JOURNAL OF RESEARCH IN EDUCATION AND SOCIETY
EĞİTİM VE TOPLUM ARAŞTIRMALARI DERGİSİ
ISSN: 2458 - 9624 (Online)

Eğitim ve Toplum Araştırmaları Dergisi/JRES, 4(1), 81-97, 2017

ÖLÇME VE ARAŞTIRMA YÖNTEMBİLİMİNDE ÇAĞDAŞ GELİŞMELER VE YENİ STANDARTLAR 2: GEÇERLİKTE ÜÇLEME (KAPSAM, ÖLÇÜT İLİŞKİLİ VE YAPI GEÇERLİKLERİ) ÖĞRETİSİNİN REDDİ VE GEÇERLİK KANITININ KAYNAKLARI

CONTEMPORARY DEVELOPMENTS AND NEW STANDARDS IN MEASUREMENT AND RESEARCH METHODOLOGY 2: REJECTION OF THE TRINITARIAN (CONTENT, CRITERION-RELATED, AND CONSTRUCT VALIDITIES) DOCTRINE IN VALIDITY AND SOURCES OF VALIDITY EVIDENCE

Vahit BADEMCİ[1]

[1] Gazi Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı. Ankara, Türkiye,
e-posta: bademci@gazi.edu.tr

Gönderim Tarihi: *06.06.2017* *Kabul Tarihi:* *13.06.2017*

Öz

Eğitimsel ve psikolojik testlerin geliştirilmesi ve değerlendirilmesine ilişkin *en* otoriter kaynak *Standards for Educational and Psychological Testing* adıyla yayımlanmaktadır ve kısaca, *Standartlar* olarak anılmaktadır. Bu çalışma geçerlikle ilgilidir. Burada, *Standartlar*dan geçerliğin evrimine ilişkin belgeler biçiminde yararlanıldı. *1966 Standartları*nda savunulan kapsam geçerliği, ölçüt ilişkili geçerlik, yapı geçerliği görüşü, *1999 Standartları*nda terk edildi. *1999 Standartları*nda, test içeriği üzerine temellenmiş kanıt, yanıt süreçleri üzerine temellenmiş kanıt, iç yapı üzerine temellenmiş kanıt, diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt ve test etmenin sonuçları üzerine temellenmiş kanıt biçiminde geçerlik kanıtının kaynakları tanımlandı. *Standartlar*ın son sürümü, 2014'de yayımlandı.

Anahtar Kelimeler: Geçerlik, geçerleme, yeni Standartlar, geçerlik kanıtının kaynakları, eğitimsel ve psikolojik test etme, Bademci'nin paradigma değişikliği.

Abstract

The *most* authoritative source regarding the development and evaluation of educational and psychological tests is published by name of the *Standards for Educational and Psychological Testing* and briefly referred to as the *Standards*. This study is related to validity. Here, the *Standards* were utilized as the documents regarding the evolution of validity. Advocated in the *1966 Standards*, the view of content validity, criterion-related validity, and construct validity was abandoned in the *1999 Standards*. In the *1999 Standards*, the sources of validity evidence were identified as evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing. The last edition of the *Standards* was published in 2014.

Atıf için Künye Bilgisi: Bademci, V. (2017). Ölçme ve araştırma yöntembiliminde çağdaş gelişmeler ve yeni standartlar 2: Geçerlikte üçleme (kapsam, ölçüt ilişkili ve yapı geçerlikleri) öğretisinin reddi ve geçerlik kanıtının kaynakları. *JRES*, *4*(1), 81-97.

Keywords: Validity, validation, new Standards, sources of validity evidence, educational and psychological testing, Bademci's paradigm shift.

Giriş: Geçerliğin Çağdaş Tanımı

Geçerlik, belirli bir evrene veya örnekleme uygulanan bir test ya da ölçme aracından elde edilen ölçümlerin kullanımlarının ve önerilen yorumlarının uygunluğunun ve yeterliğinin, kuram ve kanıt ile desteklenme derecesini ifade eder.

Geçerliğin, Bademci (2007, 2011a, 2016) tarafından yapılan çağdaş tanımı üstte bulunmaktadır ve geçerlikle ilgili bu makale de, American Educational Research Association (AERA), American Psychological Association (APA) ile National Council on Measurement in Education (NCME) tarafından *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014) adıyla yayımlanan ve *en otoriter* kaynak olarak kabul edilen ve bu çalışma içerisinde de yayımlandığı tarihle birlikte *Standartlar* (örneğin, *1999 Standartları*) olarak anılacak olan, *1999 Standartları* ve *2014 Standartları* paralelinde hazırlanmıştır.

Yüksek lisans ve doktora programlarında ölçme konularıyla ilgili yetersiz ve kalitesiz eğitimin bazı sonuçları

Geçerlik, eğitimsel ve psikolojik test etme (testing) alanındaki en önemli ve temel kavram ya da düşüncedir; ölçme ve de araştırma yöntembilimindeki bu değerine karşın geçerlik en fazla yanlış anlaşılan ya da hatalı kullanılan bir kavram olmayı da sürdürmektedir (AERA, APA, & NCME, 1999, 2014; Bademci, 2007; Frisbie, 2005; Rogers, 1995). Tam bu nokta da ifade edilebilir ki, ölçme sosyodavranışsal araştırmanın Aşil topuğudur (Pedhazur & Schmelkin, 1991) "yani, zayıf noktasıdır. Şüphesiz bu duruma, öncelikle lisansüstü programlar olmak üzere, yüksek lisans ve doktora programlarındaki ölçmeyle ilişkili konuların giderek azalmasının ve de ölçme konularıyla ilgili zayıf ve kalitesiz eğitim verilmesinin sebep olduğu söylenebilir" (Bademci, 2007, s. 88). Yaklaşık 30 yıl önce, Aiken ve arkadaşları (1990) tarafından yapılan bir araştırmada da doktora eğitim programlarındaki bu ölçme boşluğuna dikkat çekilmiştir.

> Türkiye'de de, "Doktora ve yüksek lisans programlarındaki bu ölçme boşluğu, yapılan bir araştırmayla da doğrulanmıştır: Ankara, Gazi, Hacettepe Üniversitelerinde 2000-2009 yılları arasında yapılmış olan 444 doktora ve yüksek lisans tezi *sadece güvenirlik hususu* açısından incelenmiş ve -ne yazık ki- **her 5 tezin 1'inde güvenirlik çalışması yapılmadığı** ve **10 tezin 8'inde** ise, **bilimsel olarak hatalı ve çağcıl olmayan kullanım ifadeleri bulunduğu** açıkça ortaya konulmuştur (Korkmaz, 2010)…" (Bulunduğu yer, Bademci, 2013, s. 17).

Çeşitli değişiklikler geçiren bir kavram ve kuram olarak geçerlik

Geçerlik kavramı ve de kuramı, 1920'lerden bu yana, yaklaşık 100 yılda sayısız değişiklikler geçirmiştir; yadsınamayan ve üzerinde fikir birliği meydana getirilen bu değişiklikler "evrim" (Anastasi, 1986; Shepard, 1993) ya da "başkalaşım [metamorphosis]" (Geisinger, 1992) olarak adlandırılmıştır (Angoff, 1988; Bademci, 2007, 2011a; Jonson & Plake, 1998; Rogers, 1995; Suen & Rzasa, 2004). Örneğin, 1930'larda bir geçerlik katsayısı vasıtasıyla ölçülen bir istatistik olarak geçerliğin görece basit görüşünden, bugün farklı epistemolojileri yansıtan çeşitli karmaşık modellere geçiş yapılmıştır; geçerlikle ilgili, muhtemelen hala da günümüz kitapları arasında en yaygın biçimde atıfta bulunulan ise, bazen 'üçleme' [trinitarian] (Guion, 1980) modeli ya da öğretisi olarak adlandırılandır; bu, *1966 Standartları*nda, yani 50 yıl öncede kalmış ve savunulmuş bir görüştür ve bu bakış açısında geçerlik, kapsam geçerliği, ölçüt ilişkili [yordayıcı, eşzamanlı] geçerlik ile yapı geçerliği biçiminde üç türe ayrılmıştır (APA, AERA, & NCME, 1966; Shepard, 1993; Suen & Rzasa, 2004).

Geçerlikte Üçleme (Kapsam Geçerliği, Ölçüt İlişkili Geçerlik ile Yapı Geçerliği) Öğretisinin Reddedilmesi

Standartlar, eğitimsel ve psikolojik testlerin [ölçme araçlarının] geliştirilmesi ve değerlendirilmesi ve test etme uygulamaları ve test ölçümlerinin ve önerilen yorumlarının niteliklerinin değerlendirilmesi hakkında profesyonel görüş birliğinin *en otoriter* ifadelerini ve kararlarını sağlamakta ve içermektedir (Linn, 2006; Sireci, Han, & Wells, 2008; Sireci & Parker, 2006).

Ayrıca, *Standartlar*ı, öncekinden bir sonraki sürümüne, geçerliğin evrimine ilişkin mevcut en sistemli belge ya da belgelendirme olarak da incelemek mümkündür (Jonson & Plake, 1998; Linn, 2010; Urbina, 2014). *Standartlar*ın ilk hali, *Technical Recommendations for Psychological Tests and Diagnostic Techniques* adıyla ve American Psychological Association (APA) tarafından 1954'te yayımlanmıştır; burada, *1954 Teknik Önerileri* olarak ifade edilecek olan çalışma içinde kapsam geçerliği, yordayıcı geçerlik, eşzamanlı [/uyum/uygunluk/zamandaş] geçerlik, yapı geçerliği olarak, geçerliğin dört türü listelenmiştir (APA, 1954). *Standards for Educational and Psychological Tests and Manuals* (APA, AERA, & NCME, 1966) adıyla yayımlanan *1966 Standartları*nda ise, geçerliğin üç ayrı ya da parçalı türü tanımlanmıştır; bunlar, yukarıda da belirtildiği üzere, kapsam geçerliği, ölçüt ilişkili geçerlik ve yapı geçerliğidir; burada, yordayıcı geçerlik ve eşzamanlı geçerlik birleşerek ölçüt ilişkili geçerliği meydana getirmiştir. Kısaltılmış halleriyle, APA, AERA ve NCME adlı üç

etkili profesyonel kuruluş, 'üçleme' öğretisinin (Guion, 1980; Shepard, 1993) savunulduğu *1966 Standartları*ndan bu yana, 1974'te, 1985'te, 1999'da ve 2014'te yeni standart takımları ürettiler.

*1954 Teknik Önerileri*nde ve *1966 Standartları*nda, geçerliğin testin belli amaçları başarabilme derecesini gösterdiğine vurgu yapılmıştır; *1974 Standartları*nda ise, geçerliğin bakış açıları biçiminde ölçüt ilişkili geçerlikler, kapsam geçerliği, yapı geçerliği tartışılmış, ayrıca geçerliğin test ölçümlerinden yapılan çıkarımların uygunluğuna işaret ettiği de belirtilmiştir (APA, 1954; APA, AERA, & NCME, 1966, 1974). Geçerliğin bütüncül ya da bölünmez bir kavram olduğunun yansıması ise, *1985 Standartları*dır ve bu standartlarda, geçerliğin üç türünden ziyade, kanıtın üç kategorisi (yapı ilişkili kanıt, içerik ilişkili kanıt, ölçüt ilişkili kanıt) listelenmiştir; ancak, geçerlik yine de parçalıdır (AERA, APA, & NCME, 1985; Algina & Penfield, 2009; Shepard, 1993). Ayrıca, *1985 Standartları*nda, geçerliğin test ölçümlerinden yapılan belirli çıkarımlara işaret ettiği ve geçerlenenin de, testin kendisi değil çıkarımlar olduğu ifade edilmiştir (AERA, APA, & NCME, 1985). Yine, *1985 Standartları*nda, geçerliğin bir testin özelliği olmaktan ziyade, test ölçümleri üzerine temellendirilmiş yorumların bir özelliğine doğru olan değişimini de görmek mümkündür (AERA, APA, & NCME, 1985).

*1966 Standartları*nda savunulan kapsam geçerliği, ölçüt ilişkili geçerlik ve yapı geçerliği şeklinde geçerliğin üç ayrı ya da parçalı türü olduğunu kabul eden 'üçleme' öğretisi, *1999 Standartları*nda tümüyle reddedilmiş ve terk edilmiştir; *1999 Standartları*nda, geçerliğin, testin kendisine değil, test ölçümlerinin yorumlarına işaret ettiği açıkça vurgulanırken, geçerliğin bütüncül ya da bölünmez bir kavram olduğu yine ifade edilmiş, ayrı bir önem verilerek, 1) test içeriği üzerine temellenmiş kanıt, 2) yanıt süreçleri üzerine temellenmiş kanıt, 3) iç yapı üzerine temellenmiş kanıt, 4) diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt, 5) test etmenin sonuçları üzerine temellenmiş kanıt şeklinde ve "geçerlik kanıtının kaynakları" başlığı altında, geçerlik kanıtının türleri de açıklanmıştır (APA, AERA, & NCME, 1966; AERA, APA, & NCME, 1999).

En son sürüm olan *2014 Standartları*nda, değerlendirilenin testin kendisi değil, önerilen kullanımlara yönelik test ölçüm yorumları olduğuna, *1999 Standartları*nda olduğu gibi testin geçerliği ifadesini kullanmanın doğru olmadığına, geçerliğin bütüncül veya bölünmez bir kavram olduğuna, *geçerleme* (validation) sürecinin önerilen ölçüm yorumlarına yönelik sağlam bilimsel bir temel sağlamak için ilgili kanıtı toplamayı içerdiğine, yine *1999 Standartları*ndaki gibi geçerlik kanıtının beş kaynağının 1) test içeriği üzerine temellenmiş kanıt, 2) yanıt süreçleri üzerine temellenmiş kanıt, 3) iç yapı üzerine temellenmiş kanıt, 4) diğer değişkenlerle

ilişkiler üzerine temellenmiş kanıt, 5) test etmenin sonuçları ve geçerlik için kanıt biçiminde olduğuna vurgu yapılmıştır (AERA, APA, & NCME, 2014).

Buraya kadar aktarılanlar toparlanırsa, günümüzde, artık, geçerliğin, testlerin kendileriyle değil, test ölçümlerinden yapılan yorumlarla ilgili olduğu hususunda çok geniş ve kapsamlı bir fikir birliği vardır (AERA, APA, & NCME, 1999, 2014; Cizek, 2016; Cronbach, 1971; Messick, 1989; Kane, 2006). Bundan başka, geçerlik, bütüncül ya da bölünmez bir kavramdır ve test ölçümlerinin önerilen yorumunun geçerliği kanıt ve kuram üzerine temellenmiştir; kapsam geçerliği, ölçüt ilişkili geçerlik, yapı geçerliği, şeklinde geçerliğin üç farklı ya da parçalı veya bölünmüş tipi olduğuna dair geleneksel bakış açısı çağdaş psikometristler ya da ölçme uzmanları tarafından eleştirilmiş ve reddedilmiştir ve bunun yerine de, geçerliği, çeşitli geçerlik kanıtı türlerine dayalı *bütüncül* bir kavram olarak ifade eden çağdaş görüş yerleşmiştir (AERA, APA, & NCME, 1999, 2014; Bademci, 2007, 2010; Cizek, 2016; Gronlund, 1998; Guion, 1980; Linn & Miller, 2005; Messick, 1989; Reynolds, Livingston, & Wilson, 2006; Shepard, 1993; Silva, 1993; Suen & Rzasa, 2004).

Bir diğer ifadeyle, çağdaş psikometristler ya da ölçme uzmanları 'bir test ilişkili olduğu herhangi bir şey için geçerlidir' (Guilford, 1946) görüşünü bırakmış, çok yaygın biçimde kabul edilen ve ana temasında 'geçerliğin, test ölçümlerinin kullanımlarının ve önerilen yorumlarının bir özelliği' (Kane, 2013) olan daha gelişmiş bir paradigmaya geçmiştir (Cizek, 2016; Suen & Rzasa, 2004).

Ayrıca, akıldan çıkarılmamalıdır ki, güncel sürümler olarak belirtilen *1999 Standartları* ve *2014 Standartları*, geçerliğin *bütüncül* ya da bölünmez bir kavram olduğunu ifade etmelerine rağmen, geçerliğin tümünü yapı geçerliği olarak tanımlamaktan *sakınmaktadır* (AERA, APA, & NCME, 1999, 2014; Sireci, 2009). Gözardı edilmemesi gereken bir başka husus ise şudur ki, geçerlik hakkındaki ifadeler, ölçümlerin belirtilen kullanımlarına yönelik önerilen yorumlarına işaret etmelidir (AERA, APA, & NCME, 1999, 2014).

*1954 Teknik Önerileri*nden, en son sürüm olan *2014 Standartları*na kadar, eğitimsel ve psikolojik testlerin ya da ölçme araçlarının geliştirilmesinin ve değerlendirilmesinin yanı sıra ölçme alanı bünyesinde en iyi uygulamaya dair görüş birliğini temsil eden *en otoriter* kaynak olarak kabul edilen ve de dünya çapındaki test, ölçme ve bellilendirme camiası içinde ufuk açıcı ve kilit bir rol oynayan *Standartlar*da (Linn, 2006; Moss, Girard, & Haniford, 2006; Sireci & Parker, 2006; Zumbo, 2014) geçerliğin geçirdiği değişiklikler ya da geçerliğin geçirdiği evrim, Tablo 1'de gösterilmiştir.

Tablo 1. Geçmişten Günümüze *Teknik Öneriler* ve *Standartlar*da Geçerliğin Evrimi©

1954 TEKNİK ÖNERİLERİ (APA, 1954)

Geçerliğin türleri

Kapsam geçerliği
Yordayıcı geçerlik
Eşzamanlı [/uyum/uygunluk/zamandaş] geçerlik
Yapı geçerliği

1966 STANDARTLARI (APA, AERA, & NCME, 1966)

Geçerliğin türleri

Kapsam geçerliği
Ölçüt ilişkili geçerlik (yordayıcı geçerlik ve eşzamanlı geçerlik)
Yapı geçerliği

1974 STANDARTLARI (APA, AERA, & NCME, 1974)

Geçerliğin bakış açıları

Kapsam geçerliği
Ölçüt ilişkili geçerlikler (yordayıcı geçerlik ve eşzamanlı geçerlik)
Yapı geçerliği

1985 STANDARTLARI (AERA, APA, & NCME, 1985)

Geçerlik kanıtının kategorileri

İçerik ilişkili kanıt
Ölçüt ilişkili kanıt (yordayıcı ve eşzamanlı desen/çalışma)
Yapı ilişkili kanıt

1999 STANDARTLARI (AERA, APA, & NCME, 1999)

Geçerlik kanıtının kaynakları -geçerlik, bütüncül ya da bölünmez bir kavramdır-

Test içeriği üzerine temellenmiş kanıt
Yanıt süreçleri üzerine temellenmiş kanıt
İç yapı üzerine temellenmiş kanıt
Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt
Test etmenin sonuçları üzerine temellenmiş kanıt

2014 STANDARTLARI (AERA, APA, & NCME, 2014)

Geçerlik kanıtının kaynakları -geçerlik, bütüncül ya da bölünmez bir kavramdır-

Test içeriği üzerine temellenmiş kanıt
Yanıt süreçleri üzerine temellenmiş kanıt
İç yapı üzerine temellenmiş kanıt
Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt
Test etmenin sonuçları ve geçerlik için kanıt

Yurt dışında ve Türkiye'de geçerlik ve geçerleme hususları etrafındaki bilimsel devrimin öncüleri

Yurt dışında, geçerlik kuram ve kavramının geçirdiği evrime ya da bir diğer anlatımla geçerlikteki paradigma değişikliğine ya da bilimsel devrime, başta Cronbach (1971, 1980, 1988) olmak üzere, Guion (1980), Kane (1990, 1992, 2006) ve geçerlik kuramcılarının en etkilisi olarak belirtilen Messick (1975, 1989, 1995) gibi isimlerin çeşitli çalışmalarıyla öncülük ettikleri söylenebilir (Bademci, 2011b). Türkiye'de ise, 60 yılı aşkın bir süre sonra, güvenirlikte olduğu gibi, 'geçerliğin, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliği olduğu' ana teması etrafında geçerlikte de ortaya koyduğu paradigma değişikliği ya da bilimsel devrim paralelinde, çağdaş geçerlik kavram ve kuramı ile ilgili *ilk* kuramsal makaleleri ve *ilk* bilimsel çalışmaları ve dünya çapındaki test ve ölçme camiası içinde kilit bir rol oynayan güncel *Standartlar*daki çağdaş gelişmeler üzerine -yurt içindeki- *ilk* çalışmaları da Bademci (1999, 2001a, 2001b, 2002, 2005, 2006, 2007, 2010, 2011a, 2011b, 2013, 2016; Gazi Haber, 2010) gerçekleştirmiş ve güvenirlikte olduğu gibi, geçerlikteki büyük değişime de, 20 yılı aşkın bir süredir, tek başına öncülük etmiş ve etmektedir (Bademci, 1999, 2005, 2007, 2010, 2011b, 2013, 2016; Gazi Haber, 2010).

Güncel *Standartlar* ve Geçerlik Kanıtının Kaynakları

Bir sefer daha altını çizerek belirtmek gerekirse, *Standartlar*ın güncel sürümleri olarak ifade edilen *1999 Standartları* ve en son *2014 Standartları*, geçerliğin *bütüncül* bir kavram olduğunu vurgulamalarına rağmen, geçerliğin tümünü yapı geçerliği olarak tanımlamaktan *sakınmaktadır* (AERA, APA, & NCME, 1999, 2014; Sireci, 2009). *1999 Standartları* ile *2014 Standartları*, geçerliğin "türlerine", "kategorilerine" ve "bakış açılarına" işaret etmekten ziyade, "geçerlik kanıtının kaynakları" üzerine temellendirilen bir geçerleme çerçevesini önermektedir (AERA, APA, & NCME, 1999, 2014; Sireci, 2009). *1999 Standartları* ve *2014 Standartları*, "belirli bir kullanım için test ölçümlerinin önerilen bir yorumunun geçerliğini değerlendirmede kullanılabilecek" (AERA, APA, & NCME, 2014, s.13; AERA, APA, & NCME, 1999) beş geçerlik kanıtının kaynağını ana hatlarıyla resmetmiştir: Bunlar;

1) Test içeriği üzerine temellenmiş kanıt,

2) Yanıt süreçleri üzerine temellenmiş kanıt,

3) İç yapı üzerine temellenmiş kanıt,

4) Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt ve

5) Test etmenin sonuçları üzerine temellenmiş kanıttır (AERA, APA, & NCME, 1999, 2014).

Test içeriği üzerine temellenmiş kanıt, testin içeriği ve testin ölçmeyi amaçladığı yapı arasındaki ilişkinin analizinden elde edilebilir; testin içeriği, konulara, ifade tarzına, uygulama ve ölçümlemeye ilişkin yönergelere, bir test üzerindeki sorulara ya da maddelere, görevlere, maddelerin biçimlerine ve çeşitlerine işaret eder (AERA, APA, & NCME, 1999, 2014; Reynolds, Livingston, & Wilson, 2006). *Test içeriği üzerine temellenmiş kanıt*, yapı ve testin bölümleri arasındaki ilişkiye dair uzman görüşlerinden, iş veya meslek analizlerinden, uzdaşma (alignment) çalışmalarından, vd. gelir (AERA, APA, & NCME, 1999, 2014; Sireci, 2009).

Yanıt süreçleri üzerine temellenmiş kanıt türü, sınava girenlerin fiilen meşgul olduğu yanıt veya erişimin (performance) ayrıntılı mahiyeti ve yapı arasındaki uyuma ilişkin kanıta işaret eder (AERA, APA, & NCME, 1999, 2014). *Yanıt süreçleri üzerine temellenmiş kanıt,* genellikle birey yanıtlarının analizlerinden gelmektedir; bunlar, test sorularına verdikleri yanıtları hakkında testi alanlarla görüşmeyi, test etme esnasındaki yanıt süreçlerinin niteliğine dair sesli düşünme (think-aloud) sözleşme tutanaklarını, vd. içerir (AERA, APA, & NCME, 1999, 2014; Creswell, 2012; Linn, 2010; Sireci, 2009).

İç yapı üzerine temellenmiş kanıt türü, deneyseldir ve istatistiksel analizlere işaret eder (Chatterji, 2003; Odendahl, 2011; Sireci, 2009). Bir testin iç yapısının analizleri test maddeleri ve test bileşenleri arasındaki ilişkilerin önerilen test ölçüm yorumlarının dayandırıldığı yapıya uyma derecesini gösterebilir; bir diğer ifadeyle, iç yapı analizleri, testteki farklı maddelere yönelik yanıtların ilişkilerinin önerilen test ölçüm yorumlarıyla tutarlılık derecesini ortaya koyabilir (AERA, APA, & NCME, 1999, 2014; Algina & Penfield, 2009; Linn, 2010; Reynolds, Livingston, & Wilson, 2006). *İç yapı üzerine temellenmiş kanıt*, madde ölçümlerinin etken çözümlemesini (faktör analizini), çok boyutlu ölçekleme (multidimensional scaling) işlemlerini, vd. içine alır (Algina & Penfield, 2009; AERA, APA, & NCME, 1999, 2014; Gall, Gall, & Borg, 2007; Sireci, 2009). Yapısal eşitleme modelini, Suen & Rzasa (2004) diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt kaynağı içine yerleştirirken, diğer çalışmalar ise, iç yapı üzerine temellenmiş kanıt altında göstermişlerdir (Algina & Penfield, 2009; Hubley, Zhu, Sasaki, & Gadermann, 2014; Osterlind, 2006; Urbina, 2014). Ayrıca, bu kanıt türü, yani iç yapı üzerine temellenmiş kanıt türü ile ilgili olarak, geçerlik kanıtı türetmede genellenirlik kuramı (generalizability theory) ya da test-tekrar test, değer biçiciler arası (interrater), Cronbach'ın alfası, Kuder-Richardson 20 [KR-20] gibi, diğer ölçüm güvenirlik göstergelerinden de faydalanılması önerilmektedir (Cizek, 2016; Guion & Highhouse, 2006; Osterlind, 2006; Urbina,

2014); ancak, bunların geçerlik kanıtı için ön şart biçiminde dikkatli faydalanılmaları gerektiği akılda tutulmalıdır. Zira, tam da bu noktada ve asla unutulmamalıdır ki, Cronbach'ın alfası tek boyutluluğun ya da homojenliğin bir ölçüsü *değildir*; Cronbach'ın alfası, tek boyutluluk sınandıktan ya da incelendikten <u>*sonra*</u> kullanılmalıdır (Bademci, 2014).

Birçok durumda, belirli bir kullanıma yönelik istenilen yorum, yapının bazı diğer değişkenlerle ilişkili olması gerektiğini ve sonuç olarak, test ölçümlerinin testin dışındaki değişkenlerle ilişkisinin analizlerinin bir diğer önemli geçerlik kanıtı kaynağını sağladığını kasteder; bu geçerlik kanıtının kaynağı, ***diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt*** şeklinde belirtilmiştir; başka bir anlatımla, test ölçümlerinin testin dışındaki değişkenlerle ilişkisinin analizleri, bir diğer önemli geçerlik kanıtının kaynağını sağlar (AERA, APA, & NCME, 1999, 2014). *Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt*, test-ölçüt ilişkilerini (yordayıcı ve eşzamanlı çalışmaları), yakınsak (convergent) ve ayrıçsak (discriminant) çözümlemeleri, geçerlik genelleme (validity generalization) çalışmalarını, vd. içine alır (AERA, APA, & NCME, 1999, 2014; Creswell, 2012; Gall, Gall, & Borg, 2007; Osterlind, 2006; Sireci, 2009).

Test kullanımının bazı sonuçları, testi geliştirenin istenilen kullanımlara yönelik test ölçümlerinin yorumundan doğrudan çıkmaktadır; *geçerleme* süreci, istenilen kullanımlara yönelik önerilen yorumların sağlamlığını değerlendirmek için kanıt toplar; bir başka ifadeyle, ***test etmenin sonuçları üzerine temellenmiş kanıt*** türü, bir test ya da test etme programı ile bağlantılı istenilen ve istenilmeyen sonuçların değerlendirilmesine işaret eder (AERA, APA, & NCME, 1999, 2014; Sireci, 2009). *1999 Standartları*nda "test etmenin sonuçları üzerine temellenmiş kanıt" biçiminde yer bulan ilgili bu kısım, *2014 Standartları*nda "test etmenin sonuçları ve geçerlik için kanıt" olarak başlıklandırılmıştır (AERA, APA, & NCME, 2014). *Test etmenin sonuçları üzerine temellenmiş kanıt*, bireysel/toplumsal sisteme yönelik istenilmeyen sonuçları, beklenilen faydalara dair çalışmaları, yasal haklar ve korunmalar ile uyumu ya da hukuka uygunluğu, öğretim üzerindeki etkileri, vd. içermektedir (AERA, APA, & NCME, 1999, 2014; Odendahl, 2011; Sireci, 2009; Suen & Rzasa, 2004).

Tablo 2. Geçerlik Kanıtının Kaynakları ve Geçerlik Kanıtı Türetmek İçin Kullanılan Bazı Özgün Yöntemler / Yaklaşımlar©

Test içeriği üzerine temellenmiş kanıt

İçerik / eğitim programı uzdaşma (alignment) çalışmaları
- Webb yöntemi
- Achieve yöntemi [Achieve, kar amaçsız bir örgütün adıdır]
- Yasal/uygulamadaki yetişeğin taranması/tetkik edilmesi yöntemi

İş (meslek) analizleri
Konu alanı uzmanı incelemesi
Test maddelerine ve test içerik belirtkelerine değerbiçimleme (rating)

Yanıt süreçleri üzerine temellenmiş kanıt

Sesli düşünme sözleşme tutanakları (think-aloud protocols)
Beyin-davranış ilişkilerinin bilgisi, nöropsikolojik bellilendirme (assessment) bünyesinde
Yanıtların zamanlaması (timing of responses)
Ölçümleme (scoring) ölçütlerinin uygulanmasında fikir birliği

İç yapı üzerine temellenmiş kanıt

Etken çözümlemesi (faktör analizi) ve ilişkili veri indirgeme yöntemleri,
- açınlayıcı (exploratory) ve doğrulayıcı (confirmatory) etken çözümlemeleri

Madde yanıt kuramı (item response theory)
Çok boyutlu ölçekleme
Küme analizi, temel bileşen analizi
Yapısal eşitleme modeli
Genellenirlik kuramı (generalizability theory) ya da diğer ölçüm güvenirlik göstergeleri
Farklı madde işgörüsü (differential item functioning) çalışmaları

Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt

Yakınsak (convergent) ve ayrıçsak (discriminant) kanıt,
- çoközellikli-çokyöntemli (multitrait-multimethod) çalışma

Test-ölçüt ilişkileri,
- yordayıcı (predictive) ve eşzamanlı (concurrent) çalışmalar,
- farklılaştırılmış gruplar çalışmaları

Geçerlik genellemesi (validity generalization)

Test etmenin sonuçları üzerine temellenmiş kanıt

Beklenilen faydalara dair çalışmalar
Bireysel/toplumsal sisteme yönelik istenilmeyen sonuçlar
Testi alana yönelik sonuçlar
İsabet oranı, hassaslık, özgüllük, yordayıcı güç
Yasal haklar ve korunmalar ile uyum ya da hukuka uygunluk
Öğretim üzerindeki etkiler

Geçerlik kanıtının kaynakları ve geçerlik kanıtı türetmek için kullanılan bazı özgün yöntemler veya yaklaşımlar ya da örnekler, çok ve çeşitli kaynaklardan faydalanılarak (AERA, APA, & NCME, 1999, 2014; Algina & Penfield, 2009; Bademci, 2016; Bonner, 2013; Chatterji, 2003; Cizek, 2016; Creswell, 2012; Furr & Bacharach, 2008; Gall, Gall, & Borg, 2007; Goodwin & Leech, 2003; Guion & Highhouse, 2006; Hubley, Zhu, Sasaki, & Gadermann, 2014; Kane, 2006; Linn, 2010; Nitko, 2001; Odendahl, 2011; Osterlind, 2006; Reynolds, Livingston, & Willson, 2006; Rothman, Slattery, Vranek, & Resnick, 2002; Sireci, 2009; Sireci & Faulkner-Bond, 2014; Sireci & Parker, 2006; Suen & Rzasa, 2004; Urbina, 2014), bir tablo halinde, Tablo 2'de sunulmuştur.

Sonuç Yerine

1999 Standartları ile *2014 Standartları*, belirli bir kullanıma yönelik test ölçümlerinin önerilen yorumunun geçerliğinin değerlendirilmesinde kullanılmak üzere, 1) test içeriği üzerine temellenmiş kanıt, 2) yanıt süreçleri üzerine temellenmiş kanıt, 3) iç yapı üzerine temellenmiş kanıt, 4) diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt ve 5) test etmenin sonuçları üzerine temellenmiş kanıt biçiminde, geçerlik kanıtının beş büyük kaynağına işaret eder (AERA, APA, & NCME, 1999, 2014; Linn, 2010). Bu çerçeve, bir geçerlik tartışmasında birden çok kanıt kaynağını kullanmak için teşvik eder, fakat bu, aşırı derecede kuralcı değildir, zira her geçerlik kanıtı bütün ortamlarda gerekli değildir; ancak, genellikle, belirli kullanımlar için önerilen yoruma yönelik yeterince destek birden çok kanıt kaynağını gerektirmektedir (AERA, APA, & NCME, 2014; Sireci, 2009). Araştırmanın farklı biçimleri vasıtasıyla 'üçgenleme' (triangulation) ve 'çapraz doğrulama' süreci geçerlik tartışmasını güçlendirmeye yardım etmelidir (Odendahl, 2011). Bu noktada bir örnek verilecek olursa, Ferrara & DeMauro (2006) yaptıkları bir çalışmada geçerlik kanıtının kaynaklarının tümünü, yani beşini de kullanmışlardır (Odendahl, 2011).

Kaynaklar

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. with Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist*, 45(6), 721-734.

Algina, J., & Penfield, R. D. (2009). Classical test theory. In R. Millsap, & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 93-122). Los Angeles: Sage.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association (APA) (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51, 201-238.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (APA, AERA, & NCME) (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (APA, AERA, & NCME) (1974). *Standards for educational & psychological tests*. Washington, DC: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1-15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer, & H. I. Braun (Eds.),*Test validity* (pp. 19-32). Hillsdale, New Jersey: Lawrence Erlbaum.

Bademci, V. (1999). *Türkiye'de eğitim fakülteleri ve öğretmen yetiştirme*. Panel. Düzenleyen: ESEF İşletme Araştırma Topluluğu. Ankara: G.Ü. Mesleki Eğitim Fakültesi Konferans Salonu, 21 Mayıs 1999.

Bademci, V. (2001a). *Düşünmenin öğretilmesi ve öğretimde kullanılan yöntemler-teknikler*. Konferans. Düzenleyen: TÜRMOB. Bursa: Bursa SMMM Odası Konferans Salonu, 9 Kasım 2001.

Bademci, V. (2001b). *Türkiye'deki okullar ne işe yarar?* Konferans. Düzenleyen: Ankara Türk Telekom Anadolu Teknik L. Ankara: Başkent Öğretmenevi Konferans Salonu, 9 Aralık 2001.

Bademci, V. (2002). *Türkiye'deki okullar ne işe yarar? Türkiye'nin anomi, yabancılaşma, ekonomik büyüme, demokratikleşme sorunlarına çözüm önerisi*. Konferans. Düzenleyen: ESEF Öğrenci Bilimsel Faal. Org. Kom. Ankara: G.Ü. Mesleki Eğitim Fakültesi Konferans Salonu, 30 Mayıs 2002.

Bademci, V. (2005). *Araştırmalarda ölçme ile ilgili bazı büyük hataları düzeltmek ve bir reformu başlatmak: Güvenirlik, testlerin bir özelliği değildir*. Eğitim Fakültelerinde Yeniden Yapılandırmanın Sonuçları ve Öğretmen Yetiştirme Sempozyumu. Ankara: Gazi Üniversitesi, Gazi Eğitim Fakültesi, 22-23-24 Eylül 2005.

Bademci, V. (2006). *Paradigma değişikliği: Testler güvenilir değildir*. Konferans. Düzenleyen: Gazi Üniversitesi, Endüstriyel Sanatlar Eğitim Fakültesi Dekanlığı. Ankara: G.Ü. Mesleki Eğitim Fakültesi Konferans Salonu, 28 Nisan 2006. [Konferansın bir kısmı ile ilgili haber için; *Gazi Haber*, Nisan 2006, Sayı 66, Sayfa 64.]

Bademci, V. (2007). *Ölçme ve araştırma yöntembiliminde paradigma değişikliği: Testler güvenilir değildir / Güvenirlik ve geçerlik üzerine çağdaş düşünceler: Araştırmada yöntembilimle ilgili bazı büyük hataların düzeltilmesi*. Ankara: Yenyap.

Bademci, V. (2010). *Türk eğitim ve biliminde paradigma değişikliği: Testler veya ölçekler güvenilir ve geçerli değildir*. Konferans. Düzenleyen: Gazi Üniversitesi, Endüstriyel Sanatlar Eğitim Fakültesi Dekanlığı. Ankara: G.Ü. Gazi Eğitim Fakültesi, Resim-İş Eğitimi Anabilim Dalı Konferans Salonu, 26 Nisan 2010. [Konferansın genel özeti şeklindeki ilgili haber için; *Gazi Haber*, Nisan 2010, Sayı 104, Sayfa 48-49.]

Bademci, V. (2011a). Türk eğitim ve biliminde bilimsel devrim: Testler ya da ölçme araçları güvenilir ve geçerli değildir. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 16, 116-132.

Bademci, V. (2011b). Kuder-Richardson 20, Cronbach'ın alfası, Hoyt'un varyans analizi, genellenirlik kuramı ve ölçüm güvenirliği üzerine bir çalışma. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 17, 173-193.

Bademci, V. (2013). *Yeni tez önerisi hazırlama kılavuzu*. Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Bademci, V. (2014). Cronbach's alpha is not a measure of unidimensionality or homogeneity. *Journal of Computer and Educational Research / Bilgisayar ve Eğitim Araştırmaları Dergisi*, 2(3), 19-27.

Bademci, V. (2016). *Ölçme ve araştırma yöntembiliminde çağdaş gelişmeler ve yeni standartlar 1: Geçerlik, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliğidir*. Yayına hazırlanmış makale.

Bonner, S. M. (2013). Validity in classroom assessment: Purposes, properties, and principles.In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 87-106). Los Angeles: Sage.

Chatterji, M. (2003). *Designing and using for educational assessment*. Boston: Pearson Education.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, 23(2), 212-225.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson Education.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight. In B. Schrader (Ed.), *New directions for testing and measurement. Measuring achievement: Progress over a decade* (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer, & H. I. Braun (Eds.), *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 579-621). Westport, CT: American Council on Education & Praeger.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, 24(3), 21-28.

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

Gazi Haber (2010). Türk eğitim ve biliminde paradigma değişikliği: Testler veya ölçekler güvenilir ve geçerli değildir. Sayı 104 (2010 Nisan), 48-49.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston: Pearson Education.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, 27(2), 197-222.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development*, 36(3), 181-191.

Gronlund, N. E. (1998). *Assessment in education* (6th ed.). Boston: Allyn & Bacon.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(4), 427-438.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional psychology*, 11(3), 385-398.

Guion, R. M., & Highhouse, S. (2006). *Essentials of personel assessment and selection*. New York: Lawrence Erlbaum.

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of validation practices in two assessment journals: Psychological Assessment and European Journal of Psychological Assessment. In B. D. Zumbo, & E. K. H. Chan (Eds.), *Validity, validation in social, behavioral, and health sciences* (pp. 193-213). New York: Springer.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, 58(5), 736-753.

Kane, M. T. (1990). *An argument-based approach to validation*. ACT Research Report Series, 90-13. Iowa City, Iowa: ACT.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education & Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50 (1), 1-73.

Linn, R. L. (2006). The Standards for Educational and Psychological Testing: Guidance in test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 27-37). Mahwah, New Jersey: Lawrence Erlbaum.

Linn, R. L. (2010). Validity. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education, Volume 4* (pp. 181-185). Oxford: Elsevier.

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.).Upper Saddle River, New Jersey: Pearson.

Messick, S. (1975). The Standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan Publishing Company.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741-749.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, 30(1), 109-162.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Odendahl, N. V. (2011). *Testwise*. Lanham: Rowman & Littlefield Education.

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of menta lappraisal*. Upper Saddle River, New Jersey: Pearson.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, New Jersey: Lawrence Erlbaum.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). *Measurement and assessment in education*. Boston: Pearson.

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction*. Pasific Grove, California: Brooks/Cole.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Silva, F. (1993). *Psychometric foundations and behavioral assessment*. Newbury Park, California: Sage.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The Concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema,* 26(1), 100-107.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, 13, 108-131.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, 25(3), 27-34.

Suen, H. K., & Rzasa, S. E. (2004). Psychometric foundations of behavioral assessment. In S.N. Haynes, & E. M. Heiby (Eds.), M. Hersen (Series Ed.), *Comprehensive handbook of psychological assessment, Volume 3* (pp. 37- 56). Hoboken, New Jersey: John Wiley & Sons.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Hoboken, New Jersey: Wiley.

Zumbo, B. D. (2014). What role does, and should, the test *Standards* play outside of the United States of America? *Educational Measurement: Issues and Practice*, 33(4), 31-33.