



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



*Eğitim ve Toplum Araştırmaları Dergisi/JRES, 6(2), 373-385, 2019*

## GEÇERLİK: NEDİR? NE DEĞİLDİR?

### VALIDITY: WHAT IS IT? WHAT IS IT NOT?

Vahit BADEMCİ<sup>1</sup>

<sup>1</sup>Gazi Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara, Türkiye.

e-posta: bademci@gazi.edu.tr

*Gönderim Tarihi: 14.11.2019*

*Kabul Tarihi: 20.11.2019*

#### Öz

Ölçme, bilimsel araştırmanın merkezinde, geçerlik ise, ölçmenin kalbindedir. Geçerlik, ölçmedeki en önemli kavram olarak kabul edilmektedir. Bu önemine rağmen, geçerlik, hâlâ çok yanlış anlaşılmakta ve hatalı kullanılmaktadır. Geçerlik kavram ve kuramı, son 100 yılda sayısız değişiklikler, bir diğer ifadeyle evrim geçirmiştir. En köklü değişimler ise, geçerliğin, test ölçümlerinden yapılan yorumlarla ilgili olduğunu vurgulayan çalışmalarla 1980'lerin başlarında meydana gelmiştir. Değişim sürecinin doruk noktası olan *1999 Standartlarında* ise, kapsam geçerliği, ölçüt ilişkili geçerlik ile yapı geçerliği şeklindeki geçerliğin üç parçalı türü reddedilmiş ve geçerliği, çeşitli geçerlik kanıt türlerine dayalı bütüncül bir kavram olarak ortaya koyan çağdaş görüş sergilenmiştir. 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 derecesidir. Bunun yanı sıra, güvenilirlik, belirli bir evrene veya örnekleme uygulanan bir test ya da ölçme aracından elde edilen ölçümlerin tutarlılığı veya tekrarlanabilirliğidir.

*Anahtar Kelimeler: Geçerlik, geçerleme, geçerlik kanıtının kaynakları, güvenilirlik, test etmede yeni standartlar, eğitimsel ve psikolojik test etme.*

#### Abstract

Measurement is at the core of scientific research and validity is at the heart of measurement. Validity is accepted as the most important concept in measurement. Despite this importance, validity is still misunderstood and misused. The concept and theory of validity have undergone numerous changes, in other words, evolution, in the last 100 years. The most radical changes occurred in the early 1980s, with studies emphasizing that validity is related to interpretations of test scores. In the *1999 Standards*, which was the culmination of the change process, the three-part type of validity based on content validity, criterion-related validity, and construct validity was rejected, and the contemporary view was put forward, arguing validity as a unitary concept based on various types of validity evidence. Validity is the degree to which evidence and theory support the adequacy and appropriateness of the proposed interpretations and uses of the scores obtained from the test or measurement instrument applied to a particular population or sample. Besides, reliability is the

reproducibility or the consistency of the scores obtained from the test or measurement instrument applied to a particular population or sample.

*Keywords: Validity, validation, sources of validity evidence, reliability, new standards in testing, educational and psychological testing.*

---

## Giriş

Ölçme, bilimsel araştırmanın merkezinde, geçerlik ise, ölçmenin tam kalbindedir (Bademci, 2013b, 2017a, 2019; Viswanathan, 2005). Ölçme alanı canlı ve hareketlidir ve alandaki değişim, gerçekten çok hızlıdır; son 100 yıllık sürede birçok kez değişen geçerlik ise, ölçmedeki en önemli kavram olarak kabul edilmektedir (Bademci, 2017b; Cizek, 2008; Rowe & Mahar, 2006). Öyle ki, test etme (testing) alanında geçerlikten daha önemli bir kavram bulmak da mümkün değildir (Rogers, 1995). Bu önemine rağmen, bugün, pek çok kişi tarafından, geçerlik, hâlâ çok yanlış anlaşılakta ve hatalı kullanılmaktadır (Frisbie, 2005; Osterlind, 2006). Mevcutlarının düzeltilmesi ve olası yeni yanlış anlamaların ve hatalı kullanılmaların önüne geçilmesi amacıyla, ilk önce, ölçmenin en temel kavramlarından olan güvenilirlik ile geçerliğin Bademci (2007, 2011a, 2017a, 2019) tarafından yapılan güncel tanımları aşağıda sunulmuştur.

### Güvenirlik ve Geçerliğin Çağdaş Tanımları

Güvenirlik, belirli bir evrene veya örnekleme uygulanan bir test ya da ölçme aracından elde edilen ölçümlerin tutarlılığı veya tekrarlanabilirliğidir (Bademci, 2007, 2011a, 2017a, 2019).

Geçerlik, belirli bir evrene veya örnekleme uygulanan bir test ya da ölçme aracından elde edilen ölçümlerin kullanılmalarının ve önerilen yorumlarının uygunluğunun ve yeterliğinin, kuram ve kanıt ile desteklenme derecesidir (Bademci, 2007, 2011a, 2017a, 2019).

### Geçerlik ile İlgili Gerçekler

Yukarıdaki çağdaş tanımının da dikkat çektiği ve vurguladığı üzere, geçerlik, ölçümlerin kullanılmalarının ve önerilen yorumlarının bir özelliğidir; yani, geçerlik bir testin ya da ölçme aracının kendisinin bir özelliği değildir (Bademci, 1999, 2007, 2017a, 2019; Furr & Bacharach, 2008; Kane, 2013; Koretz, 2008; Messick, 1995; Worthen, White, Fan, & Sudweeks, 1999). Bundan 45 yıl önce de etkili geçerlik kuramcılarında Guion (1974), bir testin geçerliğinden bahsetmek aptalcadır derken, öncü geçerlik kuramcılarında Cronbach (1989) ise, tam 30 yıl önce, A Testini geçerli ya da B Testini geçersiz diye adlandırmanın mantıksız olduğunu

belirtmiştir. Dolayısıyla, “testin geçerliği”, “test geçerlidir”, “ölçeğin geçerliği”, “ölçme aracının (veya yönteminin) geçerliği”, “ölçme prosedürü geçerlidir”, “bu deneyin geçerliği”, “bellilendirme (assessment) geçerliği”, “ölçümleyicilerin geçerliği”, “gözlemin geçerliği”, “sınavın geçerliği”, “test ölçümlerinin geçerliği”, “ölçümler geçerlidir” vb. hatalı ifadeler kesinlikle kullanılmamalıdır (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014; Bademci, 1999, 2007, 2017a, 2019).

Yüksek lisans ve doktora tezleri de dahil Türkiye’de yapılan neredeyse tüm çalışmalarda “testin geçerliği”ne işaret eden, örneğin, “ABC Okuma Testi geçerli midir?” benzeri sorular sorulması ve bu doğrultuda cevaplar aranılması da doğru değildir ve asla kabul edilmemelidir. Onun yerine, “okuduğunu anlamayı ölçerken ABC Okuma Testinden elde edilen ölçümlerin yorumu geçerli midir?” ve benzeri ölçüm (kullanım ve) yorum geçerliğine işaret eden daha belirli sorular sorulmalı, yanıtlar bulunulmalıdır (Bademci, 2017a; Nitko, 2001; Reynolds & Livingston, 2012).

Güvenirlilik ile ilgili çalışmalarda ‘ölçüm güvenirliliği’, geçerlik ile ilgili çalışmalarda ise, ‘ölçüm yorum geçerliği’ gibi ifadelerin, hem çalışma başlıklarında, hem de metnin içinde kullanılmaları, bu hususlardaki yanlış anlamaları önlemede etkili olacaktır (Bademci, 2013a). Akılda tutulmalıdır ki, ‘ölçüm yorum’ kavramı, geçerlik hakkındaki güncel ve çağdaş düşüncenin merkezindedir (Bademci, 2013a, 2017a; Cizek, 2012).

Günümüzde, geçerliğin, testlerin kendileriyle değil, test ölçümlerinden yapılan yorumlarla ilgili olduğu hususunda çok geniş ve kapsamlı fikir birliği vardır (AERA, APA, & NCME, 1999, 2014; Cizek, 2016; Cronbach, 1971; Kane, 2006; Messick, 1989). Bu uzlaşmanın kalbinde ise, ‘test ölçümlerinin yorumu geçerlidir’ temel düşüncesi yatmaktadır (Cronbach, 1971; Newton, 2012). Kısaca, geçerlik, test ölçümlerine bağımlı yorumların bir özelliğidir (Reynolds, Livingston, & Wilson, 2006). Tam da bu noktada, 1999 Standartları ile 2014 Standartlarında açıklığa kavuşturulmuş olan bir hususun vurgulanmasında yarar görülmektedir; test ölçüm kullanımı olmadan test ölçüm yorumuna sahip olunamayacağı gerçeği asla gözden kaçırılmamalıdır (AERA, APA, & NCME, 1999, 2014; Bademci, 2017a; Sireci, 2016). Yeri gelmişken açıklanmasında fayda var ki, American Educational Research Association, American Psychological Association ile National Council on Measurement in Education tarafından *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014) adıyla yayımlanan ve bu çalışma içerisinde de yayımlandığı tarihte birlikte

(örneğin, 1999 *Standartları*) anılan *Standartlar*, eğitimsel ve psikolojik testlerin geliştirilmesi, değerlendirilmesi ve test etme uygulamaları hakkında en otoriter kaynak olarak kabul edilmektedir (Bademci, 2017b; Linn, 2006; Sireci & Parker, 2006).

Geçerlik, bir derece meselesidir; yani, geçerlik, bir varlık-yokluk ya da hep-hiç kavramı değildir ve de test ölçümlerinin önerilen bir yorumuyla ilgili olarak ‘geçerlidir’ ya da ‘geçersizdir’ şeklindeki iki değerli kullanımlardan da kaçınılmalıdır (Bademci, 2017a; Furr & Bacharach, 2008; Koretz, 2008; Linn & Gronlund, 2000; Nunnally & Bernstein, 1994). Ölçümlerin yorum geçerliği, “yüksek geçerlik”, “orta geçerlik”, “düşük geçerlik”, “geçerlik yok” gibi belirli dereceler ile ifade edilmelidir; o, [ölçümlerin kullanımları ve önerilen] yorumlar geliştikçe ve yeni kanıt ya da kanıtlar biriktikçe zaman içinde değişebilir (Bademci, 2017a; Gronlund, 1998; Kane, 2013; Linn, 2010; Linn & Gronlund, 2000).

Güvenirlilik gibi, geçerlik de evren ya da örneklem bağımlıdır; yani, geçerlik, daima belirli bir evrene veya örnekleme özgüdür (Bademci, 2011a, 2017a; Linn & Miller, 2005). Örneğin, ABC Okuma Testi ölçümleri bir okulun okuma programını değerlendirmede kullanılırken önerilen yorum yüksek geçerliğe, bir başka okul için orta geçerliğe, bir diğer okul için ise, düşük geçerliğe sahip olabilir (Bademci, 2010, 2017a; Nitko, 2001).

Geçerlik, bir değerlendirme tartışmasıdır ve değerlendirci bir yargı içerir; bu değerlendirci yargı, önerilen yorumun uygunluğunu ve yeterliğini ve de yorumun yeterliğinin uygun (kuram ve) kanıtlarla ne derece desteklendiğini yansıtmalıdır (Bademci, 2017a; Kane, 2001; Linn & Miller, 2005; Osterlind, 2006). Bu noktada, gözden kaçırılmaması gereken bir hususun bir örnekle açıklanmasında fayda bulunmaktadır: ABC Okuma Testi ölçümleri, öğrencilerin ruh hâlleri gibi konu dışı faktörler tarafından etkilendi mi? ABC Okuma Testi ölçümleri güvenilir mi? benzeri sorular paralelinde, okuma ile ilgili tüm hususlardan uygun ve yeterli kanıtlar toplanılmadan ve gözden geçirilip birleştirilmeden, okuduğunu anlamayı ölçerken ABC Okuma Testinden elde edilen ölçümlerin yorum geçerliği hakkında bir karar ya da yargıya da asla varılmamalıdır (Bademci, 2010, 2017a, 2019; Nitko, 2001).

Geçerlik kuramının felsefi temelleri de yıllar içinde değişmiştir: 1920’lerde başlayan ve ‘geçerliğin testin bir özelliği’ olduğunu vurgulayan geçerlik üzerine geleneksel psikometrik bakış açıları olguculuğa (pozitivizme) veya olguculuk sonrasına köklendirilmiştir; ancak, 1970’lerden, özellikle 1980’lerden bu yana, ‘geçerliğin ölçümlerden yapılan yorumların bir özelliği’ olduğuna işaret eden çağdaş geçerlik kuram ve geçerleme (validation) uygulamaları

ise, yapılandırmacılık (konstrüktivizm) ya da sosyal yapılandırmacılık tarafından güçlü biçimde etkilenmiştir (Bademci, 2017a; Bonner, 2013; Gipps, 1999; Messick, 1989, 1998; Moss, 2003; Sijtsma, 2009). Bu hususta, çok etkili bir geçerlik kuramcısı olan Kane'in (1992, 2004, 2006) çalışmalarıyla, geçerliğin, gerçekçilikten (realizmden), uygulayıcılığa (pragmatizme) taşındığını belirten bir başka görüş de mevcuttur (Markus & Borsboom, 2013).

### ***1999 Standartları: Geçerliğin Kutsal Üçlününün (Kapsam Geçerliği, Ölçüt İlişkili Geçerlik, Yapı Geçerliği) Reddedilmesi ve Üçleme Öğretisinin Terkedilmesi***

1980'lerin başlarında, kapsam geçerliği, ölçüt ilişkili [yordayıcı, uyum] geçerlik ve yapı geçerliği şeklindeki geçerliğin üç parçalı türü, önemli geçerlik kuramcılarında Guion (1980) tarafından 'kutsal üçlü' metaforuyla ağır şekilde eleştirilmiş ve geçerliğin bütüncül (unitary) bir kavram olduğuna da işaret edilmiştir. Geçerlik, bütüncül ya da bölünmez bir kavramdır ve geçerlik kavramındaki bu büyük değişiklik, *1985 Standartlarında* da, kendini göstermiştir (AERA, APA, & NCME, 1985; Algina & Penfield, 2009; Bademci, 1999, 2007, 2017b, 2019).

*1966 Standartlarında* savunulan kapsam geçerliği, ölçüt ilişkili [yordayıcı, uyum] geçerlik ve yapı geçerliği şeklinde geçerliğin üç parçalı türü olduğunu kabul eden "üçleme" öğretisi ya da "geçerliğin kutsal üçlüsü" ise, *1999 Standartlarında* reddedilmiş ve terk edilmiştir (APA, AERA, & NCME, 1966; AERA, APA, & NCME, 1999; Bademci, 1999, 2007, 2017b, 2019; Guion, 1980; Silva, 1993). Geçerliği, çeşitli geçerlik kanıt türlerine dayalı bütüncül bir kavram olarak ortaya koyan çağdaş görüşü sergileyen *1999 Standartlarında* ise, "geçerlik kanıtının kaynakları" başlığı altında, geçerlik kanıtının türleri, 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 açıklanmıştır (AERA, APA, & NCME, 1999). *1999 Standartlarında* üretilen standart takımları, *2014 Standartlarında* da aynen ve iyileştirilmiş biçimde sürdürülmektedir (AERA, APA, & NCME, 2014).

### **Geçerlik Kanıtının Kaynakları**

*1999 Standartları* ve en son *2014 Standartları*, geçerliğin bütüncül bir kavram olduğunu ifade etmelerine rağmen, geçerliğin tümünü yapı geçerliği olarak tanımlamaktan sakınmakta ve "geçerlik kanıtının kaynakları" üzerine temellendirilen bir geçerleme çerçevesini önermektedir (AERA, APA, & NCME, 1999, 2014; Sireci, 2009). Belirli bir kullanım için test ölçümlerinin önerilen bir yorumunun geçerliğini değerlendirmede kullanılacak beş geçerlik kanıtının

kaynağı ise, ana hatlarıyla ve kısaca aşağıda açıklanmaya çalışılmıştır (AERA, APA, & NCME, 1999, 2014; Bademci, 1999, 2002, 2007, 2010, 2017a, 2017b, 2019).

**Test içeriği üzerine temellenmiş kanıt:** Bu geçerlik kanıtı türü, testin içeriği ve testin ölçmeyi amaçladığı yapı arasındaki ilişkinin analizinden elde edilebilir; test içeriği üzerine temellenmiş kanıt, yapı ve testin bölümleri arasındaki ilişkiye dair uzman görüşlerinden, iş (meslek) ya da uygulama analizlerinden, uzdaşma (alignment) çalışmalarından, vd. gelir (AERA, APA, & NCME, 1999, 2014; Bademci, 2017b; Sireci, 2009).

**Yanıt süreçleri üzerine temellenmiş kanıt:** Bu geçerlik kanıtı türü, sınava girenlerin fiilen meşgul olduğu yanıt veya erişimin (performance) ayrıntılı mahiyeti ve yapı arasındaki uyuma ilişkin kanıtı işaret eder; yanıt süreçleri üzerine temellenmiş kanıt, test sorularına verdikleri yanıtları hakkında testi alanlarla görüşmeyi, sesli düşünme (think-aloud) sözleşme tutanaklarını veya belgelerini, test yanıt davranışının sistematik gözlemlerini, vd. içerir (AERA, APA, & NCME, 1999, 2014; Bademci, 2017b; Linn, 2010; Sireci, 2009).

**İç yapı üzerine temellenmiş kanıt:** Bir testin iç yapısının analizleri test maddeleri ve test bileşenleri arasındaki ilişkilerin önerilen test ölçüm yorumlarının dayandırıldığı yapıya uyma derecesini gösterebilir; iç yapı üzerine temellenmiş kanıt, etken çözümlemesini (faktör analizini), çok boyutlu ölçekleme (multidimensional scaling) işlemlerini, vd. içine alır (AERA, APA, & NCME, 1999, 2014; Bademci, 2017a, 2017b; Reynolds & Livingston, 2012).

**Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt:** Birçok durumda, belirli bir kullanıma yönelik istenilen yorum, yapının bazı diğer değişkenlerle ilişkili olmasını gerektirir ve test ölçümlerinin testin dışındaki değişkenlerle ilişkisinin analizleri, bir diğer önemli geçerlik kanıtının kaynağını temin eder; diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt, test-ölçüt ilişkilerini (yordayıcı ve eşzamanlı çalışmaları), yakınsak (convergent) ve ayrıçsak (discriminant) çözümlenmeleri, geçerlik genelleme (validity generalization) çalışmalarını, vd. içine alır (AERA, APA, & NCME, 1999, 2014; Bademci, 2017a, 2017b; Reynolds & Livingston, 2012).

**Test etmenin sonuçları üzerine temellenmiş kanıt:** Bu kanıt türü, bir test ya da test etme programı ile bağlantılı istenilen ve istenilmeyen sonuçların değerlendirilmesine işaret eder; test etmenin sonuçları üzerine temellenmiş kanıtın örnekleri, öğretim ya da liseyi bırakma veya bir iş ya da yüksek okul başvurusu gibi hususlar üzerinde test etmenin etkilerinin değerlendirilmesini, vd. içermektedir (AERA, APA, & NCME, 1999, 2014; Reynolds,

Livingston, & Wilson, 2006; Sireci, 2009). Bir başka ifadeyle, test etmenin sonuçları üzerine temellenmiş kanıt, okulu bırakan öğrenci sayısının artmasını, öğretmen stresinin artmasını, öğrenci başarısının artmasını, öğrenci ve öğretmen güdülenmesinin artmasını ve diğerlerini de içine alır (Linn, 2010). *1999 Standartlarında* “test etmenin sonuçları üzerine temellenmiş kanıt” biçiminde yer bulan bu kısım, *2014 Standartlarında* “test etmenin sonuçları ve geçerlik için kanıt” şeklinde başlıklandırılmıştır (AERA, APA, & NCME, 2014).

Geçerlik kanıtının beş büyük kaynağına işaret eden bu çerçeveye, bir geçerlik tartışmasında tek kanıt kaynağı yerine birden çok kanıt kaynağının kullanılmasını teşvik eder ve genellikle, belirli bir kullanım için test ölçümlerinin önerilen bir yorumuna yönelik yeterince destek birden çok kanıt kaynağını gerektirmektedir (AERA, APA, & NCME, 2014; Odendahl, 2011; Sireci, 2009). Bununla birlikte, araştırmanın farklı biçimleri vasıtasıyla ‘üçgenleme’ (triangulation) ve ‘çapraz doğrulama’ süreci geçerlik tartışmasını güçlendirmeye yardım etmelidir (Odendahl, 2011). Örneğin, yaptıkları bir çalışmada Ferrara & DeMauro (2006) geçerlik kanıtı kaynaklarının tümünü, yani beşini de kullanmışlardır (Odendahl, 2011). Ayrıca, nitel araştırma ile ilgili bazı çalışmalar üçgenleme süreci ya da mantığının kavranılmasına da yardımcı olabilir (bkz., Denzin, 1978; Merriam, 1998; Merriam & Tisdell, 2016; Patton, 2002).

## Sonuç

Güvenirlilik gibi geçerlik de, eğitimde ve psikolojide ölçmenin merkezindeki büyük kavramlardır (Bademci, 2007, 2011b, 2013b; Bonner & Chen, 2019). Bunlardan geçerlik kavram ve kuramı ise, son 100 yılda sayısız değişiklikler, bir diğer ifadeyle evrim geçirmiştir (Angoff, 1988; Cook, Brydges, Ginsburg, & Hatala, 2015; Shepard, 1993). En köklü değişimler ise, geçerliğin, test ölçümlerinden yapılan yorumlarla ilgili olduğunu vurgulayan çalışmalarla 1980’lerin başlarında meydana gelmiştir (Cronbach, 1980; Messick, 1989; Pan, 2016; Rove & Mahar, 2006; Shepard, 1993). Değişim sürecinin doruk noktası olan *1999 Standartlarında* ise, kapsam geçerliği, ölçüt ilişkili geçerlik ile yapı geçerliği şeklindeki geçerliğin üç parçalı türü reddedilmiş, geçerliği, çeşitli geçerlik kanıt türlerine dayalı bütüncül bir kavram olarak ortaya koyan çağdaş görüş sergilenmiş ve geçerlik kanıtının türleri de, 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 biçiminde açıklanmıştır (AERA, APA, & NCME, 1999, 2014).

## Kaynaklar

- 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, 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.
- 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, 21 Mayıs). *Türkiye'de eğitim fakülteleri ve öğretmen yetiştirme: Öğretmen yetiştiren programlar nasıl olmalı?* ESEF İşletme Araştırma Topluluğu tarafından düzenlenen panelde sunulan bildiri, G.Ü. Mesleki Eğitim Fakültesi, Ankara.
- Bademci, V. (2002, 30 Mayıs). *Türkiye'deki okullar ne işe yarar? Türkiye'nin anomi, yabancılaşma, ekonomik büyüme, demokratikleşme sorunlarına çözüm önerisi*. ESEF Öğrenci Bilimsel Faal. Org. Kom. tarafından düzenlenen panelde sunulan bildiri, G.Ü. Mesleki Eğitim Fakültesi, Ankara.



- Bademci, V. (2007). *Ölçme ve araştırma yöntembiliminde paradigma değişikliği: Testler güvenilir değildir / Güvenirlilik ve geçerlik üzerine çağdaş düşünceler: Araştırmada yöntembilimle ilgili bazı büyük hataların düzeltilmesi*. Ankara: Yenyap.
- Bademci, V. (2010, 26 Nisan). *Türk eğitim ve biliminde paradigma değişikliği: Testler veya ölçekler güvenilir ve geçerli değildir*. Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dekanlığı tarafından düzenlenen konferansta sunulan bildiri, G.Ü. Gazi Eğitim Fakültesi, Ankara. [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 güvenirliliği üzerine bir çalışma. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 17, 173-193.
- Bademci, V. (2013a). *Yeni tez önerisi hazırlama kılavuzu*. Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.
- Bademci, V. (2013b). Değerbiçiciler arası (interrater) ölçüm güvenirliliğinin Cronbach'ın alfası ile kestirilmesi. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, 30, 55-62.
- Bademci, V. (2017a). Ölçme ve araştırma yöntembiliminde çağdaş gelişmeler ve yeni standartlar 1: Geçerlik, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliğidir. *JRES*, 4(1), 63-80.
- Bademci, V. (2017b). 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.
- Bademci, V. (2019, 10 Ekim). *Tarih eğitiminde ölçme: Geçerlik hakkında doğrular ve yanlışlar*. VI. Uluslararası Tarih Eğitimi Sempozyumu'nda sunulan bildiri, Abant İzzet Baysal Üniversitesi, Bolu.
- 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.

- Bonner, S. M., & Chen, P. P. (2019). *Systematic classroom assessment: An approach for learning and self-regulation*. New York: Routledge.
- Cizek, G. J. (2008). Assessing educational measurement: Omissions, omissions, opportunities. *Educational Researcher*, 37(2), 96-100.
- Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31-43.
- 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.
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560-575.
- 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. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.
- Denzin, N. K. (1978). *The research act* (2nd ed.). New York: McGraw-Hill.
- 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.
- Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, 24(1), 355-392.
- Gronlund, N. E. (1998). *Assessment in education* (6th ed.). Boston: Allyn & Bacon.

- Guion, R. M. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist, 29*(5), 287-296.
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*(3), 385-398.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170.
- 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.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, Massachusetts: Harvard University Press.
- 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* (Vol. 4, pp. 181-185). Oxford: Elsevier.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, New Jersey: Prentice-Hall.
- Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, New Jersey: Pearson.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.
- Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass.

- Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research* (4th ed.). San Francisco, CA: Jossey-Bass.
- 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.
- Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35-44.
- Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, *22*(4), 13-25.
- Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement*, *10*(1-2), 1-29.
- Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, New Jersey: Prentice-Hall.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Odendahl, N. V. (2011). *Testwise*. Lanham: Rowman & Littlefield Education.
- Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, New Jersey: Pearson.
- Pan, M. (2016). *Nonverbal delivery in speaking assessment*. Singapore: Springer.
- Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks: Sage.
- Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory & methods*. Boston: Pearson.
- 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*. Pacific Grove, California: Brooks/Cole.

- Rowe, D. A., & Mahar, M. T. (2006). Validity. In T. M. Wood, & W. Zhu (Eds.), *Measurement theory and practice in kinesiology* (pp. 9-26). Champaign, IL: Human Kinetics.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.
- Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing, 9*, 167-194.
- 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. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226-235.
- Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25*(3), 27-34.
- Silva, F. (1993). *Psychometric foundations and behavioral assessment*. Newbury Park, California: Sage.
- Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, California: Sage.
- Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Longman.