmacı bu tür çalışmalarda daha büyük ana kütle hakkında her hangi bir tahminde bulunmaz. Araştırmacının amacı incelediği vak'aya ilişkin bazı sonuçlara ulaşmak ve bu sonuçların daha sonraki başka araştırmalar için bir girdi oluşturmasını sağlamaktır. Ancak bu tür araştırmaların beklenen duyarlılıkta ve dikkatli bir şekilde yapılıp yapılmadığını belirlemek için geçerlilik analizleri gündeme gelebilir. Bazı yazarlar eylem araştırmalarında *geçerlilik* değil, *kalite* kavramının kullanılmasını önermişlerdir. Bu tür araştırmalar için önceden belirlenmiş geçerlilik kuralları yoktur.[113] Eylem araştırmalarında ölçüm verilerinin geçerliliğinden çok incelemenin yeterli derinlikte yapılması ve çok yönlü bir perspektife sahip olup olmadığı önem kazanır. Araştırmacı bu tür çalışmalarda yaptığı tercihlerin farkında olmalı ve bu tercihlerin ne gibi sonuçlar doğuracağı hakkında ön inceleme yapmalıdır. Tercihlerini belirlerken aynı zamanda ilgili literatüre atıfta bulunarak tezlerini güçlendirmeye çalışmalıdır.

Eylem araştırmalarında yazarın ulaştığı veya elde ettiği *doğrular* okuyucu için her hangi bir anlam ifade etmeyebilir. Okuyucular veya değerlendiriciler *gerçeğin* bütünüyle farklı olduğu iddiasında bulunabilirler. Yapılan tartışmaların, getirilen yorumların okuyucular için makul ve tutarlı

olup olmadığı belirsizdir. Araştırmacı tasarımını, bulgularını büyük ölçüde incelediği vakayı tanımaya yönelik olarak sergiler, fakat kullandığı yöntemin ve elde ettiği bulguların evrensel geçerliliğe sahip olduğu konusunda kanıt getiremez.

Araştırmacı eylem araştırmalarının kalitesini yükseltmek için kendisine şu soruları sormalıdır: Yaptığım çalışmayı nasıl daha iyi hale getirebilirim? Getirdiğim açıklamalar olgunun bütün yönlerini kapsıyor mu? Değerlendirmeler, ölçüm sonuçları geçmişteki uygulamalarla paralellik gösteriyor mu? Ölçümler geleceğe ait bazı konulara ışık tutuyor mu? Yorumlar, bulgular ve değerlendirmeler *gerçeği* ne ölçüde yansıtıyor?

"Eylem araştırması yapan bir kişi ya kendi davranışını değiştirmeyi düşünüyordur veya davranışlarını değiştirip değiştirmeme konusunda karar vermek isteyen kişilere yardımcı olmayı amaçlıyordur."[114] Bu çerçevede geliştirilen modelin gerçek hayattaki modele uygun olması gerekir. Bu uygunluk sağlanabildiği ölçüde vak'a araştırması geçerlidir. Ancak teknik anlamda eylem araştırmalarının geçerli olup olmadığı sorgulanmaz. Onun yerine kaliteli olup olmadığı değerlendirilir. Bir eylem araştırmasının kalitesini değerlendirmede aşağıdaki ölçütlerin kullanılabileceği belirtilmiştir.[115]

1. Araştırmanın bir amacı olmalı ve gerçek dünyada pratiği bulunmalıdır.
2. Konu açık ve aktif bir şekilde araştırılmalı, ayrıca ilgili kişilerin araştırmaya katılımı sağlanmış olmalıdır.
3. Geniş bir bilgi sağlama temelinden hareket edilmeli (sezgisel, deneysel ve kavramsal) ve bu bilgiler önceki literatür bilgileriyle ilişkilendirilmelidir.
4. Araştırma sorunu, ana kütleye veya geniş insan topluluklarına ilişkin olmaktan çok, belirli bir grubun ihtiyaçlarıyla veya sorunlarıyla ilgili olmalıdır.
5. Olayı birinci, ikinci ve üçüncü kişilerin gözünden değerlendirerek açık uçlu bir değerlendirme kapasitesi yaratılmalıdır.
6. Araştırmacı, çalışmasında sadece bulguları ve bulgularla ilgili düşüncelerini sunmakla yetinmemeli aynı zamanda eleştiri yapmalıdır.

Eylem araştırması 1960 ve 1970'li yıllarda bütünüyle ortadan kalkmışken 1980'li yılların ortasında yeniden araştırma dünyasında yaygınlaşmaya

başlamıştır. Günümüzde eylem araştırması, *problem çözme süreci* olarak değerlendirilir.

GEÇERLİLİĞİ TEHDİT EDEN FAKTÖRLER

Test veya ölçeklerin geçerliliğini tehdit eden çok sayıda faktör vardır. Bu faktörlerden bir bölümü daha önce "Araştırmanın Bir Bütün Olarak Geçerliliği" başlığı altında ele alınmıştı. Bu bölümde ise geçerliliği engelleyen faktörler daha genel bir yaklaşımla konuyu toparlayıcı bir özet olarak ele alınmıştır. Güvenilirliği tehdit eden faktörler aynı zamanda geçerliliği de tehdit eder. Literatürde geçerliliği tehdit eden faktörlerin sayısını kırka kadar çıkaran yazarlar vardır, ancak önemli olan bunların içinde en sık karşılaşılan faktörlerdir. Araştırmacı bu faktörlerin etkisini kontrol altına alabilirse ölçüm aracının geçerliliğini artırma şansına sahip olur. Geçerliliği tehdit eden faktörleri üç grupta değerlendirebiliriz. Ölçüm aracının kendisinden kaynaklanan faktörler, ölçme işleminden kaynaklanan faktörler ve katılımcılardan kaynaklanan faktörler.

ÖLÇÜM ARACINDAN KAYNAKLANAN FAKTÖRLER

Ölçüm aracından kaynaklanan faktörler, ölçek veya testin yapılandırılma sorunlarıyla ilgilidir ve bu sorunlar ölçüm aracının niteliğine göre değişir. Ölçüm aracının bilişsel bir test, duygusal bir test, duyusal, fiziksel veya psikomotor bir test olmasına göre farklı tehdit edici faktörler söz konusudur. Test aracı ayrımı gözetmeden tehdit edici faktörlerin başlıcalarını aşağıdaki gibi maddeleştirebiliriz.

1. Test maddelerinin ifadelendirme sorunları. Test maddelerinin ifadelendirme sorunları farklı kişilerin test maddesinden farklı anlam çıkarmasına ve sonuçta işaretlemelerin geçersiz olmasına yol açar.
2. Test maddelerinin kavramsal yapıyı temsil etmemesi. Test maddeleri kavramsal alanın boyutlarını yeteri ölçüde temsil etmiyorsa ölçeğin geçerliliği düşük çıkar.
3. Testin gerçek hayattaki başarı faktörleriyle ilişkisinin zayıf kalması. Test sonuçları ile başarı faktörleri arasında anlamlı bir ilişki yoksa testin geçerliliği düşük çıkar.

4. Test içeriğinin kavramsal yapıyı temsil etme özelliğinin zayıf kalması. Test içeriğine ilişkin maddelerin kavramsal yapıyı temsil etme özelliğinin düşük olmasıdır.
5. Test geliştirme prosedürüne uyulmaması. Araştırmacı test geliştirme prosedürünün aşamalarına tam olarak uygun bir şekilde hareket etmemişse test sonuçlarının geçerliliği yine düşük çıkar.

Ölçüm aracından kaynaklanan tehdit edici faktörler, bütünüyle zayıf bir ölçüm aracının kullanılıyor olmasına dayanır. Ölçüm aracının güçlü olması, güvenilirlik ve geçerlilik analizlerinin titiz bir şekilde yapılmasına bağlıdır.

KATILIMCILARDAN KAYNAKLANAN FAKTÖRLER

Katılımcılardan kaynaklanan ve geçerliliği tehdit eden faktörler örnekleme çalışmasıyla ilgilidir. Araştırmacı dikkatli bir örnekleme çalışması yapmamışsa katılımcıların kendileri geçerlilik handikabı olarak ortaya çıkar. Katılımcılardan kaynaklanan geçerliliği tehdit edici faktörler aşağıdaki gibidir:

1. Ölçek veya testin taraf tutan kişilere uygulanması.
2. Ölçek veya testin ön yargılı olanlara uygulanması.
3. Ölçek veya testin belirli bir görüşe eğilimli olanlara uygulanması.
4. Ölçek veya testin tercihi kesin olarak belirlemiş olanlara uygulanması.
5. Ölçek veya testin pilot araştırma yapılan kişilere tekrar uygulanması.

Testin sadece belirli bir olguyu çok seven kişiler tarafından, tutucu olanlar tarafından veya sadece liberal olanlar tarafından, sadece kendilerini devrimci olarak nitelendirilen kişiler tarafından, sadece belirli bir görüş doğrultusunda yetiştirilenler tarafından doldurulması sonuçların şüpheli olmasına yol açar. Bu tür sonuçlar daha geniş ana kütlelere genellenemez. Ayrıca pilot araştırma yapılan kişilere testin tekrar uygulanması bu kişilerde ilgisizlik yaratır veya öğrenme etkisi nedeniyle sonuçlar gerçek durumu yansıtmayabilir.

Geçerlilik analizleri için, *bütünleşmiş yaklaşımı* savunun Messick, de-

ğerlendirmedeki cevaplandırıcı yanlılığının yapısal geçerliliği bozacağını ön görmüştür. "Aynı kavramsal yapı farklı popülasyonlarda ne ölçüde benzer bir anlama sahiptir?" sorusunu sorarak, yanlılık varsa kavramsal yapılarla ilgili sonuçların da geçersiz olacağını belirtmiştir. Çünkü farklı gruplarda yapılan ölçümlerde, hedef kavramsal yapının dışında farklı, ilgisiz diğer kavramsal yapılar da işin içine girecektir. Söz konusu ilgisiz yapılar cinsiyet, ırk, dil farklılığı, ideolojik farklılık, sosyoekonomik statü veya engelli olma gibi belirli bir grubu tanımlayan koşullarla ilgili olabilir.[116] Geliştirilen testin/ölçeğin farklı örneklem gruplarında farklı yapıları ölçüp ölçmediğini belirlemek için varyans analizi yapılır ve gruplar arasında anlamlı bir farklılık olup olmadığı veya negatif bir farklılık bulunmadığı incelenir. Anlamlı bir farklılık varsa, soruların/ifadelerin belirli bir gruba uygun olarak hazırlandığı sonucuna varılır. Test ile geçersiz bir ölçüm yapılmaması için araştırmacı aşağıdaki soruları sormalıdır:[117]

1. Ölçüm aracının hedef kitlesi kimdir?
2. Hedef kitlenin özellikleri nelerdir?
3. Örneklem çerçevesi nedir? Bu çerçeve ana kütleyi tam olarak temsil etmekte midir?
4. Örneklem çerçevesinde pilot araştırma için seçilen bireyler hangi kriterlere göre belirlenmiştir?
5. Örneklemdeki bireylerin demografik özellikleri nelerdir?

Katılımcı kaynaklı tehdit faktörleri, iyi bir araştırma tasarımıyla önlenebilir. Bunun için bilim adamı ana kütle ve örneklem çerçevesini araştırma konusuna göre dikkatli bir şekilde oluşturmalı ve makul büyüklükteki bir ana kütleye genelleme yapmaya imkan verecek örneklemlerle çalışmalıdır.

ÖLÇME VE DEĞERLENDİRME İŞLEMİNDEN KAYNAKLANAN FAKTÖRLER

Ölçme ve değerlendirme işleminden kaynaklanan geçerliliği tehdit edici faktörler büyük ölçüde özensiz bir araştırma uygulamasına dayanır. Ölçme işleminin geçerliliği, normal bir ölçüm veya araştırma prosedürünün koşullarına tam olarak uyulmasıyla gerçekleştirilebilir. Ölçme işleminin geçerliliğini tehlikeye atmamak için bilim adamı aşağıdaki önlemleri almalıdır.[118]

1. Pilot araştırma yapılması. Ölçüm yapılması hedeflenen ana kütleden seçilecek bir örnek kütle üzerinde ön test yapılarak sadece ölçüm aracı değil, araştırma süreci ve uygulamasının da ne şekilde yapılacağına ilişkin ön bilgilerin toplanmasıdır.
2. Deneysel araştırmalarda kontrol grubunun kullanılması. Üzerinde deney yapılmayan, fakat deney yapılan kişilerle aynı özelliklere sahip başka bir grupta daha ölçüm yapılması tarih, olgunlaşma, araçsallık ve etkileşim etkisinin etkilerini azaltma imkanı sağlar.
3. Katılımcıların tesadüfî olarak belirlenmesi. Kişilerin rasgele seçilmeleri istatistiksel analiz sonuçlarının daha geniş ana kütlelerle genellenmesi imkanını sağlar. Sonuçlar ana kütleyi daha iyi temsil eder. Deney ve kontrol gruplarına kişiler rasgele atanmış olmalıdırlar.
4. İlave gruplarda araştırmalar yapılması. Bilim adamı incelediği konuyu veya ölçüm yaptığı kavramsal alanı sadece pilot araştırma ve esas araştırma yaptığı örneklem grubunda değil, bunun yanında başka gruplarda da test edebilir. Böylece sonuçlarını sadece pilot araştırma ve esas araştırma sonuçlarına değil birden fazla grupta yaptığı ölçüm sonuçlarına dayandırmış olacaktır.
5. Uygun istatistik tekniğin seçilmesi. Bilim adamı araştırma tasarımı kadar toplanan verilerin analizinde uygun istatistik tekniğin seçilmesi konusuna da önem vermelidir. Uygun istatistik teknik, istatistik testin varsayımlarının karşılanma durumunun araştırılması, bağımlı değişken verilerinin güvenilir olması test sonuçlarının geçerliliğini artırır.

Ölçme işlemi; dikkatli, titiz ve standardize edilmiş uygulamalar çerçevesinde yapıldığı ölçüde geçerlidir. Farklı gruplarda farklı yöntemlerin uygulanması, uygulama talimatlarının değişmesi veya bulunmaması, kişilerin kolayda örnekleme yöntemiyle belirlenmesi sonuçların geniş ana kütlelere genellenmesi kısıtlar ve böylece geçerliliği düşürür.

SONUÇLARIN RAPORLANMASI

Geçerlilik analizine ilişkin sonuçlar iki temel başlıkta sunulur: ölçüm aracının geçerliliği ve araştırma sürecinin geçerliliği. Tek başına ölçüm aracının geçerliliği yeterli değildir. Veriler aynı zamanda geçerli bir araştırma uygulamasıyla elde edilmiş olmalıdır.

BULGULARIN GEÇERLİLİĞİNE İLİŞKİN YORUMLAR

Kullanılın ölçek/test verilerinin geçerliliğiyle ilgili olarak araştırmacı; varsa birinci sırada önceki araştırmalardan elde edilmiş geçerlilik bilgilerini verir. Araştırmacı bu bilgileri verirken ana kütlelerin benzerliği, kullanım amacının benzerliği ve önceki ölçekte hangi oranda değişiklik yapıldığı konularını ele alır. Daha sonra araştırmacı kendi geçerlilik kanıtlarını ortaya koyar. Pilot araştırma sırasında ve esas araştırma süreci sonucunda yaptığı geçerlilik analizi bulgularını raporlar. Bu arada kendi ana kütlesinin önceki ana kütleye benzer olup olmadığı, amacının benzer olup olmadığı konularında bilgi verir. Pilot araştırma sonunda bazı konularda hâlâ tam olarak emin olunamadıysa araştırmanın sürdüğü sırada ve belli bir sayıda ankete ulaşıldığında geçerlilik analizleri tekrar yapılır. Böylece araştırma öncesinde, araştırma sırasında ve araştırma sonunda elde edilen veriler karşılaştırmalı olarak ele alınıp raporlanır. Ancak araştırma süreci içinde yapılan analizlere dayalı olarak anket formunda bir değişiklik yapılmaz. Bu analizler sadece bilgi edinmeye yöneliktir.

Geçerlilik analizi sonuçlarının raporlanmasında, *testin geçerliliği* veya *ölçeğin geçerliliği* ifadelerinin kullanılmasından kaçınılmalıdır. Geçerlilikle ilgili bulgular yorumlanırken "ölçekle elde edilen puanların iç tutarlılık katsayısı ,86 çıkmıştır" veya "test puanlarının değerlendiriciler arası Cohen kappa katsayısı ,72 çıkmıştır" ifadeleri kullanılır.[119] Bilim adamı geçerlilik kanıtlarını sunarken bu kanıtların yargısal veya istatistiksel geçerlilikle ilgili olması konusunda açıklayıcı bilgiler sunar. Yargısal geçerlilik kanıtları daha çok yüzey, içerik, nomolojik geçerlilik analizlerinde sunulur. İstatistiksel geçerlilik kanıtları ise kriter ve yapısal geçerlilik analizleriyle elde edilir.

ÖLÇÜM UYGULAMASININ GEÇERLİLİĞİNE İLİŞKİN YORUMLAR

Bilim adamı geçerlilik analizi sonuçlarını yorumlarken ikinci bölümde ölçüm uygulamasının geçerliliğine ilişkin kanıtları ele alır. Bunu yaparken; konuyu ölçümün geçerliliği, iç geçerlilik, dış geçerlilik ve istatistiksel geçerlilik başlıkları altında irdeler.

Ölçüm uygulamasının geçerliliğine ilişkin yorumlar yapılırken bilim adamı amaçladığı sonuçlara veya hedeflere ne ölçüde ulaştığı ve ne ölçüde ulaşamadığı hakkında bilgi vermelidir. Bir araştırma uygulaması, her za-

man araştırmacının istediği şekilde gerçekleşmez. Örnekleme giren bireylerdeki yanlılıklar, örneklemin seçilmesindeki kısıtlar, istatistikî analiz sonuçlarındaki istenmeyen veya beklenilmeyen bulgular açık bir şekilde okuyuculara anlatılmalı ve araştırmanın hangi kısıtlar altında gerçekleştirildiği net bir şekilde ortaya konmalıdır.

Bilim adamı klasik geçerlilik analizlerinin ötesinde olguyu modern geçerlilik anlayışıyla ele almışsa sadece içerik, kriter ve yapısal geçerlilik kanıtlarını değil, aynı zamanda test sonuçlarının nasıl kullanıldığını ve ne gibi bir fayda sağladığını veya sağlayacağını da açıklamalıdır. Modern geçerlilik anlayışında, test sonuçlarının kişi ve gruplar için olumsuz sonuçlar doğurma ihtimalinin bulunma durumu da irdelenir.

ALINTI YAPILAN KAYNAKLAR

[1] L. Bol, "Reliability and Validity of Measurement Instruments [Ölçüm Araçlarının Güvenilirlik ve Geçerliliği]," <http://courses.lib.odu.edu/eci/lbol/slides_4April.ppt> (28.06.2004).

[2] Royal Windsor Society, "Instrument Reliability [Ölçüm Aracının Güvenilirliği]," <http://www.kelcom.igs.net/ ~nhodgins/instrument_reliability.html> (24.08.2002).

[3] Institute for Semantic and Cognitive Studies, "Validity of Measures [Ölçümlerin Geçerliliği]," <http://www.issco.unige.ch/ewg95/node30.html> (01.09.2002).

[4] G. Winter, "A Comparative Discussion of the Notion of 'Validity' in Qualitative and Quantitative Research [Niceliksel ve Nitelliksell Araştırmalarda 'Geçerlilik Kavramının Tartışılması]," <http://www.nova.edu/ssss/QR/QR4-3/winter.html> (30.08.2002).

[5] H.K. Suen, "The Evolving Concept Of Validity [Geçerlilik Kavramının Gelişmesi]," <http://suen.ed.psu.edu/~hsuen/Evolve.pdf> (27.06.2004).

[6] G. Winter, "A Comparative Discussion of the Notion of 'Validity' in Qualitative and Quantitative Research [Niceliksel ve Nitelliksell Araştırmalarda 'Geçerlilik Kavramının Tartışılması]," <http://www.nova.edu/ssss/QR/QR4-3/winter.html> (30.08.2002).

[7] People Click, "What is Vality [Geçerlilik Nedir?]," t.y., <http://www.eeosource.com/topic/default.asp? TopicArea=9&MainTopicID=2> (05.08.2002).

[8] N. Wilson, "Validity and Reliability [Geçerlilik ve Güvenilirlik]," <http://epaa.asu.edu/epaa/v6n10/c16.htm> (15.08.2002).

[9] A. Brualdi, "Traditional and Modern Concepts of Validity [Geleneksel ve Modern Geçerlilik Kavramları]," <http://www.ericdigests.org/2000-3/validity.htm> (17.07.2004).

[10] A. Brualdi, "Traditional and Modern Concepts of Validity [Geleneksel ve Modern Geçerlilik Kavramı]," <http://www.ed.gov/databases/ERIC_Digests/ed435714.html> (20.09.2002).

[11] B.T. Gray, "Controversies Regarding the Nature of Score Validity: Still Crazy After All These Years [Puanların Geçerliliği İle İlgili Tartışmalar: Bunca Yıldan Sonra Süre Gelen Tartışmalar]," <http://ericae.net/ft/tamu/Valid.htm> (05.08.2002).

[12] B.T. Gray, "Controversies Regarding the Nature of Score Validity: Still Crazy After All These Years [Puanların Geçerliliği İle İlgili Tartışmalar: Bunca Yıldan Sonra Süre Gelen Tartışmalar]," <http://ericae.net/ft/tamu/Valid.htm> (05.08.2002).

[13] Gray, "Controversies Regarding."

[14] H. Suen, "Evolving Concept of Validity [Geçerlilik Kavramının Tarihsel Gelişimi]," <http://suen.ed.psu.edu/~hsuen/Evolve.htm> (02.09.2002).

[15] T. Olivarez, "Validity [Geçerlilik]," <http://www6.tltc.ttu.edu/olivarez/EPSY-5356/Chapter6notes.htm> (02.09.2002).

[16] Suen, "Evolving Concept."

[17] Winter, "A Comparative Discussion."

[18] J. Buley, "Reliability, Validity and Correlation [Güvenilirlik, Geçirlillik ve Korelasyon]," <http://com.pp.asu.edu/classes/jerryb/rvc.html> (30.08.2002).

[19] M.J. Telch, "Validity [Geçerlilik]," <http://homepage.psy.utexas.edu/homepage/class/Psy394Q/Research%20Design%20Class/Lectures/Validity%20Lecture.ppt> (09.07.2004).

[20] D. Garson, "Scales and Standard Measures [Ölçekler ve Standart Ölçüler]," t.y., <http://www2.chass.ncsu.edu/garson/pa765/standard.htm> (10.08.2002).Cohen's kappa

[21] W.G. Hopkins, "Measures of Validity [Geçerlilik Ölçümleri]," <http://www.sportsci.org/resource/stats/valid.html#nominal> (18.08.2002), Validity of Nominal Variables.

[22] EFOPlan, "Formative and Reflective Indicators and Structural Equation Modeling [Oluşturucu ve Yansıtıcı Ölçekler ve Yapısal Eşitlik Modeli]," <http://www.efoplan.bwl.uni-muenchen.de/e/content/forschung/schwerpunkte_11.asp> (11.07.2004).

[23] J.K. Johansson ve G.S. Yip, "Using PLS In Strategy Research [Strateji Araştırmalarında PLS Yönteminin Kullanılması]," <http://www.msb.edu/faculty/johanssj/web/PLS.PDF> (02.07.2004).

[24] P. Chwelo ve I. Benbasat "Empirical Test of an EDI Adoption Model [Bir EDI Uyum Modelinin Ampirik Testi]," <http://ebusiness.commerce.ubc.ca/internal/UBCBEBR2000-003.pdf> (02.07.2004).

[25] C.B. Jarvis, S.B. Mackenzie ve P.M. Podsakoff , "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research [Pazarlama ve Tüketici Araştırmalarında Yapısal Göstergelerin Eleştirel Değerlendirmesi]," <http://www.bauer.uh.edu/mark/papers/Jarvis_JCR_2003.pdf> (05.07.2004).

[26] W. Reinartz, M. Krafft, ve D.W. Hoyer, "The CRM Process: Its Measurement and Impact on Performance [CRM Süreci: Ölçümü ve Başarı Üzerindeki Etkisi]," 2004, <www.bauer.uh.edu/mark/papers/CRM%20Final%20Paper.pdf> (09.07.2004).

[27] B. Stennet, "Qpinion Survey Ratig Scales [Kanaat Araştırmalarında Ölçek Dereceleri]," <http://www.assessmentplus.com/articles/opinion_survey_rating_scales.pdf> (07.09.2002).

[28] Aynı.

[29] NCS Pearson, "Choosing the Right Scale [Doğru Ölçeği Seçme]," <http://www.pearsonncs.com/research-notes/95-03.htm> (31.07.2004).

[30] G.D. Israel, "Analyzing Survey Data [Alan Araştırması Verilerinin Analizi]," <http://edis.ifas.ufl.edu/BODY_PD007> (11.07.2004).

[31] C. Douglas, "The Structure of Inquiry and Research Design [Soruşturma ve Araştırma Tasarımı]," <http://webpages.chhs.niu.edu/douglass/AHPH%20590/590Class4and5.ppt> (11.07.2004).

[32] C. Sproule ve M. Epoca, "Physical Ability Test Validation Research for Entry-Level Corrections Officers [Giriş Seviyesindeki Düzeltme Memurları İçin Uygulanan Fiziksel Yetenek Testlerinin Geçerliliği]," <http://www.ipmaac.org/acn/oct98/physical.html> (31.07.2004).

[33] C. VerWys, "Tests of Special Abilities [Özel Yetenek Testleri]," <http://www.rpi.edu/~verwyc/Chap9tm.htm> (31.07.2004).

[34] Okunurluk ve anlaşılırlık analizi için *bk.*, H. Şencan *Bilimsel Yazım*. (İstanbul: İÜ İşletme Fakültesi Yayını, 2003).

[35] S. Vickery, "Lets Not Overlook Contant Validity [İçerik Geçerliliğini Gözden Uzak Tutmayalım]," <http://www.decisionsciences.org/Newsletter/vol29/29_4/research_29_4.pdf> (09.08.2002).

[36] A. Yu, "Reliability and Validity [Güvenilirlik ve Geçirlilik]," <http://seamonkey.ed.asu.edu/~alex/teaching/assessment/reliability.html> (02.09.2002).

[37] Vickery, "Lets Not Overlook."

[38] S.N. Haynes, "Content Validity in Psychological Assessment [Psikolojik Değerlendirmelerde İçerik Geçerliliği]," <http://www.personal.kent.edu/~dfresco/CRM_Readings/Haynes_1995.pdf> (16.07.2004).

[39] C.A. Young, "Validity Issues in Measuring Psychological Construct [Psikolojik Yapıların Ölçümünde Geçerlilik Sorunu]," <http://trochim.human.cornell.edu/tutorial/young/eiweb2.htm> (02.09.2002).

[40] Young, "Validity Issues."

[41] Kavram haritası konusunda daha fazla bilgi için *bk.*, W.M.K. Trochim, The Reliability of Concept Mapping [Kavram Haritasının Güvenilirliği], <http://trochim.human.cornell.edu/research/reliable/reliable.htm> (02.09.2002).

[42] Young, "Validity Issues."

[43] M.T. Brannick, "Steps in Test Construction [Test Oluşturmada Adımlar]," <http://luna.cas.usf.edu/~mbrannic/files/pmet/tstcon1.htm> (02.08.2002).

[44] D. Garson, "Factor Analysis [Faktör Analizli]," <http://www2.chass.ncsu.edu/garson/pa765/factor.htm> (24.08.2002).

[45] Young, "Validity Issues."

[46] Young, "Validity Issues."

[47] San Francisco State University, "Development and Testing of a Core Set of University-Wide Teaching Effectiveness Items [Üniversite Çapında Eğitimin Etkinliğini Değerleme Ölçeğinin Odak Maddelerini Geliştirme]," <http://www.sfsu.edu/~senate/tereports2002.htm> (01.09.2002).

[48] Royal Windsor Society for Nursing Research, "Content Validity [İçerik Geçerliliği]," <http://www.kelcom.igs.net/~nhodgins/content_validity.html> (09.08.2002).

[49] Vickery, "Lets Not Overlook."

[50] D. Miles, "Validity of Measurement [Ölçmün Geçerliliği]," <http://luna.cas.usf.edu/~miles/tmchpt8.htm>(02.09.2002).

[51] L.A. Becker, "Kappa," <http://web.uccs.edu/lbecker/SPSS/ctabk.htm> (02.08.2004).

[52] Doctoring Curriculum Home Page, "Kappa [kapa Formülü]," 25 KSm 2002, <http://www.musc.edu/dc/icrebm/kappa.html> (28.08.2002).

[53] M. VanDyke, "Kappa [Kappa Değerlemesi]," t.y., <http://www-class.unl.edu/psycrs/971/nonpar/nonpar/summer00/multkappa.PDF> (18.08.2002).

[54] Nova Southeastern University, "Answers [Cevaplar]," <http://www.cps.nova.edu/~cpphelp/class/psy1501/mtea-2003.html> (16.07.2004).

[55] D.L. Bolton, "Validity [Geçerlilik]," t.y., <http://courses.wcupa.edu/bolton/reliabilityvalidity/validity.htm>(30.08.2002).

[56] V. Billson, A. Clague vd., "Validity and reliability of test results [Test Sonuçlarının Geçerlilik ve Güvenilirliği]," <http://www.rcpa.edu.au/pathman/validity.htm> (30.08.2002).

[57] SIOP, "Sources of Validity Evidence [Geçerlilik Kanıtının Kaynakları]," <http://siop.org/_Principles/pages13to26.pdf> (23.06.2004).

[58] A. Steinberger, "Are Lowering Employment Standards... [İstihdam Standartlarının Düşürülmesi ...]," <http://www.ipmaac.org/acn/apr01/tightmarket.html> (27.06.2004).

[59] K. Cigularov, "Validity [Geçerlilik]," <http://psy.psych.colostate.edu/courseweb/SUM2004/Validity%20(Part%203)%20K%20Sum%2004%20online.ppt> (16.07.2004).

[60] Cigularov, "Validity."

[61] Young, "Validity Issues."

[62] J.D. Brown, "What is construct validity? [Yapısal Geçerlilik Nedir?]," <http://www.jalt.org/test/bro_8.htm> (27.06.2004).

[63] S. Wickery, "Research Issues [Araştırma Sorunları]," <http://www.decisionsciences.org/DecisionLine/VOL28/28_1/research.htm> (18.07.2004).

[64] D. Kelley, "Measurement Validity [Ölçümün Geçerliliği]," <http://www.longwood.edu/staff/kelleyds/Socl345/validity/validity.ppt> (18.07.2004).

[65] Young, "Validity Issues."

[66] C.V. King, "Factor Analysis and Negatively Worded Items [Faktör Analizi ve Negatif Bir Biçimde İfadelendirilen Maddeler]," <http://www.populus.com/tech_papers/fa&_neg_worded.pdf> (24.08.2002).

[67] A. Yu, "Reliability and Validity [Güvenilirlik ve Geçerlilik]," <http://seamonkey.ed.asu.edu/ ~alex/teaching/assessment/reliability.html> (27.06.2004).

[68] C.S. Taylor, "Evidence for the Reliability and Validity of Scores... [Puanların Güvenilirlik ve Geçerlilik Kanıtları]," <http://www.sbe.wa.gov/reports/CAA%20RPT/appendix/DRFT%20FNL%20CAA%20RPT.APPENDIX%20H%20(Taylor%20Evidence%20Rpt).doc> (27.06.2004).

[69] L.J. Cronbach ve P.E. Meehl, "Construct Validity In Psychological Tests [Psikolojik Testlerde Yapısal Geçerlilik]," 1955, <http://psychclassics.yorku.ca/Cronbach/construct.htm> (17.07.2004).

[70] Aynı.

[71] R.C. Engs, "Construct Validity And Re-Assessment Of The Reliability Of The Health Concern Questionnairec [Sağlık İlgisi Anket Formunun Yapısal Geçerliliği ve Güvenilirliğinin Yeniden Değerlendirilmesi]," <http://www.indiana.edu/~engs/quest/validity.html> (27.06.2004).

[72] Young, "Validity Issues."

[73] C.D. Stapleton, "Basic Concepts in Exploratory Factor Analysis (EFA) as a Tool to Evaluate Score Validity: A Right-Brained Approach [Açıklayıcı Faktör Analizinin Geçerlilik Puanlarının Değerlendirilmesinde Kullanılması]," <http://ericae.net/ft/tamu/Efa.HTM> (24.08.2002).

[74] B. Trochim, "Convergent and Discriminant Validity [Birleşme ve Ayrılma Geçerliliği]," <http://www.socialresearchmethods.net/kb/convdisc.htm> (28.06.2004).

[75] D. Garson, "Validity [Geçerlilik]," <http://www2.chass.ncsu.edu/garson/pa765/validity.htm> (28.06.2004).

[76] Garson, "Validity."

[77] Cronbach ve Meehl, "Construct Validity."

[78] W.M.K. Trochim, "The Nomological Network [Nomolojik Ağ]," <http://www.socialresearchmethods.net/kb/nomonet.htm> (18.07.2004).

[79] Cronbach ve Meehl, "Construct Validity."

[80] ciAd Research Manuscripts, "Limitation and Suggestion for Future Research [Kısıtlılıklar ve Daha Sonraki Araştırmalar İçin Öneriler]," <http://www.ciadvertising.org/student_account/fall_01/adv392/tai/pro_seminar/limitation.html> (18.07.2004).

[81] Garson, "Validity."

[82] W.M.K. Trochim, "Pattern Matching for Construct Validity [Yapısal Geçerlilik İçin Model Denkleştirme Tekniği]," <http://www.socialresearchmethods.net/kb/pmconval.htm> (18.07.2004).

[83] Aynı.

[84] S. Wickery, "Research Issues [Araştırma Sorunları]," <http://www.decisionsciences.org/DecisionLine/VOL28/28_1/research.htm> (18.07.2004).

[85] Cronbach ve Paul E. Meehl, "Construct Validity."

[86] Aynı.

[87] P.R. Hensel, "Reliability and Validity Issues in the ICOW Project [ICOW Projesinde Geçerlilik ve Güvenilirlik]," <http://garnet.acns.fsu.edu/~phensel/Research/isa98.pdf> (09.07.2004).

[88] T.R. O'Connor, "Measurement, Reliability and Validity [Ölçüm, Güvenilirlik ve Geçerlilik]," <http://faculty.ncwc.edu/toconnor/308/308lect04.htm> (11.07.2004).

[89] CampusProgram.com, "Scale [Ölçek]," <http://www.campusprogram.com/reference/en/wikipedia/s/sc/scale__social_sciences_.html > (11.07.2004).

[90] P. Bloch, "Basic Methods, [Temel Yöntemler]," <http://www.southernct.edu/~blochj/rm2.html> (09.07.2004).

[91] C. Calero, M. Piattini ve M. Genero, "Empirical Validation Of Referential Integrity Metrics, [Bütünleştirilimiş Referans Ölçüm Değerlerinin Ampirik Geçerliliği]," 2004, <http://alarcos.inf-cr.uclm.es/articulos/ist.pdf> (09.07.2004).

[92] Aynı.

[93] webteam@css.edu, "Design Controls In Research [Araştırmada Tasarım Kontrolleri]," <http://www.css.edu/users/dswenson/web/DESIGN.HTM> (09.07.2004).

[94] M.J. Telch, "Validity [Geçerlilik]," <http://homepage.psy.utexas.edu/homepage/class/Psy394Q/Research%20Design%20Class/Lectures/ Validity%20Lecture.ppt> (28.06.2004).

[95] J. Wasson, "Threats to Experimental Validity [Deneysel Geçerlilik Tehditleri]," <http://www.mnstate.edu/wasson/ed603/ed603lesson14.htm> (09.07.2004).

[96] Aynı.

[97] Aynı.

[98] S. Hempel, "Validity [Geçerlilik]," <http://ibs.derby.ac.uk/~susanne/PTT/lectures/PTTvalidity2003.pdf> (28.06.2004).

[99] D. Siegle, "External Validity [Dış Geçerlilik]," <http://www.gifted.uconn.edu/siegle/research/Samples/externalvalidity.html> (09.07.2004).

[100] J. Van de Weijer, "Validity [Geçerlilik]," <http://www.ling.lu.se/education/homepages/LIS140/handout05.pdf> (18.07.2004).

[101] W.M.K. Trochim, "Threats to Conclusion Validity [Sonuç Çıkarma Geçerliliğini Tehdit Eden Faktörler]," <http://www.socialresearchmethods.net/kb/concthre.htm> (18.07.2004).

[102] J.E. Jones ve W.L. Bearley, "Reliability And Validity for Training Instruments [Eğitim Araçlarının Güveenilirlik ve Geçerliliği]," <http://ous.iex.net/relval.htm> (30.08.2002).

[103] W.C. Burns, "Content Validity, Face Validity, and Quantitative Face Validity [İçerik Geçerliliği, Yüzey Geçerliliği ve Sayısal Yüzey Geçerliliği]," <http://www.burns.com/wcbcontval.htm> (02.08.2004).

[104] Biddle Consultation, "Uniform Employee Selection Guidelines Interpretation And Clarification [Standart Personel Seçim Prosedürü: Yorum ve Açıklamalar]," <http://www.uniformguidelines.com/questionandanswers.html> (18.09.2002).

[105] M.L. Kelly, "Best Practices in Employee Selection [Personel Seçiminde En İyi Uygulamalar]," <http://www.ipat.com/Pdf/selectmanual/chapter7.pdf> (18.09.2002).

[106] M.L. Kelly, "Best Practices."

[107] Biddle Consultation, "Uniform Employee."

[108] M. Beller, "Translating, Equating and Validating Scholastic Aptitude Tests [Skolastik Yetenek Testlerinin Çevirisi, Eşitlenmesi ve Geçerliliği]," <http://www.unifr.ch/ztd/ems/berichte/b2/fff/translating.htm> (01.08.2004).

[109] Guillermo Solano-Flores ve Sharon Nelson-Barber, "Cultural Validity of Assessments and Assessment Development Procedures [Ölçme ve Ölçme Prosedürlerinde Kültürel Geçerlilik]," <http://www.edgateway.net/cs/cvap/print/docs/cvap/pub_2.htm> (01.09.2002).

[110] Solano-Flores ve Nelson-Barber, "Cultural Validity."

[111] Aynı.

[112] A. Feldman, "Erzberger's Dilemma: Validity in Action Resarch [Erzberger'in İkilemi: Eylem Araştırmalarında Geçerlilik]," <http://www-unix.oit.umass.edu/~afeldman/ActionResearchPapers/Feldman1994a.pdf> (01.09.2002).

[113] Sage Publication, "Resource Page For Action Research [Eylem Araştırması Kaynak Sayfası]," <http://www.sagepub.co.uk/frame.html?http://www.sagepub.co.uk/journals/resources/jr0478.html> (01.09.2002).

[114] J.M. Newman "Validity and Action Research [Geçerlilik ve Eylem Araştırması]," <http://casino.cchs.usyd.edu.au/arow/reader/newman.htm> (01.09.2002).

[115] Sage Publication, "Resource Page."

[116] T.C.M. Lam, "Fairness in Performance Assessment [Başarı Değerlemede Hakkaniyet]," <http://ericae.net/db/edo/ED391982.htm> (02.09.2002).

[117] Utah State Office of Education, "Test Validity [Test Geçerliliği]," 19 Nis 2001, <http://www.usoe.k12.ut.us/eval/validity.htm> (02.09.2002).

[118] J.P. Key, "Experimental Research and Design [Deneysel Araştırma ve Tasarım]," <http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage2.htm> (31.07.2004).

[119] P. Snyder, "Guidelines for Reporting Results of Group Quantitative Investigations [Kantitatif Araştırma Sonuçlarının Raporlanması İçin Rehber]," <http://www.fpg.unc.edu/~jei/Quantitative.html> (24.08.2002).

EKLER

EK A. GÜVENİLİRLİK VE GEÇERLİLİK KONTROL LİSTESİ

1. Ölçüm amacı.
   a. Ölçümün veya testin amacı, açık ve net bir biçimde belirtildi ..... ☐
   b. Literatürde testin veya ölçüm aracının hangi amaçlar için kullanıldığı araştırıldı ve bu konuda gerekli bilgiler verildi. ..... ☐
2. Kavramsal yapı.
   a. Ölçümün ilgi odağı olan kavramsal yapı, açık bir biçimde tanımlandı. ☐
   b. Kavramsal yapının alt boyutları veya bileşenleri literatürden, tezlerden, İnternet'ten yararlanılarak ve yapılan ampirik araştırma bulguları verilerek açıklandı ..... ☐
   c. Kavramsal yapının basit veya karmaşık olma niteliği hakkında bilgi verildi ..... ☐
   d. Kavramsal yapının tek veya çok boyutlu olma niteliği hakkında bilgi verildi ..... ☐
   e. Kavramsal yapının daraltılarak veya geniş bir şekilde ele alınma durumu, nedenleri ve sonuçları hakkında bilgi verildi ..... ☐
   f. Kavramsal yapıyı ölçmek üzere sadece tek bir test değil, birden fazla testten veya ölçekten yararlanılarak sonuçların benzerliği kontrol edildi ..... ☐
   g. Kavramsal yapıyla ilgili olarak test şirketleriyle görüşülerek o güne kadar hazırlanmış olan testlerin listesi alındı ..... ☐
   h. Literatürde söz konusu kavramsal yapıyı ölçmeye yönelik olarak öğretim üyeleri veya test şirketleri tarafından geliştirilmiş bulunan diğer test örnekleri hakkında bilgi verildi ..... ☐
   i. Test uygulanacak hedef kitleden kişilerle görüşülerek kavramsal yapının gözden kaçabilecek ögeleri hakkında bilgi toplandı ..... ☐

3. Test maddelerinin yazımı.
   a. Test maddeleri; konu içeriğine, belirlenen göreve veya hedeflenen ölçüm amacına uygun olarak belirlendi. Maddelerin görev veya ölçüm amacıyla olan ilişkisi kontrol edildi. ☐
   b. Test maddelerinin doğru cevabı belirlendi. ☐
   c. Test maddeleri, dil bilgisi kuralları ve okuma güçlüğü açısından kontrol edildi. ☐
   d. Test maddeleri, "bir test havuzu kapsamında" oluşturuldu ☐
   e. Başlangıçta, nihaî testteki madde sayısının üç katı kadar madde geliştirildi. ☐
   f. Ölçeklerin önemleri göz önünde bulundurularak, önemine göre her bir alt ölçek/test veya testçik için gerektiği kadar madde sayısı belirlendi ☐
   g. Test maddeleri yazılırken cevaplayıcıların özellikleri göz önünde bulundurulduu ☐
   h. Cevaplayıcıların sosyoekonomik durumu, eğitimleri ve yaş gibi demografik özellikleri dikkate alındı ☐
   i. Maddelerin derece sayısı ölçüm amacına ve hedef kitlesine uygun olarak belirlendi ☐
4. Örneklem büyüklüğü.
   a. Klâsik test kuramı çerçevesinde test maddeleri için en az 1:5 kriteri temel alındı veya en az 200 kişilik bir örneklem hacmi belirlendi. ☐
   b. Klâsik test kuramı çerçevesinde test maddelerinden faktör çıkarmak için Kaiser-Meyer-Olkin örneklem büyüklüğü analizi yapıldı ☐
   c. Değişik örneklem büyüklükleri temel alınarak testin etki büyüklüğü hesaplandı ☐
   d. Modern test kuramı çerçevesinde test maddeleri için belirlenen modele uygun olarak en az 1PL için 100, 2PL için 500 ve 3PL için 1000 kişilik bir örneklem belirlendi ☐
5. Test bilgileri.
   a. Kullanılan her bir testin adı /başlığı net olarak belirlendi veya her bir teste/ölçüme bir "ad" bulundu. ☐
   b. Test veya ölçüm aracının adı, "başlık biçiminde" yazıldı ☐

c. Testin/ölçümün kuramsal temeli araştırıldı ve kuramsal araştırma ve bilgilerle geliştirilen test, yapılan ölçüm arasında köprü kuruldu. ☐

d. Testin temin edildiği yazar, kurum, organizasyon açık bir şekilde belirlendi. ☐

e. Testin geliştirilmesinde yararlanılan diğer ölçüm araçlarının adları verildi ve bu ölçüm araçlarını geliştiren bilim adamlarına atıfta bulunuldu. ☐

f. Testi yayımlayan yayınevinin adı ve adresi net olarak verildi. ☐

g. Testin alındığı tez, kitap, makale açık olarak belirtildi. ☐

h. Testin telif hakları hakkında bilgi verildi. ☐

i. Testin sürüm numarası açık bir şekilde belirtildi. ☐

j. Test bir adaptasyon ise adaptasyon çalışmalarının nasıl yapıldığı hakkında bilgi verildi. ☐

k. Testte kullanılan kavramlar açık bir şekilde tanımlandı. ☐

l. Test oluşturulurken orijinalinin en az iki, ideal olarak üç katı kadar madde geliştirildi ☐

6. Pilot araştırma.

a. Testin/ölçeğin geliştirilme aşamalarıyla ilgili olarak, yapılan pilot çalışmalar hakkında bilgi verildi. ☐

b. Pilot araştırma sırasında kullanılan örnek kütlenin büyüklüğü ve bu büyüklüğün mantığı hakkında bilgi verildi. ☐

c. Pilot araştırma, esas araştırmanın veya ölçümün yapılacağı benzer bir grupta yapıldı. ☐

d. Pilot araştırma yapılan örneklemin, istatistiksel analiz yapmaya imkan verecek bir büyüklükte olmasına dikkat edildi. ☐

7. Diferansiyel madde fonksiyonu.

a. Pilot araştırma sırasında testin diferansiyel madde fonksiyonu (DMF) en az farklı iki grupta incelendi ☐

b. DMF için örnek grupları, gizli özellik boyutunda tam olarak eşitlendi. ☐

c. DMF yanında kullanılan modele göre, DDF veya DTF analizleri yapıldı ☐

d. DMF saptanan maddelerle ilgili olarak bilgi verildi ve bu maddelerin iyileştirilmesi için hangi çalışmaların yapıldığı tartışıldı ☐

e. DMF, DDF ve DTF analizleri için uygulanan yöntem açıklandı ..... ☐

8. Testin uygulama biçimi bilgileri.

a. Testin öz değerlendirme, yüz yüze değerlendirme, aletli test, telefon görüşmesi, gözlemci değerlendirmesi vb. ölçüm yöntemlerinden hangisiyle yapıldığı açık bir şekilde tanımlandı. ..... ☐

b. Testin uygulandığı ortam, zaman ve diğer çevre koşulları hakkında bilgi verildi. ..... ☐

c. Testin standart bir biçimde uygulanması için "uygulama yönergesi" hazırlandı ve bu yönergeye sıkı bir şekilde uyuldu ..... ☐

d. Testin *uygulama yönergesi* bilimsel rapor ekine alındı.

9. Klasik veya modern test modelinden hangisinin tercih edildiği.

a. Testin/ölçümün değerlendirilmesinde tercih edilen modeller ve bu modellerin niçin tercih edildiğine ilişkin bilgi verildi..... ☐

b. Söz konusu ölçümle ilgili olarak literatürde daha çok hangi ölçüm modellerinin kullanıldığı ve son yıllardaki eğilimin ne yolda olduğuna ilişkin bilgi verildi..... ☐

c. Ölçüm modelinin tercih edilmesinde hangi mantıksal gerekçelerden hareket edildiği açıklandı ..... ☐

10. Klasik test kuramı seçilmişse.

a. Maddenin güçlük derecesinin ,20 ilâ ,80 arasında olmasına dikkat edildi ..... ☐

b. İkili kodlanan verilerde maddenin varyansının ,16'dan yüksek olmasına dikkat edildi..... ☐

c. Madde-toplam puan korelasyonu, madde-toplam puan iki serili korelasyon analizi yapıldı..... ☐

d. İki serili korelasyon analizinde katsayı işaretinin pozitif olmasına ve değerin ,30'dan yüksek olmasına dikkat edildi..... ☐

e. Madde sayısı 50'den az ise iki serili korelasyon analizinde düzeltilmiş korelasyon katsayılarının kullanılmasına dikkat edildi ..... ☐

f. Yüzde 30'un altında korelasyon katsayısına sahip maddeler ile negatif işaretli korelasyon katsayısına sahip maddeler ölçekten çıkarıldı..... ☐

g. Maddenin ayırt etme indeks değerinin pozitif işaretli ve ,30'dan yüksek olmasına dikkat edildi..... ☐

h. Ölçeğe en yüksek korelasyon katsayısına sahip önceden belirlenmiş sayıda madde alındı ☐

i. Ölçeğe alınan maddelerin özet istatistiksel analizleri yapıldı ☐

j. Madde güçlük analizi yapıldı ve ana kütlenin geniş bir kesimini temsil etmesi için düşük *p* değerine sahip maddelere de ölçekte yer verildi. ☐

k. Uç *p* değerine sahip maddeler (,10'un altında ve ,90'ın üstünde) ölçekten, testten çıkarıldı ☐

l. Alternatif ve paralel formlar kullanılmışsa bunların nereden temin edildiği veya nasıl geliştirildiği konusunda bilgi verildi. ☐

m. Alternatif veya paralel formların aritmetik ortalama, standart sapma, varyans değerleri açık bir şekilde verildi. ☐

n. Paralel formların güçlük indeksi *p* değerleri, ayırma indeksi *d* değerleri, madde-toplam puan korelasyon katsayıları karşılaştırmalı olarak verildi ☐

o. Alternatif formlar tam paralel değilse aralarında ne ölçüde farklılık olduğuna ilişkin bilgi verildi. ☐

11. Ölçeğin / testin boyutluluk analizi.

a. Ölçeğin faktöriyel yapısını belirlemek için faktör analizi yapıldı ☐

b. Veri yapısına ve ölçeğin türüne uygun faktör analizi yöntemi seçildi ☐

c. Kavramsal yapıya ait alt boyutların/faktörlerin her birinin baskın faktör içerme durumu incelendi ☐

d. Ölçeğin tek veya çok boyutlu olma durumu hakkında ayrıntılı bir şekilde bilgi verildi. ☐

12. Biçim.

a. Anket / test formu cevaplayıcıya cazip gelecek bir şekilde, 12 puntodan daha küçük olmayacak yazı büyüklüğüyle ve bir sayfaya 35 satırdan daha fazla satır sığmayacak şekilde düzenlendi ☐

b. Anket formu / test normal koşullarda cevaplayıcının 20 dakikadan daha fazla zamanını almayacak şekilde düzenlendi. ☐

c. Raporda, testin ortalama tamamlama süresi hakkında bilgi verildi ☐

d. Anket formunda renk kullanımından kaçınıldı. ☐

e. Anket formunda koyu siyah yazım biçiminden kaçınıldı. Önem verilen kelimeler italik tarzıyla yazıldı. ☐

13. Maliyet.
    a. Ölçümün / testin / anketin birim maliyet hesaplaması yapıldı. ☐
    b. Toplam maliyetin hangi kaynaklardan karşılandığı hakkında bilgi verildi. ☐
    c. Ücret karşılığında çalışan anketçilerden veya ücreti karşılığında veri toplayan araştırma kuruluşlarından yararlanılmışsa, birim ve toplam maliyetler konusunda bilgi verildi ☐
14. İyileştirme.
    a. Testin/test maddelerinin kalitesinin iyileştirilmesi için yapılan çalışmalar hakkında bilgi verildi ☐
    b. Test/ölçek iyileştirme çalışmalarında kullanılan istatistiksel analiz yöntemleri hakkında bilgi verildi ☐
    c. Test maddelerinin yetersiz kaldığı ve daha sonraki çalışmalarda iyileştirilmesi gereken konular hakkında bilgi verildi ☐
15. Test sonuçlarının yorumlanması.
    a. Test sonuçlarının yorumlanmasında; testin kullanım amacı ve gürültü faktörleri hakkında gerekli açıklamalar yapıldı ☐
    b. Sonuçların yorumlanmasında ihtiyatlı bir dil kullanıldı ☐
    c. Sonuçların tek bir ölçüme ait mi yoksa birden fazla ölçüm sonuçlarına ait mi olduğu açık bir şekilde vurgulandı ☐
16. Test sonuçlarının kodlanması.
    a. Test sonuçlarının optik okuma cihazı, el ile girilme veya elektronik olarak sayılma durumu hakkında bilgi verildi ☐
    b. Test sonuçları bilgisayara el ile girilmişse veri doğrulaması için hangi yöntemlerin uygulandığı hakkında bilgi verildi ☐
    c. Verileri bilgisayara el ile girme işlemini kimin yaptığı hakkında bilgi verildi ve ayrıca tek bir ankete/ölçeğe/ölçüme ait verilerin ortalama giriş süresi hakkında bilgi verildi ☐
    d. Verilerin kodlanmasında ağırlıklandırma yönteminden yararlanılıp yararlanılmadığı hakkında bilgi verildi ☐
    e. Test verilerinin kodlama hatalarından arındırılması için hangi tür doğrulama işleminden yararlanıldığı hakkında bilgi verildi ☐

17. Standart hata.
    a. Ölçümün standart hatası, ortalamanın standart hatası veya durumu göre koşullu standart hata hakkında bilgi verildi ☐
    b. Yararlanılan standart hata formülünün tercih edilme nedeni hakkında bilgi verildi ☐
18. Norm değerleri.
    a. Ölçüme ait belirli gruplar için norm değerleri verildi ☐
    b. Ölçümde kullanılan kesim puanları hakkında bilgi verildi ☐
    c. Norm değerleri için tercih edilen standart puanlar hakkında bilgi verildi ve bu puanların niçin tercih edildiği açıklandı ☐
19. Örneklem.
    a. Ölçümün yapıldığı örneklem hakkında bilgi verildi ☐
    b. Ölçümün yapıldığı örneklem büyüklüğü hakkında bilgi verildi ☐
    c. Örnekleme yöntemi hakkında bilgi verildi ☐
    d. Örneklem hatası hakkında bilgi verildi ☐
20. Güvenilirlik.
    a. Klasik test kuramında, hangi güvenilirlik yönteminin uygulandığı ve niçin bu yöntemin seçildiği konusunda bilgi verildi ☐
    b. Güvenilirlik katsayıları her bir grup için ayrı ayrı hesaplandı ☐
    c. Güvenilirlik analizlerinde farklı analiz yöntemlerinden yararlandı ☐
    d. Güvenilirlik sonuçları birden fazla güvenilirlik analizi yöntemiyle desteklendi ve olabildiğince ikna edici, kuşkuları azaltıcı verilere ve bulgulara ulaşılmaya çalışıldı ☐
21. Geçerlilik.
    a. Yüzey ve içerik geçerliliği için hangi çalışmaların yapıldığı hakkında bilgi verildi ☐
    b. Tahmin ve birlikte vuku bulma geçerliliği ile ilgili bilgi verildi ☐
    c. Yapısal geçerlilik analizlerinin hangi yöntem çerçevesinde yapıldığı hakkında ayrıntılı bilgi verildi ☐

EK B. GÜVENİLİRLİK DEĞERLENDİRMESİ

Düzey	*Tanımlama*
0	Güvenilirlik hesaplaması yapılmamış ve bu nedenle de güvenilirlik bilgisi raporlanmamış.
1	Güvenilirlik olgusu daha önceki çalışmalara dayandırılmış. Önceki çalışmalar öğretim üyelerinin / meslektaşların gözden geçirdiği çalışmalar olabilir veya olmayabilir. Önceki güvenilirlik çalışmaları yeterli görülmüş.
2	Güvenilirlik bilgileri raporlanmış, fakat sonuçları değerlendirmek, incelemek ve net bir kanaate varmak için veri ve rakamların yeterli olmadığı anlaşılmıştır. Bu sonuçlar bir meslektaş tarafından incelenmemiştir veya yapılan inceleme yetersizdir.
3	Güvenilirlik bilgileri belirli ölçüde ayrıntılı bir şekilde ele alınmış ve raporlanmıştır. Fakat yapılan raporlama ve sunum tam değildir. Daha fazla bilgiye ihtiyaç vardır. Güvenilirlik bilgileri bir meslektaş /öğretim üyesi tarafından incelenmiş olabilir veya olmayabilir.
4	Güvenilirlik bilgileri yeterli ölçüde ve ayrıntılı olarak ele alınmıştır. Hangi analizlerin yapıldığı, nasıl yapıldığı açıklanmış, elde edilen bulgular tablolaştırılarak yorumlanmıştır. Sonuçlar meslekten olmayan kişiler arasında tartışılmış ve değerlendirmeye tabi tutulmuştur. Duruma göre sonuçlar danışman öğretim üyesi tarafından da incelenerek gerekli düzeltmeler ve iyileştirmeler yapılmıştır (Yüksek lisans ve doktora tezleri).
5	Güvenilirlik bilgisi yeterli ayrıntıda verilmiştir. Kullanılan matematiksel ve istatistiksel analizler net bir şekilde ortaya konarak sonuçların incelenmesine olanak sağlanmıştır. Sonuçlar, danışman öğretim üyesinin dışında resmî bir prosedür çerçevesinde tarafsız hakemler ve uzmanlar tarafından da incelenerek onaylanmıştır.

EK C. GEÇERLİLİK DEĞERLENDİRMESİ

Düzey	*Tanımlama*
0	Geçerlilik analizi yapılmamış ve bu nedenle de geçerlilik bilgisi raporlanmamış.
1	Geçerlilik olgusu daha önceki çalışmalara dayandırılmış. Önceki çalışmalar öğretim üyelerinin / meslektaşların gözden geçirdiği çalışmalar olabilir veya olmayabilir. Önceki geçerlilik çalışmaları yeterli görülmüş.
2	Geçerlilik bilgileri raporlanmış, fakat sonuçları değerlendirmek, incelemek ve net bir kanaate varmak için veri ve rakamların yeterli olmadığı anlaşılmıştır. Bu sonuçlar bir meslektaş tarafından incelenmemiştir veya yapılan inceleme yetersizdir.
3	Geçerlilik bilgileri belirli ölçüde ayrıntılı bir şekilde ele alınmış ve raporlanmıştır. Fakat yapılan raporlama ve sunum tam değildir. Daha fazla bilgiye ihtiyaç vardır. Geçerlilik bilgileri bir meslektaş /öğretim üyesi tarafından incelenmiş olabilir veya olmayabilir.
4	Geçerlilik bilgileri yeterli ölçüde ve ayrıntılı olarak ele alınmıştır. Hangi analizlerin yapıldığı, nasıl yapıldığı açıklanmış, elde edilen bulgular tablolaştırılarak yorumlanmıştır. Sonuçlar meslekten olmayan kişiler arasında tartışılmış ve değerlendirmeye tabi tutulmuştur. Duruma göre sonuçlar danışman öğretim üyesi tarafından incelenerek gerekli düzeltmeler ve iyileştirmeler yapılmıştır (Yüksek lisans ve doktora tezleri).
5	Geçerlilik bilgisi yeterli ayrıntıda verilmiştir. Kullanılan yargısal, matematiksel ve istatistiksel analizler net bir şekilde ortaya konarak sonuçların incelenmesine olanak sağlanmıştır. Sonuçlar, danışman öğretim üyesinin dışında resmî bir prosedür çerçevesinde tarafsız hakemler tarafından da incelenerek onaylanmıştır.

TERİMLER SÖZLÜĞÜ

Bu bölümdeki terimlerin açıklanmasında büyük ölçüde İnternet ortamındaki sözlüklerden yararlanılmıştır. Bununla birlikte, çok sayıda kaynaktan yararlanarak derleme yapılması nedeniyle, her bir madde için ayrı ayrı kaynak vermeye gerek duyulmamıştır.

açığa çıkaran modeller (unfolding modells). Madde-yanıt kuramında ikili dereceleme dışındaki çoklu derecelerde madde parametrelerini hesaplayan analiz modelleri.

ağırlıklandırma (weighting). Bir test bataryasında yer alan testlerin önemlerinin farklı olduğu durumda her birine sübjektif olarak belirlenen kriterlere göre belirli katsayılarla veya yüzde temeline bağlı olarak tartı değerleri verme. Testlere tartı değerleri vermeden önce bataryadaki tüm test sonuçları ortak bir test değerine dönüştürülür. Bunun için daha çok *z* puanları, *T* puanları veya standart dokuz puanları kullanılır.

akademik yetenek testi (scholastic test). Okullarda sayısal, sözel ve sosyal derslere ilişkin olarak öğrenmeyi veya öğrenilen bilgiyi kullanmayı ölçen daha çok, çoktan seçmeli sorular şeklinde hazırlanan bilgi testleri.

akıcı zeka (fluid intelligence). Cattell tarafından belirlenen iki zeka türünden biridir. Zekanın eğitim-öğretimle kazanılan bölümünü değil kalıtsal yönünü oluşturur. Kişinin problem çözme ve akıllı hareket etme yeteneğini ölçer.

altın kural (golden rule). Madde yanlılığı belirlenirken bir maddenin belirli bir gruptaki doğru yanıtlama oranı diğer gruptan %15 veya daha fazla ise bu maddenin ölçekten/testten çıkarılması. Ancak çıkarma işlemine girişilmeden önce maddenin gerçek yetenek farklılığını ortaya çıkarıp çıkarmadığı incelenmelidir.

alan örneklemesi modeli (domain sampling model). Bu modele göre ölçüm hatasının temel nedeni, kavramsal yapıyı temsil eden maddeler evreninden yapılan seçimin, örneklem maddelerinin yetersiz olarak belirlenmiş olmasıdır. Alandan seçilen maddeler arasındaki korelasyon eğer yüksekse bu maddeler tek bir kavramsal yapıyı ölçüyor demektir. Böyle bir durumda alan örneklemesinin iyi yapıldığı söylenir. Düşük korelasyona sahip maddeler ise uygun bir evrenden veya alandan seçilmemiş demektir. Uygun olmayan maddelerin ölçeğe alınması, sonuçta ölçüm hatalarına neden olmakta ve ölçeğin / testin güvenilirliğini düşürmektedir.

aletsellik (instrumentation). Gözlem veya ölçüm tekniğinde yapılan herhangi bir değişikliğin ölçüm sonuçlarının geçerliliğini tehdit etmesi.

alternatif formlar (alternate forms). Başarı ve yetenek testleri oluşturulurken aynı testin aynı kavramsal yapıyı ölçecek şekilde başka bir alternatifinin veya paralel formunun daha oluşturulmasıdır. Paralel formda aynı sayıda soru, aynı düzenleme, aynı alt faktörler bulunmalı ve istatistikî analiz sonucunda maddelerin güçlükleri, aritmetik ortalamaları, standart sapmaları ve ölçeğin diğer ölçeklerle olan korelasyonları yaklaşık olarak benzer çıkmalıdır.

ampirik veri (empirical data). Alan araştırması sonucu elde edilen verilerden hareket ederek, değişkenler arasında kovaryans matrisinin oluşturulması.

ana etken (main effect). Her bir bağımsız değişkenin, diğer faktörlerin etkisi göz önünde bulundurulmadan bağımlı değişken üzerinde yaptığı etki.

anti-imaj korelasyon matrisi (anti-image correlation matrix). Faktör analizinden sonra değişkenler arasındaki kısmî korelasyon katsayıları matrisi.

apsis (abscissa). Kartezyen grafik çiziminde yatay eksen.

ardışık etkisi (sequence effect). Bir önceki sırada yer alan soru veya maddeye benzemesi nedeniyle daha sonraki maddelere verilen yanıtın önceki maddeden etkilenmesi.

artık değerler (residuals). Gözlem değerleri ile tahmin edilen değerler arasındaki fark.

artık değerler korelasyon matrisi (residual correlation matrix). (a) Gözlem değerleri korelasyon matrisi ile faktör puanlarıyla yeniden üretilen korelasyon matrisi değerleri arasındaki fark. (b) Değişkenlerin/maddelerin özge değerleri arasındaki korelasyon katsayılarını gösteren tablo.

askıntı parametreleri (nuisance parameters). Madde-yanıt kuramında yapısal parametrelerin zorunlu bir uzantısı niteliğindeki yetenek parametreleri.

gürültü değişkenleri (nuisance variables). İki değişken arasındaki ilişkiler incelenirken araya girerek bu değişkenleri etkileyen diğer değişkenler. Bir diğer adı, kontrol değişkenleri.

atanmış değer (imputed value). Bilinmeyin veya eksik bırakılan verilerin yerine geçmek üzere kullanılan tahmin değeri.

ayırma güvenilirliği (separation reliability). Rasch analiz yönteminde maddelerin ayırt etme güvenilirliği.

ayırma indeksi (discrimination index). (a) Thurstone ölçeğinde homojen olmayan maddelerin ayıklanmasını sağlayan değerlendirme yöntemi. (b) Klasik test kuramında bir maddenin yüksek puan alan öğrencilerle / kişilerle düşük puan alan öğrencileri / kişileri ayırt etme gücü. Ayırma indeksi –1,00 ilâ +1,00 arasında değişir. Diğer koşulların eşit olması şartıyla ayırma indeksinin büyük olması o maddenin ayırt etme özelliğinin yüksek olduğunu gösterir. Ayırma indeksi şu formülle hesaplanır: başarısı yüksek ilk %27 – başarısı düşük son %27 / (,5) x her iki gruptaki kişilerin toplam sayısı. Ayırma indeksinin ideal olarak pozitif işaretli ve ,30'un üzerinde olması istenir. Ayırma değeri ,20'inin altındaki maddeler zayıftır. Bilim adamları, *D* indeksinin ,30'un üzerinde olmasının o maddenin ayırt etme özelliğinin yüksek olduğu anlamına geleceğini belirtmişlerdir. Ayırma indeks değeri ,40'ın üstündeyse o maddenin parlak öğrencilere avantaj sağladığı anlamına gelir. Ayırma indeks değeri sıfır, sıfıra yakın veya negatif değerli ise tahmin veya şans faktörü çerçevesinde işaretleme yapan öğrencilerin ödüllendirildiği şeklinde yorumlanır. Soruların güçlük derecesi yükseldiği oranda ayırma indeksi sıfıra yaklaşır. Ayırma indeksi her zaman maddenin kalitesini gösteren bir gösterge olarak değerlendirilmez. Çok zor ve çok kolay maddelerin ayırma indeks değerleri düşüktür, fakat bu maddeler konunun içerik geçerliliği ve dağılım ranjı açısından gereklidir. Bir test içerik olarak farklı

konuları kapsıyor ve farklı bilişsel yetenekleri ölçüyorsa bir maddenin ayırma indeksi düşük çıkabilir. Ayırma indeksi nokta-iki serili korelasyon analizi yöntemiyle de hesaplanabilir ve aynı zamanda *madde güvenilirlik indeksi* olarak bilinir.

ayırma katsayısı (discrimination coefficient). Madde-yanıt kuramında zayıf ve kuvvetli öğrencileri/kişileri test maddesinin ayırt etme özelliğini ortaya koyan değer. Bir maddenin ayırt etme özelliğinin yüksek sayılabilmesi için ayırma katsayısının en az ,20 olması gerektiği belirtilmiştir. Bu değerin altındaki maddeler testten çıkarılır.

aykırılık (discrepancy). Rasch modelinde bir veya daha fazla beklenmedik yanıt örneği.

ayrık değerler (outliers). Hatalı girilen veya ana kütlede çok küçük bir kesimi temsil eden uç değerler.

bağımsız değişkenler (covariates). Birlikte değiştiriciler. Bağımlı değişken üzerinde etkili olan faktörler.

bağlantı (linking). Maddeleri az çok birbirleriyle karşılaştırılabilir hale getirme işlemi. Katı bir şekilde istatistiksel eşitleme yönteminden başlayarak sosyal yargılara dayalı eşitlik düzeyine kadar uzanır.

bağlantı maddeleri (linking items). Her iki paralel formda yer alan ortak maddeler.

bağlı sıralılık (tied ranks). Büyüklük sırasında bir ham değerden iki tane varsa ve bu değerler örneğin 4. ve 5. sırada yer almışsa her ikisinin de büyüklük sırası 4,5 olarak belirlenir ve bu şekilde sıralanan değerler, "bağlı sıralılığa sahip" olarak adlandırılır.

Bartlett küresellik testi (Bartlett test of sphericity). Bir korelasyon matrisi içindeki bütün korelasyonların anlamlılık testi.

başarı testi (performance - achievement test). Kişilerin eğitim, kurs veya deneyim aracılığıyla elde ettikleri bilgi ve becerilerin ne ölçüde yetkinlik düzeyine ulaştığını ölçen veya kişilerin sahip oldukları yeteneklerin onların ne ölçüde yetkin olduğunu gösteren ölçüm araçları. Başarı testi, bilgi testlerinin dışında daha çok beceri testleriyle ilgilidir. Yetenek veya yatkınlık testleri ise bilgi ve deneyimle ilgili değildir. Ancak literatürde *başarı testleri* kavramı kullanılırken yatkın-

lık, beceri ayrımı yapılmaz. Başarı testi kavramı aynı zamanda yatkınlık testleri için de kullanılır.

batarya (battery). Kişilerin değişik alanlardaki yetenek ve becerilerini ölçen, aynı ana kütle veya örneklem grubu için standartlaştırılmış bir grup test.

Bayes istatistiği (Bayesian statistics). Karar verme amacıyla kullanılan bir istatistikî analiz türü. Bu yaklaşımda önceki araştırmalardan elde edilen veya tahmine dayalı olarak belirlenen olasılık oranları mevcut araştırmadan elde edilen oranlarla karşılaştırılarak bir karara varılır.

baz ölçüm (baseline measuring). İlk temel ölçüm. Daha sonraki ölçümlerin karşılaştırılmasında kriter olarak kabul edilir.

bazal ape (basal ape). Bütün test maddelerini eksiksiz olarak başarıyla cevaplandıran kişi.

beklenen değer (expected value). Bir dizi değerin ortalama değeri. Ana kütleden çok sayıda örneklem seçildiğinde muhtemelen elde edilecek ortalama değer. Ana kütleden çok sayıda örneklemin seçilemediği durumda araştırma yapılan örneklemin ortalama değeri geçici olarak "beklenen değer" olarak ele alınır veya beklenen değer olduğu düşünülür.

benimsenmiş yanıt verme eğilimi (acquiescence response set). Kısaca *yanıt eğilimi* adı da verilen bu terim, tutum veya kişilik ölçeklerinde bireylerin belirli türdeki maddelere *Evet* veya *Hayır* verme eğilimi içinde olmalarını yansıtır.

biçimli diferansiyel madde fonksiyonu (uniform differential item functioning). İki farklı grupta ölçüm yapıldığında madde özellikleri eğrilerinin birbirlerini kesmeksizin aralarında belirli bir mesafe bırakarak paralel bir dağılıma sahip olması. Bu durum, her iki grubun gizli boyutta aynı teta değerine sahip olmadıklarını gösterir.

biçimsiz diferansiyel madde fonksiyonu (nonuniform differential item functioning). İki farklı grupta ölçüm yapıldığında madde özellikleri eğrilerinin birbirlerini çaprazlamasına kesmesi. Birinci grupta kişiler ortalamanın altında kalırlarken, ikinci gruptaki kişiler ortalamanın üstünde bir teta değerine sahip olurlar.

bileşik ölçek (composite scales). (1) Çok sayıda maddeden meydana gelen bir test veya ölçek. (2) Çok sayıda testten meydana gelen batarya.

bileşik ölçüm puanı (composite score). (1) Çok sayıda maddeden meydana gelen bir test veya ölçeğin toplam/ortalama puanı. (2) Birden fazla sayıda teste ait ham puanlar; standart *z* puanlarına dönüştürülerek, doğru yanıt oranları temel alınarak, ağırlıklandırılarak veya geliştirilen özel formüller kullanılarak kombine edilmesiyle elde edilen tek bir toplam veya ortalama puanı.

bilgi fonksiyonu (information function). Bir maddenin bilgi fonksiyonu maddenin varyansına benzer. Bilgi fonksiyonu bir indeks değeridir, fakat daha çok kemer eğrisi ile gösterilir. Belirli bir yetenek düzeyinde (θ) maddeyi cevaplayanları başarılı bir şekilde ayırt etmesi o maddenin bilgi fonksiyonunu oluşturur. Daha sonra çok sayıda *madde bilgi fonksiyonu* bir araya gelerek bu kez *test bilgi fonksiyonunu* veya "test bilgi eğrisini" oluşturur.

bilimsel notasyon (scientific notation). İstatistiksel analiz programı SPSS'te hesaplanan rakamlar ,01'den küçük çıktığı zaman bilimsel simgelerle gösterilir. Örneğin; –2,136E-04'nin anlamı –0,0002136 ve 2,136E-02'nin anlamı ise 0,02136'dır.

bilişsel alan (cognitive domain). Problem çözme, öğrenme ve bilme gibi insan davranışlarını ilgilendiren ölçüm alanı.

Birleşik Standartlar (Joint Standards). ABD'de test yazımı ve geliştirilmesinde temel kuralları belirleyen ve American Educational Research Association (AERA), American Psychological Association (APA), ve National Council on Measurement in Education (NCME) isimli kuruluşlar tarafından ortaklaşa olarak hazırlanan *Standards for Educational and Psychological Testing* isimli test standardı.

birleştirilmiş standart sapma (pooled standard deviation). Deney ve kontrol gruplarına ait standart sapma değerlerinin ortalamasının kare kökü.

kombinasyon analizi (conjoint analysis). İkili karşılaştırmaya dayanan bir ölçeğin/ölçümün maddelerine veya ögelerine bütününden daha fazla önem vererek maddelerin ikili veya çok dereceli bir boyut üzerinde değerlendirilmesiyle test içinde hangi maddenin veya ögenin daha

önemli olduğunun ortaya çıkarılması. Kombinasyon analizi yapılacak maddeler karşılaştırma matrisleri kullanılarak ölçülür.

Bloom taksonomisi (Bloom's taxonomy). Bloom tarafından önerilen sınav sorularının belirli bir sınıflandırmaya uygun olarak hazırlanması. Buna göre test soruları altı grup içinde hazırlanır: bilgi soruları, kavrama soruları, uygulama, analiz, sentez ve değerlendirme soruları.

bulaşma etkisi (contamination). Bir müdahalenin, ölçümün veya testin araya giren diğer değişkenlerden etkilenmesi ve bu nedenle sonuçların gerçek ölçüm değerlerini değil, onun yanında parazit etkileri de içermesi.

cimrilik ilkesi (parsimony principle). İki veya daha fazla teori verileri eşit ölçüde iyi bir şekilde açıklıyorsa en basit ve kolay anlaşılır olan teorinin tercih edilmesi. Faktör analizinde iki veya üç faktör yaklaşık olarak aynı miktarda varyansla modeli açıklıyorsa üç yerine iki faktörün tercih edilmesi.

cinsiyet yanlılığı (sex bias). Test veya ölçekteki maddelerin erkek veya kadınların daha avantajlı olmalarını sağlayacak şekilde oluşturulması. Sistematik hatanın cinsiyet faktörü nedeniyle ortaya çıkması.

çıpa değeri (anchor value). Bir ölçüm sisteminde kriter olarak alınan çıpa maddenin/maddelerin sahip olduğu, önceden belirlenmiş değer veya değerler.

çıpa madde (anchor item). Bir ölçüm sisteminde maddelerin iyileştirilmesinde temel alınan kriter madde (maddeler) veya metin. Bu maddeler veya metin birer "mihenk taşı" olarak hizmet görür. Termometrede suyun kaynama noktası ve donma noktasının diğer termometre derecelerinin belirlenmesinde çıpa olarak görülmesine benzer şekilde bazı maddeler diğer maddeleri iyileştirmek için temel alınabilir. Örneğin, en basit bir soru alt düzey yeteneği ve en zor bir soru ise üst yeteneği göstermek üzere belirlenebilir. Çıpa madde iki veya daha fazla testte yer alır. Çıpa maddeler özellikleri ve davranışları iyi bilinen maddelerdir. Bu maddeler testin yeni bir versiyonu oluşturulurken teste alınır ve testte kişilerin bu maddelere nasıl yanıt verdiğine bakılır.

çok dereceli maddeler (polythomous items). İkiden fazla dereceye, şıkka veya yanıta sahip maddeler. Çok dereceli maddeler ile çok kategorili

maddeler birbirinden farklıdır. Çok dereceli maddelerde yanıtlar sıralı büyüklüğe sahipken çok kategorili maddelerde böyle bir sıralama söz konusu değildir.

çoklu veriler (polythomous data). İkiden fazla dereceye, şıkka veya yanıta sahip veriler. Çoklu madde ile aynı anlamda.

çoklu doğrusallık (multicollinearity). *Bk.*, Koşutluk.

değişmezlik (invariance). Madde-yanıt kuramında kullanılan bir kavramdır. Madde parametrelerinin seçilen alt grup yanıtlayıcılarında farklı değerler vermeyeceği ve yine kalibre edilmiş maddeler uygulandığında kişilerin yetenek parametre değerlerinin değişmeyeceği düşüncesi.

dış kriter (external criterion). Mevcut test sonuçlarının güvenilirlik ve geçerliliğinin karşılaştırmalı olarak incelenebilmesi için temel alınan başka bir test, değerlendirme uygulaması veya başka bir ölçüm. Başarı testlerinde dış kriter, "zeka testi" veya "yıl sonu başarı notları" olabilir. Yetenek testlerinde nezaretçinin yaptığı "performans değerlendirme puanları" dış kriterdir. Test sonuçlarının geçerliliğini saptamak için dış kriter puanlarıyla test puanları arasında korelasyon analizi yapılarak sonuçlar arasındaki ilişkinin gücüne bakılır.

diferansiyel deste fonksiyonu (differential bundle functioning). Diferansiyel deste fonksiyonu tek bir madde yerine bir grup maddenin kombine olarak değişik gruplarda farklı bir şekilde çalışıp çalışmadığının incelenmesidir. *Bk.*, testçik.

diferansiyel madde fonksiyonu (differential item functioning). Maddelerin değişik gruplarda farklı çalışması ve bu nedenle ayrımcılığa neden olma işlevi. Test maddelerinin aynı yetenek düzeyindeki veya aynı bilgi düzeyindeki kişiler arasında cinsiyet faktörüne göre, ırk faktörüne göre veya sosyoekonomik faktörlere göre farklı şekillerde çalışması. Ölçüm yapılan odak grup puanlarının diğer grupların puanlarından daha yüksek veya daha düşük çıkması.

doğrusallık (linearity). İki değişken arasındaki ilişki veya değişkenlere ait noktaların beklenen dağılım eğrisine uygun düşmesi.

duygusal alan (affective domain). Tutumlar, inançlar, hisler ve değerlerle ilgili alan.

emik ve etik maddeler (emic and etic items). Belirli bir alandaki maddelerin benzerliğini ölçmenin iki yöntemi vardır. Bu iki yöntem emik ve etik maddeler olarak adlandırılır. Maddeleri benzerliğine göre yığma emik yöntemi tanımlar. Emik yöntem *direkt* sınıflandırma temeline dayanır. Emik maddeler kendilerini hemen ele veren ifadelerdir. Emik uygulamada birbirine benzer olan maddeler bir yığın haline getirilirken benzemeyenler ayrı bir yığın olarak belirlenir. Maddeler bu şekilde temel bir sınıflandırmaya tâbi tutulduktan sonra emik maddeler etik yöntemle tekrar incelenir. Etik yöntem, dolaylı veya *endirekt* sınıflandırmayı tanımlar. Bu maddelerin ölçüm alanıyla ilgisi daha dolaylıdır. Emik kelimesi İngilizcedeki "(phon)emic" ve etik kelimesi ise "(phon)etic" sözcüklerinden türetilmiştir. Kavramlar, maddede kullanılan dilin madde özelliklerini açıklamaya hizmet etmesi nedeniyle tercih edilmiştir.

ranj (range). Sınırlandırılmış alandaki belirli bir puan aralığı.

eşik değer (threshold value). Madde-yanıt modelinde *b* parametresi değeri. Başarı testlerinde bir maddeye doğru yanıt vermek için gerekli olan gizli değişkenin miktarı. Tutum ölçeklerinde bir maddeyi onaylamak için gerekli olan gizli değişkenin miktarı.

eşik parametresi (threshold parameter). Madde güçlük parametresi.

eşitleme (equating). İki veya daha fazla alternatif forma ait puanların birbirlerinin yerine kullanılabilmeleri için puanların ayarlanmasını sağlayan istatistiksel işlem.

etki büyüklüğü (effect size). Bir tedavinin, revizyonun veya müdahalenin ne ölçüde etkili olduğunu gösteren standart değer.

etkileşim etkisi (interaction effects). İki veya daha fazla bağımsız değişkenin birbirlerini etkileyerek (etkileşim içine girerek) bağımlı değişkeni birlikte etkilemesi.

faktör (factor). (a) Etken. (b) Psikolojik kavramsal yapı. (c) Psikometride sözel, sayısal veya düzlemsel yeteneklerden her hangi biri. (ç) İstatistik analiz programı SPSS'te General Linear Model mönüsünde kategorik veya bağımsız değişken. (d) Alt test puanları arasındaki kovaryans. (e) Temel bileşenler analizinde orijinal değişkenlerin doğrusal kombinasyonu. (f) Faktör analizinde gizli yapı.

faktör yükleri (factor loadings). Orijinal değişkenlerle faktörler arasındaki ilişkileri gösteren standardize edilmiş regresyon katsayıları.

faktör matrisi (factor matrix). Faktör yüklerini gösteren tablo.

faktöriyel tasarım (factorial design). Bilim adamının iki veya daha fazla bağımsız değişkenin (faktörün) bağımlı değişken üzerindeki etkisini veya bağımsız değişkenlerin birbirleriyle olan etkileşimleri çerçevesinde bağımlı değişken üzerindeki etkisini araştırmak için kurduğu araştırma modeli.

Faktör temelli tasarım (factorial design). Ölçümlerin çoğunda tek bir faktörün etkisi belirlenmeye çalışılır. Örneğin, iş tatminsizliğinin düşük örgütsel bağlılığa yol açması gibi. Faktör temelli tasarımda ise, ölçülen bağımlı değişkendeki değişkenlikte birden fazla faktörün rol oynayıp oynamadığı araştırılır. Örneğin, düşük örgütsel bağlılıkta; iş tatminsizliği, kıdem, cinsiyet ve yönetim düzeyi faktörlerinin birlikte etkili olup olmadığının araştırılması faktöriyel tasarımı gerektirir. Bu tasarımda, her bir bağımsız faktörün etkisini tek tek ve ayrıca etkileşimli olarak birlikte görmek mümkündür. Faktörler araştırmacının kontrolündedir veya tesadüfî olarak ortaya çıkmış olabilir.

fark ettirilmeden toplanan gözlem verileri (unobtrusive measures). Kişilerin kendi üzerlerinde gözlem yapıldığını fark etmedikleri ve bu nedenle herhangi bir rahatsızlık yaratmadan toplanan ölçüm verileri.

fazlalıklar, gürültü etkenleri (nuisances). Maddede istenmeyen varyansa neden olan can sıkıcı etkenler.

genel faktör (general factor). Bütün değişkenleri temsil eden faktör. Bütün değişkenlerle yüklü olan faktör.

gönüllü katılım beyan formu (informed consent form). Ölçüme katılan kişilerin ölçümün/araştırmanın amacını ve ölçme işleminin ne şekilde yapılacağını önceden bildiklerini gösteren ve ölçüme kendi rızalarıyla katıldıklarını bildiren ad ve soyadlarıyla birlikte imzalarını içeren form.

göstergeler (indicators). Bir ölçekteki maddeler. Ayrıca *bk.*, "ölçüm ajanları".

gözlenen korelasyon matrisi (observed correlation matrix). Ham verilere dayalı olarak elde edilen korelasyon matrisi.

güçlü (robust). Normal dağılım özelliği göstermeyen verilerde dahi bu durumdan etkilenmeyip doğru hesaplama yapabilme yeteneğine sahip olma.

güçlük indeksi (difficulty Index). Klasik ölçüm kuramında bir maddeye doğru yanıt veren kişilerin oranıdır ve *p* simgesiyle gösterilir. Oranın yüksek çıkması test maddesinin daha kolay olduğunu gösterir. (Bazen *q* simgesi kullanılarak yanlış cevap veren kişilere işaret edilir.) Maddelerin güçlük indeks değerlerinin ,15 ilâ ,85 gibi değerler arasında dağılması arzu edilir. Bu değerlerden yüksek ve düşük olan maddeler testten çıkarılır. Kriter referanslı testlerde ise bu kurala uyulmaz. Ayrıca *bk.*, "madde güçlüğü".

gürültü (noise). Sistematik olmayan, tesadüfî hata. Ölçümün dikkatli, titiz ve özenli bir biçimde yapılmaması nedeniyle araya giren veya ölçüme karışan yabancı etkenler.

ham puanlar (raw scores). Herhangi bir ölçümden sonra elde edilen ilk puanlar. Ham puanlar değerlendirme yapmak için uygun görülmemiştir. Değerlendirme ve yorum yapma kolaylığı açısından bu puanlar "gelişme puanları" adı verilen *sınıf eşlik* puanlarına veya *standart puanlara* dönüştürülür. Bir diğer dönüştürme işlemi "statü puanları" olarak da isimlendirilen *yüzdelik* veya *standart dokuz* puanlarıdır.

hiyerarşik yapı (hierarchical structure). Ölçülen kavramsal yapının hiyerarşik bir yapı göstermesi. Maddeler önce kendileriyle ilgili birinci düzeydeki kavramsalı yapıları/faktörleri ortaya çıkarır ve daha sonra faktörler de genel kavramsal yapıyı ortaya koyar.

homojen alt ana kütle (homogeneous sub-population). Madde-yanıt kuramında test maddesi çok geniş bir kitleye uygulandığında belirli bir logit değerin üzerinde, altında veya belirli logit değerleri arasında puan alan kişilerden oluşan kesim. Maddeye doğru yanıt verme olasılığı, homojen alt ana kütle içinde araştırılır.

IQ ve IQ eşlik değerleri (IQ and IQ Equivalent Scores). Zekâ katsayısı ve zekâ eşlik değerleri. Zekâ katsayısı, zihinsel yaş kronolojik yaşa bölünmek ve 100 ile çarpılmak suretiyle bulunur. Zeka eşlik değeri ise, *z* puanlarının 15 ile çarpılması ve 100 ile toplanması suretiyle bulunur.

ikili indeks (index). Gizli bir değişkeni ölçtüğü varsayılan bir dizi madde. Tek göstergeli maddelere ve içinde birden fazla gizli değişken bulunan ölçeklere göre tek gizli değişkenli indekslerdeki maddeler birbirleriyle daha fazla ilişkilidir.

ikili normal dağılım (bivariate normal distribution). Pearson momentler çarpımı korelasyon analizinde her iki değişkenin/maddenin de normal dağıldığı varsayımı. Böylece her iki değişken, bütünleşik tek bir normal dağılım eğrisine sahiptir.

ikili yanıtlar (dichotomous item responses, binary data). Cevap şıkları *evet-hayır* veya *doğru-yanlış* şeklinde olan ve 1-0 şeklinde kodlanan yanıtlar. Kişilik ölçeklerinde ve diğer sosyal ölçümlerde "puanlama anahtarına" göre *var-yok* şeklinde yine ikili olarak kodlanan değerler. Kişilik ölçeklerinde boyutlar *içe dönük, dışa dönük* örneğinde olduğu gibi çift kutuplu ise birinci kutup 1 ve ikinci kutup 0 olarak kodlanır. Çoktan seçmeli sorularda verilerin doğru veya yanlış şeklinde kodlanması. İkili yanıtlar, sorudaki şık sayısıyla değil, sorunun ikili olarak kodlanmasıyla ilgilidir.

izlem çizgileri (trace lines). Madde-yanıt kuramında ayrı grafikler şeklinde çizilen veya aynı grafik üzerinde gösterilen "madde özellikleri eğrisi" çizgileri. Madde özellikleri eğrisinin bir diğer adı. Çizgi, en düşük yetenek düzeyinden en yüksek yetenek düzeyine doğru eğik bir S harfi şeklinde gözükmesi nedeniyle "izlem çizgisi" olarak adlandırılmıştır.

kabiliyet (aptitude). Doğuştan gelen ve sonradan kazanılan özelliklerin kombinasyonu. Kişiye belirli ölçüde eğitim verilmesi ve egzersiz yaptırılmasıyla belirli yetenek ve becerilere sahip olma gücü.

kalan (residual). Bir kişiye ait gözlem puanları ile formülün hesapladığı değerler arasındaki fark.

kalibrasyon (calibration). Birkaç anlamda kullanılır. Gözlemci veya değerlendiricilerin yaptıkları puanlamalarda homojenliği sağlayarak güvenilirliği artırma anlamındadır. Birden fazla değerlendiriciden yararlanıldığında bu kişilere geri besleme yapılarak benzer puanlar vermelerini sağlama kalibrasyon çalışmasıdır. Puanlama kuralları üzerinde hakemlerin/gözlemcilerin/değerlendiricilerin mutabakat geliştirmeleriyle birlikte kalibrasyon gerçekleşmiş demektir. Kalibrasyonun ikinci anlamı, test maddelerinin güvenilirliğini sağlamaktır. Kla-

sik test kuramında maddelerin toplam puanla yüksek korelasyon katsayısına sahip olmasını sağlamak iken modern test kuramında bir testin logaritmik ölçek derecelerini belirleme sürecidir. Madde-yanıt kuramında, farklı testlerdeki güvenilir ve istikrarlı maddelerin (çıpa maddeler) seçilerek karşılaştırma kriteri olarak kullanılması ve böylece ortak güçlük derecesine sahip yeni bir test veya ölçek (teta ölçeği) oluşturulması kalibrasyondur. Rasch modelinde ise, ölçüm ajanı boyutu üzerindeki bir değişkenin güçlük dereceleridir.

kalibrasyon örneklemi (calibration sample). Madde-yanıt kuramının temel alındığı ölçümlerde maddelerin ilk kez sınandığı örnek kütle.

karışıcı değişken (confounding variable). İki değişken arasındaki nedensellik ilişkileri incelenirken araya giren ve her iki değişkeni de etkileyen üçüncü bir değişken. İlişki aranılan *neden* ve *sonuç* değişkenlerinin her ikisinin de üzerinde etkili olan ve bu nedenle kontrol altına alınmadığı durumda şaşırtıcı, kafa karıştırıcı bulgulara yol açan faktör. Örneğin, "içki – akciğer kanseri" ilişkileri incelemesinde sigara karışıcı değişkendir, çünkü hem akciğer kanseri üzerinde ve hem de içki içme davranışı üzerinde etkili olur. Karışıcının etkisi her iki değişkende de aynı yönde ise "pozitif karışıcı", farklı yönde ise "negatif karışıcı" olarak adlandırılır.

karmaşık yapı (complex construct). Ölçülmek istenen kavramsal yapının tek bir özellik yerine birden fazla özelliğin, faktörün veya alt boyutun birleşmesinden meydana gelmiş olması.

katışık değişkenler (confounded variables). İki değişken birlikte hareket ediyor ve birlikte değişkenlik gösteriyorsa, hangi değişkenin diğeri üzerinde daha fazla etkili olduğu belirlenemiyorsa bu iki değişkene verilen addır. Örneğin, *p* "anlamlılık" değerleri katışıktır. Değer, iki değişken arasında *istatistiksel olarak anlamlı bir ilişki olduğunu* gösterirken bu ilişkide hangi değişkenin daha etkili olduğunu belirtmez.

kaynak test (seed test). Asıl test. Paralel formlar uygulamasında madde bankasında yer alan maddeleri uyarlamak için kullanılan veya karşılaştırma kriteri olarak temel alınan test.

kemer eğrisi (ogive function). Lojistik fonksiyon adı da verilen bu eğrinin çizimi değişik açılarda düzenlenen kemere benzer. En geniş açılı kemer, "kaş-kemer"dir. Bunun dışında kemer eğrisi sivri, basık, geniş veya dar olabilir. Kemer eğrisinin bir diğer adı, "bilgi fonksiyo-

nu" eğrisidir. Logaritmik değerler alınarak çizilen kemer eğrisi, ham puanlar temel alınarak çizilen "normal dağılım" eğrisiyle karıştırılmamalıdır.

kesikli değişkenler (discrete variables). Cevaplama şıkları az sayıda değerden oluşan maddeler.

kesinlik (precision). (1) Ölçümde tesadüfî hatanın bulunmaması. Ölçümün tesadüfî hatalardan arındırılması. (2) tahmin edilen değerin belirli bir olasılık düzeyinde yer alabileceği aralığın genişliği.

kesişimsizlik (asymptotic, asymptote). Kesişim noktasına yakın bir biçimde sonsuza uzanma. Matematikte x veya y doğrusuna yaklaşan bir eğrinin doğruyu kesmemesi, doğruya sonsuzda eşitlenmesi veya doğruyu sonsuzda kesmesi.

kesit araştırması (cross-sectional survey). Belirli bir zaman diliminde yapılan alan araştırması. Çalışmada gözlem, mülakat veya anket yöntemi uygulanmış olabilir.

kısmî kredili maddeler (partial credit item). Birden fazla doğru şıkkın bulunduğu sorularda sadece bir şıkkın doğru işaretlenmesi halinde ilgili kişiye kısmî bir puan verilmesi.

kısmî kredili model (partial credit model). Hollandalı matematikçi Rasch tarafından geliştirilen bu yaklaşımda, çok dereceli birden fazla madde ile tek bir kavramsal yapıyı ölçen tutum ölçekleri yerine, her bir maddenin kendi dereceleme yapısına sahip olduğu tek maddeli ölçekler kullanılır. Kısmî kredi modeli, çoktan seçmeli sorulardan türetilmiştir. Çoktan seçmeli sorularda cevabı yanlış olmakla birlikte kişinin bazı şeyler bildiğini gösteren maddelere kısmi bir kredi (puan) verilir. Kısmî doğruluk derecesi maddeler arasında farklılık gösterebilir.

mihenk değer, kıyas değeri (benchmark). Karşılaştırma yapmak amacıyla kullanılmak üzere belirli bir meslekî örgüt, kurum veya uzman tarafından resmî veya gayrı resmî bir şekilde belirlenmiş standart bir değer veya standart değerler dizisi. Mihenk değer, bir ölçek veya testte başarılı olunduğunun göstergesi olarak kabul edilebileceği gibi ilişkiyi test eden korelasyon analizi sonucunun gücünü belirlemeye yönelik olarak da oluşturulabilir. Başarıyı değerlendirmek için ölçüm

sonucunda elde edilen gözlem değerleri söz konusu mihenk değerlerle karşılaştırılır.

ki-kare testi (chi-square test). Karşılaştırılan verilerin ki-kare dağılımına sahip olup olmadığını belirlemek için kullanılan herhangi bir test. Mantel-Haenszel iyi bilinen bir ki-kare testidir.

konjenerik model (congeneric model). Ortak gen modeli. Bu modelde ölçüm maddeleri arasında hata paylarının eşit olması, maddelerin gizli değişkenle eşit ölçüde ilişkili olması gibi bir gereklilik söz konusu değildir.

konu içeriği uzmanları (subject matter experts). Konuyla ilgili görevliler, yükümlüler, nezaretçiler, konuyu öğreten eğitimciler, öğretim elemanları, uzmanlar ve konuyu iyi bilen hakemler gibi ölçüm yapılan kavramsal yapı hakkında yeterli bilgi, deneyim ve becerisi olan kişiler. İlgili kişilerin çok sayıda olması halinde kısmî bilgiye sahip olanların yerine en iyilerinin seçilmesine çalışılır. Bu kişiler aynı zamanda asgarî yetenek düzeyine sahip bireylerin bir soruyu hangi oranda çözebilecekleri konusunda fikir yürütürler. Bir işle ilgili olarak bilgi, yetenek ve becerilerin belirlenmesinde "konu içeriği uzmanlığı" görevini yerine getirecek en iyi kişiler çalışanların ilk amirleridir.

koşutluk (multicolinearity). Test maddelerinin ikişerli olarak birbirleriyle yüksek derecede ilişkili olması (r >,90 olması). Aynı zamanda "çoklu doğrusallık" olarak isimlendirilir. Bir değişkenin başka bir değişkenin yerine geçebilecek kadar benzer olması.

kovaryans (covariance). İki değişkenin birlikte (kombine olarak) değişkenlik gösterme derecesini temsil eden istatistikî değer. Standardize edilmemiş korelasyon katsayısı.

köprü maddesi (link item). Çıpa olarak kullanılan madde.

kriter referanslı test (criterion-referenced test). Okullarda öğrencilerin, iş hayatında çalışanların belirli üst yetenek düzeylerine sahip olup olmadıklarını belirlemek amacıyla kullanılır. Bu testlerde kişilerin başarı puanları önceden belirlenmiş sınır/kesim puanlarıyla karşılaştırılarak bu puanı aşan kişilerin "yetkin" olduğuna karar verilir. Sınır puanı, "normal başarıyı" (%50-%69 yüzdelik dilim), "normalin üzerindeki" başarıyı (%70-%84 yüzdelik dilim), "olağanüstü başarıyı"

(%85-%94 yüzdelik dilim) veya "süper başarıyı" (%95+ yüzdelik dilim) ortaya çıkaracak şekilde belirlenebilir. Yüzde 95'lik dilimin üzerinde kalanlar üstün yetenekli kişilerdir. Kesim puanı, üst düzey başarıyı göstermek üzere +2 standart sapma şeklinde de belirlenmiş olabilir. Ortalamadan +1 standart sapma yönündeki başarı "normal" olarak tanımlanır. Yetkinlik normalin üzerindeki başarıyla ilgilidir. Testin kendisi değil, sonuçların değerlendirilme biçiminin farklı olması nedeniyle kriter referanslı olarak adlandırılmıştır. Norm referanslı testler, genel amaçlı değerlendirmelere konu olurken, kriter referanslı testler kişilerin spesifik yetenek alanlarında yetkin olma durumlarını belirlemeye yönelik olarak kullanılır.

logit (logit). Gizli özellik modellerinden biri olan Rasch yönteminde kullanılan eşit aralıklı ölçüm birimidir. Bir testteki doğru–yanlış oranının doğal logaritmaya dönüştürülmesi suretiyle elde edilir. Testin uzunluğuna ve örneklem büyüklüğüne göre gerekli düzeltmeler yapılarak kullanılır.

lojistik stokastik model (logistic stochastic model). Ölçümde yapılan derecelendirmelerin değişik yönlerden (güçlük, ayırt edicilik, şans gibi) etkilendiğini öngören yaklaşım. Bu modelde, ham puanlar belirli formüllerle metrik puanlara çevrilerek eşit aralıklı ölçek boyutu üzerindeki her bir derecenin işaretleme olasılığı hesaplanır.

MA-aralık (H-spread). Menteşeler arası aralık. Birinci çeyrek dilimle üçüncü çeyrek dilim arasındaki mesafe. Üçüncü çeyrek dillimden birinci çeyrek dilim çıkarılarak bulunur.

madde ağırlıklandırması (differential weighting of items). Özellikle başarı testlerinde araştırmacının ölçüm sonuçlarının ayırt edicilik özelliğini arttırmak için bazı maddelerin puan baremini arttırmasıdır. Örneğin, 10 soruluk bir testte maddelerin puan baremleri eşit olarak 10 puan yerine bazıları 5, bazıları 10 ve diğerleri 20 olacak şekilde düzenlenir.

madde ayırt edicilik parametresi *a* (item discrimination parameter *a*). Madde-yanıt kuramında kullanılan bir terimdir. Maddenin belirli bir yetenek düzeyinin altında ve üstünde kalan kişileri ayırt etme özelliği. Madde özellikleri eğrisinin "eğimi" ile incelenir. Dik eğimli olan maddelerde ayırt etme özelliği yüksek, yatık olanlarda ise ayırt etme özelliği düşüktür.

madde bankası (item bank). Madde bankası genel bir terimdir. Kalibre edilmiş ve kalibre edilmemiş maddelerin her ikisi için de kullanılabilir. Yanlış anlamaya meydan vermemek için kalibre edilmemiş maddelerden oluşan soru/madde havuzuna "ham madde bankası" kalibre edilmiş sorulardan/maddelerden oluşanına ise "kalibre edilmiş madde bankası" adı verilir. Kalibre edilmiş madde bankasının oluşturulması test kuramlarına göre farklılık gösterir. Klasik test kuramında güçlük derecesi, çeldiricilik ve ayırt etme özelliği belirlenmiş, ayrıca güvenilirlik ve geçerlilik analizleri yapılarak kalitesi doğrulanmış maddelerin bir araya getirilmesinden oluşmuş madde havuzudur. Madde-yanıt kuramında ise, güçlük, ayırt etme ve şans parametrelerinden en az biri, duruma göre ikisi veya üçü de hesaplanarak belirli bir kavramsal boyutu değişik yetenek/özellik düzeylerinde ölçtüğü saptanmış olan maddelerdir.

madde bilgi fonksiyonu (information function). Bir maddenin bilgi fonksiyonu maddenin varyansına benzer. Maddenin belirli bir yetenek düzeyinde başarılı ve başarısız olanları ayırt etme gücü (Ayrıca *bk.*, "bilgi fonksiyonu" ve "test bilgi fonksiyonu").

madde etkisi (item impact). Farklı gruplara ait kişilerin bir maddeye doğru yanıt verme / bir maddeyi onaylama olasılıkları farklı ise böyle bir durumda madde etkisinden söz edilir. Madde etkisinde, gizli özelliğin / yeteneğin ölçümü açısından gruplar arasında gerçek bir farklılık vardır.

madde güçlüğü (item difficulty). Madde güçlüğü, maddelerin niteliğine göre farklı şekillerde ele alınır. Çoktan seçmeli bilgi testlerinde cevaplayıcıların bir maddeye doğru yanıt verme oranıdır. Psikoteknik testlerde kişilerin bir soruyu veya bir test ögesini doğru yanıtlama oranıdır. Likert tipi ölçeklerde yanıtların tüm derecelere yaklaşık olarak eşit oranda dağılmasıdır. Thurstone ve Guttman tipi ölçeklerde ise, bir maddenin ne ölçüde kolay onaylandığıdır. Guttman ölçeklerinde bir numaralı madde en kolay ve beş numaralı madde ise en zor olanı gösterir (Ayrıca *bk.*, "güçlük indeksi"). Madde güçlük indeks değerinin maksimum ayırt edicilik değerine sahip olması için ,50 civarında olması arzu edilir.

madde güçlük parametresi *b* (item difficulty parameter *b*). Madde-yanıt kuramında gizli özelliği temsil eden teta boyutu üzerindeki herhangi bir nokta. Madde özellikleri eğrisinin başlangıç/çıkış/eşik noktası.

Maddenin güçlük derecesi ile bir kişinin yeri/konumu aynı boyut üzerinde gösterilir.

madde haritası (item map). Rasch ve tek parametreli MYK modellerinde maddelerin ayırt etme güçlerinin eşit olduğu kısıtlaması söz konusudur. Madde haritası, maddelerin konumlarının bu kısıtlama çerçevesinde teta boyutu üzerinde konuşlandığı grafiğe verilen addır.

madde özellikleri eğrisi (item characteristic curve). Madde-yanıt kuramında seçilen modele göre sadece maddenin güçlük derecesini, güçlük derecesiyle birlikte ayırt etme özelliğini veya bu özelliklerin yanında şans faktörünü de gösteren monotonik artış eğilimi gösteren eğri. Bir diğer adı, "madde yanıt fonksiyonu".

madde parametreleri tahmin değerleri (item parameter estimates). Seçilen modele göre maddelerin güçlük, ayırma ve şans faktörünü açıklayan özelliklerin tahminî olarak metrik ölçek birimleriyle ifade edilmesi.

madde tutamaçları (item anchors). Maddeye ait yanıt şıklarını gösteren ve cevaplayıcıların tercih yapmalarını kolaylaştıran dereceleme noktaları. İki dereceli ölçeklerde iki tutamaç ve beş dereceli ölçeklerde ise, beş tutamaç vardır. Ölçeğin çipa noktaları.

madde yanıt fonksiyonu (item response function). *Bk.*, "madde özellikleri eğrisi".

metrik yetenek ölçeği (ability scale metric). Ölçüm yapılan kişilerin yeteneklerinin eşit aralıklı bir ölçek üzerinde sıralanması ve dolayısıyla kişiler arasındaki farkların doğru olarak belirlenmesi. Madde-yanıt kuramında teta boyutu.

mihenk taşı, altın standardı (gold standard). Altının gerçek olup olmadığını veya kaç ayar bir altın olduğunu belirlemek için belirli kimyasal sıvılarla birlikte kullanılan test taşı. Psikometride ise, bir testin yapısal geçerliliğini / iç tutarlılık güvenilirliğini test ederken kullanılan daha önceden geçerlilik ve güvenilirliği sınanmış bir başka test. Mihenk taşı test de başka bir testin mihenk taşı olarak temel alınmasıyla geliştirilmiş olabilir. Çok değişik nitelikteki yapısal geçerlilik analizi yöntemleri arasında araştırmacıların sık başvurdukları üç yöntemden biri (mihenk taşı yöntemi). Diğer ikisi *faktör analizi* ile *birleşme ve ayrılma geçerliliği* yöntemleridir.

monoton türdeşlik (monotone homogeneity). Matematikte bir fonksiyonun doğrusal olarak artış göstermesi. Madde-yanıt kuramında madde özellikleri eğrisinin eğimi.

moment (moment). Kısa mesafe. İki boyutlu bir düzlemde x ve y eksenlerinin 0 simgesiyle gösterilen birleşim merkezinden uzaklığı.

momentler çarpımı korelasyonu (product moment correlation). Kovaryansın (S_{XY}), momentler çarpımına (S_X S_Y) olan oranı. Kovaryans hesaplamasında X ve Y değişkenlerin ortalamalarından sapmaları gösteren x ve y değerlerinin her biri momenti gösterir.

mutlak sıfır (absolute zero). Fizik bilimlerinde "hiç yok" şeklinde mutlak bir anlamı olan değer. Zihinsel yetenekleri, tutumları, değerleri ve kişiliği ölçen sosyal bilimlerde mutlak sıfırın bulunmaması bazı araştırmacılar ve meslekten olmayan kişiler tarafından eleştirilmiş ve bu bilimlerin bir kusuru olarak görülmüştür.

nesnellik (objectivity). Hakemlerin, gözlemcilerin benzer puanlar vermeleri nedeniyle puanların birbiriyle tutarlı ve güvenilir olmasıdır.

nominal yanıt modeli (nominal response model). Madde-yanıt kuramında iki dereceli maddeler yerine çok dereceli maddelerde parametre tahminini yapmak için geliştirilmiş bir hesaplama yaklaşımı.

normal eğri eşlik değerleri (normal curve equivalents). Normalleştirilmiş standart puanlar. Bu puanların ortalaması 50 ve standart sapması ise, 21,06'dır. Standart sapmanın 21,06 olarak seçilmesinin nedeni normal eğri eşlik değerlerinin (NEED) "yüzdelik" değerlerinde olduğu gibi 1 ilâ 99 arasında değişimini sağlamak içindir. Her bir standart dokuz puanına karşılık gelen yaklaşık 11 NEED vardır.

normlar (norms). Norm grubu olarak adlandırılan belirli bir grupta yapılan ölçümler sonucunda elde edilen aritmetik ortalama, medyan veya yüzdelik sırası puanlarıdır. Normlar başarı standartları değildir, fakat test puanlarının yorumlanmasında referans çatısı olarak kullanılır.

norm grubu (norm group, norming group). Test uygulanması amaçlanan hedef kitleyi temsil etme özelliğine sahip, ölçülen özellikler açısından normal dağılım özelliği gösteren ve test standartlarının belirlendiği büyük hacimli ölçüm grubu veya ölçüm örneklemi. Bir diğer adı, standardizasyon örneklemi. Norm grubu cinsiyet, yaş, kıdem, eği-

tim, deneyim ve meslek gibi faktörlere göre belirlenebilir. Bir ülke içinde ırk temeline bağlı olarak norm grubu oluşturulması hem yasal hem de etik açıdan doğru değildir. İş hayatında personel seçim testleri için norm grubu oluşturulurken iş için başvuran kişilerin cinsiyet, eğitim, yaş dağılımı ve deneyim özellikleri birlikte göz önünde bulundurulur. Diğer bir deyişle başvuran adayların özellikleriyle norm grubunun özellikleri birbirine benzer olmalıdır.

normlar ve standartlar (norms and standards). Normlar tek başına standartlar anlamına gelmez. Standartlar geçici yargılardır. Grup ortalaması veya grup normu standart olarak kabul edilebileceği gibi ortalamanın üzerindeki daha yüksek bir değer de kriter olarak belirlenebilir.

odak grubu (focal group). Ne şekilde etkilendiği merak edilen ve bu nedenle araştırma ilgisinin yöneltildiği grup.

olasılık oranı (odds ratio). Bir modeldeki bağımsız değişkenlerin her birinin yüzde cinsinden vuku bulma oranı. İlişkinin gücü.

olumsuz etki (adverse impact). ABD'de testlerle ilgili olarak kullanılan hukukî bir terim. Test sonuçlarına dayalı olarak kişilerin ırk, cinsiyet veya etnik köken temelinde orantısız bir şekilde işe alınmaları, terfi ettirilmeleri veya sınavı kazanmış sayılmaları. Olumsuz etki, testin yanlı olduğu anlamına gelmez veya yanlı olduğunun kanıtı olarak gösterilmez.

ortak maddeler (common items). İki veya daha fazla paralel formda test eşitliğini sağlamak amacıyla tutulan aynı maddeler. Ortak maddeler çıpa test oluşturmak amacıyla da kullanılabilir.

ortak ölçek (comman scale). Rasch modelinde bütün ajanların, nesnelerin temsil edildiği paylaşılan ölçüm birimi.

kombinasyon ölçüm modeli (conjoint measurement model). Kavram fizik bilimlerinden psikolojiye geçmiştir. Fizik bilimlerinde hacim ve ağırlığın birlikte ölçülmesine benzer şekilde psikolojide bağımlı değişkenin iki veya daha fazla bağımsız değişkenle birlikte ölçülmesi anlamındadır.

ölçek (scale). (1) genel anlamda, belirli bir özelliği veya tutumu ölçmek üzere oluşturulmuş birden fazla madden oluşan bileşik ölçümlerdir. (2)

teknik anlamda, birden fazla maddenin aralarında hiyerarşik bir sıra olmak üzere belirli bir ölçek boyutunda sıralanmasıyla oluşturulmuş ölçüm araçlarıdır. Maddeler arasında büyüklük küçüklük sırası söz konudur. Thurstone, Mokken, Guttman, Bogardus başlıca ölçek türleridir. Likert yönteminin kullandığı ölçüm araçları teknik anlamda ölçek değil, indekstir. Çünkü Likert indeksinde maddeler arasında büyükten küçüğe veya küçükten büyüğe doğru bir sıralanma söz konusu değildir.

ölçeklenme (scaling). Bileşik ölçümlerde maddelerin (derecelerin değil) tek bir boyutu ortaya çıkaracak şekilde düzenlenmesi. Öte yandan maddeler arasında hiyerarşik bir sıra düzeni olmalı ve maddeler ölçüm boyutu üzerinde eşit aralıklı bir dağılıma sahip bulunmalıdır.

ölçeklenmiş puan (scaled score). Ham puanların matematiksel formüller kullanılarak eşit aralıklı ölçek puanlarına dönüştürülmesi. Bu puanlar zaman içinde farklı puanları karşılaştırırken yararlıdır. Standartlaştırılmış başarı testlerinin hepsinde ölçeklenmiş puanlar kullanılır.

ölçüm ajanı (agent of measurement). Rasch modelinde ölçüm birimleri (maddeler, sorular veya ifadeler). Bir değişkeni, bir nesnenin pozisyonunu veya bir değişken üzerinde yer alan bir kişinin konumunu tanımlamak için kullanılır.

ön tanımlı (default). Bir yazılımda önceden tanımlanmış ögeler, değerler.

örneklem çerçevesi (sampling frame). Örneklemin çekildiği ana kütlenin sistematik bir şekilde ve genelleme yapmaya imkan sağlayacak oranda daraltılmış şeklidir. Örneklem çerçevesindeki her bir birey eşit seçilme şansına sahiptir.

özge varyans (unique variance). Bir değişkende ortak faktörlerle açıklanamayan varyans ($1 - h^2$).

özgün nesnellik (spesific objectivity). İki maddenin güçlük parametresinin her hangi bir örneklemde birbirinden bağımsız olmasıdır. Bu nedenle iki kişinin yetenek düzeyinin karşılaştırılması, kendilerine verilen test maddeleriyle ilgili değildir. Wilson (1991) özgün nesnelliği şu şekilde açıklamıştır: İlgili sınıfa ait olması koşuluyla, siz bir kere ölçüm yaptıktan sonra başka kimin ölçtüğünün veya hangi ölçüm aracının kullanıldığının önemli olmaması.

panel çalışması (panel study). Aynı kişilerin farklı zaman dilimlerinde birkaç kez gözlenmesi, kendileriyle bir konunun tartışılması veya kendilerine birkaç kez anket uygulanması.

parametre (parameter). (1) Bir değişkenin ana kütle için ölçüm değeri, (2) madde-yanıt kuramında maddenin güçlük, ayırt etme ve şans eseri veya tesadüfen işaretlenmiş olma özellikleri.

parametre cimriliği (parameter parsimony). Bir ölçüm çalışmasında maddenin özelliklerini tanımlayan çok sayıda parametre yerine daha az sayıda parametre ile çalışma.

paydaşlık oranı (communality). Orijinal değişkende faktörlerin neden olduğu toplam değişkenlik.

portföy değerlendirmesi (portfolio assment). Bireylerin standart testler yerine daha uzun bir dönem içinde gerçekleştirdiği değişik etkinliklerin göz önünde bulundurulmasıyla yapılan değerlendirme. Etkinlikler bireysel veya grup halinde yapılmış olabilir. Sözel, yazılı, sunum şeklinde veya gösteri şeklinde yapılan ödevler kişinin bir bütün olarak değerlendirilmesi için kullanılır.

profil grafiği (profile). Bireye veya gruba uygulanan başarı / yetenek testlerine ait puanların nispî büyüklüğünü görmeye imkan sağlayan grafik. Test maddelerine değil bir bataryadaki test puanlarının ortalamalarına veya alt testlerin/ölçeklerin ortalama/toplam puanlarına uygulanır.

ranj kısıtlaması (restricted range). Ölçüm yapılan alanda, sahada, seride, dizi verilerde, tür veya sınıflarda belirli bir kısıtlamaya gidilerek verilerin bir bölümünün araştırmaya alınmasıdır. Kısıtlama olmasından dolayı veriler normal dağılım özelliği göstermez. Örneklem varyansı ana kütle varyansından daha düşüktür.

Rash modeli (Rasch Model). Rash modeli, tek parametreli gizli değişkeni ölçme yaklaşımıdır. Bu modelde ölçek maddelerinin zorluk derecesi ve bireylerin yetenekleri birlikte ele alınır. Maddelerin kalibrasyonunu yapmak için olasılık tahminlerinden yararlanılarak kişi-madde etkileşiminin belirli modele ne ölçüde uygun düştüğüne bakılır.

referans grubu (reference group). Üzerinde araştırma/ölçüm yapılan gruptaki değişikliklerin anlamlı olup olmadığını belirlemek için karşılaştırma kriteri olarak kullanılan grup.

regresyon etkisi (regression effect). Çekilme etkisi. Yeniden test veya son test puanlarının ilk test puanlarına göre ortalamaya daha yakın bir değere gelmesi. İlk test uygulamasında yüksek veya düşük puan alan kişiler son test uygulamasında çekilme etkisi nedeniyle ortalamaya yakın değerler elde ederler.

rota analizi (path analysis). Rota analizi, değişkenler arasında nedensellik ilişkisi kuran modellerin açıklama gücünü ortaya koyan çoklu regresyon analizi prosedürünü tanımlar. Cimri modellerin açıklama gücünün üstün olduğuna inanıldığından değişkenler arasındaki ilişkileri tahmin etmek için daha az açıklayıcı rota çizgilerinden yararlanılır.

sabit etki modeli (fixed effects model). Sabit etki yaklaşımında araştırmacının kontrolü altında tutulan müdahale türü, araştırma çevresi, araştırma grubu veya bireysel özellikler gibi faktörlerin bağımlı değişkende bir etki büyüklüğü yaratmayacağı varsayımından hareket edilir. Farklı çalışmalardaki etki büyüklüğünün sadece örnekleme hatasından kaynaklandığı düşünülür. Sabit etki modelinde analizi yapılmak istenen birim veya kişilerin tamamı araştırmacının belirlemiş olduğu koşullara tabi tutulur. Örneğin, bir araştırmada katılımcıların önce soğuk ve üşüdükleri bir salonda sınava alınması ve daha sonra ısıtılmış bir salonda sınava alınmaları sabit etki modelini tanımlar

sapma (deviation). Gözlem değerinden aritmetik ortalamanın çıkarılmasıyla elde edilen değer ($X - \overline{X}$).

sapmaların çarpımı (cross-product deviations). Kovaryans hesaplamasında X ve Y değişkenlerin ortalamalarından olan sapmaların x ve y olarak birbiriyle çarpılması.

semptomsuz kovaryans matrisi (asymptotic covariance matrix). Örneklem verilerine dayalı korelasyon ve koveryans matrislerinden hareket ederek semptomsuz (asymptotic) ana kütle korelasyon ve kovaryans matrisi tablolarının hesaplanması. Terimdeki *semptomsuz* sözcüğü "yansız" veya "hastalık belirtisi olmayan" anlamlarında kullanılmıştır. Semptomsuz kovaryans matrisi verileri ana kütledeki ilişkilerin resmini daha gerçekçi bir biçimde yansıtır. Semptomsuz kovaryans

matrisinin hesaplanabilmesi için örneklem büyüklüğünün $k(k + 1)/2$ formülüyle hesaplanan sayı kadar olması gerekir. Formüldeki k değişken sayısıdır.

senaryo (script). Bir testte daha çok açık uçlu sorulara verilen el yazısıyla yazılmış yanıtlar. Senaryo cevaplar.

sınıf eşlik değerleri (grade-equivalent scores). Norm referanslı puanlardır. Öğrenci başarılarına ait puanlar sınıf düzeylerine göre yeniden belirlenir. Yıl içinde alınan öğrenci notları, bir önceki yıl sonu sınıf ortalaması veya medyan değeri temel alınarak yeniden hesaplanır. Örneğin, üçüncü sınıftaki bir öğrencinin matematik dersinden birinci yazılı sınav notu 4 ise bu not sınıf eşlik değeri formülüyle yeniden belirlenerek 3,6 şeklinde başka bir değere dönüştürülür. Bu notun anlamı, öğrencinin norm gruba göre sınıf içindeki düzeyinin üçüncü yılın altıncı ayında olduğudur. Angoff, 1972 yılında sınıf eşlik değeri puanlarının entelektüel gelişmeyi ölçmede önemli ölçüde yetersiz kaldığını söyleyerek eleştirmiştir.

sınır çizgileri (boundary curves). Madde-yanıt kuramında "madde özellikleri eğrisi"nin bir diğer adı.

sınır etkisi (boundary effect). Bir testin çok kolay veya çok zor olması nedeniyle elde edilen puanların çok yüksek veya çok düşük çıkmasıdır. Tavan-taban etkisinin görüldüğü bu tür test ve ölçümler güvenilir değildir.

standart (standard). (1) Kesim/eşik puanı. Belirli bir testten başarılı sayılabilmek için alınması gereken minimum puan. (2) yönerge uygulama veya kuralları.

standart testler (standardized tests). Belirli bir akademik veya meslekî alanda akademik/meslekî başarıyı veya bilgiyi ölçmek amacıyla geliştirilmiş, test uygulanacak kitleyi temsil eden norm/referans grubu temel alınarak ham puanların dönüştürme formülleri aracılığıyla yüzdelik veya diğer standart puanlara dönüştürüldüğü ve başarı için standart değerlerin elde edildiği ölçüm araçları. Bazı standart testler günümüzde artık madde-yanıt kuramı bulguları çerçevesinde analiz edilmektedir. Standart testlerin karşıtı, kişilerin günlük davranışlarını ve iş yapış biçimlerini temel alan ve portföy değerlendirmelerine dayanan "performans testleri"dir. Fakat performans testlerinin güvenilirliği düşüktür. Standart testler sürekli olarak aynı şekilde uygulanır,

aynı şekilde puanlanır ve aynı şekilde yorumlanır. Standart testler norm referanslı ve kriter referanslı olmak üzere iki grupta değerlendirilir. Uygulamada standart testleri geliştirmek oldukça güç bir iştir. Çünkü bu testlerin evrensel olarak kabul edilmesinde değişik sorunlarla karşılaşılır. Ölçümlerin sınırlı sayıda kişide ve belirli örneklem gruplarında yapılmış olması genel kamuoyunu tam olarak tatmin etmez. Test sonuçlarına ilişkin güvenilirlik ve geçerlilik analizlerinin tam olmaması, kriter geçerliliğinin nesnel bir dayanağının olmaması, ölçümün standart hatasının değişik örneklemlerde farklı çıkması nedeniyle gerçek anlamda standart bir test elde etmek oldukça zordur. Öte yandan standart testler, deney yapma, gerçek dünyanın matematik problemlerini çözme, araştırma raporu yazma, roman ve hikaye okuyup bunları analiz etme ve sunum yapma gibi değişik alanlardaki bilgi ve yetenekleri ölçmeye uygun olmaması nedeniyle eleştirilmiştir. Standart testler bu tür öğrenme biçimlerini ölçme konusunda yetersizdir. Standart testler daha üst düzeydeki düşünme, analiz etme, sentez etme, değerlendirme, yaratıcılık gibi uzun süre içinde sergilenebilecek başarıyı belirleme açısından sınava ayrılan sürenin kısıtlı olması nedeniyle kişiyi bütün olarak değerlendiremez.

standart dokuz sistemi (stanine system). Test puanlarına 1 ilâ 9 arasında yeni değerler atamak için kullanılan standart puanlama sistemi. Sistem ABD Hava Kuvvetleri tarafından geliştirilmiştir. Standart dokuz puanlarının aritmetik ortalaması 5 ve standart sapması 1,96'dır. Standart dokuz, sisteminde ilk üç puan "ortalamanın altında" ikinci üç puan "orta" ve üçüncü üç puan ise "ortalamanın üstünde" ifadeleriyle tanımlanır. Ancak bu puanlama sistemi dikkatli bir şekilde kullanılmadığı zaman bazı yanlış değerlendirmelere neden olabilir. Örneğin Mehmet ve Selim'in yüzdelik sırası 22 ve 24 olsun. Bu puanlama sisteminde Mehmet 3 standart dokuz puanı ile "ortalamanın altında" ve Selim ise 4 standart dokuz puanı ile "orta düzeyde" bir öğrenci sayılacaktır. Standart dokuz sistemi, küçük örneklemlerde istenen sonucu vermez. Ayrıca hesaplamalarda dokuzun üstündeki ve birin altındaki değerler dokuz ve bire eşitlenir.

standart hata (standard error). Örneklem hatası ölçüsü. Farklı örneklem ortalamalarının standart sapmasıdır. Değişik örneklem ortalamaları arasında ne kadar bir hata olabileceği hakkında fikir verir. Aynı zamanda örneklem verilerinden hareket edilerek ana kütle varyansı hakkında bilgi veren bir değerdir.

standart puanlar (standard scores). Yorumlama ve karşılaştırma kolaylığı sağlaması nedeniyle ham puanların *z* veya *T* puanlarına dönüştürülmesidir. Literatürde bazı yazarlar sadece *z*, *T* ve WISC III gibi ortalama ve standart sapma değerlerinin temel alındığı puanları standart puan kavramı kapsamında kullanırlarken diğer yazarlar belirli bir matematiksel işlemden geçirilerek dönüştürülen tüm değerleri standart puan kavramıyla ifade ederler. Buna göre yüzdelik değerleri, *z* puanları, *T* puanları, normal eğri eşlik değerleri (normal curve equivalents), standart dokuz ve standart on puanları, sınıf eşlik değerleri (grade equivalents), yaş eşlik değerleri standart puanlardır. Standart puanları da kapsayacak şekilde dönüştürülmüş puanların tümünü tanımlamak için "türetilmiş puanlar" terimini kullanmak daha doğru olur. Dönüştürmedeki amaç, puanları "normalize etmek" değil, yorumlanabilir hale getirmektir.

standart yaş puanları (standard age scores). Bir test bataryasındaki her bir testin belirli yaş grupları için standardize edilmiş puanları. Puanların aritmetik ortalaması 100 ve standart sapması ise 15'tir. Standart yaş puanları şu formülle elde edilir: SYP = 15 *z* + 100. Yaş gruplarının belirlenmesinde kesin bir kural yoktur, ancak literatürde daha çok –19, 20/24, 25/34, 35/44, 45/54, 55/64, 65+ sınıflaması kullanılır.

tahmin düzeltmesi (correction for guessing). Çoktan seçmeli sorularda cevaplayıcının bilerek değil, tahminde bulunarak cevapladığı soruların gerçek değerinin yeniden hesaplanması.

taşıma etkisi (carry over effects). Önceki test uygulamasından kişilerin bazı soruların yanıtlarını hatırlamaları.

tau eşitliği (tau equivalent). İki ölçümde gerçek puanların eşit olması, fakat hata varyanslarının eşit olmamasıdır. Ölçümlerin her birinin farklı oranda tesadüfî hata içermesi nedeniyle, gözlem puanlarının varyans değerleri aynı değildir.

tekillik (singularity). Değişken çiftleri arasındaki korelasyon katsayısının 1,0 olması.

test bilgi fonksiyonu (test information function). Madde bilgi fonksiyonlarının birleşiminden/toplamından meydana gelir ve bir testin bütün olarak performansını gösterir.

test bataryası (test battery). *Bk.*, "Batarya".

test eşitleme (test equating). (a) Aynı yeteneği/özelliği ölçen birden fazla testin birbirine eşit olduğunu belirleme. Süreç sonunda ortak yetenek/özelliği ölçen iki test oluşturma. (b) madde özellikleri eğrileri benzer olan maddeleri bir araya getirip (toplamda test bilgi fonksiyonu benzer çıkacaktır) iki paralel form oluşturma.

test gölgelenmesi (test compromise). Testin ününe gölge düşmesi. Teste ait maddelerin güvenliğinin sağlanamaması nedeniyle maddelerin bilinir olması ve bu nedenle sonuçların güvenilirliğinin kuşkulu hale gelmesi.

testçik (testlet). İlgili maddelerin bir araya getirilmesinden oluşmuş küçük bir ölçek veya test. Test maddeleri destesi. (1) Matematik, fen bilgisi gibi belirli disiplinlerde ünite bazında hazırlanan ve 3-15 sorudan oluşan test maddeleri. (2) Belirli bir paragrafa veya ortak soru köküne bağlanmış 3-5 sorudan oluşan soru grubu. (3) Belirli bir vak'aya bağlanmış ve vak'ayı analiz etmeye yönelik olarak hazırlanmış belirli sayıda soru grubu, (4) Test broşürü veya küçük test kitapçığı.

test yansızlığı (test fairness). Testin yansız ve uygulama prosedürlerinin kurallara uygun olması, sonuçlarının değişik gruplar arasında ayırımcılığa yol açmaması, güçlü ve zayıf olduğu yönler hakkında testin teknik el kitabında kullanıcılara bilgi verilmiş olması.

test sonuçlarını kendi içinde yorumlama (ipsative test interpretation). Bir bataryadaki test sonuçlarını grup puanlarıyla karşılaştırmaksızın kendi içinde yorumlama.

teta (theta). Madde-yanıt kuramında bir test veya ölçekle ölçülmeye çalışılan gizli özellik. Yapı ve özellikler eşit aralıklı, sürekli veri niteliğindeki bir ölçek/boyut üzerinde ölçülür. Madde-yanıt kuramında kişilerin yetenekleri/özellikleri ile maddenin güçlüğü aynı boyut üzerinde gösterilir.

tetrakorik korelasyon analizi (tetrachoric correlation). İki sürekli verinin kukla değişken olarak ikili veri şeklinde kodlanması sonucunda bu değişkenler arasında yapılan korelasyon analizidir. İkili verilerden hareket edilerek sürekli veriler arasındaki korelasyon tahmin edilmeye çalışılır.

topuk yükseltme, çoğaltma (bootstrapping). Güven aralığını daraltmak ve daha kesin değerler elde etmek için sunî bir şekilde örneklem büyüklüğünün artırılması. Örneğin, 30 gibi bir örneklem büyüklüğü ile çalışılıyorsa daha sağlıklı bir güven aralığı elde etmek için bu örneklem büyüklüğünün değişik sayıda katları alınarak veya kullanılan yazılımın "tesadüfî rakam üretme" modülü kullanılarak örneklem hacminin büyütülmesi ve güven aralığının yeniden hesaplatılması. Ancak bu şekilde üretilen rakamlar ilk 30 birimlik örnekleme dayalı olduğundan ana kütleyi tam olarak temsil etme özelliğine sahip olmaz.

***T*-Puanı (*T*-Score).** Standart puanı hesaplama yöntemlerinden biridir. *t*-Testi'ndeki *t* harfi küçük yazılırken *T*-puanındaki T harfi büyük yazılır. T- puanının *t* değeri ile karıştırılmaması gerekir. Ortalaması 50 ve standart sapması 10 olan puandır. *T*-puanı standart z puanlarından hareket edilerek de hesaplanabilir [*T*-puanı = (*z*·10) + 50].

türdeşsellik, sabit varyans (homoscedasticity). İkili serilerde (*X* serisi ve *Y* serisi) verilere ait varyansların aynı veya türdeş olmasıdır. Verilerin varyansları aynı olunca doğrusal ilişkinin her hangi bir noktasında hata varyansının da aynı olduğu varsayılır. Verilerin tamamının regresyon doğrusu üzerine düşmesidir. Veriler, regresyon doğrusu boyunca paralel kümelenmiş bir bulut oluşturmuşsa türdeşsellik sağlanmıştır denilir. Hataların tüm ölçümlerde yeknesak olması ölçüm verilerinin güvenilir olduğunu gösterir.

uyarlı test (adaptive testing). Bu uygulamada kişiye verilen test maddeleri onun test sorularını çözme kabiliyetine göre değişkenlik gösterir. Kişi aldığı ilk birkaç soruyu başarıyla çözmüşse kendisine daha zor sorular sorulur. Çözememişse bu kez daha kolay sorulara geçilir. Kişiye verilen sorular onun yetenek düzeyine ve başarı seviyesine uygundur. Kişiler farklı düzeylerden başlayarak soruları farklı düzeylerde bitirirler.

uygulanabilirlik (applicability). Gözlem / ölçüm sonuçlarının diğer düzlemlerde de doğru çıkması. Kavram aynı zamanda şu terimlerin eş anlamlısıdır: dış geçerlilik, genellenebilirlik, ilgililik, aktarılabilirlik.

uyum testi (adjustment test). Bireyin içinde yaşadığı topluma uyum gösterme ve kişisel gereksinimlerini karşılama kabiliyetini ölçen kişilik testi.

uyuşma indeksi (index of concordance). Gözlemcilerin veya değerlendiricilerin uyuştukları madde sayısının toplam değerlendirme sayısına olan oranı.

uyuşma istatistiği (fit statistic). Beklenen ile gerçekleşen gözlem verileri arasındaki uyuşmazlığın/uyuşma derecesinin özeti.

üçleme (triangulation). Belirli bir yapı veya olguyu incelerken resmi daha doğru ve gerçekçi bir şekilde ortaya koymak için birden fazla yöntemden veya kaynaktan birlikte yararlanmak. (Örneğin, alan araştırmasının yanında, gözlem ve mülâkat yöntemlerini de kullanarak bilgi toplamak ve bu bilgileri önceki literatür bulgularıyla karşılaştırmak.) Burada üç rakamı genel bir terim olarak kullanılmıştır. Bilim adamı gerek duyması halinde üçten fazla kaynaktan da yararlanabilir.

varsayım ihlalleri (assumption violations). Bir istatistiksel teste ait ön kabullerin veya ön koşulların sağlanamaması.

varyans (variance). Standart sapmanın karesi.

varyansların homojenliği (homogeneity of variance). İki değişkene ait değerlerle ilgili varyansların eşit olması.

veri taraması (data screening). Yapılması düşünülen temel istatistikî analizlerden önce verilerin kalitesinin değerlendirilmesi. Bilim adamı bu amaçla ön inceleme niteliğinde bir dizi istatistiksel analiz tekniğini uygular. Örneğin, verilerin dağılım özelliği, eksik veri analizi, çarpıklık ve basıklık katsayılarının saptanması, ayrık değerlerin tespiti, türdeşsellik ve doğrusallık özelliği, koşutluk (çoklu doğrusallık) veya tekillik özelliği, verilerin dönüştürülmesi, aritmetik ortalama ve standart sapma ve varyasyon katsayısı gibi teknikleri kullanır. Bu tür hesaplamalardan aritmetik ortalama, standart sapma, medyan minimum ve maksimum gibi tekniklere aynı zamanda *özet istatistik analizleri* adı verilir.

veri uyarlaması (data augmentation). Gizli değişkenlere dayalı olarak tekrarlamalı algoritmalar oluşturmaktır.

yanıt eğilimi (response set). Test alan bir adayın sorunun cevabını bilmediği durumda hep aynı şıkları işaretlemesi. Örneğin, hep *c* şıkkını veya hep uzun ifadeler içeren şıkları işaretlemesi.

puan verme şablonu (rubric). Test alan adaylara yanıtların nasıl işaretleneceğine ilişkin olarak yapılan açıklamalar veya verilen talimatlar. İşaretleme kılavuzu.

yanıtlama oranı (response rate). Fiili olarak tamamlanan ve geri gönderilen anketlerin toplam anket sayısına olan nispeti.

yanıtsızlık yanlılığı (nonresponse bias). Ana kütledeki veya örneklem planındaki kişilerin bir bölümünün mevcut olmaması veya yanıt vermemesi nedeniyle ortaya çıkan yanlılık.

yanlı madde (biased item). Bir maddenin örneklem olarak seçilen bir gruba uygulandığında elde edilen değerin ana kütleye uygulandığında elde edilebilecek değerden daha yüksek veya daha düşük çıkması.

yanlılık (bias). (a) Testlerde ölçüm hatasının tesadüfî olarak değil, sistemli bir biçimde dağılmasıdır. (b) Bir grubun testlerde diğer gruplara göre daha fazla avantaj elde etmesi. Bir maddenin aynı yetenek düzeyindeki kültürel, etnik, cinsiyet, dinî, sosyoekonomik farklılığa sahip ve kırsal kesimden gelen gruplarda içerdiği bazı özellikler nedeniyle farklı sonuçlar vermesi. (c) Parametre tahmininde eğer beklenen değer gerçek değere eşitse parametre yansızdır denir ($E = \hat{\theta} - \theta$). Yanlılık örneklem hacmi n, büyüdükçe azalma eğilimi gösteriyorsa buna kesişimsiz yansızlık (asymptotic unbiasedness) adı verilir.

yanlış pozitif hatası (false positive error). Yanlış kabul veya hatalı kabul. Yanlış "sorun yok" teşhisi. Tip II hatasının yapılması.

yanlış negatif hatası (false negative error). Yanlış ret veya hatalı ret. Yanlış "sorun var" teşhisi. Tip I hatasının yapılması.

yansız etki büyüklüğü (unbiased effect size). Fisher z' ağırlıklı ortalama puanlarının tekrar r değerine dönüştürülmesi.

yapay ikili (artificial dichotomoy). Bilim adamının kendi tasarımına göre ikili değerler oluşturması.

yapay sonuçlar (statistical artifact). Aldatıcı yapay istatistiksel sonuçlar. Bilim adamının yetersiz örneklemle çalışarak anlamlı bir sonuç elde etmesi veya güvenilir ve geçerli olmayan bir ölçekle anlamlı bir sonuç elde etmesi.

yapı (construct). İnsan davranışlarını açıklamak için kullanılan, verileri düzenli veya sistemli bir şekilde açıklayan hipotetik bir teori veya kavram. Davranışsal anlamda kişilik özelliği, vasıf, kuram veya etken.

yeniden üretilebilirlik katsayısı (coefficient of reproducibility). Guttman ölçeklerinde tutarlı işaretleme sayısının toplam işaretleme sayısına bölünmesiyle bulunur.

yeniden üretilen korelasyon matrisi (reproduced correlation matrix). Faktörlerden üretilmiş korelasyon matrisi.

yerel bağımsızlık (local independence). Bir ölçekte gizli özelliği (yapıyı) ölçme koşulu sağlandıktan sonra bir maddeye verilen yanıtların diğer maddelere verilen yanıtlardan bağımsız olması.

yetenek tahmin değerleri (ability estimation values). Madde-yanıt kuramında test alan kişilerin ölçekteki maddelerin bütününde gösterdikleri performansa bağlı olarak belirlenen ve teta yetenek boyutu üzerinde konuşlandırılan metrik puan değerleri.

yetenek testleri (ability tests). Yetenek testleri, belirli zihinsel ve psikomotor alanlarda kişilerin mevcut başarı düzeylerini ölçerek gelecekteki başarıları hakkında tahmin yapmaya imkan sağlayan araçlardır.

yeterli, yetkin ve usta (proficient, competent, master). Kişinin bilgi, yetenek ve beceri açısından belirlenmiş bir standarda göre değerlendirilmesi ve bu değerlendirme sonucunda söz konusu standardı aşması. Kriter referanslı testlerde kesim puanını aşan kişiler. Yetkinlik üç düzeyde değerlendirilebilir: yeterli, yetkin, usta. Kesim puanı ,70'i aşanlar *yeterli*; ,84'ü aşanlar *yetkin* ve ,94'ü aşanlar ise *usta* olarak değerlendirilir. Kriter olarak kullanılan kesim puanını belirlemek için belirli bir gruptan elde edilen ve normal dağılım eğrisine dayanan grup norm puanları temel alınır. Günlük yaşamda "yetkin" sözcüğü genelde derece ayrımı gözetilmeksizin yeterli, yetkin ve usta kavramlarının her üçünü de içerecek şekilde kullanılmaktadır.

Yığılımsallık, değişken varyans (heteroscedasticity). İkili serilerde, *X* serisi ve *Y* serisine ait verilerin varyanslarının farklı olmasıdır. Değişkenlerden biri veya her ikisi de sağa veya sola çarpık ise bu özellik görülür. Noktalar kümesi regresyon doğrusunun üst bölümünde veya alt bölümünde bir bulut halinde yığılmış, kümelenmiştir.

Yığılımsallıkta korelasyon katsayıları ya çok yüksek veya çok düşük çıkar. *Yeknesak olmayan hata* olarak da isimlendirilir. Ölçüm verilerinin bazıları diğerlerinden daha fazla güvenilirdir. Yığılımsallığı test etmek için tek bir örneklemde yapılan ölçümler yeterli olmayabilir, bu nedenle araştırmacı konuyu birden fazla örneklemde test etmelidir. Ana kütledeki farklı grupların farklı varyanslara sahip olmasıdır.

Y-kesim noktası (Y-intercept). Regresyon çizgisinin *Y* eksenine veya ordinatına değdiği / denk geldiği nokta. Madde-yanıt kuramında *c* parametresi. Başarı testlerinde tahmin/şans faktörü. Tutum ölçekleri ve kişilik testlerinde ayırt edici olmayan yanıt olasılığı veya sosyal beğeni faktörüne göre verilen yanıtlar.

yönlendirme (coaching). Test uygulamasına geçmeden önce test alan kişilerin; testin süresi, yanıtların işaretlenme biçimi, yanlış cevapların doğru cevabı götürüp götürmediği, zaman kullanma, test gerilimi gibi konularda aydınlatılması ve kendilerine en yüksek puanı elde etmelerine imkan sağlayacak diğer gerekli bilgilerin verilmesi.

yüksek ödüllü ölçekler (high-stakes scales). Çok sayıda kişiye uygulanan ve sonuçta kişilerin daha sonraki yaşamlarını veya meslekî konumlarını etkileyen, onları ödüllendiren ölçüm araçları, sınavlar, testler.

yüzdelik dilimi (percentile). Belirli bir puanın altında kalan diğer puanların yüzdesidir.

zayıflığı düzeltme formülü (correction for attenuation formule). Düşük korelasyon katsayılarının bir formül aracığıyla düzeltilerek yükseltilmesidir. Değişik korelasyon katsayıları arasında karşılaştırmalar yapılırken güvenilirliği sabit tutmak için kullanılır. Hesaplama yapmak için, korelasyon katsayısı güvenilirlik katsayısının karekökünе bölünür.

zayıflık (attenuation). Korelasyon ve regresyon analizlerinde ortaya çıkan bir olgudur. İlişki aranılan her iki değişkenin mükemmel bir güvenilirliğe sahip olmaması nedeniyle ilişki katsayısı belirli oranda düşük çıkar. Zayıflık bu düşüklüğe verilen addır.

zayıflık yanıltmacı (attenuation paradox). Olgu ilk kez Gulliksen tarafından açıklanmış ve "test varyansının maksimize edilmesi (güvenilirlik) kriterinin uç noktalara kadar götürülemeyeceği" anlamında kul-

lanılmıştır. Testin varyansı, katılımcıların yarısı sıfır ve diğer yarısı da 1 puan verirse maksimum seviyeye çıkar. Kuşkusuz bu tür bir dağılım bilinen nedenlerle kabul edilmez, ancak KTK'de bu tür bir dağılımı ret etmek için herhangi bir neden bulunmadığı ileri sürülmüştür.

SEÇİLMİŞ KAYNAKLAR

Bu bölüme, genel olarak yararlanılan kitaplar ile güvenilirlik-geçerlilik konusunda önemli bilgiler içeren İnternet Ağ kümelerinin adresleri alınmıştır. Doğrudan yararlanılan kaynaklar hakkında bilgi edinmek isteyen okurlar, bölüm sonlarındaki "Alıntı Yapılan Kaynaklar" listesine başvurmalıdırlar.

A. BASILI KAYNAKLAR

American Educational Research Association, American Psychological Association ve National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, 1999.

Arık, İ. Alev. *Psikolojide Bilimsel Yöntem.* Genişletilmiş 2'nci b. İstanbul: Çantay, 1998.

Bracht, G.H.; K.D. Hopkins ve J.C. Stanley. *Perspectives in Educational and Psychological Measurement* [Eğitimsel ve Psikolojik Ölçümlerde Perspektifler]. Englewood Cliffs, N.J.: Prentice-Hall, inc, 1972.

Cronbach, L.J. *Essentials of Psychological Testing* [Psikolojik Testlerin Temelleri]. New York: Harper and Row, 1970.

Erdoğan, İlhan. *İşletmelerde Kişi Değerlendirmede Psikoteknik.* İstanbul: İşletme Fakültesi Yayını, 1990.

Ghiselli, E.E. *Theory of Psychological Measurement* [Psikolojik Ölçüm Kuramı]. New York: McGraw-Hill Book, 1964.

Hambleton, R.K.; H. Swaminathan ve H.J. Rogers, *Fundementals of Item Response Theory* [Madde Yanıt Kuramının Temelleri], London: Sage, 1991.

Kerlinger, F.N. *Foundations of Behavioral Research* [Davranışsal Araştırmaların Temelleri]. 2'nci b. London: William Clowes and Sons, 1973.

Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems* [Madde-Yanıt Kuramının Pratik Test Sorunlarına Uygulanması]. Hillsdale, NJ: Erlbaum, 1980.

Runyon, R.P. ve A. Haber. *Fundementals of Behavioral Statistics*. 7'nci b. New York: McGraw-Hill, Inc., 1991.

Thorndike, R.L. ve E.P. Hagen, *Measurement and Evaluation in Psychology and Education* [Psikoloji ve Eğitimde Ölçme ve Değerlendirme]. New York: John Wiley and Sons, 1969.

Viteles, M.S. *Industrial Psychology* [Endüstriyel Psikoloji]. New York: W.W. Norton and Company, 1932.

B. ELEKTRONİK KAYNAKLAR

Becker, LA. "Reliability and Validity [Güvenilirlik ve Geçerlilik]," <http://web.uccs.edu/lbecker/Psy590/relval_II.htm> (10.08.2004).

Educational Testing Service, "ETS Standards For Quality and Fairness [Kalite ve Adalet için ETS Standartları]," <ftp://ftp.ets.org/pub/corp/standards.pdf> (19.08.2003).

Garson, D. "Reliability [Güvenilirlik]," <http://www2.chass.ncsu.edu/garson/pa765/reliab.htm> (10.08.2004).

School of Library, Archival and Information Studies. "Measurement, Validity and Reliability [Ölçme Geçerlilik ve Güvenilirlik]," <http://www.slais.ubc.ca/resources/research_methods/measurem.htm> (10.08.2004).

StatSoft, Inc. "Reliability and Item Analysis [Güvenilirlik ve Madde Analizi]," 2003, <http://www.statsoft.com/textbook/streliab.html> (10.08.2004).

Trochim, William M.K. "Measurement [Ölçüm]," <http://www.socialresearchmethods.net/kb/measure.htm> (10.08.2004).

Trochim, William M.K. "The Web Center for Social Research Methods [Sosyal Araştırma Yöntemleri İçin Ağ Merkezi]," <http://omni.cornell.edu/index.html> (10.08.2004).

Vehkalahti, K. Reliability of Measurement Scales [Ölçeklerin Güvenilirliği]. 2000, <http://www.google.com.tr/search?hl=tr&ie=UTF-8&q=reliability+Measurement+Scales&btnG=Arama> (10.08.2004).

DİZİN

Kitap Hakkında

Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik kitabı Türkçede kendi alanında yayımlanmış bir ilktir. Bu kitapta, davranış bilimlerinde, psikolojide, eğitim bilimlerinde, sosyolojide, işletme bilimlerinde ve diğer sosyal bilim disiplinlerinde kullanılan ölçüm araçlarının, yöntemlerinin ve ölçüm verilerinin güvenilirlik ve geçerliliğini sağlamak için sık başvurulan matematiksel ve istatistiksel teknikler ele alınmıştır. Kitapta iki temel konu üzerinde odaklanılmıştır: güvenilirlik ve geçerlilik. Birinci konuda, ölçüm verilerinin *olmazsa olmazı* olan doğruluk ve kesinlik teması işlenirken, yazar bu bölümde geniş bir literatür taraması yapmış ve güvenilirlik analizi yöntemlerini veri yapılarıyla ilişkili olarak ele almıştır. Kitapta güvenilirlik yöntemleri örnek uygulamalarla açıklanmış ve verilerin bilgisayara tanıtım biçimleri üzerinde durulmuştur. İkinci konu, ölçüm verilerinin ölçüm konusuyla olan bağıntısını belirleyen geçerliliktir. Bu bölümde ise yargısal ve istatistiksel geçerlilik yöntemleri ele alınmıştır. Kitap büyük ölçüde yüksek lisans ve doktora öğrencileri için hazırlanmış olsa da, ölçme konusuna ilgi duyan diğer araştırmacıların da kolaylıkla yararlanabilmeleri için basit bir dille yazılmış; matematiksel aksiyomların ispatlarından ve kapsamlı istatistiksel formüllerin verilmesinden kaçınılmıştır.

Yazar Hakkında

Prof. Dr. Hüner Şencan, İÜ İşletme Fakültesi, Davranış Bilimleri Ana Bilim Dalı'nda öğretim üyesidir. Şencan, ana bilim dalında Örgütsel Davranış, İletişim Teknikleri, Araştırma Yöntem Bilimi, Davranışsal Araştırmalarda Bilgisayar Uygulamaları, Davranışsal İstatistik ve İşletmelerde Psikoteknik Uygulamalar gibi dersleri vermektedir.

Seçkin Yayıncılık A.Ş.
Sağlık Sokak No: 19/B 06410 Sıhhiye - Ankara
Tel: (0.312) 435 30 30 Faks: (0.312) 435 24 72
İnternet Adresi: www.seckin.com.tr
E-posta: seckin@seckin.com.tr

ISBN 975 347 884 4