3. BÖLÜM

ÖLÇME ARAÇLARININ VE SONUÇLARININ SAHİP OLMASI GEREKEN NİTELİKLER

ESSENTIAL QUALITIES OF MEASUREMENT INSTRUMENTS, SCORES, AND SCORE INTERPRETATIONS

Prof. Dr. Mustafa İLHAN

ÖLÇME ARAÇLARININ VE SONUÇLARININ SAHİP OLMASI GEREKEN NİTELİKLER

Prof. Dr. Mustafa İLHAN
Dicle Üniversitesi

Öz

Ölçme sonucunda elde edilen puanların doğruluğunu ve bu puanlara dayalı olarak bireyler hakkında yapılan değerlendirmelerin isabet derecesini belirlemeye yönelik çalışmalar, ölçmede geçerlik ile güvenirliğin konusunu oluşturur. Ölçme sonuçlarına dayanarak herhangi bir karar vermeden önce ölçümlerin güvenilir ve ölçümlerden hareketle yapılacak yorumların geçerli olduğuna ilişkin kanıtları ortaya koymak gerekir. Bununla beraber, bir ölçme aracının taşıması gereken nitelikler geçerlik ve güvenirlikten ibaret değildir. Ölçme aracının uygulanabilirliğini, kullanım kolaylığını ve farklı koşullarda ne ölçüde tercih edileceğini etkileyen bir diğer kritik özellik kullanışlılıktır. Bu bölümde sıralanan bu kavramlar üzerinde durulmuştur. Önce ölçmede hatadan bahsedilmiş; hata kavramı, hata kaynakları ile türleri başlıklarına yer verilmiştir. Daha sonra güvenirlik başlığı ele alınmış, bu başlık altında klasik test kuramında kullanılan güvenirlik belirleme yöntemleri (test-tekrar test güvenirliği, paralel formlar güvenirliği, iç tutarlılık güvenirliği ve puanlayıcı güvenirliği), güvenirliği raporlarken dikkat edilmesi gereken hususlar, güvenirliği artırmak için alınabilecek önlemler ile ölçmenin standart hatası açıklanmıştır. Devamında geçerlik konusu işlenmiş; kapsama, içyapıya, ölçme aracından alınan puanların diğer değişkenlerle ilişkisine, cevaplama süreçlerine ve ölçme sonuçlarına dayalı geçerlik kanıtları başlıklar halinde sunulmuş, sıralanan başlıkları takiben geçerliği artırmak için alınabilecek önlemlere değinilmiştir. Ardından kullanışlılık başlığı altında kullanışlılığın ne anlama geldiği tanımlanmış; geçerlik, güvenirlik ile kullanışlılık dengesinden söz edilerek bölüm sonlandırılmıştır.

Anahtar Kelimeler

Psikometrik Özellikler, Ölçme Hatası, Güvenirlik, Geçerlik, Kullanışlılık

ESSENTIAL QUALITIES OF MEASUREMENT INSTRUMENTS, SCORES, AND SCORE INTERPRETATIONS

Prof. Dr. Mustafa İLHAN
Dicle University

Abstract

The accuracy of measurement scores and the extent to which decisions and inferences based on these scores appropriately reflect individuals are central concerns of reliability and validity in educational and psychological measurement. Before measurement results are used to support any decision-making process, sufficient evidence should be provided to demonstrate that the instrument produces reliable scores and supports valid interpretations of those scores. Nevertheless, the qualities expected of a measurement tool are not limited to reliability and validity alone. Another essential attribute that influences the applicability of an instrument, its ease of use, and its suitability across different contexts is practicality. This chapter examines these fundamental concepts in detail. It begins with a discussion of measurement error, including the concept of error, its sources, and its types. The chapter then addresses reliability by presenting the reliability estimation methods commonly employed within classical test theory, including test–retest reliability, parallel-forms reliability, internal consistency reliability, and rater reliability. In addition, issues that should be considered when reporting reliability, strategies for improving reliability, and the standard error of measurement are discussed. Next, the chapter turns to validity. Validity evidence based on test content, internal structure, relationships between test scores and other variables, response processes, and the consequences of testing is presented under separate headings. This discussion is followed by practical recommendations for enhancing validity. Subsequently, under the heading of practicality, the concept of practicality is defined, and the chapter concludes with a discussion of the balance among validity, reliability, and practicality.

Keywords

Psychometric Properties, Measurement Error, Reliability, Validity, Practicality

1. GİRİŞ

Tüm ölçme işlemlerinde amaç; ölçülen özelliğin gerçek değerini belirlemek, hatasız sonuçlara ulaşmaktır. Ancak gerçek puan platonik bir kavramdır, ne kadar uğraşırsak uğraşalım ölçülen özelliğin gerçek değerini tespit etmek olanaksızdır (Lord & Novick, 2008). Özellikle sosyal bilimlerde zekâ, başarı, yetenek, tutum, kaygı, kişilik gibi soyut kavramlarla ilgilenildiğinden ölçmeye bir miktar hata karışması kaçınılmazdır. Bundan dolayı ölçme sonucunda elde ettiğimiz puan, ölçülen özelliğin gerçek değeri ve hata olmak üzere iki bileşenden oluşur. Klasik test kuramı bunu $X = T + E$ eşitliği ile gösterir. Bu eşitlikte X gözlenen puanı, T (true) ölçülen özelliğin gerçek değerini, E (error) ise hata bileşenini yansıtır (Kline, 2005). Hatalı ölçmeler bireyler hakkında yanlış kararlar almamıza neden olacağından ölçmeye hata karışmasına yol açan faktörleri bilmemiz ve hatayı asgari düzeye çekecek tedbirleri hayata geçirmemiz gerekir. Ayrıca hatayı tamamıyla ortadan kaldırmak mümkün olmadığından yaptığımız ölçmelere ne kadar hata karıştığını ve ulaştığımız sonuçlara ne düzeyde itimat edebileceğimizi saptamamız zorunluluk arz eder. Ölçümlerin ve bu ölçümlerden hareketle yapılan çıkarımların isabet derecesini kestirmek amacıyla gerçekleştirilen işlemler ölçmede güvenirlik ile geçerliğin konusunu oluşturur. Ölçme sonuçlarını bireyler hakkında kararlar almak için kullanmadan önce bu iki psikometrik nitelik mutlaka test edilmelidir. Geçerlik ve güvenirlik dışında; ölçme aracının işlevselliğini etkileyen, özellikle hangi durumda ne sıklıkta işe koşulduğu üzerinde belirleyici rol oynayan üçüncü faktör ise kullanışlılıktır. Bu bölümde sıralanan bu kavramlar üzerinde durulacaktır. Önce ölçmede hatadan bahsedilecek; daha sonra güvenirlik, geçerlik ve kullanışlılık başlıklarına yer verilecektir.

2. ÖLÇMEDE HATA

Bir özelliğin ölçülen değeri ile gerçek değeri arasındaki fark hata olarak ifade edilir. Örneğin, bir kişinin ağırlığı gerçekte 70 kg iken baskülde görülen değerin 72 kg olması ölçme işlemine 2 kg'lık bir hata karıştığını gösterir. Baskülün aynı kişinin ağırlığını 67 kg olarak belirlemesi ise ölçme işlemine –3 kg'lık bir hata karıştığı anlamına gelir. Yani hatanın bir miktarı, bir de yönü vardır. Ölçme sonuçlarına farklı miktarlarda ve yönde hata karışabilmektedir.

2.1. Hata Kaynakları

Ölçme sonuçlarına ölçme aracından, ölçmenin yapıldığı ortamdan, ölçmeyi yapan kişiden ve ölçülen bireylerden kaynaklanan hatalar karışabilmektedir. Hata, bu kaynakların sadece birinden veya aynı anda birkaçından dolayı oluşabileceği gibi sözü edilen değişkenler arasındaki etkileşimler de hataya kaynaklık edebilmektedir. Bu hata kaynakları aşağıda örnekler üzerinden açıklanmıştır.

2.1.1. Ölçme Aracından Kaynaklanan Hatalar

Ölçme aracının yeterli duyarlılıkta olmaması hatalı ölçümlere yol açan faktörlerden ilkidir. Duyarlılık ölçme aracının ölçülen özellikteki farklılıkları ortaya koyabilme derecesini ifade eder. Ölçme aracının birimi küçüldükçe ölçülen özellikteki farklılıkları ortaya çıkarma gücü artar, buna bağlı olarak ölçmeye karışan hata azalır. Örneğin, gerçek ağırlığı 503 gram olan bir cismi hassas tartı ile tarttığımızda 503 değerine ulaşabiliriz. Ancak aynı cismi eşit kollu terazi ile tarttığımızda ulaşacağımız sonuç 500 gram olur. Eşit kollu terazinin duyarlılığının düşük olması ölçme işlemine 3 gramlık bir hata karıştırır. Benzer durum eğitimde kullanılan ölçme araçları için de geçerlidir. Bir başarı testinde yeterli sayıda maddeye yer verilmezse testin duyarlılığı düşer ve ölçme işlemine karışan hata artar. Yine Likert tipi ölçeklerde cevap kategorilerinin sayısı yeteriz olursa duyarlılık azalır ve hata artar. Yanlış madde türünün tercih edilmesi, ölçme aracının kapsaması gereken konuların bir kısmıyla ilgili maddeler içermemesi, hedef kitleye uygun olmayan bir yazı puntosunun kullanılması, baskının kalitesizliği, maddelerin silik çıkması, madde yanlılıkları (sorunun bir gruba haksız avantaj sunması), yönergenin ve/veya maddelerin net olmaması ölçme aracından kaynaklanan hatalara verilebilecek diğer örneklerdir.

2.1.2. Ölçmenin Yapıldığı Ortamdan Kaynaklanan Hatalar

Baskülün eğri bir zeminde bulunması ağırlığın hatalı ölçülmesine neden olacağı gibi sınavların uygulandığı ortamlar da hataya sebebiyet verebilmektedir. Ortamdaki havalandırmanın ya da aydınlatmanın yetersizliği; ortamının çok sıcak, çok soğuk veya gürültülü olması, sınava girdiği sıranın boyutunun adayın fiziksel özelliklerine uygun olmaması, çevrimiçi yapılan sınavlarda internetin yavaş çalışması ve/veya bilgisayarda teknik problemlerin bulunması ölçme hatalarını tetikleyebilecek ortamdan kaynaklı özelliklerdir.

2.1.3. Ölçülen Kişilerden Kaynaklanan Hatalar

Bireyin sınav günü hasta olması, sınav anında yüksek kaygı/heyecan duyması, uykusuz, yorgun veya karnı aç bir halde sınava girmesi, sınava yakın tarihte travmatik bir olay yaşaması, kopya çekmesi ya da cevaplarını optik forma geçirirken kaydırma yapması ölçme sonuçlarında hataya yol açabilen ölçülen kişiden kaynaklı faktörlerdir.

2.1.4. Ölçmeyi Yapan Kişiden Kaynaklanan Hatalar

Uygulanması uzmanlık gerektiren bir sınavın bu konuda yeterli bilgisi/deneyimi bulunmayan bir kişi tarafından yapılması, öğretmenin puanlamayı yaparken taraflı davranması ya da dikkatsizlik/yorgunluk sebebiyle bazı öğrencilere gereğinden fazla ya da eksik puan vermesi ölçmeyi yapan kişiden kaynaklı hatalara örnek olarak verilebilir.

2.2. Hata Türleri

Ölçmeye karışan hatalar sabit, sistematik ve tesadüfi olmak üzere üçe ayrılır. Bu hata türleri Tablo 1'de başlıklar halinde açıklanmıştır.

Tablo 1

Sabit, Sistematik ve Tesadüfi Hatalar

Hata Türü	Örnekler
Sabit Hata Kaynağı bilinen, tüm ölçümlere aynı miktarda ve yönde karışan, puanların standart sapması ile varyansını değiştirmeyen hatadır (Başol, 2018).	• Öğretmenin sınav kâğıtlarını puanlarken tüm öğrencilere 5 puan fazla vermesi ya da her öğrencinin notuna sınıf ortalamasının %10'u kadar puan eklemesi, • Bir madde silik çıktığı ya da net olmadığı için sınava giren hiç kimse tarafından doğru cevaplanamaması, • Bir baskülün her nesneyi 100 gram eksik tartması, • Bir saatin sürekli olarak 5 dakika geri göstermesi, • İlk *x* santimi kırık bir cetvelin bir defa kullanıldığı uzunluk ölçümleri
Sistematik Hata Ölçülen büyüklüğe, ölçmeyi yapan kişiye ya da ölçme koşullarına göre miktarı ve/veya yönü değişen; kaynağı belli olan hatalardır (Turgut, 1983).	• Öğretmenin yazısı güzel olanlara 5 puan fazla vermesi ya da her öğrenciye sınavdan aldığı notun %10'u kadar puan eklemesi, • Bir maddenin içeriği itibariyle aynı yetenek düzeyindeki farklı alt gruplardan birine avantaj sağlar nitelikte olması, • Çoktan seçmeli bir maddede yanlış seçeneğin doğru cevap olarak belirlenmesi, • Bir saatin her 60 dakikada bir 5 dakika geri kalması, • İlk *x* santimi kırık bir cetvelin birden fazla kez kullanıldığı uzunluk ölçümleri
Tesadüfi Hata Kaynağı bilinmeyen; ölçme sonuçlarına nerede, nasıl ve hangi miktarda karıştığı belli olmayan hatalardır (Başokçu, 2009).	• Öğretmenin farkında olmadan bazı öğrencilere fazla ya da eksik puan vermesi, • Öğrencinin cevaplarını optik forma geçirirken kaydırma yapması, • Şans başarısı (maddenin kör tahminle doğru cevaplanması), • Maddenin cevaplayıcılar tarafından farklı şekillerde anlaşılması, • Sınav ortamının sıcak, soğuk, gürültülü olmasından dolayı ya da sınav sırasında kaygılı/hasta olduğu için öğrencinin gerçek performansını ortaya koyamaması • Sınavda kopya çekilmesi

Tablo 1'deki hata türleriyle ilgili vurgulanması gereken diğer noktalar şunlardır: Sistematik hata bir grubun lehine, diğer grubun aleyhine sonuçlar ürettiğinden yanlılık olarak da isimlendirilir. Tesadüfi hata için ise gelişigüzel/rastgele etiketleri de kullanılır. Tesadüfi hata güvenirliği düşürücü etkiye sahipken sabit ve sistematik hata güvenirliği etkilemez. Geçerlik ise her üç hata türünden de etkilenir.

3. GÜVENİRLİK

Güvenirlik, aynı özellik için elde edilen iki ya da daha fazla ölçme sonucu arasındaki tutarlılığı ifade eder. Örneğin, bir basküle arka arkaya birkaç kez çıktığımızda her seferinde aynı sonucu görürsek baskülün tutarlı ölçmeler yaptığını düşünür, bunu güvenirlik kanıtı olarak yorumlarız (Banks, 2012). Ancak basküle her çıkışımızda başka bir değer ile karşılaşırsak baskülün güvenilir ölçmeler yapmadığını anlarız. Benzer şekilde, kan basıncımızı biri analog diğeri dijital iki ayrı tansiyon aleti ile ölçtüğümüzde aynı sonuca ulaşırsak ölçümlerin güvenilir olduğuna kanaat getiririz. Analog ve dijital tansiyon aletinin gösterdiği sonuçlar arasında tutarsızlık olması halinde ise tansiyon aletlerinden en az birinin hatalı ölçme yaptığı çıkarımına varır, elde ettiğimiz sonuçlara güvenemeyiz. Öğrencilerin matematik dersine yönelik tutumlarını, cevap kategorileri "Katılıyorum, Biraz Katılıyorum, Katılmıyorum" şeklinde etiketlenen Likert tipi bir ölçek aracılığıyla ölçtüğümüz bir başka örnek ele alalım. Diyelim ki öğrencilerden biri bu ölçekteki "Matematik dersini seviyorum" ve "Matematik dersinden nefret ediyorum" maddelerinin ikisine de *Katılıyorum* cevabını verdi. Bu durumda öğrencinin iki maddedeki cevapları tutarsızdır ve bir güvenirlik sorunu var demektir.

Ölçme sonuçlarındaki tutarsızlıklara ölçme işlemine karışan tesadüfi hatalar yol açar. Dolayısıyla güvenirlik, ölçme sonuçlarının tesadüfi hatalardan arınıklık derecesi olarak da tanımlanır. Yani güvenirlik ile tesadüfi hata arasında negatif bir korelasyon vardır. Hata arttıkça güvenirlik azalırken hata düştüğünde güvenirlik yükselir. Ölçme sonuçlarına; ölçme aracı, ölçmenin yapıldığı ortam, ölçülen ve ölçmeyi yapan kişi gibi farklı kaynaklardan gelen hatalar karışabildiğinden güvenirliği test ederken olası hata kaynaklarını göz önünde bulundurmak gerekir. Bu süreçte kullanılan yöntemler güvenirlik kestiriminde hangi ölçme kuramının esas alındığına bağlı olarak değişir. Daha açık ifade etmek gerekirse ölçme kuramları bölümünde yer verilen klasik test kuramı (KTK), madde tepki kuramı (MTK) ile genellenebilirlik (G) kuramı güvenirliği farklı şekillerde kavramsallaştırır ve analiz eder. Örneğin, MTK'ye göre aynı testi alan farklı yetenek düzeyindeki kişiler için ölçme sonuçlarına karışan hata miktarı aynı değildir, yetenek düzeyine göre farklılaşır. Haliyle farklı yetenek düzeyindeki bireyler için elde edilen güvenirlik değerleri birbirinden farklı olur. Öte yandan KTK'de hata miktarı ve güvenirlik değeri yetenek düzeyinden bağımsız olup testin uygulandığı gruptaki tüm kişiler için özdeştir. Güvenirlik kestiriminde G kuramı ile KTK'nin birbirinden ayrıştığı temel nokta ise şudur. Genellenebilirlik kuramı, potansiyel hata kaynaklarının tamamını aynı anda analiz edip güvenirliğe ilişkin tek bir katsayı verir. Buna karşın KTK her bir hata kaynağı için ayrı bir güvenirlik çalışması yürütmeyi gerektirir. Bu bölümde güvenirlik konusu KTK çerçevesinde işlenmiş, KTK'de güvenirliği hesaplarken kullanılan yöntemler Tablo 2'de sunulmuştur.

Tablo 2

KTK'de Kullanılan Güvenirlik Hesaplama Yöntemleri[*]

Hata Kaynağı	Güvenirlik Belirleme Yöntemi	Taşıdığı Anlam
Testin uygulandığı bireylerde veya ölçme koşullarında zaman içinde meydana gelen değişiklikler	Test-tekrar test güvenirliği	Kararlılık
Alternatif formlar arasındaki farklılıklar	Paralel formlar güvenirliği	Denklik/Eşdeğerlik
Ölçme aracındaki maddeler	İç tutarlılık güvenirliği	Maddelerin homojenliği / heterojenliği
Puanlayıcılar/gözlemciler/hakemler	Puanlayıcı güvenirliği	Objektiflik

[*] Han'dan (2021) uyarlanmıştır.

Tablo 2'deki güvenirlik belirleme yöntemlerinin her biri tutarlılığın başka bir yönüne odaklanır. Test-tekrar test güvenirliğinde analiz edilen zamana karşı tutarlılıktır. Paralel formlar güvenirliğinde formlar arasındaki tutarlılık incelenir. İç tutarlılık güvenirliğinde maddelerin birbiriyle tutarlılığına bakılır. Puanlayıcı güvenirliği ise aynı özellik için farklı kişilerin ya da aynı kişinin farklı zamanlarda yaptığı puanlamalar arasındaki tutarlılığı irdeler.

3.1. Güvenirlik Belirleme Yöntemleri

Güvenirlik, tutarlılık kavramı ile eşleştirilse de (Popham, 2025) KTK'de güvenirliği test etmek için tutarlılığa; kararlılık, eşdeğerlik, iç tutarlılık ve objektiflik gibi başka anlamlar yükleyen farklı yöntemler ileri sürülmüştür. KTK'de kullanılan güvenirlik kestirim yöntemleri aşağıda başlıklar halinde açıklanmıştır.

3.1.1. Test-Tekrar Test Güvenirliği

Test-tekrar test güvenirliğinde ölçme aracı aynı gruba belli aralıklarla iki defa uygulanır. Daha sonra katılımcıların ölçme aracının birinci ve ikinci uygulamasından aldığı puanlar arasındaki korelasyon hesaplanır. Elde edilen korelasyon katsayısı *kararlılık* anlamında güvenirlik olarak yorumlanır. Peki, ölçümlerin kararlı olduğunu iddia edebilmek için hesaplanan korelasyon katsayısı en az kaç olmalıdır? Bu konuda net bir alt sınır ileri sürmek güçtür. Crocker ve Algina'ya (2008) göre kişilik, ilgi, tutum vb. bilişsel olmayan özellikler ölçülüyorsa 0,80'i aşan test-tekrar test katsayıları, ölçülmesi hedeflenen yapının iyi tanımlandığına ve ölçme aracının iyi yapılandırıldığına işaret eder. Yetenek

Test-tekrar test güvenirliği, birinci ve ikinci uygulama arasında geçen sürede bireylerin ölçülen özelliğe ilişkin gerçek puanlarında bir değişim olmadığı varsayımına dayanır (Furr & Bacharach, 2008). Bu varsayımın ihlal edildiği biliniyorsa, örneğin iki uygulama arasında bir öğretim müdahalesi söz konusuysa, test-tekrar test güvenirliğini hesaplamak yanlış olur.

testlerinde ise daha yüksek bir test-tekrar test güvenirliğine ulaşılması beklenir. Elbette, ölçülen özelliğin yanı sıra iki uygulama arasında geçen süre de alt sınır olarak kabul edilmesi gereken değeri etkiler (Bandalos, 2018).

"Birinci ve ikinci uygulama arasında ne kadar süre bulunmalıdır?" test tekrar güvenirliğine ilişkin bir başka kritik sorudur. Güvenirlik için esas alınması gereken alt sınır gibi, bu sürenin ne kadar olması gerektiğine dair de net bir şey söylemek mümkün değildir. Çünkü ideal süre ölçülen özelliğe ve ölçme aracının uygulandığı gruptaki kişilerin gelişim özelliklerine bağlı olarak farklılaşır (Mitchell ve ark., 2000). Akademik başarı gibi kısa vadede değişebilen bir özellik ölçülüyorsa iki uygulama arasındaki süre kısa tutulmalıdır. Diğer taraftan zekâ, siyasi görüş ya da tutum gibi değişmesi uzun zaman alan bir özellik üzerinde duruluyorsa iki uygulama arasındaki süre daha uzun olabilir. Bu anlamda birinci ve ikinci uygulama arasındaki süreye karar verirken dikkat edilmesi gereken temel nokta şudur: Aradaki zaman dilimi katılımcılarda olgunlaşma ya da öğrenmeye bağlı değişimler yaşanmayacağı kadar kısa ama ilk uygulamada verilen cevapların hatırlanmayacağı kadar da uzun olmalıdır (Crocker & Algina, 2008). Test-tekrar test güvenirliğine ilişkin uygulamalarda gözetilmesi gereken diğer kritik hususlar şöyle sıralanabilir:

- ⇨ İki uygulama arasında ne kadar süre olduğu mutlaka belirtilmelidir.
- ⇨ Ölçme aracının birinci ve ikinci uygulamasındaki koşullar olabildiğince benzer olmalıdır. Mesela test ilk uygulamada kâğıt kalem formatında, ikinci uygulamada çevrimiçi uygulanırsa katılımcıların puanlarında uygulama biçiminden kaynaklı farklar ortaya çıkabilir. Bu da test-tekrar test güvenirliği hakkında yanıltıcı sonuçlar doğurur.
- ⇨ Birinci ve ikinci uygulamaya ait veriler aynı sınıftan geldiğinden, diğer bir deyişle aynı metrikte olduğundan iki uygulamadan elde edilen puanlar arasındaki korelasyon hesaplanırken eğer varsayımları (örneğin, normallik) karşılanıyorsa sınıf-içi korelasyon katsayısına başvurulmalıdır.
- ⇨ Test-tekrar test yöntemi her bir katılımcının birinci ve ikinci uygulamada yanıtladığı ölçme araçlarını eşleştirmeyi gerektirir. Bunun için ölçme aracı isim/rumuz alınarak katılımcılara uygulanır. Bu işlem zorlayıcı olabildiğinden test-tekrar test güvenirliğine yönelik uygulamalarda büyük örneklemlerden veri toplamak zordur. Ancak yine de yeterli büyüklükte bir örneklemden veri toplandığından emin olunmalıdır. Ulaşılması gereken örneklem büyüklüğü hakkında farklı araştırmacılar değişik görüşler ileri sürmüştür. Koçak ve arkadaşları (2022) en az 30 katılımcıdan oluşan bir örneklemden veri toplanmasını önermiştir. Kennedy (2022) yaptığı ampirik araştırmada 20-30 kişilik örneklemlerin test-tekrar test güvenirliğine ilişkin doğru sonuçlar vermediğini bildirmiş, 40 ila 50 katılımcı içeren örneklemleri kısmen kabul edilebilir bulsa da bu örneklem büyüklüklerinde 0,70 eşiğine ulaşmanın zor olduğunu rapor etmiştir. Test-tekrar test güvenirliğine ilişkin daha güçlü kestirimlere ulaşabilmek için en az 100 kişiden oluşan bir örnekleme

ulaşılmasını önermiştir. Charter (1999) ise güvenilirlik tahmininin kesinliğini örneklem büyüklüğünün bir fonksiyonu olarak nitelendirip test-tekrar test güvenilirliği için örneklemin en az 400 kişiden oluşması gerektiğini ifade etmiştir.

Örnek teşkil etmesi açısından Tablo 3'te, test-tekrar test güvenirliğine ilişkin hipotetik olarak üretilmiş küçük bir veri seti sunulmuştur. Verilerin, çoktan seçmeli altı maddeden oluşan bir testin 10 kişilik bir öğrenci grubuna 3 hafta arayla iki kez uygulanarak elde edildiği varsayılmaktadır. Maddelere ait doğru yanıtlar 1 olarak puanlanmış, yanlış yanıtlanan ve boş bırakılan maddelere ise 0 puan verilmiştir. Tablodaki "Toplam" etiketli sütunlar öğrencilerin testten aldıkları toplam puanı bildirmektedir. Tabloya bakıldığında birinci ve ikinci uygulamadaki toplam puanların genel olarak birbirine benzer olduğu görülmektedir. Buna bağlı olarak test-tekrar test güvenirliği için 0,93 gibi yüksek bir korelasyon katsayısı hesaplanmıştır.

Tablo 3

Test-Tekrar Test Güvenirliğine İlişkin Örnek Veri Seti

Öğrenciler	Birinci Uygulama							İkinci Uygulama						
	M1	M2	M3	M4	M5	M6	Toplam	M1	M2	M3	M4	M5	M6	Toplam
Ö1	1	1	1	1	0	1	5	1	1	1	1	1	1	6
Ö2	1	1	1	1	1	1	6	1	1	1	1	1	1	6
Ö3	1	0	0	0	0	0	1	1	0	1	0	0	0	2
Ö4	1	1	0	0	0	1	3	1	1	1	0	0	1	4
Ö5	1	0	0	0	0	1	2	1	1	0	0	0	1	3
Ö6	1	1	0	1	1	1	5	1	1	0	1	0	1	4
Ö7	1	1	1	0	0	0	3	1	0	1	0	0	0	2
Ö8	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Ö9	1	1	1	1	1	1	6	1	1	1	1	0	1	5
Ö10	1	1	0	1	0	1	4	1	1	1	1	0	1	5
İki uygulamadan alınan puanlar arasındaki sınıf-içi korelasyon katsayısı: 0,93 [%95 GA (0,753; 0,984)]														

* *GA, güven aralığını ifade etmektedir. Görüldüğü gibi korelasyon katsayısına ilişkin kestirim, güven aralığının alt ve üst sınırının orta noktasında değildir. Örneklem küçükse, korelasyon katsayısı mutlak değerce 1'e yakınsa (Cumming & Fidler, 2015) ya da verilerde ranj daralması söz konusu ise (Fidler & Cumming, 2019) korelasyon katsayısına ilişkin güven aralığı asimetrik olmaktadır.*

Test-tekrar test güvenirliğinde ilk uygulamanın ardından belli bir süre geçmiş olsa bile katılımcılar ölçme aracındaki maddeleri ve maddelere verdikleri cevapları hatırlayabilmekte, bu da ikinci uygulamadaki test puanlarını etkileyebilmektedir (carry over effect) (Price, 2017). Bu sınırlılık özellikle maksimum performansın (başarı, zekâ, yetenek vb.) ölçüldüğü durumlarda kendini göstermektedir.

3.1.2. Paralel (Eşdeğer) Formlar Güvenirliği

Bu güvenirlik türünde aynı özelliği ölçen ancak farklı maddelerden oluşan iki ölçme aracı aynı gruba uygulanır. Ardından iki formdan alınan puanlar arasındaki korelasyon hesaplanır ve bulunan korelasyon katsayısı *eşdeğerlik* anlamında güvenirlik olarak yorumlanır. Paralel formlar güvenirliğinde en önemli konu iki formun içerik itibariyle gerçekten denk olup olmadığıdır. İki formu denk olarak nitelendirilebilmemiz için bu formlar aynı kapsamı ölçmelerinin yanında madde sayıları ve türleri açısından da özdeş olmalı, benzer güçlük ve ayırt edicilikte maddeler içermelidir. Biri çoktan seçmeli, diğeri açık uçlu maddelerden meydana gelen iki matematik testi düşünelim. Bu testleri paralel form olarak ele almak uygun değildir. Çünkü madde türündeki farklılıklar katılımcıların test puanlarında farklılık yaratacak, böyle bir durumda iki forma ait puanlar arasındaki korelasyon paralel form güvenirliğine ilişkin doğru bir sonuç sunmayacaktır.

Formların katılımcılara nasıl uygulanacağı paralel form güvenirliğindeki bir diğer önemli hususdur. Formlar çok sayıda maddeden oluşuyorsa aynı anda uygulanmaları katılımcılarda yorgunluğa ve dikkatsizliğe yol açabilir. Buna mani olmak için formları farklı oturumlarda uygulama yoluna gidilmelidir. Ancak bu sefer de uygulama sırasından kaynaklı etkiler oluşabileceği unutulmamalı, söz konusu etkiyi kontrol altına alabilmek için karşıt dengelenmiş desen (counter balanced design) takip edilmelidir. Buna göre örneklem iki gruba ayrılmalı, grubun yarısına önce A sonra B formu uygulanırken diğer yarısında tersi bir sıra izlenmelidir. Formlar tek oturumda uygulanabilecek kadar kısa olduğunda ise iki form katılımcılara aynı anda sunulmalı, tabi burada da yine karşıt dengelenmiş desene bağlı kalınmalıdır. Grubun yarısı önce A arkasından B formundaki maddeleri cevaplarken diğer yarısı B formundaki maddeler ile başlayıp sonrasında A formundaki maddelere geçmelidir.

Şimdi paralel form güvenirliğine örnek oluşturacak bir veri seti ele alalım. Elimizde; içerdikleri maddelerin güçlük ve ayırt edicilik indekslerinin benzer olduğu öngörülen, vücudumuzdaki sistemler öğrenme alanıyla ilgili iki biyoloji testi bulunduğunu varsayalım. Testlerin ikisinde de dolaşım, solunum, sindirim, boşaltım, üreme, iskelet ve sinir sistemleriyle ilgili birer madde olmak üzere yedi madde (M1–M7) yer alıyor olsun. Diyelim ki bu testleri 10 öğrenciye uyguladık ve uygulama sonucunda Tablo 4'te sunulan verilere ulaştık.

Tablo 4

Paralel Formlar Güvenirliğine ilişkin Örnek Veri Seti

Öğrenciler	A Formu								B Formu							
	M1	M2	M3	M4	M5	M6	M7	Toplam	M1	M2	M3	M4	M5	M6	M7	Toplam
Ö1	1	1	1	0	1	1	1	6	1	1	0	0	1	1	1	5
Ö2	1	1	0	0	1	1	1	5	1	1	1	0	1	1	1	6
Ö3	1	0	0	0	1	0	1	3	1	1	0	0	1	1	1	5
Ö4	1	0	0	0	0	1	1	3	1	1	0	0	0	1	1	4
Ö5	1	1	1	1	1	1	1	7	1	1	1	1	1	1	1	7
Ö6	1	0	0	0	0	0	0	1	1	2	0	0	0	0	1	4
Ö7	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Ö8	1	1	1	0	0	0	1	4	1	1	1	0	0	1	1	5
Ö9	0	1	1	1	1	0	1	5	1	1	1	0	1	1	1	6
Ö10	1	1	1	1	1	0	1	6	1	1	1	0	1	1	1	6

A ve B formundan alınan puanlar arasındaki Spearman sıra farkları korelasyonu: 0,864 [%95 GA (0,515; 0,967)]

Tablo 4'te görüldüğü üzere maddeler 1-0 şeklinde puanlanmış, öğrenciler A ve B formu olarak isimlendirilen iki biyoloji testinden oldukça benzer puanlar almıştır. İki formdan elde edilen puanlar arasındaki korelasyon 0,864 bulunmuştur. Popham (2000) standart test geliştiricilerinin genelde 0,80 ila 0,90 civarında bir paralel form güvenirliğine ulaşmayı hedeflediklerini belirtmektedir. Ancak iki form farklı zamanlarda uygulandığında arada geçen süreden kaynaklı etkilerin de ölçme sonuçlarına karışabileceğini ifade etmekte, bu tür durumlarda 0,70 değerinin ölçüt olarak alınmasını önermektedir. Sıralanan referans değerlere göre, Tablo 4'teki korelasyon katsayısının eşdeğerlik anlamında yüksek bir güvenirliğe işaret ettiği anlaşılmaktadır.

Test-tekrar test güvenirliği için geçerli olan cevapların hatırlanması gibi bir risk taşımasa da paralel form güvenirliğinin de önemli bir sınırlılığı vardır. Aynı konu alanını örnekleyen, madde sayıları ile istatistikleri bakımından benzer olan iki form geliştirmenin zor ve zaman alıcı olması bu yöntemin temel sınırlılığıdır. Bu sebeple paralel formlar güvenirliği çok fazla kullanılmamaktadır.

3.1.3. İç Tutarlılık Güvenirliği

Ölçme aracındaki maddelerin birbiriyle ve ölçme aracından alınan toplam puanla ilişkisine dayalı olarak hesaplanan bir güvenirlik türüdür. Ölçme aracındaki tüm maddeler aynı örtük özelliği ölçmeye yönelik olduğunda, diğer bir deyişle maddeler ölçtükleri özellik bakımından homojen bir yapı gösterdiğinde iç tutarlılık güvenirliği yüksek çıkar. Maddeler ölçtükleri özellik açısından heterojen bir yapıya sahip olduğunda ise düşük bir iç tutarlılık katsayısı hesaplanır.

Bir Türkçe testini ele alalım. Testte 15 madde bulunuyor olsun. Bu testteki 14 madde okuduğunu anlama becerisini ölçmeye yönelikken bir maddenin dilbilgisi ile ilgili olduğunu varsayalım. Dilbilgisi içerikli madde testteki diğer maddelerden daha başka bir özelliği ölçtüğünden iç tutarlılık güvenirliğini düşürücü etki yaratacaktır. İç tutarlılık güvenirliğine ilişkin kanıt elde edebilmek için eşdeğer yarılar (test yarılama) yöntemi kullanılabilir ya da verilerin yapısı göz önünde bulundurularak KR-20, KR-21, Cronbach alfa, McDonald omega, tabakalı alfa ve hiyerarşik omega katsayılarından uygun olanı hesaplanır.

> Tek uygulamaya dayandığından ve birden fazla forma ihtiyaç duyulmadığından iç tutarlılık güvenirliği test-tekrar test ve paralel (eşdeğer) formlar güvenirliğine kıyasla çok daha sık başvurulan bir güvenirlik belirleme yöntemidir.

3.1.3.1. Eşdeğer Yarılar (Test Yarılama) Yöntemi

Bu yöntemde test bir bütün olarak katılımcılara uygulanır ama uygulamadan sonra testteki maddeler olabildiğince eş iki yarıya ayrılır. Bu ayırma işlemi genelde tek ve çift numaralı maddeler şeklinde yapılır. Tabi bu noktada "Testi neden ilk ve ikinci yarı şeklinde değil de tek ve çift numaralı maddeler şeklinde ikiye bölüyoruz?" sorusunu cevaplamamız gerekir. Öncelikle katılımcıların sınav anındaki motivasyonlarını olumsuz etkilememek için testteki maddeler genelde kolaydan zora doğru sıralanır. Dolayısıyla testi ilk ve ikinci yarı şeklinde bölmemiz halinde iki yarıdaki maddelerin güçlük düzeyleri birbirinden farklı olur. Bu durumda iki yarı eşdeğer kabul edilemez. Ayrıca bir testte aynı konu alanına yönelik maddeler çoğunlukla bir arada sunulur. Örneğin, bir ölçme ve değerlendirme testinde en başta genellikle temel kavramlarla ilgili maddeler yer alır. Bu maddeleri sırasıyla ölçme araç ve sonuçlarında bulunması gereken özellikler, eğitimde kullanılan geleneksel ve tamamlayıcı ölçme araçları, test geliştirme, madde analizi ile ölçme sonuçları üzerinde istatistiksel işlemler konularına yönelik maddeler takip eder. Durum böyleyken testi ilk ve ikinci yarı şeklinde bölmek eşdeğer olmayan iki yarıyla sonuçlanır. Örneğin, matematiksel hesaplamaların yer aldığı ölçme konularının tamamı ikinci yarıda bulunur. İlk ve ikinci yarı şeklinde yapılacak bir yarılamadan kaçınmayı gerektiren bir başka faktör de öğrencilerin bir kısmının testin sonundaki maddeleri yetiştirememiş olma ihtimalidir.

Testi en uygun şekilde ikiye ayırdıktan sonra her bir öğrencinin testin iki yarısından aldıkları puanlar belirlenir. Daha sonra iki yarıdan alınan puanlar arasındaki korelasyon katsayısı hesaplanır. Bu korelasyon katsayısı testin tamamına değil sadece bir yarısına ait güvenirliği bildirir. Bunu şöyle düşünebiliriz. Test-tekrar test yönteminde 10 maddelik bir testten alınan puanlar ile bu 10 maddelik testin ikinci uygulamasından elde edilen puanlar arasındaki korelasyon hesaplanır ve bulunan korelasyon 10 maddelik teste ait bir güvenirlik indeksi verir. Aynı şekilde eşdeğer formlar yönteminde 10 maddelik bir testten alınan puanlarla 10 madde içeren eşdeğer forma ait puanlar

arasındaki korelasyon hesaplanır ve ulaşılan korelasyon değeri 10 maddelik test için bir güvenirlik kanıtı sunar. Diğer taraftan test yarılama yönteminde test yapay olarak iki yarıya ayrılır ve tek numaralı maddeler (5 madde) ile çift numaralı maddelerden (5 madde) alınan puanlar arasındaki korelasyon hesaplanır ve bu korelasyon katsayısı sadece 5 sorunun güvenirliğini gösterir (Aybek, 2021). Testin tamamının güvenirliğine ulaşabilmek için hesaplanan korelasyon katsayısına Sperman-Brown düzeltmesi uygulanır (bkz. Eşitlik 1).

$$r_T = \frac{2 \times r_Y}{1 + r_Y}$$

r_Y = İki yarı arasındaki korelasyon

r_T = testin tamamına ait güvenirlik

Eşitlik 1

Her bir maddenin 1 ile 5 arasında puanlandığı açık uçlu 8 maddeden oluşan bir testin 10 öğrenciye uygulandığını ve uygulama sonucunda Tablo 5'te sunulan verilere ulaşıldığını farz edelim. Öğrencilerin testin tek numaralı maddelerinden (M1, M3, M5 ve M7) aldığı toplam puan "Tek" etiketli sütunda, çift numaralı maddelerinden (M2, M4, M6 ve M8) aldığı toplam puan ise "Çift" başlıklı sütunda verilmiştir. Tabloda görüldüğü gibi testin tek ve çift numaralı maddelerinden alınan puanlar arasındaki Spearman sıra farkları korelasyonu (r_Y) 0,793 çıkmıştır. Testin yarısına ait güvenirliği ifade eden bu korelasyon katsayısı, Spearman-Brown formülünde yerine konulduğunda testin bütününe ait iç tutarlılık katsayısı 0,885 olarak kestirilmiştir.

Tablo 5

Test Yarılama Yöntemiyle İç Tutarlılık Katsayısının Hesaplanması

Öğrenciler	Maddeler								Toplam	Tek	Çift
	M1	M2	M3	M4	M5	M6	M7	M8			
Ö1	5	4	3	4	4	4	5	4	33	17	16
Ö2	2	4	1	3	2	2	2	3	19	7	12
Ö3	5	5	4	5	5	5	5	4	38	19	19
Ö4	5	5	5	5	5	5	5	5	40	20	20
Ö5	3	2	1	2	4	2	3	2	19	11	8
Ö6	2	3	2	1	2	3	2	2	17	8	9
Ö7	5	4	3	4	3	4	4	5	32	15	17
Ö8	4	4	3	3	4	4	5	4	31	16	15
Ö9	3	2	1	2	3	2	3	2	18	10	8
Ö10	5	4	3	3	5	4	5	4	33	18	15

r_Y = 0,793 olarak elde edilir ve bu korelasyon katsayısına göre r_T = [(2×0,793)/(1+0,793)] = 0,885 bulunur.

Dikkat edilirse testin tamamına ait güvenirlik katsayısı testin bir yarısına ait güvenirlikten yüksektir. Bu sonuç, madde sayısı ile güvenirlik arasındaki ilişkiye kanıt oluşturmaktadır. Şöyle ki madde sayısı artırıldığında testin duyarlılığı artacağından güvenirlik de yükselir. Ancak bu aşamada, teste eklenen her maddenin muhakkak

güvenirlik katsayısında artış sağlayacağı gibi bir yanılgıya düşülmemelidir. Eğer teste eklenen yeni bir madde testteki mevcut maddelerden daha başka bir özelliği yokluyorsa güvenirliği artırmayacak, tam tersi güvenirlik katsayısında düşüşe neden olacaktır. Ayrıca teste eklenen maddelerin testin güvenirliğini belli bir noktaya kadar artıracağı, bu noktadan sonra madde sayısındaki artışın güvenirliği pek fazla etkilemeyeceği hatta madde sayısının çok fazla artırılmasının öğrencilerde yorgunluğa neden olarak güvenirliği düşüreceği gözden kaçırılmamalıdır.

3.1.3.2. KR-20 ve KR-21 Katsayıları

KR-20 ile KR-21 Chicago Üniversitesi'nde görev yapan G. Frederic Kuder ve Marion Richardson (1937) tarafından maddelerin iki kategorili olarak (0–1) puanlandığı ölçme araçları için geliştirilmiş iç tutarlılık katsayılarıdır. Daha açık bir anlatımla çoktan seçmeli, doğru-yanlış ya da eşleştirme türünde maddelerden oluşan testler ile var/yok, evet/hayır gibi iki kategorili puanlanan performans görevleri iç tutarlılığın KR-20 ve KR-21 katsayıları ile kestirilebileceği ölçmelerdir.

Kuder ve Richardson 1937 tarihli "*Test güvenirliğini kestirim teorisi [The theory of the estimation of test reliability]*" başlıklı makalelerinde 22 ayrı eşitlik sıralamışlardır. KR-20 ile KR-21 katsayıları, ilgili makaledeki 20 ve 21 numaralı eşitlikler kullanılarak hesaplanmaktadır. İşte KR-20 ve KR-21 kısaltmasındaki *K* ile *R* harfleri bu formülleri geliştiren araştırmacıların soyadlarından, 20 ve 21 sayıları ise sözü edilen makaledeki eşitlik numaralarından gelmektedir.

KR-20 katsayısı ölçme aracındaki maddelerin güçlük düzeyleri birbirinden farklı olduğunda kullanılırken KR-21 madde güçlüklerinin aynı olduğunu varsayar. Ölçme aracındaki maddelerin güçlük düzeyleri bilinmiyorsa tüm maddelerin güçlük düzeyleri eşit kabul edilir ve iç tutarlılık KR-21 ile kestirilir. Maddelerin güçlük düzeylerinin birbirine yakınlığı ölçüsünde KR-21, KR-20'ye yaklaşır. Ancak KR-21'in KR-20'den büyük olması söz konusu değildir. Yani bu iki katsayı arasında KR-20 ≥ KR-21 şeklinde bir matematiksel ilişki vardır.

Günümüzde ölçümlerin güvenirliğini tespit etmeye yönelik analizler istatistiksel yazılımlar aracılığıyla gerçekleştirildiğinden KR-20 ve KR-21 katsayılarının elle hesaplandığı bir örnek verilmemiştir. Ancak güvenirliği etkileyen faktörleri anlayabilmek için bu katsayıların formüllerini bilmek önemlidir. Bundan dolayı KR-20 ile K-21 katsayılarının hesaplanmasında kullanılan formüller aşağıda sunulmuştur.

$$\text{KR-20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{S^2}\right)$$

Eşitlik 2

k = Testteki madde sayısı

S^2 = Test puanlarının varyansı

p_i = i. maddeyi doğru cevaplayan kişi sayısının tüm cevaplayıcılara oranı (madde güçlük indeksini ifade eder)

q_i = i. maddeyi yanlış cevaplayan kişi sayısının tüm cevaplayıcılara oranı ("*1-p*" değerine eşittir)

$$\text{KR-21} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\bar{X}(k-\bar{X})}{kS^2}\right)$$

k = Testteki madde sayısı
$\bar{X}$ = Test puanlarının ortalaması

Eşitlik 3

Yukarıdaki eşitliklere bakıldığında (k/k-1) oranın iki formüldeki ortak terim olduğu görülmektedir. Bu terim, madde sayısının güvenirlik üzerindeki etkisini azaltmaya dönük bir düzeltme işlemi olarak formüllere dâhil edilmiştir. KR-20 formülündeki $\sum_{j=1}^{k} p_j q_j$ terimi, testteki maddelerin varyansları toplamına karşılık gelmektedir. Test puanlarına ilişkin varyans ise madde varyansları ile maddeler arasındaki kovaryansların toplamıdır. Maddeler arasında ilişki yoksa daha teknik bir dille ifade edecek olursak maddeler arasında kovaryans bulunmuyorsa test puanlarının varyansı madde varyanslarının toplamına eşit, yani $S^2 = \sum_{j=1}^{k} p_j q_j$ olur. Bu durumda Eşitlik 2'nin sağındaki ikinci parantezin değeri ve beraberinde KR-20 katsayısı 0 çıkar. Maddeler birbiriyle ne kadar ilişkiliyse maddeler arasındaki kovaryans o kadar fazla olur. Buna paralel olarak $\sum_{j=1}^{k} p_j q_j / S^2$ oranı küçülür ve güvenirlik katsayısı 1'e yaklaşır (Kaplan & Saccuzzo, 2018).

KR-21 madde güçlük indekslerinin eşitliği varsayımına dayandığından Eşitlik 3, madde güçlükleriyle ilgili bir terim içermemektedir**.** Fakat uygulamada bir testteki tüm maddelerin güçlük indekslerinin eşit olması neredeyse imkânsızdır. Bu nedenle KR-21 katsayısını hesaplamak çoğu zaman yanlış bir tercihtir. Unutmamak gerekir ki KR-21, güvenirlik hesaplamalarının kâğıt-kalem ile yapıldığı dönemlerde KR-20'ye kıyasla daha kolay hesaplanabilen bir katsayı olarak ileri sürülmüştür ve sadece iç tutarlılığın alt sınırına ilişkin bir kestirim sunmaktadır (Crocker & Algina, 2008). Artık matematiksel olarak en karmaşık güvenirlik hesaplamaları bile bilgisayarlar aracılığıyla kolayca yapılabildiğinden KR-20 yerine KR-21'i hesaplamayı haklı kılabilecek koşullar ortadan kalkmıştır.

3.1.3.3. Cronbach Alfa ile McDonald Omega

Bu iki katsayı, KR-20 ile KR-21'den farklı olarak, maddelerin ikili puanlandığı ölçme araçlarında kullanılabildiği gibi çoklu puanlandığı (örneğin; Likert tipi ölçekler, dereceleme ölçeği veya rubrikler aracılığıyla puanlanan açık uçlu maddeler) ölçme araçlarında da kullanılabilmektedir. Bu anlamda Cronbach alfa ile McDonald omega katsayıları, KR-20 ve KR-21'in puanlamadaki kategorisi sayısının ikiden fazla olduğu durumlara genelleştirilmiş hali olarak nitelendirilmektedir (İlhan ve ark., 2024).

Cronbach alfanın McDonald omegadan ayrıldığı nokta ise daha katı varsayımlara (tau denkliği, maddelerin hata varyanslarının ilişkisizliği) sahip olmasıdır. McDonald omega (ω) ile Cronbach alfa (α) arasında "ω ≥ α" şeklinde bir matematiksel ilişki bulunmaktadır (McDonald, 1999). Bu iki iç tutarlılık katsayısının hesaplanmasında kullanılan formüller aşağıda sunulmuştur.

> Maddelerin iki kategorili (1-0) olarak puanlandığı bir test için Cronbach alfa hesaplanacak olursa elde edilecek katsayı bu teste ait KR-20 iç tutarlılık katsayısına eşit çıkar.

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k} S_i^2}{S^2}\right)$$

k = Ölçme aracındaki madde sayısı

S_i^2 = i. maddeye ait varyans

S^2 = Ölçme aracından alınan puanlara ilişkin varyans

*Eşitlik 4**

Tüm maddeler ortalaması 0, standart sapması 1 olacak biçimde standartlaştırılırsa α aşağıdaki formülle de hesaplanabilir.

$$\alpha = \frac{k\bar{r}}{1+(k-1)(\bar{r})}$$

k = Ölçme aracındaki madde sayısı

$\bar{r}$ = maddeler arası olası bütün ikili kombinasyonlar için hesaplanan korelasyon değerlerinin ortalaması

*Eşitlik 5**

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \left(\sum_{i=1}^{k} (\Psi_i)\right)}$$

λi = i. maddeye ait standartlaştırılmış faktör yükü

Ψ_i = i. maddeye ait hata varyansı

*Eşitlik 6***

* Hogan (2019), ** McDonald (1999)

Eşitlik 4'e yakından bakarsak Cronbach alfa katsayının KR-20'nin genel bir formu olduğunu kolayca fark edebiliriz. KR-20 ile Cronbach alfanın formülü arasındaki tek fark, madde varyansının hesaplanma biçimidir. İki kategorili (1-0) puanlanan maddelerde, madde varyansı maddenin güçlük indeksi üzerinden hesaplanır. Buna karşın çoklu puanlanan maddelerde madde varyansını hesaplamak için daha başka bir yol takip edilir ve bu yol iki kategorili puanlanan maddelere de uygulanabilmektedir. Eşitlik 5 incelendiğinde ise Cronbach alfanın hem maddeler arası ilişkilerden hem de ölçme aracındaki madde sayısından etkilendiği açık şekilde görülmektedir. Son olarak, McDonald omeganın formülü bu noktaya kadar değinmediğimiz faktör yükü kavramını içermektedir. Şu aşamada, McDonald omeganın bir iç tutarlılık katsayısı olduğunu ve Cronbach alfanın görece yanlı iç tutarlılık katsayıları ürettiği durumlarda da kullanılabildiğini bilmemiz yeterlidir. Faktör yükü kavramı bu bölümün kapsamı dışında kalmakta, konu hakkında bilgi edinmek isteyen okuyucuların faktör analizi kitaplarına ve/veya faktör analizini bir bölüm olarak içeren kaynaklara başvurmaları önerilmektedir.

Örnek olması amacıyla aşağıda biri dört, diğer sekiz maddelik iki ölçme aracı için hesaplanan maddeler arası korelasyon değerleri ile Cronbach alfa ve McDonald omega katsayıları sunulmuştur. Tablolardaki "madde-kalan korelasyonu (item-rest corelation)" ifadesiyle herhangi bir maddeye ait puanlar ile bu madde hariç ölçme aracındaki diğer maddelerden alınan toplam puan arasındaki ilişki kastedilmektedir. Tablo 6 ve Tablo 7 maddeler arası korelasyonlar açısından karşılaştırıldığında dört maddelik ölçme aracındaki değerlerin daha yüksek olduğu göze çarpmaktadır. Buna rağmen sekiz

maddelik ölçme aracı için hesaplanan Cronbach alfa ve McDonald omega katsayıları daha yüksek çıkmıştır. Çünkü Cronbach alfa ile McDonald omega maddeler arası ilişkilere dayanır ama aynı zamanda madde sayısının bir fonksiyonudur. Ölçme aracına benzer özellikte maddeler eklendikçe iç tutarlılık katsayıları yükselme eğilimi gösterir, bu eğilim madde sayısı belli bir eşiğe ulaşıncaya dek devam eder.

Tablo 6

Dört Maddelik Veri Seti için Maddeler Arası Korelasyonlar ile Cronbach Alfa ve Mcdonald Omega Katsayıları

Maddeler	M1	M2	M3	M4	Madde-kalan korelasyonu
M1	–				0,594
M2	0,457	–			0,580
M3	0,513	0,504	–		0,626
M4	0,474	0,457	0,488	–	0,581
Cronbach alfa: 0,786 [95% GA (0,734; 0,831)]; McDonald omega: 0,787 [95% GA (0,739; 0,836)]					

Tablo 7

Sekiz Maddelik Veri Seti için Maddeler Arası Korelasyonlar ile Cronbach Alfa ve Mcdonald Omega Katsayıları

Maddeler	M1	M2	M3	M4	M5	M6	M7	M8	Madde-kalan korelasyonu
M1	–								0,500
M2	0,308	–							0,512
M3	0,308	0,313	–						0,503
M4	0,358	0,325	0,308	–					0,488
M5	0,348	0,358	0,358	0,300	–				0,520
M6	0,328	0,330	0,370	0,308	0,330	–			0,503
M7	0,328	0,340	0,325	0,305	0,348	0,318	–		0,511
M8	0,318	0,368	0,323	0,340	0,333	0,320	0,375	–	0,521
Cronbach alfa: 0,799 [95% GA (0,753; 0,838)]; McDonald omega: 0,799 [95% GA (0,757; 0,841)]									



- Hız testlerinin güvenirliğini kestirirken iç tutarlılığa dayalı yöntemler kullanılmamalıdır (Kurpis & Stafford, 2006).
- KR-20, KR-21, Cronbach alfa ve McDonald omega tek boyutluluğa dayalı güvenirlik katsayılarıdır. Fakat bu katsayıların yüksek çıkması tek boyutluluğa kanıt oluşturmaz.
- Ölçme aracının iki yarısı aynı uzunlukta ise Cronbach alfa katsayısı, iki yarı yönteminden elde edilebilecek olası tüm güvenirlik değerlerin ortalamasına eşit olur (Cronbach, 1951; Warrens, 2015). Daha açık ifade etmek gerekirse bir ölçme aracı çok sayıda farklı şekilde iki yarıya ayrılabilir. Oluşan yarıların her biri ayrı bir güvenirlik değeri üretir ve bu güvenirlik değerleri yaklaşık olarak normal dağılım sergiler. İlgili güvenirlik değerlerinin ortalaması alınırsa bulunan değer aynı veri setine ait Cronbach alfa katsayısına eşit olur. Bundan dolayı, iki yarı güvenirliğini Cronbach alfa (ya da McDonald omega) ile birlikte sunmak yersizdir. Cronbach alfa verildiyse iki yarı güvenirliğini sunmaya gerek yoktur.
- Alanyazın incelendiğinde Cronbach alfanın McDonald omegaya kıyasla çok daha fazla kullanıldığı görülmektedir. Ancak gerektirdiği varsayımların karşılanması oldukça zayıf bir ihtimalken ve McDonald omega gibi varsayım açısından daha esnek bir alternatifi mevcutken Cronbach alfayı kullanma konusunda ısrarcı davranmak mantıklı değildir (Hayes & Coutts, 2020). Özellikle günümüzde istatistiksel yazılımların hızla geliştiği ve hemen her yazılımın güvenirlik analizi sekmesine Cronbach alfayla birlikte McDonald omeganın da dâhil edildiği dikkate alındığında hesaplama kolaylığı Cronbach alfa için tercih nedeni olmaktan çıkmalıdır.

3.1.3.4. Tabakalı Alfa ve Hiyerarşik Omega

Kimi ölçme araçlarında birden fazla alt boyut bulunur ama bu boyutlar bir araya gelip daha genel bir yapıyı oluşturur. Ölçme aracının alt boyutlar içerdiği ve bu boyutların daha üst bir yapıyı meydana getirdiği durumlarda toplam puanlara ait iç tutarlılığı kestirirken tabakalı alfaya ya da hiyerarşik omegaya başvurulur (Kalkbrenner, 2023; Tan, 2009). Örneğin, bir matematik testinde cebir ve geometri şeklinde iki boyut yer alabilir. Bu durumda cebir ve geometri boyutlarının iç tutarlılığı Cronbach alfa ya da McDonald omega ile hesaplanır. Ancak cebir ve geometri boyutlarının bileşiminden oluşan matematik testine ait iç tutarlılığı belirlerken Cronbach alfa ya da McDonald omegayı kullanmak doğru olmaz, yanlı güvenirlik kestirimleri ile neticelenir. Burada izlenmesi gereken yol toplam puanlara ait iç tutarlılığı tabakalı alfa veya hiyerarşik omega ile hesaplamaktır.

Yukarıda sıralanan iç tutarlılık katsayılarının tamamı, katılımcıların ölçme aracındaki maddelere verdikleri yanıtlar arasındaki tutarlılığı test eder. Maddelerin ölçtükleri özelliğin homojen bir yapı göstermesi, yanıtlar arasında yüksek bir tutarlılık olmasını sağlar ve iç tutarlılık güvenirliğini yükseltir. Maddelere verilen yanıtlar arasında tutarsızlığa yol açacak faktörler ise iç tutarlılığı düşürür. Bu faktörlerden ilki maddelerin birbirinden farklı özellikleri ölçmesi, yani heterojen olmasıdır. Hedef yapıyla ilişkisi düşük maddelerin yanı sıra birden fazla cevabı olan, kafa karıştırıcı ve/veya cevaplayıcıların aşina olmadığı ifadeler içeren maddeler de iç tutarlılığı azaltır. Mesela olumsuz maddeler, bireylerde kafa karışıklığına yol açarak iç tutarlılığı düşürücü etki yaratabilmektedir.

Alanyazında iç tutarlılık için en sık referans alınan ölçütün 0,70 olduğu bilinmektedir. Ancak her koşulda bu ölçüte bağlı kalmak uygun değildir. İç tutarlılığı etkileyen önemli bir değişken ölçme aracındaki madde sayısı olduğundan az sayıda madde içeren ölçme araçlarında 0,70'in altındaki değerler de makul karşılanabilir. Öyle ki 3 ila 5 maddeden oluşan bir ölçme aracında iç tutarlılığın 0,60'ı hatta 0,50'yi aşması yeterli olabilir (İlhan & Çetin, 2021). Bir işe ya da eğitim kurumuna seçilecek kişileri belirlemek amacıyla yapılan bir sınavda ise 0,70 ölçütü düşük kalacaktır. Bireyler için hayati sonuçları olabilen, dolayısıyla çok daha yüksek bir iç tutarlılığa ulaşmayı gerektiren yüksek riskli testlerde 0,90 civarı değerler ölçüt kabul edilmelidir. İç tutarlılık katsayısı 0,95'i aştığında da temkinli olmak gerekir. Çünkü iç tutarlığın bu denli yüksek çıkması ölçme aracındaki artık maddelerden kaynaklanıyor olabilir. Bu bağlamda, 1'e çok yakın iç tutarlılık katsayıları elde edildiğinde ölçme aracında tamamıyla aynı içeriği ölçen ve çıkarıldığında hedef kapsamı daraltmayacak maddelerin olup olmadığıyla ilgili kontroller yapılmalı, varsa tekrar niteliğindeki maddeler ölçme aracından çıkarılmalıdır (İlhan ve ark., 2024).

3.1.4. Puanlayıcı Güvenirliği

Açık uçlu maddelerden oluşan testlerin ya da performans görevlerinin kullanıldığı ölçmelerde puanlayıcı kaynaklı faktörler de ölçme sonuçlarında hataya yol açabilmektedir. Dolayısıyla puanlayıcı yargılarına dayanan bir ölçme işlemi yürüttüğümüzde puanlayıcı güvenirliğini de hesaplamamız gerekir. Puanlayıcı güvenirliği objektiflik anlamında bir güvenirlik ölçüsü olup *puanlayıcılar arası* ve *puanlayıcı içi* güvenirlik şeklinde iki başlıkta incelenir.

Puanlayıcılar arası güvenirlikte aynı özelliği/performansı puanlayan iki ya da daha fazla puanlayıcı arasındaki uyum test edilirken puanlayıcı içi güvenirlikte aynı puanlayıcının aynı özellik/performans için iki farklı zamanda yaptığı puanlamalar arasındaki uyuma bakılır. Burada bahsi geçen uyumun mutlak uyum ve göreli uyum (tutarlılık) olmak üzere iki yönü vardır. Aynı öğrenci grubunu puanlayan iki puanlayıcı düşünelim. İki puanlayıcı öğrencilere birbirinden farklı notlar vermiş olmasına rağmen bu notlara dayalı olarak

öğrenciler arasında sıralama yapıldığında benzer bir tablo ortaya çıkıyorsa iki puanlayıcı arasındaki mutlak uyum düşüktür ama aralarında yüksek bir göreli uyum var demektir. Tablo 8'deki veriler bu durumu örneklendirmektedir.

Tablo 8

Puanlayıcılar Arası Güvenirliğe ilişkin Örnek Veri Seti

Öğrenciler	Puanlayıcı1	Puanlayıcı2	Puanlayıcı3
Ö1	10	10	10
Ö2	6	4	3
Ö3	3	3	1
Ö4	7	7	4
Ö5	8	8	6

Tabloya bakıldığında birinci ve ikinci puanlayıcının beş öğrenciden dördüne aynı puanı verdiği görülmektedir. Yani bu iki puanlayıcı arasında yüksek bir mutlak uyum bulunmaktadır. Üçüncü puanlayıcı ise birinci ve ikinci puanlayıcıdan farklı puanlar vermiştir. Buna göre üçüncü puanlayıcı ile diğer iki puanlayıcı arasındaki mutlak uyumun düşük olduğu anlaşılmaktadır. Bununla birlikte, öğrenciler en başarılıdan en başarısıza doğru sıralandığında her üç puanlayıcı için de Ö1, Ö5, Ö4, Ö2 ve Ö3 şeklinde bir sonuç ortaya çıkmaktadır. Bu da üçüncü puanlayıcı ile birinci ve ikinci puanlayıcı arasında yüksek bir göreli uyum (tutarlılık) olduğunu yansıtmaktadır. Eğer bir sıralama sınavı yapıldıysa ya da bağıl bir değerlendirme uygulanacaksa puanlayıcılar arasında göreli uyumunun sağlanması yeterli kabul edilebilir. Ancak bir yeterlik sınavı yapıldıysa, standart belirleme çalışması yürütülecekse ya da mutlak değerlendirme uygulanacaksa puanlayıcılar arasında göreli uyumun sağlanmış olması yetmez mutlak uyuma odaklanmak gerekir.

Puanlayıcı güvenirliğini incelerken kullanılabilecek farklı teknikler vardır. Bunun için en sık başvurulan tekniklerden biri, iki puanlayıcının aynı kişilere (ya da tek bir puanlayıcının aynı kişilere iki farklı zamanda) verdiği puanlar arasındaki korelasyonu hesaplamaktır. Bu hesaplamada aynı sınıfa ait ölçümler arasındaki ilişki inceleneceğinden Pearson momentler çarpımı ya da Spearman sıra farkları korelasyonu yerine sınıf-içi korelasyon katsayısı tercih edilmelidir (Mathews, 2010). Uyum yüzdesi, Kappa istatistikleri, Kendall uyuşum değeri ile Krippendorf alfa katsayısı da puanlayıcı güvenirliğini test ederken başvurulabilecek teknikler arasında yer almaktadır.

3.2. Güvenirliğin Raporlanması

Güvenirlik ölçme aracına ait sabit bir özellik değildir, gruba bağımlıdır. Aynı ölçme aracı bir grupta güvenirliği yüksek ölçümler verirken bir başka grupta düşük güvenirlikte ölçümler üretebilir. Bu da her ölçme çalışmasının ardından güvenirliğin raporlanmasını zorunlu kılar. Diğer bir deyişle güvenirliği ölçme aracının yalnızca ilk uygulamasında hesaplamak yeterli değildir. Ölçme aracının her kullanımında ilgili uygulamadaki

güvenirliği ortaya koymak gerekir. Ölçümlere ait güvenirliğin raporlanmaması alanyazında *güvenirlik tümevarımı (reliability induction)* olarak adlandırılmaktadır (Vacha-Haase ve ark., 2000). Güvenirlik tümevarımı ölçme aracını uygulayıp elde edilen sonuçları raporlarken güvenirlikten hiç bahsedilmemesi ya da ölçme aracının önceki uygulamalarında ulaşılan güvenirlik değerinin sunulması ama eldeki verilere ait güvenirlik katsayısının bildirilmemesi şeklinde gerçekleşebilmektedir (Sánchez-Meca ve ark., 2021). Hangi biçimde gerçekleşmiş olursa olsun güvenirlik tümevarımı kaçınılması gereken hatalı bir yaklaşımdır.

Güvenirliği raporlarken dikkat edilmesi gereken ikinci bir husus, güvenirlik kestiriminde kullanılan yöntemin ve yazılımın ayrıntılı bir şekilde sunulmasıdır. Örneğin, "Ölçümlere ait güvenirlik katsayısı 0,75 bulunmuştur" demek yerine "Ölçümlere ait test-tekrar test güvenirliği 0,75 bulunmuştur" ifadesi tercih edilmelidir. Dahası test-tekrar test güvenirliğine ait bu katsayısının hangi korelasyon tekniğiyle elde edildiği de açıklanmalıdır. Benzer şekilde iç tutarlılık güvenirliği hesaplandıysa bunun hangi iç tutarlılık katsayısı (Cronbach alfa, McDonald omega vb.) olduğu söylenmelidir. Güvenirlik analizinde kullanılan yazılım da sadece adıyla değil sürümüyle birlikte (örneğin, ITEMAN 4.4) belirtilmelidir.

Hesaplanan güvenirlik katsayısını nokta kestirimi şeklinde sunmayıp güven aralığı ile birlikte vermek güvenirliği raporlarken gözetilmesi gereken ilkelerden bir diğeridir. Ölçme sonuçlarına karışan tesadüfi hata miktarı tam olarak ortaya konulamadığından güvenirliğin değerini kesin olarak bilmek olanaksızdır. Ancak eldeki verilerden hareketle güvenirlik katsayısı hakkında bir tahmin yapılabilmektedir (Köse, 2012). Güvenirliği raporlarken güvenirlik kestirimine ilişkin bu süreç hesaba katılmalıdır. Elde edilen katsayı herhangi bir hata payı içermiyormuşçasına kesin bir değer olarak verilmemeli, aşağıdaki örnekte görüldüğü gibi yanında güven aralığı yer alacak tarzda rapor edilmelidir.

"Ölçümlere ait Cronbach alfa iç tutarlılık katsayısı 0,745 olarak bulunmuştur."	
"Ölçümlere ait Cronbach alfa iç tutarlılık katsayısı 0,745 olarak bulunmuştur [%95 GA (0,710; 0,780)]."	

3.3. Güvenirliği Artırmak için Alınabilecek Önlemler

Daha önce de belirtildiği gibi güvenirlik ile tesadüfi hata arasında negatif bir korelasyon vardır. Dolayısıyla güvenirliği artırmanın yolu ölçme işlemine karışabilecek tesadüfi hataları minimize etmekten geçer. Ölçme sonuçlarına ölçme aracının kendisinden, uygulama koşullarından, ölçmeyi yapan kişiden veya ölçülen gruptan kaynaklanan

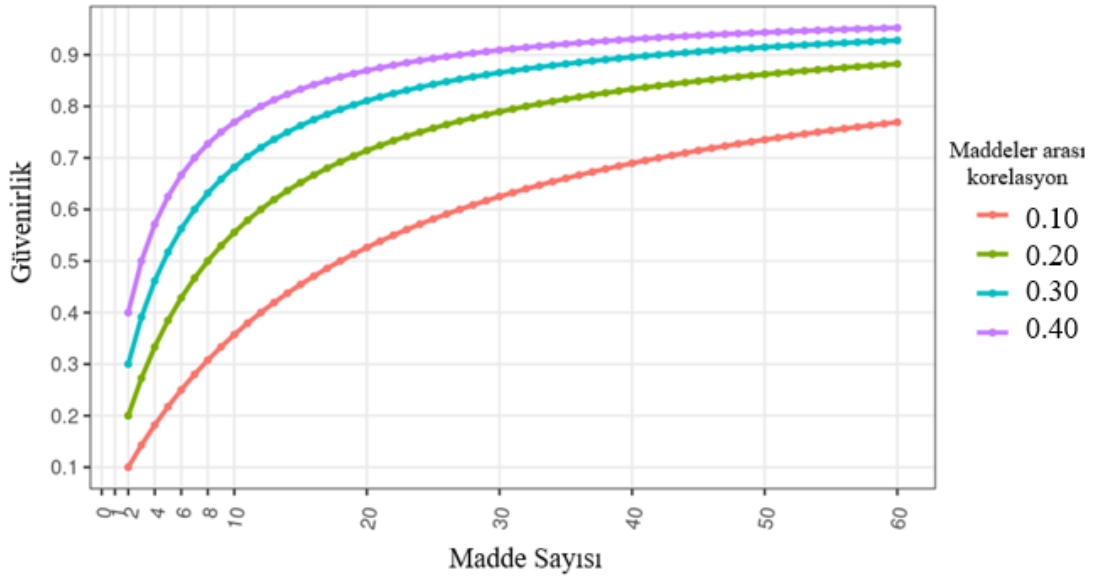
hatalar karışabildiğinden güvenirliği artırmak için sıralanan faktörlerin her biriyle ilgili gözetilmesi gereken belli başlı hususlar vardır. Güvenirliği artırmak için dikkat edilmesi gereken noktalar aşağıdaki gibi sıralanabilir:

- ⇨ Ölçme aracının başında; ölçme aracının uygulama süresi, puanlama şekli ve cevaplamada dikkat edilmesi gereken hususlar gibi temel bilgileri içeren açık bir yönerge sunulmalıdır.
- ⇨ Ölçme aracındaki maddeler net olmalı, tüm cevaplayıcılar tarafından aynı şekilde anlaşılmalı, farklı kişiler tarafından farklı yorumlanabilecek ifadeler içermemelidir. Bunu sağlayabilmek için maddenin anlaşılırlığını negatif yönde etkileyecek olumsuz ifadelerden mümkün olduğunca kaçınılmalı, özellikle de çift olumsuz yargılardan uzak durulmalıdır. Ayrıca maddenin cevap hakkında ipucu verebilecek herhangi bir ifade içermediğinden emin olunmalıdır.
- ⇨ Ölçme aracında yeterli sayıda maddeye yer verilmelidir. Madde sayısı az olduğunda duyarlı ölçmeler yapılamayacak, dolayısıyla katılımcılar birbirinden yeterince iyi ayırt edilemeyecek ve güvenirlik düşük çıkacaktır. Şöyle düşünelim: Elimizdeki testte tek bir madde olduğunda testin uygulandığı gruptaki kişiler "maddeyi doğru cevaplayanlar" ile "yanlış cevaplayanlar/boş bırakanlar" olarak iki gruba ayrılabilecektir. Madde sayısını ikiye çıkardığımızda testin uygulandığı kişiler "her iki maddeyi de doğru cevaplayanlar", "maddelerden yalnızca birini doğru cevaplayanlar" ve "iki maddeyi de yanlış yanıtlayanlar/boş bırakanlar" olmak üzere üç gruba ayrılabilecektir. Eklenen yeni maddelerin testte var olan maddeler ile arasında kabul edilebilir düzeyde korelasyonlar bulunduğu varsayımı altında, madde sayısı artıkça bir noktaya kadar güvenirlik yükselecektir (bkz. Şekil 1).

> "*Ölçme aracında kaç madde olmalıdır?*" sorusuna verilebilecek sihirli bir yanıt yoktur (Aybek, 2021). İdeal madde sayısı ölçülmesi hedeflenen kapsama, ölçme sonuçlarının hangi amaçla kullanılacağına, cevaplayıcıların gelişim özelliklerine vb. faktörlere bağlı olarak değişir.

Şekil 1

Maddeler Arası Belli Korelasyon Değerleri için Madde Sayısı ile Güvenirlik Arasındaki İlişki (Evans, 2021)



Şekil 1'e bakıldığında madde sayısındaki artışa bağlı olarak güvenirlik ilk başta hızlı bir şekilde yükselmekte madde sayısı artmaya devam ettikçe güvenirlikte gözlenen artış giderek daha düşük olmaktadır. Yani eşdeğer yarılar yöntemi başlığında da ifade edildiği üzere belli bir noktadan sonra madde sayısındaki artış güvenirlikte önemli bir yükselme sağlamamakta, dahası madde sayısının gereğinden fazla artırılması cevaplayıcıların yorulmasına/sıkılmasına neden olarak güvenirliği ters yönde etkilemektedir.

- ⇨ Açık uçlu maddeler ya da performans görevleri kullanılarak yapılan ölçmelerde, puanlamada kullanılan anahtarın cevaplardaki olası farklılıkları açığa çıkarılabilecek kadar kategori içerdiğinden emin olunmalıdır. Söz gelimi açık uçlu bir matematik sorusuna verilen cevapları 1-0 şeklinde puanladığımızda öğrencileri sadece iki kategoriye ayırmış oluruz. Oysaki bazı öğrenciler doğru çözüm yolunu kullanmış, sadece işlem hataları sebebiyle yanlış bir sonuç bulmuş olabilir. Ya da uygun şekilde başlangıç yapmış ama çözümün devamını getirememiş öğrenciler bulunabilir. Bu koşullar altında ilgili maddeyi 1 (*tam doğru*), 0,75 (*neredeyse tamamen doğru*), 0,50 (*yarı yarıya doğru*), 0,25 (*kısmen doğru*) ve 0 (*yanlış/boş*) şeklinde puanlamak öğrencileri daha iyi ayırt etmemizi sağlar ve daha yüksek bir güvenirlikle neticelenir.
- ⇨ Ölçme aracındaki maddelerin güçlük düzeyi uygulama yapılacak grubun yetenek düzeyine uygun olmalıdır. Testteki maddeler grup için çok kolay olduğunda gruptaki kişilerin tamamı maddelerin hepsini doğru yanıtlar. Böyle bir durumda katılımcılar arasındaki farklılıklar ortaya konulamaz, haliyle de güvenirlik düşük

çıkar. Aynı şekilde, testteki maddelerin tamamının grup için çok zor olması halinde katılımcılar testteki hiçbir maddeyi doğru yanıtlayamaz. Buna bağlı olarak, katılımcılar arasındaki farklılıklar ortaya çıkarılamaz ve yüksek bir güvenirlik değerine ulaşılamaz. Güvenirliği artırabilmek için katılımcıların yanıtlarında değişkenlik yaratacak maddeler teste dâhil edilmelidir. Bunun için de madde güçlükleri, ölçülen grubun yetenek düzeyleri açısından gösterdiği dağılıma göre ayarlanmalıdır. Buradan şu sonuca da varmak mümkündür: Aynı ölçme aracı heterojen bir grupta yüksek güvenirlikte ölçümler üretirken daha homojen bir grupta düşük güvenirlikte ölçümler verebilir.

- ⇨ Açık uçlu maddeler ya da performans görevleri gibi doğası itibariyle sübjektif nitelik taşıyan ölçme araçları kullanıldığında puanlama sürecini daha nesnel hale getirecek uygulamalara yer verilmelidir. Bu doğrultuda puanlamaların puanlama anahtarına dayalı olarak yapılması sağlanmalı, eğer mümkünse puanlama işlemi öncesinde puanlayıcılara eğitim verilmeli ve puanlamalar birden fazla puanlayıcı tarafından gerçekleştirilmelidir.
- ⇨ Şans başarısını (maddenin kör tahminle doğru yanıtlanma olasılığını) azaltacak uygulamalara yer verilmelidir. Örneğin; çoktan seçmeli maddelerde düzeltme formülü uygulanabilir, yani seçenek sayısının bir eksiği kadar yanlış cevabın bir doğruyu götürdüğü bir puanlama yapılabilir. Yine madde çoktan seçmeli ise çeldiricilerin (doğru cevap dışındaki seçeneklerin) iyi yazılması, diğer bir deyişle maddenin ölçtüğü özelliğe sahip olmayan ya da düşük düzeyde sahip olan öğrencilerin bile kolayca eleyeceği çeldiricilere yer verilmemesi şans başarısını düşürür. Böylece tesadüfi hata olasılığı azalır ve güvenirlik yükselir.
- ⇨ Kopya çekilmesini engelleyecek önlemler alınmalıdır. Alternatif kitapçıkların oluşturulması, sınav salonundaki gözetmen sayısının artırılması ve sınavın kamerayla kayıt altına alınması yüz yüze yapılan sınavlarda kopya ihtimalini azaltacak düzenlemelerdir. Elektronik ortamda yürütülen sınavlarda ise başka sekmeler açmaya, kopyala/yapıştır işlemine ya da ekran görüntüsü almaya mani olacak güvenli tarayıcı sistemlerinin devreye sokulması ve yüz tanıma uygulamalarının kullanılması bu amaçla hayata geçirilebilecek uygulamalardır.
- ⇨ Sınav ortamı iyi ayarlanmalıdır. Ortam çok sıcak, çok soğuk ya da gürültülü olmamalı, yeterli havalandırma ve ışıklandırma sağlanmalı, sınava girdiği sıranın/masanın boyutu öğrenciye uygun olmalıdır. Elektronik sınav yapılıyorsa sistemin uygun hızda ve kesintisiz çalışmasını sağlayabilecek bir ağ alt yapısı ile sunucu kullanılmalıdır.
- ⇨ Ölçme aracını cevaplamaları için katılımcılara tanınacak süre iyi ayarlanmalıdır. Gereğinden az ya da fazla süre verilmesinin güvenirliği düşüreceği unutulmamalıdır. Uygun süreye karar verirken uygulama yapılacak grup göz önünde bulundurulmalı, gruptaki kişilerin ölçülen özellik bakımından ne düzeyde olduğu ve ne kadar heterojen bir yapı gösterdiği hesaba katılmalıdır.

- ⇨ Ölçme aracının uygulandığı kişilere ait değişkenlerin (kaygı, motivasyon, heyecan, hastalık, açlık, susuzluk vb.) güvenirliği önemli ölçüde etkileyeceğinin farkında olunmalı (Boztunç Öztürk & Şahin, 2021) ve yapılabiliyorsa bu değişkenlerle ilgili önlemler alınmalıdır. Örneğin, testten aldığı puanın birey için önemli olduğu sınavlarda kişi zaten yüksek bir motivasyonla maddeleri cevaplar. Öte yandan kimi ölçme araçlarında kaç puan aldığı birey için bir anlam ifade etmez. Bu durumda bireyin motivasyonu düşük olabilir ve rastgele yanıtlanmış ölçme araçlarıyla karşılaşılabilir. Böyle yanıtlar güvenirliği olumsuz etkileyeceğinden katılımcıların motivasyonunu artıracak yollar aranmalıdır. Katılımcılara ölçme aracının neden uygulandığına dair bilgi verilmesi, güvenilir sonuçlara ulaşmanın uygulayıcı için taşıdığı öneme dair ikna edici açıklamalar yapılması ve mümkünse cevaplayıcılara küçük hediyeler verilmesi bireylerin motivasyonlarını artırıp gerçeğe yakın yanıtlar alınmasını sağlar. Bu da güvenirliği yükseltir.

3.4. Ölçmenin Standart Hatası

Güvenilirlik katsayısı ölçümlerin psikometrik özellikleri hakkında son derece önemli bilgiler sunmakla birlikte, ölçme hatasının ne boyutta olduğunu doğrudan ortaya koyamaz. Şöyle ki güvenirlik katsayısı [0-1] aralığında uzanan bir metrikte ifade edilir. Dolayısıyla bir testi bir grup bireye uyguladığımızda güvenirliğin, ölçme sonuçlarında gözlenen ortalama hatanın ne büyüklükte olduğuna dair söyledikleri test puanlarıyla aynı birimde olmaz (Furr & Bacharach, 2008). Oysa bireyin gözlenen puanından yola çıkarak gerçek puanı hakkında tahminde bulanabilmek için ölçme sonuçlarına karışan hatayı test puanlarıyla aynı birimde ifade etmek gerekir. Bu noktada ölçmenin standart hatası kavramı karşımıza çıkar. Ölçmenin standart hatası test puanlarının standart sapması ile güvenirliğine dayalı olarak hesaplanır ve hesaplamada Eşitlik 7'deki formül kullanılır.

$$S_E = (S_x) \times \sqrt{1 - r}$$

S_E = Ölçmenin standart hatası
S_x = Test puanlarının standart sapması
r = Güvenirlik katsayısı

Eşitlik 7

Ölçmenin standart hatası güvenirlik ile ters bir ilişki içerisindedir. Daha yüksek güvenirlik katsayıları ölçmenin standart hatasına ilişkin daha düşük değerler ile sonuçlanır. Ölçmenin standart hatasından yararlanarak bir kişinin ölçülen özelliğe ilişkin gerçek puanının hangi aralıkta yer alacağı belirli bir olasılık dâhilinde tespit edilebilmektedir. Hatırlarsak ölçme sonucunda elde ettiğimiz puanı gözlenen puan olarak nitelendirmiş, ölçme işlemi ne kadar dikkatli yapılırsa yapılsın gözlenen puanlarda bir miktar hata bulunmasının kaçınılmaz olduğunu vurgulamıştık. Bu sebeple bir testi bir bireye tekrar tekrar uyguladığımızda araya öğrenme, unutma gibi gerçek puanlarda farklılık yaratacak başka değişkenler girmese dahi tesadüfi hataların etkisiyle bireyin testten alacağı puanlar bir uygulamadan diğerine değişecektir. Bu değişimin

hangi aralıkta olacağını kestirirken ölçmenin standart hatasına başvurulmaktadır. Ölçmenin standart hatası gözlenen puandan çıkarıldığında gerçek puana ilişkin alt sınır, gözlenen puana eklendiğinde ise gerçek puana ilişkin üst sınır tahmin edilebilmektedir. Bu alt ve üst sınırın oluşturduğu aralık ise güven aralığı adıyla anılmaktadır. Ölçme sonuçları raporlanırken genellikle %68, %95 ya da %99 olasılık güven aralıkları sunulmaktadır. Gözlenen puan, diğer bir deyişle bireyin testten aldığı puan *X* ve ölçmenin standart hatası S_E olmak üzere %68, %95 ve %99 olasılıklı güven aralıkları aşağıdaki gibi hesaplanmaktadır.

	Güven aralığının alt sınırı	**Güven aralığının üst sınırı**
% 68 olasılıklı güven aralığı	$X - S_E$	$X + S_E$
% 95 olasılıklı güven aralığı	$X - 2S_E$	$X + 2S_E$
% 99 olasılıklı güven aralığı	$X - 3S_E$	$X + 3S_E$

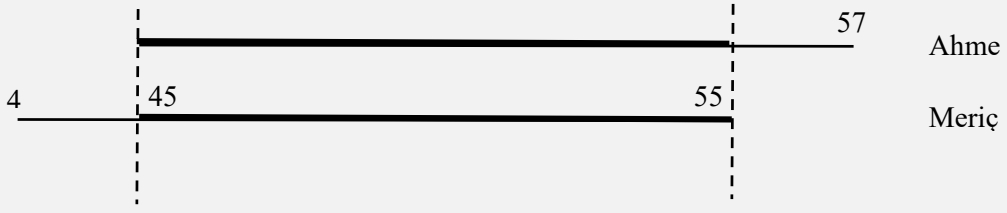
Standart normal dağılım eğrisi altındaki alanları gösteren Z tablosundan yararlanarak %68, %95 ve %99 dışında daha başka olasılıklar için de güven aralıklarını hesaplamak mümkündür.

Anlaşılacağı üzere ölçme sonucunda bireyin gerçek puanına ulaşmak mümkün olmasa bile ölçmenin standart hatası sayesinde bireyin gerçek puanının hangi aralıkta değişeceği tahmin edilebilmektedir. Bir örnek üzerinde sıralanan kavramları biraz daha somutlaştıralım. Diyelim ki Enes isimli bir öğrenci uyguladığımız testten 72 puan aldı. Test puanlarının standart sapması 5 ve güvenirlik katsayısı 0,84'tü. Buna göre ilgili ölçme işlemine ait standart hata 2'dir $[(5)\times\sqrt{1-(0{,}84)}]$. Dolayısıyla Enes'in gerçek puanı %68 olasılıkla 72±2, %95 olasılıkla 72±4 ve %99 olasılıkla 72±6 aralığındadır. Yani Enes bu sınava aynı koşullar altında 100 defa girerse

- 68'inde 70–74 arasında, 32'sinde ise bu aralığın dışında bir puan alır,
- 95'inde 68–76 aralığında bir puan elde eder ama diğer 5 sınavda puanı bu aralığın dışında kalır,
- 99'unda 66–78 aralığında bir puan alır ama kalan 1 sınavda puanı bu aralığın dışında olur.

Ölçmenin standart hatası küçükse aynı test bireye tekrarlı olarak uygulandığında elde edilecek gözlenen puanlar bireyin gerçek puanlarına yakın, dar bir aralıkta kümelenir. Öte yandan ölçmenin standart hatasına ilişkin yüksek değerler, aynı testin bireye tekrarlı olarak uygulanması halinde bireyin gözlenen puanlarında belirgin dalgalanmalar olacağını bildirir. Böyle bir durumda bireyin gerçek puanını tahmin etmek zorlaşır ve standart hatanın büyüklüğü ölçüsünde güven aralığı genişler.

Geçme notunun 50 ve ölçmenin standart hatasının 3'e karşılık geldiği bir sınavda Meriç 49, Ahmet 51 puan almış olsun. Ölçmenin standart hatasını dikkate almazsak Ahmet dersi geçecek Meriç ise dersten kalacaktır. Hâlbuki gözlenen puanlara ilişkin güven aralıklarına göre %95 olasılıkla Ahmet'in gerçek puanı 45-57 ve Meriç'in gerçek puanı 43-55 arasında yer alacaktır. Aşağıdaki şekilde net biçimde görüldüğü gibi iki öğrencinin gerçek puanını içeren aralıklar büyük oranda kesişmektedir. Bu da ölçme sonuçlarına dayalı olarak karar verirken, özellikle de belli bir kesme puanına göre geçti/kaldı, başarılı/başarısız türünde sınıflandırmalar yaparken, ölçmenin standart hatasını hesaba katmanın daha adil/savunulabilir sonuçlara ulaşabilmek için ne denli önemli olduğunu gözler önüne sermektedir.



4. GEÇERLİK

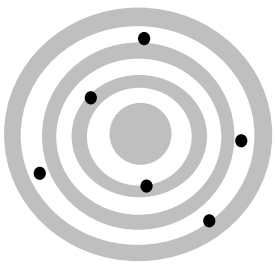
Geçerlik, ölçme ve değerlendirme alanının en yanlış anlaşılan ve en hatalı kullanılan terimlerinden biridir (Frisbie, 2005). Yanlış anlaşılmaya ve yorumlanmaya müsait olması sebebiyle geçerlik için geçmişten günümüze çeşitli tanımlar ileri sürülmüştür. 1940 ile 1950'lerin başında yapılan ilk tanımlarda test puanlarının nasıl kullanıldığı dikkate alınmadan geçerlik ölçme aracına ait mutlak bir özellik olarak karakterize edilmiştir (Oerman & Gaberson, 2021). Ölçme aracının "amaca hizmet etme derecesi" ve "ölçmeye çalıştığı yapıyı başka değişkenler ile karıştırmadan ölçebilmesi" bu bakış açısının ürünü olan tanımlardır. Ancak geçerlik ölçme aracının kendisine veya ölçümlere ait bir nitelik değildir, ölçme aracından elde edilen puanların ne kadar doğru yorumlandığıyla ilgili bir özelliktir. Bu da vurgunun ölçme aracında ya da ölçümlerde olduğu geçerlik tanımlarını gözden geçirme ihtiyacını doğurmuştur. Bu bağlamda *Eğitsel ve Psikolojik Test Standartları*'nda geçerlik "test puanlarına dayalı olarak yapılan yorumların, bu yorumların doğruluğunu test etmek için toplanan kanıtlar ve ölçülen özelliğin teorik çerçevesi ile uyumu" şeklinde kavramsallaştırılmıştır (AERA, APA & NCME, 2014). Mesela bir fizik testindeki maddeleri cevaplayabilmek için ağır matematiksel işlemler yapmak gerekiyorsa öğrencilerin test puanları fizik bilgilerinin yanı sıra matematik bilgilerini de yansıtır. Diğer bir deyişle ölçülmesi hedeflenen özellik fizik başarısı iken matematik başarısı da ölçme sonuçlarına karışmış olur. Bu durumda test puanlarını sadece fizik başarısının göstergesi olarak yorumlamak bir geçerlik sorununu gündeme getirir. Buna karşın tüm cevaplayıcıların matematik yeterliklerinin oldukça yüksek olduğu bir başka grupta ilgili testten alınan puanlar fizik başarısının göstergesi olarak kabul edilebilir. Dolayısıyla aynı test ikinci grupta amaca daha fazla hizmet eder ve çok daha geçerli sonuçlar verir. Benzer şekilde, lise öğrencilerinin kimya

bilgisini yoklamaya yönelik Almanca hazırlanmış bir test düşünelim. Bu test Almanya'daki liselerde uygulandığında test puanları öğrencilerin kimya bilgilerinin göstergesi olabilir. Oysa aynı test Türkiye'deki liselerde uygulandığında test puanlarını kimya bilgisinin bir ölçüsü olarak yorumlamak hatalı olur. Çünkü kullanılan test Türkiye örnekleminde kimya bilgisi kadar Almanca bilgisini de ölçer ve test puanlarını yalnızca kimya bilgisiyle ilişkilendirmek bir geçerlik problemi yaratır (İlhan ve ark., 2024).

Geçerlik, güvenirlik ile yakından ilişkilidir. O kadar ki Salkind ve Frey (2023) geçerlik ile güvenirliği birinci dereceden kuzenler olarak nitelendirir. Ölçümler güvenilir değilse bu ölçmelerden hareketle yapılan yorumların geçerli olması imkânsızdır. Bununla beraber güvenilir ölçümler, ölçümlere dayalı çıkarımların da mutlaka geçerli olacağı anlamına gelmez. Güvenirlik, geçerlik için gereklidir fakat yeterli değildir. Bunun tersi ise doğrudur. Test puanlarına bağlı olarak yapılan yorumlar geçerliyse söz konusu puanlar güvenilir demektir. Domino ve Domino (2006) geçerlik ve güvenirlik arasındaki ilişkiyi şöyle örneklendirmektedir: Bir mezura aracılığıyla başınızın çevresini ölçüp bulduğunuz değeri 6,3 ile çarpın. Daha sonra ulaştığınız sonuca sol elinizdeki parmak sayısının üç katı ile göz sayınızın altı katını ekleyin. Bu işlem sizi IQ puanınıza götürür. Bir sınıftaki öğrencilerden bunları yapmasını ve sıralanan işlemleri birkaç kez tekrarlamalarını istedik diyelim. Öğrencilerin baş ebatları değişmediği ve tekrarlı ölçmeler arasında vücut uzuvlarında bir kayıp yaşamadıkları sürece hesaplanan IQ puanları aynı kalır. Bu da elde edilen puanların son derece güvenilir olduğunu bildirir. Ancak ulaşılan ölçümler gerçekte IQ puanını yansıtmayacağından sonuçlar geçerli olmaz. Güvenirliğin geçerlik için ön koşul niteliği taşıdığı ama geçerliği garanti etmeyeceğini açıklamak için sıklıkla Şekil 2'deki görselden yararlanılır. Şekil 2'ye bakıldığında bir numaralı görselde atışların ne tutarlı ne de isabetli olduğu görülmektedir. Bu sonuç, güvenirliğin ve geçerliğin düşük olduğunu ima etmektedir. İki numaralı görselde atışlar tutarlıdır ama isabetli değildir. Buna göre güvenirlik yüksek ama geçerlik düşüktür. Üç numaralı görselde ise atışların hem tutarlı hem isabetli olması güvenirliğin de geçerliğin de yüksek olduğunu yansıtmaktadır.

Şekil 2

Geçerlik ve Güvenirlik Arasındaki İlişki (Fraenkel ve ark., 2012; Suter, 2012)



(I) Düşük Güvenirlik, Düşük Geçerlik



(II) Yüksek Güvenirlik, Düşük Geçerlik



(III) Yüksek Güvenirlik, Yüksek Geçerlik

Geçerlik bazı yönleriyle güvenirliğe benzerken kimi yönleriyle güvenirlikten ayrılır. Öncelikle geçerlik de güvenirlik gibi doğrudan gözlenemez, dolaylı olarak tespit edilebilir. Geçerliğin güvenirliğe benzeştiği ikinci nokta; derecelerinin olması, var-yok biçimde değerlendirilememesidir. Bu bakımdan "geçerlidir" ya da "geçerli değildir" gibi kalıplar kullanmak doğru değildir. Bunun yerine "geçerliği düşüktür", "geçerliği yüksektir" gibi ifadeler benimsenmelidir. Geçerlik ve güvenirliği birbirinden ayıran başlıca özellikler ise şunlardır: Güvenirlik *tutarlılık* kavramı ile eşleştirilirken geçerlik *doğruluk* kavramı ile eşleştirilebilir. Güvenirlik tesadüfi hatalardan etkilenirken geçerlik sabit, sistematik ya da tesadüfi fark etmeksizin tüm hata türlerinden etkilenir. Ayrıca güvenirliğin kararlılık, iç tutarlılık, eşdeğerlik ve objektiflik gibi farklı anlamları bulunsa da bunlar tek bir istatistiksel değer ile belirlenebilir. Söz gelimi ölçümlerin iç tutarlılığı Cronbach alfa katsayısı ile saptanabilir. Ancak geçerliğe gelince işin rengi değişir, tek bir katsayıya dayanarak geçerlik hakkında yorum yapmak mümkün olmaz. Geçerliğin test edilmesi, ölçme aracından alınan puanlara dayanarak ne tür çıkarımlar yapılabileceğine ya da hangi çıkarımlarda bulunmanın yanlış olacağına dair bir kanıt oluşturma sürecini içerir. Bu süreçte bir suçluyu ortaya çıkarmak için delil peşinde koşan dedektifler gibi hareket etmek, test puanlarının doğru yorumlandığı konusunda güçlü kuramsal ve ampirik kanıtlar sunmak gerekir (Coaley, 2010). Geçerliği değerlendirme işleminin statik değil dinamik bir yapıya sahip olduğu da unutulmalıdır (Sireci, 2007). Buna göre başlangıçta suçsuz bulunan bir sanığın toplanan yeni delillerin ardından suçlu olduğu anlaşılabileceği gibi geçerlik hakkında yeni kanıtlar toplandıkça önceki kanıtların işaret ettiğinden daha farklı bir sonuçla karşılaşılabilir. Elbette, her zaman böyle olmak zorunda değildir. Nasıl ki yeni deliller kimi zaman sanığın suçluluğuna ya da suçsuzluğuna yönelik kanaati güçlendirebiliyorsa yeni kanıtların geçerliğe ilişkin önceki değerlendirmeleri desteklemesi de mümkün olabilir. Sözün özü, "Test puanlarına ilişkin yorumların geçerli olduğu doğrulanmıştır, daha başka geçerlik kanıtları toplamaya gerek yoktur" gibi bir iddiada bulunmak olanaksızdır (Swanson, 2012).

4.1. Geçerlik Kanıtı Türleri

Geçerlik hakkında toplanabilecek kanıtlar oldukça çeşitlidir. Bununla birlikte, ilgili kanıtlar belli başlıklar altında gruplandırılmaktadır. *Eğitsel ve Psikolojik Test Standartları*'nın 1985 yılındaki basımında geçerlik kanıtları; kapsamla, yapıyla ve ölçütle bağlantılı kanıtlar olmak üzere üçe ayrılmıştır. Kitabın 1999 yılında yayınlanan basımıyla birlikte ise geçerliğe ilişkin kanıtlar Şekil 3'te sunulan beş başlıkta kategorize edilmiştir (McDonald, 2018). Aynı kategorizasyon kitabın 2014 tarihli son baskısında da benimsenmiştir. Bu aşamada şunun da altını çizmekte fayda vardır. Şekil 3'te yer alan ve günümüzde kabul gören başlıklar geçerlik türlerini değil, geçerlik kanıtı türlerini ifade eder. Geçerlik tekil (üniter) bir kavram olup farklı geçerlik türleri yoktur, geçerliğe ilişkin farklı kanıtlar mevcuttur. Bu anlamda, 1950'li yılların ilk yarısından 1985'e kadar hâkimiyetini sürdüren geleneksel geçerlik teorisinde var olan "geçerlik türleri" anlayışı (Messick, 1989) tamamıyla terk edilmelidir.

Şekil 3

Geçerlik Kanıtı Türleri



4.1.1. Kapsama Dayalı Kanıtlar

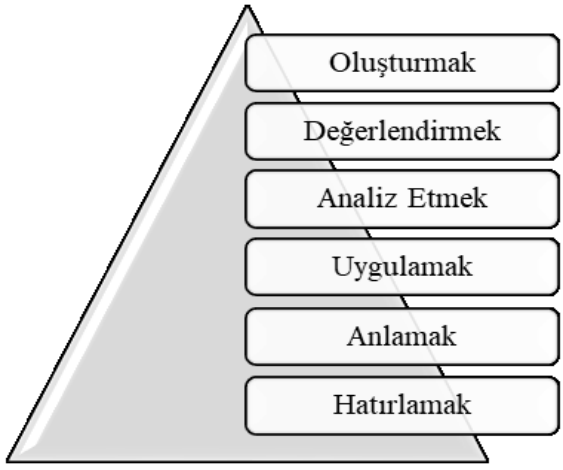
Kapsama dayalı geçerlik kanıtları, ölçme aracının ölçmeye konu özelliği ne kadar iyi temsil edebildiği sorusuna yanıt arar. Teorik olarak bir yapıyı ölçmeye yönelik sonsuz sayıda madde oluşturulabilir. Ancak uygulamada bu mümkün olmadığından belli sayıda madde seçilir ve ölçme aracına dahil edilir (Baumgartner ve ark., 2021). Hedef yapıyı tam anlamıyla temsil eden bir madde örneklemi seçebilmek için ilk yapılması gereken ayrıntılı bir test planı hazırlamaktır. Bu aşama, bir mimarın inşa süreçleri öncesinde kâğıt üzerinde çizimler yapmasına benzer (Violato, 2019). İnşaat sonucunda ortaya çıkacak binanın sağlam ve beklentileri karşılar nitelikte olması mimarın sürecin başında doğru çizimler yapmasına bağlı olduğu gibi ölçme sonucunda geçerli ölçümlere ulaşabilmenin yolu da uygun bir test planıyla işe başlamaktan geçer. Test planı sayesinde ölçülecek kapsama ilişkin analitik bir çözümleme yapılır. Çözümlemede, hedef içeriğin hangi bileşenlerden oluştuğunu betimlemeye yarayan ve ölçme literatüründe "belirtke tablosu (table of specification/test blueprint)" olarak adlandırılan bir matris kullanılır.

Belirtke tablosu satırlarında hedef davranışları, sütunlarında ise bu hedef davranışların bilişsel düzeylerini listeler. Ölçülecek içeriğin böyle bir tabloya yerleştirilmesi; bazı hedef davranışların ölçme aracında temsil edilmemesi, gereğinden az ya da fazla temsil edilmesi gibi riskleri minimuma indirir. Ayrıca tablo hedef davranışların bilişsel düzeylerini içerdiğinden hangi madde türünü/türlerini kullanmak gerektiğine dair de yol gösterici olur. Bu anlamda belirtke tablosu, hem geçerliği sağlamaya hizmet eder hem de geçerliğe kanıt oluşturur. Test geliştiricisi belirtke tablosunu hazırlarken önce hedef davranışları, diğer bir deyişle ölçülecek öğrenme çıktılarını tespit etmeli ve devamında

öğrenme çıktılarının bilişsel düzeylerini saptamalıdır. Alanyazında bu amaçla kullanılabilecek Bloom, SOLO, Fink ve Dettmer gibi farklı taksonomiler bulunmaktadır. Bunlar arasında en fazla kullanılanı ise Bloom taksonomisidir (Arı, 2013).

Bloom taksonomisi doğrudan bu bölümün kapsamına giren bir başlık değildir. Bununla birlikte aşağıda Bloom taksonomisinden ana hatlarıyla söz edilmiş, taksonomi hakkında sunulan bilgiler belirtke tablosunu örneklendirmeye imkân tanıyacak bir içerikle sınırlandırılmıştır. Dolayısıyla yalnızca taksonomiye ait bilişsel süreçler üzerinde durulmuştur.

Bloom taksonomisine göre bilişsel süreçler; *hatırlamak*, *anlamak*, *uygulamak*, *analiz etmek*, *değerlendirmek* ve *oluşturmak* şeklinde basitten karmaşığa doğru ilerleyen altı düzeyden oluşur. Bu düzeyler, alt basamaklar geçilmeden üst basamaklara erişmenin mümkün olmadığı hiyerarşik bir yapıya sahiptir (Anderson & Krathwohl, 2001). Düzeylerin temel özellikleri Şekil 4'te açıklanmıştır.



Şekil 4

Bloom Taksonomine göre Bilişsel Süreç Basamakları [Krathwohl'dan (2002) Yararlanılarak Oluşturulmuştur]

Hatırlamak: Görünce tanıma, ezberden tekrar etme ile ilgilidir. Örneğin; öğrenciden bir tanım yapması istendiğinde kitaptaki tanımı aynen aktarması, benzer şekilde bir kavramı örneklendirmesi talep edildiğinde verdiği örneklerin derste sunulanlardan ibaret kalması hatırlama basamağına karşılık gelir. Yine bir olayın tarihini söyleme, bir antlaşmanın maddelerini sıralama hatırlama düzeyine ilişkin örneklerdir.

Anlamak: Bu basamakta öğrenci öğrendiklerini kendi cümleleriyle ifade edebilir, farklı temsil biçimleri arasında dönüşümler yapabilir, ayrıca sınıflama, yorumlama, açıklama ve özetleme işlemlerini gerçekleştirebilir. Örneğin; öğrenci bir kavramı kitapta geçtiği şekliyle tanımlamak yerine kendi cümleleriyle açıklayabilir, kavrama yönelik kitapta/derste sunulanların ötesinde başka örnekler verebilir, okuduğu bir parçayı özetleyip parçanın ana fikrini belirleyebilir, aktif bir cümleyi edilgen cümleye çevirebilir, sözel ifadeleri grafiğe aktarabilir ya da bir grafiği sözel olarak yorumlayabilir.

Uygulamak: Bireyin öğrendiklerini karşılaştığı problem durumlarında ve/veya gerçek hayatta kullanılabildiği basamaktır. Konu hakkındaki bilgilerinden yola çıkarak bir matematik problemini çözmek veya kompozisyon yazarken noktalama işaretlerini doğru şekilde kullanmak uygulama basamağına örnek olarak sunulabilir.

Analiz etmek: Bir bütünü bileşenlerine ayırma, parçaların birbiriyle ve bütünle ilişkisini saptama, parçalar arasındaki ilişkilerden hareketle bütüne ulaşma analiz etme düzeyinin karakteristik özellikleridir. Söz gelimi, merkezi sınavlarda sıkça yer verilen "Bu parça iki paragrafa ayrılacak olursa ikinci paragraf hangi cümleyle başlar?" tarzındaki sorular analiz etme düzeyine karşılık gelir.

Değerlendirme: Bir ölçütle karşılaştırıp yargıya varma söz konusudur. Mesela bir test yönergesini taşıması gereken özellikler açısından gözden geçirip yönergenin ne kadar iyi hazırlandığına dair yargıya varmak değerlendirme basamağıyla eşleşen bir işlemdir.

Oluşturmak: Yeni, tutarlı ve özgün bir ürünün ortaya konulabildiği basamaktır. Tez yazmak, proje hazırlamak, beste yapmak, bir öğretim materyali tasarlamak oluşturma basamağında gerçekleştirilebilecek eylemlerdir.

Bloom taksonomisinin hatırlama, anlama ve uygulama basamakları alt; analiz, değerlendirme ve oluşturma basamakları ise üst düzey bilişsel süreçler olarak kabul edilir. Alt bilişsel süreçlerden üst bilişsel süreçlere doğru gidildikçe yanıtı yapılandırmayı gerektiren ölçme araçlarına (örneğin; yazılı ve sözlü sınavlar, performans görevleri) duyulan ihtiyaç artar. Özellikle oluşturma basamağına gelindiğinde yanıtı seçmeyi gerektiren ölçme araçları (doğru-yanlış testleri, çoktan seçmeli testler vb.) amaca hizmet edemez, öğrenme çıktıları ancak yanıtı yapılandırmayı gerektiren araçlarla ölçülebilir.

Bloom taksonomisinin temel çerçevesini öğrendiğimize göre artık örnek bir belirtke tablosu inceleyebiliriz. Aşağıda özel dörtgenler konusuna ilişkin bir belirtke tablosu sunulmuştur (bkz. Tablo 9). Tablo 9'a göre ölçülmesi gereken öğrenme çıktılarının sayısı 25'tir ve geliştirilecek test beş ayrı dörtgen türüyle ilgili maddeler içermelidir. En fazla madde yamukla ilgili olmalı (%28), bunu kare (%20) ve dikdörtgene (%20) ilişkin madde sayıları izlemeli, paralelkenar (%16) ve eşkenar dörtgene (%16) ilişkin madde sayıları ise görece daha az olmalıdır. Tablo 9'daki dağılımdan testte anlama, uygulama ile analiz etme düzeylerine karşılık gelen maddelere yer verilmesi ve maddelerin yaklaşık yarısının uygulama düzeyinde olması gerektiği anlaşılmaktadır. Belirtke tablosu tam olarak bunu, öğrenme çıktılarını kuşbakışı görebilmeyi sağlar (Earnest & Gupta, 2022). Böylece hedef içeriğin bazı kritik bileşenlerinin test geliştirme sürecinde gözden kaçması ihtimalini azaltır.



Sınıf içinde uygulanan sınavlar, konu ya da ünite testleri genellikle belirtke tablosunda listelenen her bir öğrenme çıktısıyla ilgili maddeler içerecek tarzda hazırlanır. Diğer taraftan geniş ölçekli sınavlarda bir veya birkaç öğretim yılı boyunca işlenen kapsamlı bir içeriğin ölçülmesi amaçlandığından bunu yapmak olanaksızdır. Bu gibi durumlarda kritik hedef davranışlara odaklanılır ve ölçme aracı bu doğrultuda organize edilir.

Tablo 9

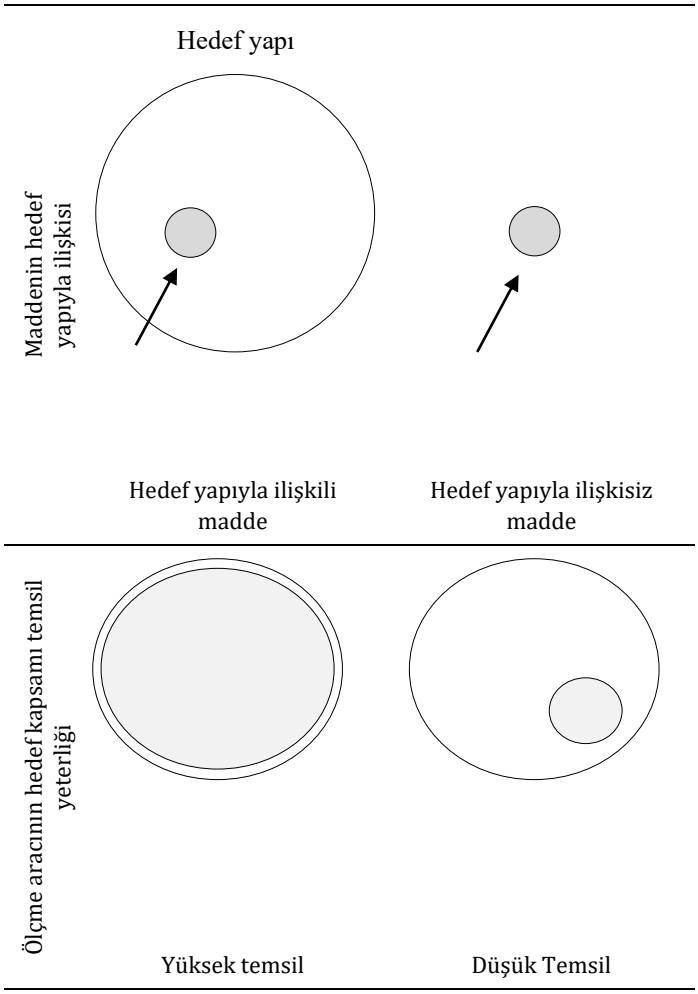
Bloom Taksonomisine göre Hazırlanmış Belirtke Tablosu Örneği

	Bilişsel Düzeyler						
Konular	**Hatırlamak**	**Anlamak**	**Uygulamak**	**Analiz Etmek**	**Değerlendirmek**	**Oluşturmak**	**Toplam**
Dikdörtgen	–	1 (%4)	2 (%8)	2 (%8)	–	–	5 (%20)
Kare	–	1 (%4)	2 (%8)	2 (%8)	–	–	5 (%20)
Paralelkenar	–	1 (%4)	2 (%8)	1 (%4)	–	–	4 (%16)
Eşkenar dörtgen	–	1 (%4)	2 (%8)	1 (%4)	–	–	4 (%16)
Yamuk	–	2 (%8)	3 (%12)	2 (%8)	–	–	7 (%28)
Toplam	–	6 (%24)	11 (%44)	8 (%32)	–	–	25 (%100)

Belirtke tablosunu düzenleyip ölçme aracını bu tablodaki dağılım doğrultusunda hazırlamak kapsama dayalı önemli bir geçerlik kanıtı sağlasa da bu kanıtı uzman görüşleriyle desteklemek son derece önemlidir. Bunun için taslak ölçme aracı alan uzmanlarına sunulmalı, uzmanlar hem her bir maddenin hedef yapıyla ilişkisini (item relevance) değerlendirmeli hem de maddelerin bu yapıyı ne kadar kapsadığına (content coverage) dair görüş bildirmelidir. Çünkü maddelerin tamamı ölçülen yapıyla ilişkili olduğu halde ölçme aracı hedef yapıyı yeterince kapsamıyor olabilir. Örneğin, obsesif-kompulsif semptomları değerlendirmeye yönelik bir ölçek için yazılan maddelerin tamamı ritüel davranışlar (el yıkama alışkanlığının varlığı ve sıklığı gibi) ile ilişkiliyse madde bazındaki incelemeler geçerliğe ilişkin olumlu bir tablo ortaya koyar. Ancak maddelerin obsesif kompülsif bozukluğun ritüel davranışların dışında kalan semptomlarını içermemesi ölçme aracının hedef yapıyı tam anlamıyla temsil edemediğini yansıtır (Reynolds ve arkadaşları, 2021). Bu durumda madde düzeyindeki incelemelerin aksine ölçme aracının geneline ilişkin değerlendirmeler geçerliğe dair olumsuz bir sonuç bildirir. Maddenin ölçülen özellikle ilişkisi ve ölçme aracının hedef kapsamı temsil yeterliği arasındaki bu ayrım Şekil 5'te resmedilmiştir.

Şekil 5

Maddenin Ölçülen Özellikle İlişkisinin ve Ölçme Aracının Hedef İçeriği Temsil Kabiliyetinin Görsel Sunumu (Reynolds ve ark., 2021)



İlhan ve arkadaşları (2024) maddeler hakkında en az iki alan uzmanından görüş alınmasını önermektedir. Ancak uzman sayısının odak nokta haline getirilmemesi gerektiğini savunmakta, yetkinliği yüksek az sayıda uzmanın düşük yetkinlikteki çok sayıda uzmana kıyasla maddeler hakkında daha önemli geri bildirimler sunabileceğini belirtmektedir. Tabi alan uzmanlarının yanında maddelerin madde yazım ilkelerine uygunluğunu değerlendirmesi amacıyla ölçme ve değerlendirme uzmanlarından; yazım, imla ve noktalama kuralları açısından inceleme yapması amacıyla da dil uzmanlarından görüş talep edilmelidir.

Uzmanlardan görüş alırken uzmanların maddelerle ilgili düşüncelerini nicel olarak ifade edebilecekleri bir form kullanılması, kapsama dayalı geçerlik kanıtlarını daha somut hale getirir. Bundan dolayı uzman görüşlerine başvururken Şekil 6'daki tarzda bir form hazırlanması önerilir.

Şekil 6

Örnek Uzman Görüş Formu [Özer Özkan (2022) ile İlhan ve arkadaşlarından (2024) uyarlanmıştır]

Sayın Uzman;

Aşağıda yer alan maddeler …….…….. amacıyla hazırlanmış olup ölçme aracının ……..…….. kişilerden oluşan bir örnekleme uygulanması planlanmaktadır. Sizden taslak ölçme aracındaki her bir maddenin belirtilen amaca ne kadar hizmet ettiğini değerlendirmeniz rica edilmektedir. Madde;

- Hedef yapıyla ilişkili ve doğrudan kullanılabilir nitelikte ise *Gerekli-Aynen Kullanılabilir*
- Hedef yapıyı ölçebilir özellikte ama revize edilmesi gereken kısımları varsa *Gerekli-Düzeltilmeli*
- Hedef yapıyla ilişkili ancak gerekli değilse *Yararlı fakat Gerekli Değil*
- Hedef yapıyı temsil etmiyorsa *Gereksiz*

seçeneklerini işaretleyiniz. Maddelere ilişkin görüşlerinizi en sağ sütunda detaylandırabilirsiniz. Özellikle revize edilmesi gerektiğini düşündüğünüz maddeler için görüşlerinizi ayrıntılı biçimde sunmanız ilgili düzeltmelerin yapılmasını kolaylaştıracaktır. Katkılarınızdan dolayı teşekkür ederim/ederiz.

Ölçme aracını geliştiren kişinin/kişilerin;
Unvanı, Adı ve Soyadı
İletişim bilgileri

Maddeler	Ölçtüğü öngörülen boyut/kazanım	Gerekli		Yararlı fakat Gerekli Değil	Gereksiz	Görüş ve Öneriler
		Aynen Kullanılabilir	Düzeltilmeli			
……	……					
……	……					
……	……					
……	……					
……	……					

Ölçme aracının hedef kapsamı temsil yeterliğini 1 ile 10 arasında puanlayacak olursanız kaç puan verirsiniz? ……. *(puan)*

Eklemek istedikleriniz:

……
……
……
……

Uzmanlardan görüş alındıktan sonra her bir madde için kapsam geçerlik oranı (KGO) hesaplanır. KGO uzman görüşlerini nicel verilere dönüştürerek kapsama dayalı geçerlik kanıtlarına nesnel bir kimlik kazandırır (Yurdugül & Bayrak, 2012). KGO'yu hesaplamak için Lawshe'nin (1975) geliştirdiği Eşitlik 8'deki formülden yararlanılır.

$$KGO = \frac{N_G}{N/2} - 1$$

N_G = Gerekli görüşünü bildiren uzman sayısı
N = Görüşüne başvurulan toplam uzman sayısı

Eşitlik 8

Eşitlik 8'e göre, bir maddeyi görüşüne başvurulan uzmanlardan hiçbiri gerekli bulmazsa KGO -1 çıkar. Madde için tüm uzmanların gerekli şeklinde görüş bildirmesi durumunda ise KGO +1 olarak hesaplanır. Yani KGO'nun alabileceği değerler ±1 aralığında uzanır (Özer Özkan, 2022). Burada akla gelecek temel soru ise taslak ölçme aracındaki bir maddenin ölçme aracının nihai formuna dâhil edilebilmesi için KGO'sunun en az kaç olması gerektiğidir. Bununla ilgili ilk tanımlama Lawshe (1975) tarafından yapılmıştır. Lawshe (1975), *p*=0,05 anlamlılık düzeyinde uzman sayısına göre KGO'nun eşik değerlerini içeren bir tablo oluşturmuş, bu tablodaki kritik değerler daha sonra farklı araştırmacılar (Ayre & Scally, 2014; Wilson ve ark., 2012) tarafından da test edilmiştir. Ayre ve Scally (2014) yaptıkları incelemede Lawshe'nin (1975) belirlediği kritik değerlerin güvenilir olduğunu rapor etmiş, bununla beraber daha kesin olduğunu iddia ettikleri yeni sınır değerler ileri sürmüşlerdir. Tablo 10, KGO için Lawshe (1975) ile Ayre ve Scally (2014) tarafından tanımlanan kritik değerleri içermektedir.

Tablo 10

KGO için Tanımlanan Kritik Değerler

Lawshe (1975)		**Ayre ve Scally (2014)**			
Uzman Sayısı	**Kritik Değer**	**Uzman Sayısı**	**Kritik Değer**	**Uzman Sayısı**	**Kritik Değer**
5	0,99	5	1,00	23	0,391
6	0,99	6	1,00	24	0,417
7	0,99	7	1,00	25	0,440
8	0,75	8	0,750	26	0,385
9	0,78	9	0,778	27	0,407
10	0,62	10	0,800	28	0,357
11	0,59	11	0,636	29	0,379
12	0,56	12	0,667	30	0,333
13	0,54	13	0,538	31	0,355
14	0,51	14	0,571	32	0,375
15	0,49	15	0,600	33	0,333
20	0,42	16	0,500	34	0,353
25	0,37	17	0,529	35	0,314
30	0,33	18	0,444	36	0,333
35	0,31	19	0,474	37	0,297
40	0,29	20	0,500	38	0,316
		21	0,429	39	0,333
		22	0,455	40	0,300

Hangi maddelerin ölçme aracında kalması gerektiğine karar verirken her bir madde için hesaplanan KGO, Tablo 10'daki kritik değer ile karşılaştırılır. Bu karşılaştırma doğrultusunda iki kutuplu bir karar alınır. KGO'su kritik değeri aşıyorsa madde ölçme aracının nihai formuna alınır, yok eğer eşik değerin altında kalıyorsa madde testten

çıkarılır. Son adımda ise nihai forma seçilen maddelere ait KGO değerlerinin ortalaması alınıp ölçme aracının geneline ilişkin kapsam geçerlik indeksi bulunur. Çok boyutlu ölçme araçlarında bu son adım her boyut için ayrı ayrı yapılır (Yurdugül, 2005). Mesela ölçme aracında A ve B şeklinde iki boyutlu bir yapı varsa A boyutuna seçilen maddelerin KGO'larının ortalaması alınarak bu boyutun kapsam geçerlik indeksi elde edilir ve aynı işlem B boyutu için de tekrarlanır. Örnek teşkil etmesi açısından Tablo 11'de 10 maddelik bir test için 15 uzmandan görüş alındığı varsayılmış, uzmanların değerlendirmeleri doğrultusunda maddelere ait KGO değerleri ile ölçme aracının geneline ilişkin kapsam geçerlik indeksinin hesaplanmasına yönelik süreç özetlenmiştir.

Tablo 11

KGO İçin Örnek Uygulama

Maddeler	Gerekli diyen uzman sayısı	KGO	Ölçme Aracının Geneline ait Kapsam Geçerlik İndeksi
M1	14	[14/(7,5)] – 1 = 0,87	15 uzman için kritik KGO değeri Lawshe'ye (1975) göre 0,49 iken Ayre ve Scally'ye (2014) göre 0,60'tır. Bu iki ölçütten hangisi dikkate alınırsa alınsın ölçme aracından çıkarılması gereken maddelerin M5, M7 ve M9 olduğu anlaşılmaktadır. Kalan yedi madde üzerinden nihai test için kapsam geçerlik indeksi hesaplanacak olursa $\frac{(0,87+1,00+0,73+1,00+0,60+0,73+1,00)}{7} = 0,72$ bulunur.
M2	15	[15/(7,5)] – 1 = 1,00	
M3	13	[13/(7,5)] – 1 = 0,73	
M4	15	[15/(7,5)] – 1 = 1,00	
M5	7	[7/(7,5)] – 1 = –0,07	
M6	12	[12/(7,5)] – 1 = 0,60	
M7	10	[10/(7,5)] – 1 = 0,33	
M8	13	[13/(7,5)] – 1 = 0,73	
M9	9	[9/(7,5)] – 1 = 0,20	
M10	15	[15/(7,5)] – 1 = 1,00	

4.1.2. İçyapıya Dayalı Kanıtlar

Bu geçerlik kanıtında ölçme aracını oluşturan maddelerin ve boyutların ölçülmesi amaçlanan gizil yapı ile ne kadar örtüştüğü belirlenmeye çalışılır (Holmboe ve ark., 2018). Kapsama dayalı geçerlik kanıtlarına benzemekle birlikte burada maddeler bireysel olarak değerlendirilmez, maddeler arası ilişkilere odaklanılır (Mertler, 2022). Örneğin, bir test tek bir yetenek ya da özelliği ölçmek amacıyla tasarlandıysa testin her bir maddesinden alınan puanlar ile geneline ait puanlar arasında güçlü korelasyonlar bulunmalıdır. Öte yandan ölçmeye konu özellik çeşitli faktörlerden oluşuyor ve test bu faktörler doğrultusunda hazırlandıysa maddeler, ölçmeye çalıştığı faktörden alınan toplam puan ile yüksek korelasyonlar göstermeli; diğer faktörlerden ve testin genelinden elde edilen puanlar ile görece daha zayıf bir ilişki içerisinde olmalıdır (Morrow ve ark., 2023). Kısacası maddeler ve boyutlar arasında gözlenen ilişkiler, hedef yapının teorik çerçevesi ile paralellik taşımalıdır.

Konuyu biraz daha işevuruk hale getirmeye çalışalım. Kişiliği ölçmek amacıyla beş faktörlü kişilik modeline dayalı bir envanter geliştirdiğimizi varsayalım. Büyük beşli olarak da adlandırılan beş faktörlü modele göre kişilik; "dışadönüklük, uyumluluk, sorumluluk, nevrotiklik ve deneyime açıklık" boyutlarından oluşur (McCrae & Costa, 2003; McCrae & John, 1992). Bu durumda hazırladığımız ölçme aracı sıralanan beş boyutlu yapı ile uyumlu olmalıdır. Bu uyumu test etmek için ölçme aracı yeterli büyüklükte bir gruba uygulanır ve toplanan veriler üzerinden çeşitli analizler gerçekleştirilir. Yapılan analizler maddeler arası ilişkilerin referans alınan beş boyutlu yapıyla eşleştiğini gösterirse bu sonuç geçerlik için içyapıya dayalı kanıt teşkil eder.

İçyapıya dayalı geçerlik kanıtlarına ulaşmak için en çok başvurulan teknik faktör analizidir (Reynolds ve ark., 2021). Faktör analizi ölçme aracındaki maddeleri aralarındaki korelasyonların örüntüsü ile uyumlu, doğrudan gözlenmesi mümkün olmayan bir veya birkaç faktör altında toplamaya dönük bir veri indirgeme tekniğidir (İlhan ve ark., 2024). Kapsamı gereği bu bölümde faktör analizine yönelik ayrı bir başlık açılmamıştır. Bununla beraber, faktör analizinin en azından mantığının kavranmasına yardımcı olacağı düşüncesiyle Tablo 12'de örnek bir korelasyon matrisi sunulmuştur. Hipotetik bir veri seti üzerinden elde edilen bu korelasyon matrisi, 6 maddelik (M1-M6) bir testte katılımcıların madde puanları arasındaki ilişkileri göstermektedir. Testin ilk üç maddesinin sözel, devamındaki üç maddesinin ise sayısal yeteneği ölçtüğü öngörülmektedir.

Tablo 12

İki Boyutlu Olduğu Öngörülen Altı Maddelik Bir Test için Korelasyon Matrisi

Maddeler	M1	M2	M3	M4	M5	M6
M1	–					
M2	0,610	–				
M3	0,427	0,481	–			
M4	0,144	0,159	0,163	–		
M5	0,166	0,177	0,124	0,617	–	
M6	0,183	0,161	0,177	0,414	0,560	–

Tablo 12'deki korelasyon katsayılarına göre M1, M2 ve M3'ün birbiriyle ilişkisi, testin devamındaki maddeler ile olan ilişkilerine kıyasla daha güçlüdür. Benzer şekilde M4, M5 ve M6'nın kendi aralarındaki ilişkiler testin ilk üç maddesiyle olan ilişkilerinden belirgin derecede yüksektir. Yani korelasyon matrisi, maddelerin beklendiği gibi iki boyut altında kümelendiğine işaret etmektedir. İşte faktör analizi, maddeler arası ilişkiler üzerine temellendirilen bir dizi istatistiksel işlemle ölçme sonucunda gözlenen yapının kuramsal olarak öngörülen bileşenleri destekleyip desteklemediğini ortaya koyar. Bu yolla geçerliğe ilişkin içyapıya dayalı kanıtlar sunar.

4.1.3. Ölçme Aracından Alınan Puanların Diğer Değişkenler ile İlişkisine Dayalı Kanıtlar

Ölçülmesi hedeflenen yapıya benzer olduğu bilinen ya da bu yapıdan ayrı bir özelliği temsil ettiğine inanılan dışsal bir değişken ölçüt olarak alınır. Ölçme aracından elde edilen puanlar ile ölçüt değişkene ait puanlar arasındaki ilişki incelenerek geçerlik hakkında değerlendirme yapılır. Ulaşılan sonuçların hedef yapıyla ilgili kuramsal çerçeveyi ve/veya daha önce yapılmış ampirik araştırmaları destekler nitelikte olması geçerliğe kanıt sağlar. Kettler (2019) ölçme aracının diğer değişkenlerle ilişkisine dayalı kanıtları; eş zamanlı ve yordayıcı kanıtlar, yakınsak ve ıraksak kanıtlar ile karşıt gruplara dayalı kanıtlar şeklinde kategorilere ayırmıştır. Bu bölümde Kettler'in (2019) yaptığı sınıflama dikkate alınıp ölçme aracının diğer değişkenlerle ilişkine dayalı kanıtlar üç başlıkta sunulmuştur.

Ölçmeye konu hiçbir özellik diğer yapılardan bütünüyle izole değildir. Aksine her bir yapı, içerisinde başka yapıların da bulunduğu bir sistemin parçasıdır. Bu sistemde birbirine benzer özellikleri ölçen yapılar bulunabildiği gibi farklı özellikleri ölçen yapılar da yer alabilir. Geçerliğe kanıt ararken bunun farkında olmak ve hedef yapının nomolojik ağını incelemek önemlidir (Suen & Rzasa, 2004). Geçerlik çalışmalarında ölçme aracının diğer değişkenlerle ilişkisini analiz etmenin arkasında yatan felsefe budur.

4.1.3.1. Eş Zamanlı ve Yordayıcı Kanıtlar

Geçerliği sorgulanan ölçme aracından elde edilen puanlar ile onunla ilişkisi olması beklenen bir başka değişkene ait veriler arasındaki ilişkiye bakılır. Eş zamanlı ve yordayıcı kanıt ayrımının altında, ölçüt değişkene ait verilerin ne zaman elde edildiği yatar. Eş zamanlı kanıtlarda ölçüt değişkene ait veriler, hedef yapıya ait ölçümler ile aynı anda toplanır. Buna karşılık yordayıcı kanıtlarda ölçüt değişkenin verileri daha ileri bir zaman dilimine aittir. Akademik erteleme eğilimi ölçeğinin geçerliğine ilişkin bir çalışma yürütüyor olalım. Bu kapsamda üniversite son sınıf öğrencilerine akademik erteleme eğilimi ölçeği ile zaman yönetimi becerileri ölçeğini aynı anda uyguladık diyelim. Ayrıca öğretim yılı sonunda, öğrenci işleri daire başkanlığından çalışmaya katılan öğrencilerden kaçının mezun olmaya hak kazandığı ve kaçının dönem uzattığı ile ilgili veri elde ettiğimizi kabul edelim. Bu çalışmada akademik erteleme eğilimi ölçeği ile zaman yönetimi becerileri ölçeğinden alınan puanlar arasında negatif yönlü anlamlı korelasyon saptanması geçerliğe yönelik eş zamanlı kanıt sağlar. Akademik erteme eğilimi ölçeğinden alınan puanların mezun olan ve dönem uzatan öğrencileri büyük ölçüde doğru sınıflaması, yani ölçekten düşük puan alan öğrenciler genelde mezun olurken dönem uzatanların çoğunlukla ölçekten yüksek puan alanlar arasından olması geçerlik hakkında yordayıcı kanıt sunar.



Yordayıcı kanıtlara ulaşmak eş zamanlı geçerlik kanıtlarına kıyasla daha fazla vakit alır ve daha çok çaba gerektirir (Johnson & Christensen, 2017). Bu sebeple literatür incelendiğinde geçerliğe ilişkin yordayıcı kanıtların daha az raporlandığı görülür. Ancak unutmamak gerekir ki özellikle seçme ve yerleştirmeye yönelik sınavlarda, yordayıcı kanıtların incelenmesi sınavın amaca ne kadar hizmet ettiğine dair son derece önemli bilgiler verir.

4.1.3.2. Yakınsak ve Iraksak Kanıtlar

Geçerliği test edilen ölçme aracından alınan puanlar ile aynı özellik için farklı yöntemlerden ve/veya daha önce geçerliği kanıtlanmış başka bir ölçme aracından elde edilen puanlar arasında pozitif yönlü güçlü korelasyonlar bulunması yakınsak geçerlik kanıtı sağlar (Dougherty, 2019; Whitley & Kite, 2013). Iraksak geçerlik kanıtlarında ise hedef yapıdan farklı bir özelliği ölçtüğü bilinen dışsal bir değişken vardır. Geçerli sonuçlar verdiği daha önce ispatlanmış olan bu dışsal değişkene ait ölçümler ile hedef yapı için elde edilen puanlar arasında ilişki bulunmaması ya da zayıf ilişkiler olması geçerliğe dair ıraksak kanıtlar sunar (Aiken, 1996). Yakınsak ve ıraksak geçerlik kanıtları genellikle Campbell ve Fiske (1959) tarafından ileri sürülen çoklu özellik çoklu yöntem yaklaşımı ile incelenir. Yakınsak ve ıraksak kanıtları aynı anda değerlendirmeye imkân veren bu yaklaşımda birden fazla özellik (örneğin, genel zihinsel yetenek ve motivasyon) birden fazla yöntem (örneğin, öz bildirim ve akran değerlendirme) kullanılarak ölçülür. Daha sonra ölçümler arasındaki korelasyonlar hesaplanır ve bir korelasyon matrisine yerleştirilir (Cascio & Aguinis, 2019). Aynı özellik için farklı yöntemlerden, farklı özellikler için aynı yöntemden ve farklı özellikler için farklı yöntemlerden elde edilen korelasyon katsayılarını içeren bu matrisin esas köşegeni üzerine ise ölçümlere ait güvenilirlik katsayıları yazılır (bkz. Tablo 13).

Tablo 13

*Çoklu Özellik Çok Yöntem Yaklaşımı için Örnek Korelasyon Matrisi**

		Yöntem1		**Yöntem2**	
		Özellik-A	**Özellik-B**	**Özellik-A**	**Özellik-B**
Yöntem1	Özellik-A	r_{A1}			
	Özellik-B	IG_1	r_{B1}		
Yöntem2	Özellik-A	YG_1	IG_3	r_{A2}	
	Özellik-B	IG_2	YG_2	IG_3	r_{B2}

Korelasyon matrisinin farklı özellikler için aynı yöntemden elde edilen ölçümler arasındaki ilişkileri ortaya koyması, istenmeyen sistematik bir varyans olan yöntem etkisini de tespit etmeyi mümkün kılar (Dimitrov, 2012).

Rowe ve Mahar'dan (2006) uyarlanmıştır. YG: Yakınsak geçerlik, IG: Iraksak geçerlik

İki ayrı özellik iki farklı yöntem ile ölçüldüğünden Tablo 13'te esas köşegen üzerinde dört ayrı güvenirlik değeri (r_{A1}, r_{B1}, r_{A2} ve r_{B2}) mevcuttur. Tablo yorumlanırken önce güvenirlik kestirimleri kontrol edilmeli, ölçümlerin güvenirliğinden emin olunduktan sonra korelasyon katsayılarına bakılmalıdır. Hesaplanan korelasyonların geçerliğe kanıt oluşturabilmesi için (a) aynı özellik için farklı yöntemlerden elde edilen ölçümler arasında pozitif yönlü güçlü ilişkiler gözlenmeli, (b) farklı özellikler aynı yöntemle ölçülmüş olsa dahi bu özellikler arasında zayıf ilişkiler bulunmalıdır (Sullivan & Feldman, 1979).

Çoklu özellik çoklu yöntem yaklaşımı hakkında yukarıda sıralanan özellikleri bir örnek üzerinde pekiştirelim. İlkokul öğrencilerinde saldırganlık ile dijital bağımlılığı öz bildirim ve davranışsal gözlem şeklinde iki farklı yöntemle ölçtüğümüzü kabul edelim. Bu iki yapı için iki ayrı yöntemden elde ettiğimiz puanlar arasında Tablo 14'teki korelasyon katsayılarını hesapladığımızı varsayalım.

Tablo 14

*Saldırganlık ve Dijital Bağımlılık Verileri için Çoklu Özellik Çoklu Yöntem Yaklaşımı Korelasyon Matrisi**

Yöntem	Özellik	Öz Bildirim		Davranışsal Gözlem	
		Saldırganlık	Dijital Bağ.	Saldırganlık	Dijital Bağ.
Öz Bildirim	Saldırganlık	(0,91)			
	Dijital Bağ.	*0,44*	(0,89)		
Davranışsal Gözlem	Saldırganlık	**0,81**	0,25	(0,88)	
	Dijital Bağ.	0,28	**0,77**	*0,41*	(0,85)

* Örneğin oluşturulmasında ve yorumlanmasında Urbina (2004, s. 172) ile Finch ve French (2019, s. 210) referans alınmıştır.

Tablo 14'te esas köşegen üzerinde parantez içerisinde sunulan katsayılar iç tutarlılık yöntemiyle hesaplanan güvenirlik değerlerini göstermekte, tüm katsayıların 0,80'i aşması ölçümlerin yüksek güvenirlikte olduğunu simgelemektedir. Buna göre ölçümlere ait güvenirliğin değişkenler arasındaki korelasyonda düşüşe yol açması ihtimali zayıftır, tablodaki korelasyon katsayıları güvenle yorumlanabilir. Tablo 14'te daha koyu yazılan korelasyon değerleri, aynı özellik için iki farklı yöntemden elde edilen puanların ilişkisini ifade etmektedir. Bu korelasyonların pozitif yönlü ve güçlü olması beklenmekte, tablodaki katsayılar bu beklentinin karşılandığını yansıtmaktadır. İtalik biçimde sunulan korelasyon katsayıları farklı psikolojik yapılar için aynı yöntemden elde edilen puanlar arasındaki ilişkileri belirtmektedir. Bu korelasyonlar, beklendiği gibi, aynı özelliğin farklı yöntemlerle ölçüldüğü durumdaki katsayılardan (koyu yazılan değerlerden) daha düşüktür. Son olarak düz metin formatındaki korelasyon katsayıları farklı özellikler için farklı yöntemlerden gelen puanların ilişkisini yansıtmakta, matristeki en düşük değerlerin bunlara ait olması beklenmektedir. Tablo 14'teki korelasyonlar bu beklentiyle

de uyumludur. Bütün bulgular bir arada değerlendirildiğinde değişkenler arasındaki ilişkide hedef yapıdan kaynaklanan varyansın yüksek, yönteme atfedilebilecek varyansın ise düşük olduğu anlaşılmaktadır. Dolayısıyla çoklu özellik çoklu yöntem matrisindeki korelasyon katsayılarının geçerliğe ilişkin sunduğu yakınsak ve ıraksak kanıtlar oldukça güçlüdür denilebilir.



- Çoklu özellik çoklu yöntem yaklaşımı geçerlik hakkında oldukça kritik bilgiler sağlayabilmektedir. Bu potansiyeline rağmen veri toplama sürecindeki güçlükler nedeniyle nadiren kullanılmaktadır (Göçer Şahin, 2021).
- Geçerliğe ilişkin yakınsak kanıtlar ile eş zamanlı ve yordayıcı kanıtların hareket noktası benzerdir. Üçü de hedef yapıya ait kavramsal çerçevenin ölçme aracından alınan puanlarla ilişkili olduğuna işaret ettiği, geçerliği önceden ispatlanmış dışsal bir değişkeni ölçüt olarak alır. Yakınsak kanıtlarda bu dışsal değişken geçerliği araştırılan ölçme aracıyla *aynı* özelliği ölçen başka bir araçtan/yöntemden elde edilen puanlardır. Eş zamanlı ve yordayıcı kanıtlarda ise geçerliği sorgulanan ölçme aracıyla aynı değil *benzer* özelliği ölçen bir dışsal değişken vardır (Brain, 2016; Brown, 2023).

4.1.3.3. Karşıt (Bilinen) Gruplara Dayalı Kanıtlar

Bu yöntemde ölçme aracından alınan puanlarla ilişkisi incelenen değişken sınıflama düzeyindedir. Ölçme aracının bu sınıflama değişkeninin farklı kategorilerinde yer alan bireyleri birbirinden ayırt edip edemediği incelenir. Bu incelemelerde genelde bağımsız örneklemler *t*-testi ve varyans analizi gibi grup ortalamalarını karşılaştırmaya yönelik analizler kullanılır. Analiz sonuçlarının grupların puanları arasında istatistiksel açıdan anlamlı fark bulunduğunu ortaya koyması ve bu farkın kuramsal yapının işaret ettiği yönde çıkması geçerliğe kanıt oluşturur. Daha istatistiki terimler ile ifade edecek olursak bu yöntemde yönlü bir alternatif hipotez test edilir. Yapılan analizler yokluk hipotezinin reddedilip yönlü alternatif hipotezin kabulü yönünde bir kararla neticelenirse bu sonuç geçerlik için karşıt gruplara dayalı kanıtlar sağlar. Örneğin; Gezer ve İlhan (2020) iklim değişikliği bilgi testini öğretmen adaylarından oluşan bir örneklem üzerinde Türkçeye uyarladıkları araştırmada, "Coğrafya öğretmenliği programına kayıtlı öğretmen adaylarının test puanları, diğer bölümlerde okuyan öğretmen adaylarınınkine kıyasla anlamlı derece daha yüksektir" hipotezini test etmişlerdir. Bu amaçla tek yönlü varyans analizi uygulamış, analiz çıktılarının hipotezlerini doğrulamasını test puanlarının geçerliğine kanıt olarak göstermişlerdir.

Biçer'in (2024) dikkat eksikliği tanılamasında kullanılmak üzere materyal destekli test bataryası geliştirdiği çalışma, test puanlarının geçerliğinin karşıt gruplara dayalı kanıtlarla desteklendiği bir başka araştırmadır. Biçer (2024) çalışmasında üç grubun dikkat eksikliği envanterinin alt testlerinden aldıkları puanları karşılaştırmıştır. İlk

grupta çocuk psikiyatristleri tarafından dikkat eksikliği tanısı konmuş çocuklar yer almıştır. İkinci grup henüz dikkat eksikliği tanısı konmamış fakat sınıf öğretmenlerinin kendilerinde dikkat eksikliği bulunduğu izlenimini edindiği öğrencilerden oluşmuştur. Üçüncü grup ise dikkat eksikliği ile ilgili herhangi bir tanı almamış öğrencilerden meydana gelmiştir. Araştırma sonucunda beklendiği gibi dikkat eksikliği tanısı konmuş ya da tanı konmamış olmasına rağmen öğretmenlerinde bu yönde bir kanaat oluşmuş öğrenciler, dikkat eksikliği tanısı almamış öğrencilere kıyasla alt testlerin çoğunda daha düşük performans göstermiştir. Araştırmacı bu sonucu geliştirilen test bataryasının geçerliğine kanıt olarak yorumlamıştır.

4.1.4. Cevaplama Süreçlerine Dayalı Kanıtlar

Ölçme aracındaki maddelerin ölçmeyi hedeflediği yapı ile bireyin bu maddeleri cevaplarken kullandığı bilişsel süreçler arasındaki uyumdan hareketle incelenen bir geçerlik kanıtıdır (Engelhard & Winel, 2018). Birey cevaplama sırasında maddeyi yanıtlayabilmek için işe koşulması gerektiği öngörülen bilişsel süreçleri kullanıyorsa bu durum geçerliğe kanıt sağlar. Öte yandan madde, cevaplanabilmesi için gerekli olduğu düşünülen zihinsel süreçlere başvurulmadan, daha başka yollar kullanılarak da doğru yanıtlanabiliyorsa geçerliğin sağlanmadığı anlaşılır. Dolayısıyla cevaplama süreçlerini incelemeden maddeyi doğru yanıtlayan öğrencilerin cevaba ulaşabilmek için gereken bilişsel süreçleri kullandığını düşünmek bir varsayımın ötesine geçemez (Cizek, 2020).

Matematiksel akıl yürütme becerisini ölçmeye yönelik bir testi öğrencilere uyguladığımızı farz edelim. Öğrencilerin cevaplarını incelediğimizde akıl yürütme süreçlerinin işletildiğini görürsek testin amacına hizmet ettiğini, test puanlarının öğrencilerin akıl yürütme becerisini yansıttığını söyleyebiliriz. Ancak maddelerin ezber bilgiler, formüller ve bilindik algoritmalar takip edilerek cevaplandığını saptarsak test puanlarını akıl yürütme becerisinin göstergesi olarak yorumlayamayız ve bir geçerlik sorunu olduğu kanaatine varırız (Banks, 2012). Başka bir örnek üzerinden konuyu biraz daha somutlaştıralım. Çoktan seçmeli bir matematik testinde "4x – 3 = 2x + 17 ise x kaçtır?" maddesine yer vermiş olalım. Bu madde "Birinci dereceden bir bilinmeyenli denklemleri çözer" hedefini ölçmeye yöneliktir. Gelen cevapları incelediğimizde çoğu öğrencinin maddeyi denklem çözme adımlarını takip ederek değil seçeneklerdeki değerleri denklemde yerine koyarak yanıtladığını görürsek geçerliğin düşük olduğu çıkarımına varırız. Bu da ilgili maddeyi açık uçlu formatta ya da cevabı gizlenmiş çoktan seçmeli madde biçiminde (soru kökünde "*x kaçtır?*" yerine "*x'in rakamları toplamı kaçtır?*" demek gibi) sunduğumuzda geçerliği daha yüksek ölçümler elde edebileceğimizi gösterir.

Cevaplama süreçlerine dayalı geçerlik kanıtlarını toplamak için öğrencilerin cevaplarını oluştururken kullandıkları stratejileri açık biçimde görebilmek gerekir. Bu kapsamda bireylerden maddeleri yanıtlarken sesli düşünmeleri istenebilir (protokol analizi), cevaplayıcılar ile retrospektif görüşmeler gerçekleştirilebilir (Messick, 1989), göz izleme yöntemi kullanılabilir ve maddelerin cevaplanma süresi kontrol edilebilir (Bandalos,

2018). Ek olarak bireylerin sınav kâğıdı üzerinde yaptıkları işlemler ile tuttukları notlar analiz edilebilir ve açık uçlu maddelerden meydana gelen bir sınav uygulandıysa puanlamayı yapan kişilerden de görüş alınabilir (İlhan ve ark., 2024).

4.1.5. Ölçme Sonuçlarına Dayalı Kanıtlar

Testler çeşitli faydalar sağlayacağı düşüncesiyle geliştirilir ve bireylere uygulanır. Örneğin, bir işe alınacak kişilere karar vermek amacıyla yapılan bir sınavın kalifiye elemanların seçilmesini sağlayacağı ve vasıfsız kişilerin işe girmesini engelleyeceği varsayılır (Reynolds ve ark., 2021). Ancak uygulamada durum her zaman böyle olmaz, süreç yeterliliği düşük kişilerin işe alımıyla da sonuçlanabilir. Daha açık bir anlatımla, ölçme işlemleri istendik sonuçlar ortaya çıkarabileceği gibi beklenmedik bir takım çıktılar da doğurabilir. Bunlar geçerlikle ilgili önemli geri bildirimler sağlar (Miller ve ark., 2011). Ölçme sonuçlarına dayalı geçerlik kanıtları bu geri bildirimleri analiz edip ölçme işleminin kendisinden beklenen faydaları ne ölçüde sağladığını ve istenmeyen sonuçlara yol açıp açmadığını saptamaya çalışır (Sandilands & Zumbo, 2014). Geçerlik için ölçme sonuçlarına dayalı kanıtlar toplanırken aşağıdaki sorular üzerinden ilerlenir (Goodwin, 2002, akt: McDonald, 2018):

- Testten beklenen faydalar ne ölçüde gerçekleşiyor?
- Test sonucunda beklenmedik çıktılar (pozitif veya negatif) ne ölçüde meydana geliyor?
- Testi alan farklı alt gruplar için ne düzeyde farklı sonuçlar gözleniyor?

İlhan ve arkadaşları (2024) sıralanan soruları referans alarak ölçme sonuçlarına dayalı geçerlik kanıtlarını tıp alanından bir örnekle açıklamıştır. Buna göre nasıl ki bir aşının etkili olduğunu söyleyebilmek için hastalığı büyük oranda engellemesi ve yan etkilerinin olumlu etkilerinden ağır basmaması gerekiyorsa bir ölçme aracının geçerli ölçümler ürettiğini iddia edebilmek için kendisinden beklenen faydaları sağlaması, sebep olduğu herhangi bir olumsuz sonucun bulunmaması ya da olumsuz sonuçlarının minimumda kalması gerekir. Ayrıca aşının ar-ge süreçlerinde farklı gruplar (bebekler, hamileler, emziren kadınlar vb.) üzerindeki etkisini tanımlamak gerektiği gibi bir test uygulandığında da bu testin sınava giren farklı gruplarda ne kadar benzer çalıştığı sorgulanmalıdır.

Ölçme sonuçlarına dayalı geçerlik kanıtları tüm sınavlara uygulanabilir. Bununla birlikte özellikle seçme, yerleştirme, terfi edecekleri belirleme gibi amaçlarla yapılan sınavlarda bu geçerlik kanıtıyla ilgili incelemelerde bulunmak daha önemlidir (Reynolds ve ark., 2021). Çünkü ölçme sonuçlarına dayalı geçerlik kanıtları, sınavın geçerliği ile mevcut sistemin geçerliği arasında bir köprü vazifesi görür (Sireci, 2016).

Buraya kadar ele alınan diğer geçerlik kanıtları tüm ölçme uzmanları tarafından kabul görmektedir. Eğitsel ve psikolojik test standartlarına 1999 yılında giren bu geçerlik kanıtı (Gliner ve ark., 2017) hakkında ise psikometrisyenler farklı görüşler ileri sürmüştür. Uzmanların bir kısmına göre kapsama, içyapıya ve ölçme aracının diğer değişkenlerle ilişkisine dayanan geleneksel kanıtlar sağlanıyorsa geçerlik hakkında yeterince güçlü argümanlar mevcuttur ve daha farklı kanıtlar incelemeye gerek yoktur. Kimi uzmanlar ise buna karşı çıkmakta, diğer geçerlik kanıtları sağlansa bile bazı geçerlik problemlerinin hala ihtimal dâhilinde olabileceğini belirtmekte, ölçme sonuçlarının birey ve toplum üzerindeki etkileri ile sosyal ve politik sonuçlarına da bakmak gerektiğini savunmaktadır (Lovett, 2023).

Gerçekten de geçerlik hakkında değerlendirme yaparken ölçme sonuçlarına dayalı kanıtları dikkate almak son derece elzemdir. Şöyle bir örnek düşünelim. Öğretimi iyileştirmek, eksik öğrenmeleri gidermek ve hatalı öğrenmeleri düzeltmek amacıyla biçimlendirici değerlendirmeler yaptık, bu kapsamda öğrencilere belli aralıklarla konu ve tarama testleri uyguladık diyelim. Bu testlerden elde edilen sonuçların öğrenmeyi desteklemesi ve öğretimde iyileşme sağlaması biçimlendirici değerlendirme amacına hizmet eder ve geçerliğe kanıt oluşturur. Fakat yaptığımız testler öğretimi iyileştirmekten çok öğrencilerde strese neden oluyorsa ne kadar geçerli olduklarını sorgulamamız gerekir. Çünkü bu testler ölçmeye konu içeriği iyi temsil etse (kapsama dayalı yanıtlar), istatistiksel özellikleri bakımından gayet iyi olsa (içyapıya dayalı kanıtlar) ve ilişkili çıkması beklenen dışsal değişkenlerle anlamlı korelasyonlar gösterse (diğer değişkenlerle ilişkiye dayalı kanıtlar) dahi amacımıza hizmet etmemiştir. Bu da ortada bir geçerlik sorunu var demektir.



Cizek ve arkadaşları (2010) bilimsel kaynaklarda ölçme sonuçlarına dayalı geçerlik kanıtlarının ne sıklıkta sunulduğunu belirlemek amacıyla 2408 makale ile farklı kurumlar (AERA, APA ve NCME) tarafından yayınlanan 326 raporu incelemiştir. Çalışmanın sonuçlarına göre ne makalelerde ne de kuramsal raporlarda bu geçerlik kanıtına yer verilmiştir. Cizek ve arkadaşları (2010) bunu, ölçme sonuçlarına dayalı geçerlik kanıtlarını toplamanın zor olması ve araştırmacıların daha kolay elde edilebilecekleri geçerlik kanıtlarına yönelmesiyle açıklamıştır. Elbette tek neden bu değildir. Bir ölçme işleminin ya da testin toplumsal etkileri ile sosyal ve politik sonuçlarının görünür hale gelmesinin genellikle uzun sürmesi ve test geliştiricilerin bu kadar uzun süre bekleyecek vakitlerinin bulunmaması, test raporlarında ölçme sonuçlarına dayalı geçerlik kanıtlarından bahsedilmemesine yol açan bir başka faktördür (Lovett, 2023). Diğer kanıtlara kıyasla ölçme sonuçlarına dayalı geçerlik kanıtlarının nasıl toplanabileceğine dair kılavuzların sınırlı oluşu da (McIntire & Miller, 2007) bu geçerlik kanıtının göz ardı edilmesine neden olabilmektedir.

4.2. Geçerliği Artırmak için Alınabilecek Önlemler

Güvenirlik geçerliğin ön koşulu olduğundan güvenirliği destekleyen uygulamaların tamamı geçerliği de artırır. Ancak geçerlik, güvenirliğin aksine sabit ve sistematik hatadan da etkilendiğinden güvenirliği artıran uygulamalar arasında doğrudan yer almayan bazı düzenlemeler, ölçme aracının hedeflenen yapıyı daha doğru temsil etmesini sağlayarak geçerliği destekleyebilir. Aşağıda güvenirliği artırmaya yönelik önlemler arasında listelenmeyen ama geçerliği güçlendirecek uygulamalar maddeler halinde sıralanmıştır.

- ⇨ Hedef yapıyı ölçmeye hizmet edebilecek ölçme araçları kullanılmalıdır. Söz gelimi yazma becerisi çoktan seçmeli bir test ile yoklanırsa ulaşılan sonuçların geçerliği düşük olur.
- ⇨ Madde sayısı ve beraberinde sınav süresi ölçülecek kapsam dikkate alınarak belirlenmelidir. Mesela bir dersin vize sınavında 20-25 hatta bazen daha az sayıda madde ile geçerli sonuçlara ulaşılabilir. Ancak yükseköğretim kurumlarına yerleştirilecek öğrencileri belirlemek amacıyla yapılan bir sınavda oldukça geniş bir kapsam yoklandığından öğrenciler hakkında doğru kararlar alabilmek için sınavda daha fazla sayıda maddeye yer vermek gerekir.
- ⇨ Ölçme aracında herhangi bir grubun lehine çalışabilecek yanlı maddeler bulunmadığından emin olunmalıdır. Bir matematik sınavında olasılık konusuyla ilgili "*Bir satranç takımındaki taşlardan rastgele bir tane seçilmiştir. Bu taşın piyon olma olasılığı nedir?*" maddesine yer verildiğini varsayalım. Bu soru satrançta toplam kaç taş bulunduğunu, kaç tane piyon olduğunu bilmeyen öğrencileri dezavantajlı duruma getirir ve geçerliği düşürür. Dolayısıyla aynı yetenek düzeyindeki farklı alt gruplardan birine avantaj sağlayabilecek maddelerden uzak durulmalıdır.
- ⇨ Ölçülmesi amaçlanan özellik dışındaki faktörler puanlama işlemine karıştırılmamalıdır. Örneğin, açık uçlu maddelerden oluşan bir tarih sınavında yazı güzelliği/kâğıt düzeni gibi faktörler puanlamayı etkilememelidir.
- ⇨ Değerlendirmede kullanılan ölçüt doğru belirlenmelidir. Örneğin, bir yeterlik sınavında bağıl ölçüt kullanılması geçerliği düşük çıkarımlar ile sonuçlanır. Yine ölçme aracının diğer değişkenlerle ilişkisinin araştırıldığı geçerlik çalışmalarında ölçüt değişken doğru belirlenmiş ve yeterli güvenirlikte ölçülmüş olmalıdır. Ayrıca ölçülen özellikte ya da ölçüt değişkende ranj daralması sorunu bulunmamalı, diğer bir deyişle verilerin toplandığı örneklem hem ölçmeye konu özellik hem de ölçüt değişken açısından heterojen bir yapı göstermelidir.
- ⇨ Bilgisayar ortamında yapılan sınavlarda adayların bilişim teknolojilerine ilişkin okuryazarlıklarının test puanlarını etkilemeyeceği bir sistem oluşturulmalıdır. Örneğin, bir matematik sınavı çevrimiçi yapılıyor ve öğrencilerin matematiksel ifadelerin girildiği ara yüzü kullanma becerileri test puanlarını etkiliyorsa bu durum geçerliği düşürür (Wu ve arkadaşları, 2016). Bu anlamda çevrimiçi sınavlarda kullanımı özel bilgi ya da yeterlikler gerektirmeyen ara yüzler tercih edilmelidir.

5. KULLANIŞLILIK

Kullanışlılık ölçme aracının hazırlanması, çoğaltılması, uygulanması, puanlanması ve elde edilen puanların yorumlanması bakımından pratik ve ekonomik olması demektir (Tekin, 2009). Kullanışlılık üzerinde etkili olan bir dizi değişken bulunduğundan bir ölçme aracı bir yönüyle kullanışlıyken bir başka yönüyle kullanışlı olmayabilir. Bunu çoktan seçmeli testler ve yazılı sınavlar üzerinden örneklendirelim. Çoktan seçmeli maddelerden oluşan bir test geliştirirken çok sayıda madde yazmak, her bir madde için madde kökünü oluşturup seçeneklere karar vermek ve çeldiricileri (doğru cevap dışındaki seçenekleri) ölçülen özelliğe düşük düzeyde sahip olan kişileri kendisine çekebilecek şekilde yapılandırmak gerekir. Bu işlemler uzun zaman almakta ve çoktan seçmeli soru yazma deneyimi az olan madde yazarlarını zorlayabilmektedir. Yazılı sınavlarda ise daha az sayıda maddeye yer verilmekte ve bu maddeler için çeldirici yazma gibi bir iş yükü bulunmamaktadır. Tüm bunlar hazırlanışı itibariyle yazılı sınavları çoktan seçmeli testlere kıyasla daha kullanışlı bir araç yapmaktadır. Bu iki ölçme aracı puanlanışı açısından karşılaştırıldığında ise tersi bir durum söz konusudur. Çoktan seçmeli testler optik okuyucular tarafından otomatik olarak oldukça kısa sürede puanlanabilmektedir. Diğer taraftan yazılı sınavlarda çoktan seçmeli testlerdeki kadar gelişmiş bir otomatik puanlama sistemi bulunmamakta, puanlama işlemi daha uzun sürmekte ve eğer puanlayıcılara ücret ödeniyorsa maliyet de yükselmektedir. Bu sebeple çoktan seçmeli maddeler puanlanışı açısından yazılı sınavlara göre çok daha kullanışlıdır.

Uygulama koşulları kullanışlılık üzerinde etkili olan bir başka önemli unsurdur. Ölçme aracının aynı anda çok sayıda kişiye hızlıca uygulanabilmesi, uygulama süreçlerinin büyük ölçüde standart olması ve uzmanlık gerektirmemesi kullanışlılığını pozitif yönde etkiler. Öte yandan cevaplanmasının uzun sürmesi, bir seferde tek bir kişiye uygulanabilmesi ve uzman uygulayıcılara ihtiyaç duyulması ölçme aracının kullanışlılığını düşürür. Bu anlamda, grup testleri bireysel testlere kıyasla daha kullanışlıdır. Ayrıca uygulama süresi dikkate alındığında özellikle kalabalık gruplarda, çoktan seçmeli testler yazılı sınavlara, yazılı sınavlar da sözlü sınavlara göre daha kullanışlı olmaktadır.

Gerektirdiği bütçe de ölçme aracının kullanışlılığını etkilemektedir. Örneğin; üç sayfalık bir sınav kâğıdı yazı puntosu küçülterek, satır aralığı azaltılarak, şekil/grafik/tablo boyutları küçültülerek veya başka işlemlerle okunurluğu zedelenmeden önlü arkalı basılabilecek şekilde iki sayfaya sığdırılabiliyorsa basım/fotokopi maliyeti azalacak, bu durum ölçme aracını daha kullanışlı hale getirecektir. Yine kimi sınavlar elektronik ortamda uygulandığında maliyet düşebilmekte, dolayısıyla bazen elektronik uygulama basılı sınava kıyasla daha kullanışlı bir seçenek olabilmektedir. Ölçme aracında renkli basılması gereken maddeler/görseller yoksa gri tonlamalı basımın tercih edilmesi de maliyeti düşürüp kullanışlılığı artırılabilmektedir.

Madde sayısı kullanışlılık üzerinde kilit bir role sahip değişkenlerden biridir. Madde sayısını azaltmak uygulama süresini ve basım maliyetini aşağı çekerek kullanışlılığı artırır. Ancak kullanışlılık ile ilgili diğer tüm işlemler gibi madde sayısında da temel ilke, ölçme aracından çıkarılan maddelerin geçerlikte ve güvenirlikte kayba yol açmamasıdır. Evet, kullanışlılık ölçme araç ve sonuçlarının sahip olması gereken bir niteliktir ama geçerlik ve güvenirliğin önüne konulacak bir özellik değildir. Öncelikli hedef geçerli ve güvenilir ölçümler sağlayacak bir araca ulaşmak olmalı; kullanışlılık, geçerlik ile güvenirlikten taviz verilmeden artırılmaya çalışılmalıdır. İç tutarlılık katsayısı 0,92 olan ve 25 madde içeren bir ölçme aracı düşünelim. Bu ölçme aracındaki madde sayısını 20'ye düşürdüğümüzde hedef yapıyı yine tüm yönleriyle yoklayabiliyor ve iç tutarlılık güvenirliğinde belirgin bir düşüş gözlemiyorsak kullanışlılığı artırmak amacıyla madde sayısını azaltabiliriz. Fakat çıkardığımız maddelerin ardından bazı hedef davranışlar ölçme aracında temsil edilmiyorsa ve/veya güvenirlikte ciddi bir kayıp yaşanıyorsa madde sayısını azaltmamız yanlış bir karar olur. Kısacası işin odağında geçerli ve güvenilir sonuçlara ulaşmak bulunmalı, geçerlik ve güvenirlik üzerinde negatif bir etki yaratmadığı sürece daha az emek ve maliyet gerektiren ölçme araçları tercih edilmelidir (Akbaş, 2022).

Son olarak kullanışlılıkla ilgili özelliklerin teknolojik yeniliklerden etkilenmeye açık olduğuna değinmekte fayda vardır. Örneklendirmek gerekirse optik okuyucular geliştirilmeden önce çoktan seçmeli testler puanlanışı yönüyle bugünkü kadar kullanışlı bir araç değildi. Günümüzde, yapay zekâ teknolojileriyle birlikte açık uçlu maddelerin otomatik puanlanmasına yönelik önemli gelişmeler olduğu (bkz. Aydın ve ark., 2025; Demir, 2023; Sayın & Melanlıoğlu, 2024; Tate ve ark., 2024; Yiğiter & Boduroğlu, 2025) dikkate alındığında yakın tarihte açık uçlu maddelerden oluşan ölçme araçlarının puanlanışı itibariyle daha kullanışlı hale gelmesi muhtemel görünmektedir. Benzer şekilde yapay zekâ araçlarının madde üretme ve çeldirici yazma konusunda hızla gelişme kaydettiği (bkz. Barış Pekmezci & Şengül Avşar, 2023; Kıyak ve ark., 2024; Lin & Chen, 2024) göz önünde bulundurulduğunda teknolojideki ilerlemelerin hazırlanışı boyutuyla da ölçme araçlarının kullanışlılığını etkilemesi beklenmektedir.

6. SONUÇ

Ölçme araçları ile sonuçlarının niteliği hakkında değerlendirme yaparken güvenirlik, geçerlik ve kullanışlılık bütüncül olarak ele alınmalıdır. Güvenirlik ölçümlerin tutarlılığını ve hatalardan arınıklığını ifade ederken geçerlik ölçme sonuçları dikkate alınarak bireyler, kurumlar, öğretim programları ve/veya eğitim sistemleri hakkında verilen kararların ne kadar doğru olduğunu yansıtır. Kullanışlılık ise ölçme araçlarının ve sonuçlarının sahip olması gereken bu özelliklerin uygulanabilir bir zemine taşınmasıyla ilgilidir. Bu anlamda psikometrik açıdan güçlü ölçme sonuçları ile uygulamada tercih edilen ölçme araçlarına ulaşmanın yolu güvenirlik, geçerlik ve kullanışlılığın dengeli biçimde bir araya gelmesinden geçer. Böylece hem bilimsel araştırmalarda hem de uygulama ortamlarında sağlam temellere dayanan, daha doğru ve savunulabilir kararlar ile neticelenen ölçme ve değerlendirme süreçleri yürütülebilir.

KAYNAKÇA | REFERENCES

Aiken, L. R. (1996). *Assessment of intellectual functioning* (2. baskı). Springer.

Akbaş, U. (2022). Ölçme araç ve sonuçlarında bulunması gereken nitelikler. B. Çetin (Ed.), *Eğitimde ölçme ve değerlendirme* (2. baskı, s. 45–84) içinde. Anı.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (Ed.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, L.W. & Krathwohl, D.R. (2001). *A taxonomy for learning teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Arı, A. (2013). Bilişsel alan sınıflamasında yenilenmiş Bloom, SOLO, Fink, Dettmer taksonomileri ve uluslararası alanda tanınma durumları. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, *6*(2), 259–290. https://doi.org/10.12780/UUSBD164

Aybek, E. C. (2021). Ölçme araçlarının temel özellikleri: Güvenirlik. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 75–98) içinde. Nobel.

Aydın, B., Kışla, T., Tan Elmas, N. & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, *133*, 103784. https://doi.org/10.1016/j.system.2025.103784

Ayre, C. & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. The Guilford.

Banks, J. A. (Ed.) (2012). *Encyclopedia of diversity in education (Vol. 4).* Sage. https://doi.org/10.4135/9781452218533

Banks, S. R. (2012). *Classroom assessment: Issues and practices* (2. baskı.). Waveland.

Barış Pekmezci, F. & Şengül Avşar, A. (2023, 8–11 Haziran). *Madde yazımında yeni bir yol: Yapay zekâ teknolojisinin kullanımıyla otomatik madde üretimi* [Kongre sunum özeti]. 10. Uluslararası Eğitim Araştırmaları Kongresi, TED Üniversitesi, Ankara, Türkiye.

Başokçu, O. T. (2009). Ölçme sonuçlarının nitelikleri: Ölçme hatası, güvenirlik, geçerlik ve kullanışlılık. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 31–74) içinde. Pegem Akademi.

Başol, G. (2018). *Eğitimde ölçme ve değerlendirme* (5. baskı). Pegem Akademi.

Baumgartner, T. A., Hensley, L. D., Zhu, W. & Kulinna, P. H. (2021). *Conducting and reading research in kinesiology* (6. baskı). Jones & Bartlett Learning.

Biçer, M. (2024). *Dikkat eksikliği tanılamasında yeni yöntem: Materyal destekli test bataryası geliştirme* (Tez No. 890786). [Yüksek Lisans Tezi, Gaziantep Üniversitesi]. YÖK Ulusal Merkezi.

Brain, C. (2016). *Edexcel psychology for A level book 2*. Hodder Education.

Boztunç Öztürk, N. & Şahin, M. G. (2021). Bilimsel araştırmaya temel oluşturan ölçme kavramları. B. Çetin, M. İlhan & M. G. Şahin (Ed.), *Eğitimde araştırma yöntemleri: Temel kavramlar, ilkeler ve süreçler* (s. 1–34) içinde. Pegem Akademi.

Brown, C. (2023). *The evidence-based practitioner: Applying research to meet client needs* (2. baskı). F. A. Davis Company.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cascio, W. F. & Aguinis, H. (2019). *Applied psychology in talent management* (8. baskı). Sage.

Charter, R. A. (1999) Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1076/jcen.21.4.559.889

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cizek, G. J., Bowen, D. & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*(5), 732–743. https://doi.org/10.1177/0013164410379323

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. Sage.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Cumming, G. & Fidler, F. (2015). Confidence intervals better answers to better questions. *Journal of Psychology*, *217*(1), 15–26. https://doi.org/10.1027/0044-3409.217.1.15

Demir, S. (2023). Investigation of ChatGPT and real raters in scoring open-ended items in terms of inter-rater reliability. *International Journal of Turkish Educational Sciences*, *11*(21), 1072–1099. https://doi.org/10.46778/goputeb.1345752

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. American Counseling Association.

Domino, G. & Domino M. L. (2006). *Psychological testing: An introduction* (2. baskı) Cambridge University.

Dougherty, J. (2019). Collecting evidence. N. A. Schmidt & J. M. Brown (Ed.), *Evidence-based practice for nurses: Appraisal and application of research* (4. Baskı, s. 253–283) içinde. Jones & Bartlett Learning.

Earnest, J. & Gupta, S. K. (2022). *Outcome-based curriculum in engineering education*. Phi.

Engelhard, G., Jr. & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge/Taylor & Francis Group.

Evans, C. (2021, April 9). *Chris (Evans) R SAFAQ: Spearman-Brown formula*. https://www.psyctc.org/R_blog/posts/2021-04-09-spearman-brown-formula/

Fidler, F. & Cumming, G. (2019). Effect sizes and confidence intervals. G. R. Hancock, L. M. Stapleton & R. O. Mueller (Ed.), *The reviewer's guide to quantitative methods in the social sciences* (2. Baskı, s. 72–85) içinde. Routledge.

Finch, W. H. & French, B. H. (2019). *Educational and psychological measurement*. Routledge.

Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2012). *How to design and evaluate research in education* (8. baskı.). McGraw-Hill.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, *24*(3), 21–28. https://doi.org/10.1111/j.1745-3992.2005.00016.x

Furr, R. M. & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Sage.

Gezer, M. & İlhan, M. (2018, 18–22 Nisan). *İklim değişikliği bilgi testinin Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması*. 27. Uluslararası Eğitim Bilimleri Kongresi [Sözlü sunum], Antalya, Türkiye.

Gliner, J. A., Morgan, G. A. & Leech, N. C. (2017). *Research methods in applied settings: An integrated approach to design and analysis* (3. baskı.). Routledge.

Göçer Şahin, S. (2021). Ölçme araçlarının temel özellikleri: Geçerlik ve kullanışlılık. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 99–121). Nobel.

Hayes, A. F. & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. but…. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hogan, T. P. (2019). *Psychological testing: A practical introduction* (4. Baskı). Wiley.

Holmboe, E. S., Durning, S. J. & Hawkins, R. E. (2018). *Practical guide to the evaluation of clinical competence* (2. baskı). Elsevier.

İlhan, M. & Çetin, B. (2021). Nicel verilerin toplanması ve analizi. B. Çetin, M. İlhan & M. G. Şahin (Eds.), *Eğitimde araştırma yöntemleri: Temel kavramlar, ilkeler ve süreçler* (s. 161–200) içinde. Pegem Akademi.

İlhan, M., Güler, N. & Taşdelen Teker, G. (2024). *Puanlayıcı yargılarına dayalı ölçmelerde geçerlik ve güvenirlik*. Pegem Akademi.

Johnson, R. B. & Christensen, L. (2017). *Educational research: Quantitative, qualitative, and mixed approaches* (6. baskı). Sage.

Kalkbrenner, M. T. (2023). Alpha, omega, and H internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77–88. https://doi.org/10.1080/21501378.2021.1940118

Kaplan, R. M. & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications & issues* (9. baskı). Cengage Learning.

Kennedy, I. (2022). Sample size determination in test-retest and Cronbach alpha reliability estimates. British *Journal of Contemporary Education*, *2*(1), 17–29. https://doi.org/10.52589/BJCE-FY266HK9

Kettler, R. J. (2019). *Research methodologies of school psychology: Critical skills (Foundations of school psychology research and practice).* Routledge.

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ. & Uluoğlu, C. (2924). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, *80*, 729–735. https://doi.org/10.1007/s00228-024-03649-x

Kline, T. J. B. (2005). *Psychological testing: A practical aproach to design and evaluation*. Sage.

Koçak, D., Gülleroğlu, H. D. & Çokluk Bökeoğlu, Ö. (2022). *Ölçme aracı uyarlama el kitabı*. Pegem Akademi.

Köse, İ. A. (2012). Ölçmede güvenirlik. N. Çıkrıkçı Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 91–115) içinde. Edge Akademi.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Kurpis, S. E. R. & Stafford, M. E. (2006). *Testing and measurement: A user-friendly guide*. Sage.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. http://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lin, Z. & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, *123*, 103344. https://doi.org/10.1016/j.system.2024.103344

Lord, F. M. & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age.

Lovett, B. L. (2023). *Practical psychometrics: A guide for test users*. Guilford.

Mathews, P. (2010). *Sample size calculations: Practical methods for engineers and scientists*. Mathews Malnar & Bailey.

McCrae, R. R. & Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective* (2. baskı). The Guilford.

McCrae, R. R. & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McDonald, M. E. (2018). *The nurse educator's guide to assessing learning outcomes* (4. Baskı). Jones and Bartlett.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

McIntire, S. A. & Miller, L. A. (2007). *Foundations of psychological testing: A practical approach* (2. baskı). Sage.

Mertler, C. A. (2022). *Introduction to educational research*. Sage.

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* (3. baskı, s. 13–103) içinde. American Council on Education and Macmillan.

Miller, L. A., McIntire, S. A. & Lovler, R. L. (2011). *Foundations of psychological testing: A practical approach* (3. baskı). Sage.

Mitchell, J. E., Crosby, R. D., Wonderlich, S. & Adson, D. E. (2000). *Elements of clinical research in psychiatry*. American Psychiatric

Morrow Jr, J. R., Mood, D. P., Zhu, W. & Kang, M. (2023). *Measurement and evaluation in human performance* (6. baskı.). Human kinetics.

Oerman, M. H. & Gaberson, K. B. (2021). *Evaluation and testing in nursing education* (6. baskı). Springer.

Özer Özkan, Y. (2022). Ölçek geliştirme sürecinde ön inceleme maddeleri gözden geçirme ve düzeltme. M. Acar Güvendir & Y. Özer Özkan (Ed.), *Tüm yönleriyle ölçek geliştirme süreci* (s. 27–36) içinde. Pegem Akademi.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6. baskı.). Allyn & Bacon.

Popham, W. J. (2025). *Classroom assessment: What teachers need to know* (10. Baskı). Pearson.

Price, L. R. (2017). *Psychometric methods: Theory into practice*. The Guilford.

Reynolds, C. R., Altmann, R. A. & Allen, D. N. (2021). *Mastering modern psychological testing: Theory and methods*. Springer. https://doi.org/10.1007/978-3-030-59455-8

Rowe, D. A. & Mahar, M. T. (2006). Validity. T. M. Wood & W. Zhu (Ed.), *Measurement theory and pratice in kinesiology* (s. 9–26) içinde. Human Kinetics.

Salkind, N. J. & Frey, B. B. (2023). *Tests & measurement for people who (think they) hate tests & measurement*. Sage.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Hidalgo-Montesinos, M. D., Rosa-Alcázar, Á. I., Gómez-Conesa, A., López-Soler, C. & Gómez-García, F. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, 12, 407–570. https://doi.org/10.1002/jrsm.1487

Sandilands, D. & Zumbo, B. D. (2014). (Mis)alignment of medical education validation research with contemporary validity theory: The Mini-CEX as an example. Zumbo, B. & Chan, E. (Ed), *Validity and validation in social, behavioral, and health sciences* (s. 289–310) içinde. Springer. https://doi.org/10.1007/978-3-319-07794-9_17

Sayın, A. & Melanlıoğlu, D. (2024). Üniversite öğrencilerinin yazılı üretimlerinin OpenAI GPT ile değerlendirilmesi. *Yükseköğretim Dergisi*, *14*(3), 121–134. https://doi.org/10.53478/yuksekogretim.1418870

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Sireci, S. G. (2016). A theory of action for test validation. H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement* (s. 253–271) içinde. Information Age.

Suen, H. K. & Rzasa, S. E. (2004). Psychometricfoundations of behavioral assessment. S. N. Haynes & E. M. Heiby (Ed.), *Comprehensive handbook of psychological assessment* (s. 37–56) içinde. John Wiley & Sons.

Sullivan, J. L. & Feldman, S. (1979). *Multiple indicators: An introduction*. Sage.

Suter, W. N. (2012). *Introduction to educational research: A critical thinking approach* (2. baskı). Sage.

Swanson, J. L. (2012). Measurement and assessment in counseling psychology. E. M. Althamier & J.-I. C. Hansen (Ed.), *The Oxford handbook of counseling* (s. 208–236) içinde. Oxford University.

Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Education and Science*, *34*(152), 101–112.

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W. & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence, 7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Yargı.

Turgut, M. F. (1983). *Eğitimde ölçme ve değerlendirme metotları* (2. baskı). Saydam.

Urbina, S. (2004). *Essentials of psychological testing*. John Wiley & Sons.

Vacha-Haase, T., Kogan, L. R. & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*(4), 509–522. https://doi.org/10.1177/00131640021970682

Violato, C. (2019). *Assessing competence in medicine and other health professions*. CRC.

Warrens, M. J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. *Quantitative Psychology Research*, *89*, 293–300. https://doi.org/10.1007/978-3-319-07503-7_18

Whitley, Jr, B. E. & Kite, M. E. (2013). *Principle of research in behavioral science* (3. baskı) Routledge. https://doi.org/10.4324/9780203085219

Wilson, F. R., Pan, W. & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210. https://doi.org/10.1177/0748175612440286

Wu, M., Tam, H. P. & Jen, T. H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer. https://doi.org/10.1007/978-981-10-3302-5

Yiğiter, M. S. & Boduroğlu, E. (2025). Öğrencilerin el yazısıyla yanıtladığı açık uçlu maddelerin puanlanmasında yapay zekâ performansının incelenmesi. *Eğitim ve Bilim*, *50*(Ek Sayı 1), 1–18, https://doi.org/10.15390/EB.2025.14119

Yurdugül, H. (2005, 28–30 Eylül). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması* [Sözlü sunum]. XIV. Ulusal Eğitim Bilimleri Kongresi, Denizli, Türkiye.

Yurdugül, H. & Bayrak, (2012). Ölçek geliştirme çalışmalarında kapsam geçerlik ölçüleri. Kapsam geçerlik indeksi ve Kappa istatistiğinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *Özel Sayı 2*, 264–271.

YAZAR HAKKINDA | ABOUT AUTHOR

PROF. DR. MUSTAFA İLHAN

Dicle Üniversitesi | mustafailhan21[at]gmail.com

ORCID: 0000-0003-1804-002X

Prof. Dr. Mustafa İlhan, lisans eğitimini 2009 yılında Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, İlköğretim Bölümü, İlköğretim Matematik Öğretmenliği programında tamamlamıştır. 2011 yılında Dicle Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitim Bilimleri Anabilim Dalı'nda yüksek lisans derecesini; 2015 yılında Gaziantep Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitim Bilimleri Anabilim Dalı'nda doktora derecesini ve 2020 yılında Eğitimde Ölçme ve Değerlendirme alanında doçent unvanını almıştır. 2010 ile 2016 yılları arasında Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, Matematik ve Fen Bilimleri Eğitimi Bölümü, Matematik Eğitimi Anabilim Dalı'nda araştırma görevlisi olarak çalışmıştır, 2016 ile 2025 yıllarında arasında aynı bölümde öğretim üyesi olarak görevini sürdürmüştür. 2025 yılı Aralık ayından bu yana Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda profesör olarak görev yapmakta ve Eğitimde Ölçme ve Değerlendirme Anabilim Dalı Başkanlığı görevini yürütmektedir.

PROF. DR. MUSTAFA İLHAN

Dicle University | mustafailhan21[at]gmail.com

ORCID: 0000-0003-1804-002X

Prof. Dr. Mustafa İlhan received his bachelor's degree in Elementary Mathematics Education from Dicle University in 2009. He earned his master's degree in Educational Sciences from Dicle University in 2011 and his Ph.D. in Educational Sciences from Gaziantep University in 2015. He was awarded the title of Associate Professor in Educational Measurement and Evaluation in 2020. He served as a research assistant in the Department of Mathematics and Science Education, Division of Mathematics Education, Ziya Gökalp Faculty of Education, Dicle University, from 2010 to 2016. From 2016 to 2025, he served as a faculty member in the same department. Since December 2025, he has been a professor in the Department of Educational Sciences, Division of Educational Measurement and Evaluation, Ziya Gökalp Faculty of Education, Dicle University. He also serves as the Head of the Division of Educational Measurement and Evaluation.