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

Eğitim ve Toplum Araştırmaları Dergisi/JRES, 4(1), 63-80, 2017

ÖLÇME VE ARAŞTIRMA YÖNTEMBİLİMİNDE ÇAĞDAŞ GELİŞMELER VE YENİ STANDARTLAR 1: GEÇERLİK, ÖLÇÜMLERİN KULLANIMLARININ VE ÖNERİLEN YORUMLARININ BİR ÖZELLİĞİDİR

CONTEMPORARY DEVELOPMENTS AND NEW STANDARDS IN MEASUREMENT AND RESEARCH METHODOLOGY 1: VALIDITY IS A PROPERTY OF THE PROPOSED INTERPRETATIONS AND USES OF SCORES

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

Testler ya da ölçme araçları geçerli değildir; çünkü, geçerlik, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliğidir. Böylece, "test geçerlidir" veya "ölçümler geçerlidir" benzeri ifadeler kullanılmamalıdır. Geçerlik, son 70 yılda evrim geçirmiştir. Bu süre zarfında, kapsam geçerliği, ölçüt ilişkili [yordayıcı, eşzamanlı] geçerlik ile yapı geçerliği türleri reddedildi. Geçerliğin bu türlerinin yerini, *1999 Standartları*nda, geçerlik kanıtının kaynakları aldı. Güncel *Standartlar*da, geçerlik kanıtının kaynakları, 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 şeklinde belirtilmektedir. Bununla birlikte, geçerlik, bütüncül bir kavramdır.

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

Abstract

Tests or measurement instruments are not valid, because validity is a property of the proposed interpretations and uses of scores. Thus, such statements as "the test is valid" or "the scores are valid" should not be used. Validity has evolved in the last 70 years. Meanwhile, the types of content validity, criterion-related [predictive, concurrent] validity, and construct validity were rejected. These types of validity were replaced with the sources of validity evidence in the *1999 Standards*. In current *Standards*, the sources of validity evidence are specified 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. Besides, validity is a unitary concept.

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

Atıf için Künye Bilgisi: Bademci, V. (2017). Ölçme ve araştırma yöntembiliminde çağdaş gelişmeler ve yeni standartlar 1: Geçerlik, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliğidir. *JRES*, *4*(1), 63-80.

Giriş

Geçerlikle ilgili bu çalışma, 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 içinde ölçüm güvenirliği ve test ölçümlerinin yorumlarının geçerliğine ilişkin yönlendirici ilkelerden, eğitim ve psikolojideki testlerin ve diğer ölçme araçlarının geliştirilmesi, uygulanması ve değerlendirilmesine yönelik ölçütler sağlamaya kadar, psikometrinin içeriği ile de bağlantılı çok çeşitli standartlar barındıran ve bu makale ortamında da yayımlandığı tarihle birlikte *Standartlar* (örneğin, *1999 Standartları*) olarak anılacak olan, *1999 Standartları* ve *2014 Standartları* paralelinde hazırlanmıştır.

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 (Bademci, 2007, 2011a).

Geçerlik, Ölçümlerin Kullanımlarının ve Önerilen Yorumlarının Bir Özelliğidir

Geçerlik, ölçmenin ve zihinsel test etmenin (testing) merkezinde yer almakta ve psikometrideki en temel ve önemli bilimsel kavram ya da husus olarak kabul edilmektedir (AERA, APA, & NCME, 1999, 2014; Angoff, 1988; Osterlind, 2006). Bu önemli kavramın, bu satırların da yazarı Bademci (2007, 2011a) tarafından yapılan çağdaş ve güncel tanımı yukarıda bulunmaktadır; geçerliğin bu güncel tanımı, önemli içermelere sahiptir ve bunların, psikometri disiplini içindeki ve dışındaki okuyuculara yönelik biçimde de açıklanmasında fayda bulunmaktadır: *Birincisi*, geçerlik, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliğidir; bir diğer ifadeyle, geçerlik bir testin ya da ölçme aracının kendisinin bir özelliği *değildir*; dolayısıyla, bir testin veya ölçme aracının kendisi, ne geçerlidir, ne de geçerli değildir (Bademci, 1999, 2007, 2011a; Furr & Bacharach, 2008; Kane, 2009, 2013; Koretz, 2008; Messick, 1995; Worthen, White, Fan, & Sudweeks, 1999).

"Testin geçerliği" veya "test geçerlidir" ya da "ölçümler geçerlidir" diye ifade etmek yanlıştır

Daha 1970'li yılların başında, yani bundan 40 yılı aşkın bir süre önce de, etkili geçerlik kuramcılarından ve *1974 Standartları*nın da (APA, AERA, & NCME, 1974) başyazarı olan Guion (1974), bir testin geçerliğinden bahsetmek aptalcadır ve uzun süredir de bu böyle

bilinmektedir ifadelerini kullanmıştır. Yaklaşık 30 yıl önce, öncü geçerlik kuramcılarından Cronbach (1989) ise, A Testini geçerli ya da B Testini geçersiz diye adlandırmanın mantıksız olduğunu vurgulamıştır. Böylece, "test geçerlidir", "testin geçerliği", "ölçeğin geçerliği", "ölçme aracının (veya yönteminin) geçerliği", "ölçme prosedürü geçerlidir", "bu deneyin geçerliği", "bellilendirme (assessment) geçerliği", "ölçümleyicilerin geçerliği", "gözlemin geçerliği", "sınavın geçerliği" ve benzeri ifadelerin kullanılması yanlıştır ve kesinlikle *kullanılmamalıdır* (AERA, APA, & NCME, 1999, 2014; Bademci, 2007, 2011a, 2013a, 2013b). Yine, benzer olarak, bir grup test maddesinin kendileri de veya örneğin, 30 test maddesinden elde edilen ölçümler de, ne geçerlidir, ne de geçerli değildir; bir diğer ifadeyle, "test ölçümlerinin geçerliği" veya "ölçümler geçerlidir" gibi anlatımlar da *hatalıdır* (Bademci, 2007, 2011a, 2013b; Furr & Bacharach, 2008; Gall, Gall, & Borg, 2007; Kane, 2001, 2013).

Yurt dışında ve Türkiye'de geçerlik ve geçerleme hususları etrafında ölçme ve araştırma yöntembilimindeki bilimsel devrimin öncüleri

Geçerlik, evrim geçirmiş ve geçirmeye devam eden bir kavram ya da düşünce veya özelliktir; dolayısıyla, son 70 yılı aşkın bir sürede, geçerlik kuram ve kavramı ve içerdikleri de birçok kez değişmiştir (Bademci, 2011a; Angoff, 1988; Messick, 1989; Sireci, 2007). Yurt dışında, 'güvenirliğin test ölçümlerinin bir özelliği' olduğu şeklinde güvenirlikteki paradigma değişikliğine ya da bilimsel devrime, Crocker & Algina (1986), Ebel & Frisbie (1991), Gronlund & Linn (1990), Suen (1990), Thompson (1994, 2003; Thompson & Vacha-Haase, 2000) gibi isimlerin ve çalışmalarının katkısı büyüktür (Bademci 2011b); geçerlik kuram ve kavramının geçirdiği evrime ya da bir diğer anlatımla geçerlikteki paradigma değişikliğine ise, 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).

Türk eğitim ve biliminde "Vahit Bademci'nin paradigma değişikliği ya da Vahit Bademci markası: Testler veya ölçekler güvenilir ve geçerli değildir" (Diri, 2014, s. 34; Gazi Haber, 2010, s. 48; Korkmaz, 2010, s. 21).

Türkiye'de ve Türk eğitim ve bilim topluluğunda ise, 1940'lardan bu yana, 60 yılı aşkın bir süredir kullanılan "test güvenilirdir" veya "test geçerlidir" gibi hatalı ifade etme biçimlerine, *hem güvenirliğin* ve *hem de geçerliğin* hatalı yorumlanış ve uygulama şekillerine -en azından 1994 yılından beridir- sürdürdüğü çalışmalarıyla Bademci (2001a; 2001b; 2002; 2004; 2005a; 2005b; 2005c; 2006a; 2006b; 2006c; 2007; 2008; 2010; 2011) *tek başına* ve açık biçimde karşı çıkmış, bir *paradigma değişikliği*

> gerekliliğini vurgulayarak, bir *yeni* paradigmayı da *bilimsel kanıtlarıyla* Türk eğitim ve bilim topluluğunun gündemine taşımıştır (Bademci, 2011b, s. 177).

Türkiye'de, 60 yılı aşkın bir süre sonra, güvenirlikte olduğu gibi, 'geçerliğin, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliği olduğu' ana teması etrafında geçerlikte meydana getirdiği *yeni* paradigma ya da bilimsel devrim paralelinde de, geçerlikle ilgili *ilk* kuramsal makaleler ve *ilk* bilimsel çalışmalar yine Bademci (1999, 2001a, 2001b, 2002, 2005, 2006, 2007, 2010, 2011a, 2011b, 2013a) tarafından gerçekleştirilmiştir.

Türkiye'deki neredeyse tüm ölçme ve değerlendirme ile araştırma yöntemleri kitapları ve tamamlanmış lisansüstü tezleri ile bilimsel yayınlarda olduğu gibi, geçerliğin, hala, 1930'ların sonlarındaki tanımlanma biçimleriyle, 'bir testin ya da ölçme aracının ölçtüğünü iddia ettiği şeyi ölçtüğünü gösterdiğini' (Garrett, 1939) ya da ölçme araç ve yöntemlerinin özellikleri başlıkları altında 'testin geçerliği' ifadeleriyle geçerliğin bir testin doğasında veya kendisinde olan bir özellik olduğunun düşüncesini savunmak ya da geçerliğin (kapsam, ölçüt ilişkili, yapı geçerlikleri gibi) farklı türlerinin varlığını ileri sürmek, 1954'ten beri var olan *Teknik Öneriler* (APA, 1954) ve ardışık sürümlerindeki *Standartlar*ın uzlaşılarının yanı sıra, geçerlik kuramı ve geçerleme (validation) üzerine yapılan araştırmaların *en az* 70 yılını tümüyle göz ardı etmek demektir (Bademci, 2007, 2011b; Sireci, 2009).

Günümüzde, artık, geçerliğin, testlerin kendileriyle değil, test ölçümlerinden yapılan yorumlarla ilgili olduğu hususunda çok geniş ve kapsamlı bir mutabakat ya da uylaşma veya fikir birliği vardır (AERA, APA, & NCME, 1999, 2014; Cizek, 2016; Cronbach, 1971; Messick, 1989; Kane, 2006). Bir başka söyleyişle ve kısaca, geçerlik test ölçümlerine bağımlı yorumların bir özelliğidir (Reynolds, Livingston, & Wilson, 2006). Tam da bu noktada ifade edilmelidir ki, *1999 Standartları* ile *2014 Standartları*ndaki geçerlik ile ilgili tanımlamalardan da görüleceği üzere, test ölçüm kullanımı olmadan test ölçüm yorumuna sahip olamayacağımız ve test ölçüm kullanımı ile test ölçüm yorumunun etle tırnak gibi ayrılmaz olduğu gerçekleri de, asla gözden kaçırılmamalıdır (AERA, APA, & NCME, 1999, 2014; Sireci, 2016).

Geçerlik hakkındaki ifadeler, ölçümlerin belirtilen kullanımlarına yönelik önerilen yorumlarına işaret etmelidir (AERA, APA, & NCME, 1999, 2014). Test ölçümlerinin önerilen bir yorumu uygun kanıtla desteklenmişse yüksek geçerliğe, yeterli gerekçeye dayanmamışsa veya haklı çıkmamışsa düşük geçerliğe sahiptir denilebilir; bir diğer aktarımla, ölçümlerin *önerilen* (kullanımı ve) *yorum*u, uygun kanıtla desteklenmişse yüksek geçerlidir, yeterli kanıtla desteklenmemişse veya kullanılan kanıtla çelişmişse düşük geçerlidir, şeklinde ifade edilebilir

(Cronbach, 1971; Messick, 1989; Kane, 2013, 2016). Kısaca, geçerliği onaylanan, test ya da test ölçümü değil, çıkarımları ve kararları içeren *yorum*dur (Cronbach, 1971; Kane, 2001, 2006).

"Testin geçerliği"ne işaret eden "ABC Okuma Testi geçerli midir?" sorusunu sormak da kesinlikle *doğru değildir*, bu ve benzeri ifadeler de asla *kullanılmamalıdır*. Onun yerine, "okuduğunu anlamayı ölçerken ABC Okuma Testinden elde edilen ölçümleri yorumlamak geçerli midir?" ya da "zekayı yansıtırken XYZ Zeka Ölçeğindeki ölçümlerin yorumu geçerli midir?" ve benzeri ölçüm (kullanım ve) *yorum* geçerliğine işaret eden daha belirli sorular sorulmalıdır (Nitko, 2001; Reynolds & Livingston, 2012). Bilimsel araştırma raporu, makale, bildiri, yüksek lisans tezi, doktora tezi gibi bilimsel etkinliklerde, yeni dilin yerleşmesi amacıyla, ilgili başlıklar da çalışmaya göre "geçerlik", "ölçüm yorum geçerliği", "ölçüm kullanım ve yorum geçerliği" ve benzeri olabilir.

Bir test ya da ölçme aracından elde edilen ölçümler, birçok şekilde yorumlandığında, her yorumun geçerliğinin denetlenmesi gerekir veya her yorum geçerlenmelidir. Örneğin, bir matematik başarı testi ölçümleri, sırasıyla, 1) bir öğrenciyi uygun bir öğretim programına yerleştirmede, 2) lise diplomasını verme ya da onaylamada veya 3) üniversiteye kabul kararını bildirmede kullanılabilir: Bu kullanımların her biri matematik başarı testi ölçümlerinin kısmen farklı bir yorumunu ima eder; bu yorumlar, sırasıyla, 1) öğrenci belli bir öğretimden yararlanacaktır, 2) öğrenci belirtilen eğitim programını başarıyla tamamlamıştır veya 3) öğrenci üniversite seviyesinde ödev ya da çalışmalarda muhtemelen başarılı olur; görüldüğü üzere, bu kullanımların her biri farklı yorumları içerir ve her yorumun geçerliği sağlanmalı ya da değerlendirilmelidir (AERA, APA, & NCME, 1999, 2014; Reynolds, Livingston, & Wilson, 2006). Bir diğer söyleyişle, belirli kullanıma yönelik önerilen her bir yorum geçerleme gerektirir (AERA, APA, & NCME, 2014).

Geçerlik, Varlık-Yokluk *Değil,* Derece Meselesidir

Yukarıda verilen geçerlik tanımındaki içermelerin *ikincisi* şu ki, geçerlik bir derece meselesidir; geçerlik, bir varlık-yokluk ya da hep-hiç kavramı değildir ve test ölçümlerinin önerilen bir yorumuyla ilgili olarak, iki değerli biçimdeki 'geçerlidir' ya da 'geçersizdir' düşüncesinden ve söyleminden ve de ifadelerinden *kaçınılmalıdır* (Furr & Bacharach, 2008; Linn & Gronlund, 2000; Messick, 1989; Nunnally & Bernstein, 1994; Reynolds, Livingston, & Wilson, 2006; Zumbo, 2007). Geçerlik, bir süreklilik içinde ya da üzerinde yer alır ve test ölçümlerinin belirtilen kullanımlarına yönelik önerilen yorumlarının göreli geçerliğine, "yüksek geçerlik",

"orta geçerlik" ve "düşük [ya da "yok"] geçerlik" şeklinde derece belirten sınıflamalar cinsinden atıfta bulunulur; bir başka söyleyişle, ölçümlerin belirtilen kullanımlarına yönelik her bir yorum, geçerliğin farklı derecelerine sahiptir (Gronlund, 1998; Kane, 2013; Linn, 2010; Linn & Gronlund, 2000; Reynolds, Livingston, & Wilson, 2006). Kısaca, ölçümlerin yorum geçerliği, yüksek, orta, düşük gibi dereceler ile ifade edilir ve bu, yorumların gelişmesiyle ve de yeni kanıt ya da kanıtların toplanmasıyla zaman içinde değişebilir (Gronlund, 1998; Kane, 2013; Linn, 2010).

Geçerlik, Daima Belirli Bir Evrene ya da Örnekleme Özgüdür

Makalenin başında verilen geçerliğin çağdaş tanımındaki *üçüncü* içerme; geçerlik, daima, sınava girenlerin belirli bir evrenine veya örneklemine dair belirtilen kullanıma yönelik ölçümlerin belirli yorumuna özgüdür; kısaca, güvenirlik gibi, geçerlik de evren ya da örneklem bağımlıdır ya da bir başka ifadeyle, geçerlik daima belirli bir evrene veya örnekleme ya da gruba özgüdür (Bademci, 2011a; Linn & Miller, 2005). Örneğin, ABC Okuma Testi ölçümleri bir okulun okuma programını değerlendirmede kullanılırken önerilen yorum yüksek geçerliğe, bir başka okul için ise, düşük geçerliğe sahip olabilir; yine, örneğin, hesaplama becerisini gösterirken XYZ Matematik Testi ölçümlerinin önerilen yorumu, dördüncü sınıf öğrencileri için geçerliğin yüksek bir derecesine, üçüncü sınıf öğrencileri için ise, geçerliğin orta veya düşük bir derecesine sahip olabilir (Bademci, 2010; Linn & Miller, 2005; Nitko, 2001).

Geçerlik Kavram ve Kuramının Değişen Felsefi Temelleri

Geçerlik tanımındaki *dördüncü* içerme, geçerliğin testin kendisinin değil, test ölçümlerinden yapılan yorumun özelliği olduğuna, kısaca ölçümlerin önerilen *yorum*una vurgu yaptığına, dolayısıyla geçerliğin felsefi esaslarının da zaman içinde değişiklik gösterdiğine işaret etmektedir. Açıktır ki, yıllar içinde geçerlik kuramının felsefi temelleri de değişmiştir; -çeşitli yazar ya da geçerlik kuramcılarının değişik görüşlerine göre- 1920'lerde başlayan ve 'geçerliğin testin ya da gözlemin bir özelliği' olduğunu vurgulayan geçerlik ve geçerleme üzerine geleneksel psikometrik bakış açıları olguculuğa (pozitivizme) veya mantıksal olguculuğa ya da olguculuk sonrasına köklendirilmiştir; fakat, 1970'lerden, özellikle 1980'lerden bu yana, 'geçerliğin ölçümlerden yapılan yorum ya da yorumların bir özelliği' olduğuna işaret eden çağdaş geçerlik kuramı ve geçerleme uygulamaları ise, yapılandırmacılığa (konstrüktivizme) ya da yapılandırmacı-gerçekçi görüşe taşınmış veya yorumlayıcı yaklaşım [bkz., Neuman, 2006] ya da sosyal yapılandırmacılık [sosyokültürel yapılandırmacılık] tarafından güçlü biçimde etkilenmiş ve bu doğrultuda tartışmalar ve çalışmalar da

sürdürülmektedir (Bonner, 2013; Brookhart, 2003; Shepard, 1993; Markus, 1998; Messick, 1989, 1998; Moss, 2003; Sijtsma, 2009). Bir başka görüşe göre ise, geçerlik ve geçerleme, son dönemlerin etkili geçerlik kuramcısı Kane'in (1992, 2004, 2006) geçerliğe tartışma temelli yaklaşımı ve çalışmaları ile birlikte, gerçekçilikten (realizmden), uygulayıcılığa (pragmatizme) doğru taşınmış ya da yön değiştirmiştir (Markus & Borsboom, 2013).

Geçerlik, Bir Değerlendirme Tartışmasıdır

Geçerlik tanımındaki içermelerin *beşincisi*, geçerliğin tümüyle değerlendirici bir yargı içerdiğidir ve bu değerlendirici yargı, önerilen yorumun uygunluğunu ve yeterliğini ve de yorumun yeterliğinin uygun (kuram ve) kanıtla ne derece desteklendiğini yansıtmalıdır; bir diğer söyleyişle geçerlik, bir değerlendirme tartışmasıdır ve test ölçümlerinin önerilen bir yorumunun kanıtlarıyla ve gerekçeleriyle akla ve mantığa uygunluğunun ve yeterliğinin ve inandırıcılığının bir değerlendirmesini kapsar; zira, geçerlenen, [çıkarımları ve kararları içeren] *yorum*dur (Linn & Miller, 2005; Kane, 2001, 2006; Osterlind, 2006). Hatırlanmalıdır ki, bir yorum uygun kanıtla desteklenmişse geçerliğin farklı bir derecesine sahip olduğu ifade edilir, önerilen yorum eğer gerekçelendirilmemişse yorumun geçerli olmadığı söylenir (Kane, 2016; Linn & Miller, 2005). Bir başka anlatımla, geçerlik, test ölçümlerinin kullanımlarının ve yorumlarının destekleyici kanıtlar tarafından ve bu kullanımların ve yorumların sonuçları bakımından ne derece gerekçelendirildiklerinin bir değerlendirmesini gerektirir ve bu değerlendirme içindeki etkinlikleri yerine getirmek, testi geliştirenler ile testi kullananların da ortak sorumluluğundadır (AERA, APA, & NCME, 1999, 2014; Linn & Miller, 2005; Osterlind, 2006).

Geçerlik, Bütüncül ya da Bölünmez Bir Kavramdır

Bu çalışmanın başında verilen geçerliğin tanımındaki *altıncı* içerme şöyle ki; 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 eleştirilmiş, çökmüş ve atılmış ve bunun yerine de, geçerliği, çeşitli geçerlik kanıtı türlerine dayalı *bütüncül* bir kavram olarak ifade eden çağdaş görüş yerleşmiştir (AERA, APA, & NCME, 1985, 1999, 2014; Bademci, 2007, 2010; Gronlund, 1998; Guion, 1980; Linn & Miller, 2005; Messick, 1989; Reynolds, Livingston, & Wilson, 2006; Shepard, 1993; Silva, 1993).

1999 Standartları: Kapsam Geçerliği, Ölçüt İlişkili Geçerlik ile Yapı Geçerliğinin Reddedilmesi ve Geçerlik Kanıtının Beş Kaynağının Tanımlanması

Yaklaşık bundan 20 yıl önce *1999 Standartları*nda ve sonrasındaki sürüm olan *2014 Standartları*nda da, "kutsal üçlü" (the holy trinity) (Guion, 1980) olarak ifade edilen ve eleştirilen, kapsam geçerliği, ölçüt ilişkili geçerlik [yordayıcı ve uyum -/eşzamanlı /uygunluk /zamandaş - geçerlikleri] ile yapı geçerliği türleri *terk edilmiş*, bunların yerini 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 şeklinde ve "geçerlik kanıtının kaynakları" başlığı altında, geçerlik kanıtının türleri almıştır (AERA, APA, & NCME, 1999, 2014); geçerleme üzerindeki çağdaş psikometrik görüşü yansıtan ve *1999 Standartları* ile *2014 Standartları*nda da ortaya konulan geçerlik kanıtının türleri, Türkiye'de ilk defa yine Bademci'nin (2001a, 2001b, 2002, 2007, 2010, 2013a; Gazi Haber, 2010) bilimsel çalışmalarıyla gündeme taşınmıştır.

Güncel *Standartlar* ve Geçerlik Kanıtının Kaynaklarına Kısa Bir Giriş

Bu çalışmada *Standartlar*ın güncel sürümleri olarak belirtilen *1999 Standartları* ve en son *2014 Standartları*, geçerliğin *bütüncül* bir kavram olduğunu ifade etmelerine 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 belirtmiştir (AERA, APA, & NCME, 1999, 2014). Tekrar etmek gerekirse, 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.

Test içeriği üzerine temellenmiş kanıt

Bu geçerlik kanıtı türü, 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).

Bu noktada açıklanması gerekir ki, *yapı* terimi, *Standartlar*da, bir testin ölçmek için tasarlandığı niteliği ya da kavramı kastetmek için kullanılmaktadır; zeka, matematik erişisi, bunalım, özsaygı, hissiyat, tutum, yaratıcılık, rüya, ilgi, benlik saygısı, öğrenme, halihazırda kullanılan yapı örnekleridir ve bunlar, doğrudan gözlenemezler (AERA, APA, & NCME, 1999, 2014; Crocker & Algina, 1986; Furr & Bacharach, 2008).

Test içeriği üzerine temellenmiş kanıt, yapı ve testin bölümleri arasındaki ilişkiye dair uzman görüşlerinden, iş [veya meslek] ya da uygulama analizlerinden, uzdaşma (alignment) çalışmalarından, vd. gelir (AERA, APA, & NCME, 1999, 2014; Sireci, 2009).

Yanıt süreçleri üzerine temellenmiş kanıt

Bu geçerlik kanıtı türü, sınava girenlerin fiilen meşgul olduğu yanıt veya erişimin (performance: erişim) ayrıntılı mahiyeti ve yapı arasındaki uyuma ilişkin kanıta işaret eder (AERA, APA, & NCME, 1999, 2014).

Yanıt süreçleri üzerine temellenmiş kanıt, test sorularına verdikleri yanıtları hakkında testi alanlarla görüşmeyi, test etme esnasındaki yanıt süreçlerinin niteliğine dair sesli düşünme (think-aloud) sözleşme tutanaklarını veya belgelerini, test yanıt davranışının sistematik gözlemlerini, vd. içerir (AERA, APA, & NCME, 1999, 2014; Creswell, 2012; 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; 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, etken çözümlemesini (faktör analizini), çok boyutlu ölçekleme (multidimensional scaling) işlemlerini, vd. içine alır (AERA, APA, & NCME, 1999, 2014; Sireci, 2009)

Diğer değişkenlerle ilişkiler üzerine temellenmiş kanıt

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 ya da belirtir; bir başka söyleyişle, 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).

Test etmenin sonuçları üzerine temellenmiş kanıt

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 bu 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" şeklinde başlıklandırılmıştır (AERA, APA, & NCME, 2014).

*Test etmenin sonuçları üzerine temellenmiş kanıt*ın örnekleri, liseyi bırakma ve iş başvuruları ya da bir yüksek okul başvuruları gibi hususlar üzerinde test etmenin etkilerinin değerlendirilmesini, öğretim üzerinde test etmenin etkilerinin değerlendirilmesini, yan etkiyi, vd. içermektedir (AERA, APA, & NCME, 1999, 2014; Reynolds, Livingston, & Wilson, 2006; Sireci, 2009).

Sonuç Yerine

1999 Standartları ile *2014 Standartları* belirli bir kullanıma yönelik test ölçümlerinin önerilen yorumunun geçerliğini değerlendirmekte kullanabilecek beş büyük ya da ana kanıt kaynağını tanımlamaktadır; bunlar, 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 kanıttır (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). Örneğin, Ferrara & DeMauro (2006) yaptıkları bir çalışmada geçerlik kanıtının kaynaklarının tümünü, yani beşini de kullanmışlardır (Odendahl, 2011).

Türkiye'de Ölçme, Seçme ve Yerleştirme Merkezi'nce (ÖSYM) ve Milli Eğitim Bakanlığı'nca (MEB) yapılan ulusal sınavlar üzerine çok önemli ve kısa bir not: Geçerlik kanıtının beş kaynağı da kullanılmalıdır

Türkiye'de Ölçme, Seçme ve Yerleştirme Merkezi'nce (ÖSYM) ve Milli Eğitim Bakanlığı'nca (MEB) yapılan çok çeşitli ulusal sınavlarda türlü testler kullanılmaktadır; burada, belirtilen kullanımlarına yönelik test ölçümlerinin önerilen yorumlarını desteklemek için kanıtın kaynaklarının tümü de, yani geçerlik kanıtının beş kaynağı da mutlaka kullanılmalıdır. Başkaca, test ölçümleri birden fazla yorumlandığında ise, önerilen her yorum mutlaka geçerlenmelidir ve bu süreçte, yine kanıtın kaynaklarının tümü, yani beşi de çalıştırılmalıdır.

Teşekkür

1999'dan bu yana 18 yıllık yoğun emeğin sentezi olan iki makalemin beşinci kez en başından yazımını, 2016 yılında tamamlayarak kişisel arşivime koydum. Geçerlikteki çağdaş gelişmeler ve yeni Standartlarla ilgili bu iki makalemi konunun önemine binaen en kısa sürede gündeme taşımam konusundaki nazik ve içten konuşmalarından dolayı Gazi Üniversitesi Eğitim Bilimleri Enstitüsü Müdürü Prof. Dr. Ülkü ESER ÜNALDI'ya teşekkür ederim. 1 ve 2 olarak birbirlerini tamamlayan iki makalenin de altıncı kez en başından yazdığım hali buradadır ve 'Bademci'nin paradigma değişikliği'nin ana çatısını tamamlamayı amaçlamaktadır.

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 (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) (1974). *Standards for educational and psychological tests*. 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). *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. (2013a). *Yeni tez önerisi hazırlama kılavuzu*. Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Bademci, V. (2013b). Değerbiçiciler arası (interrater) ölçüm güvenirliğinin Cronbach'ın alfası ile kestirilmesi. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, 30, 55-62.

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. 86-106). Los Angeles: Sage.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 22(4), 5-12.

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.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theor*y. Fort Worth: Holt, Rinehart and Winston.

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.

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.

Diri, M. (2014). *Sıra dışı sanatsal yaratıcılık ile sıra dışı bilimsel yaratıcılık üzerine bir araştırma: Bingül Başarır ile Vahit Bademci durum (örnek olay) çalışmaları* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, New Jersey: Prentice Hall.

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.

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

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

Garrett, H. E. (1939). *Statistics in psychology and education* (2nd ed.). New York: Longmans, Green.

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

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

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

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. (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. (2001). Currents 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. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The Concept of validity: Revisions, new directions, and applications* (pp. 39-64). Charlotte, NC: Information Age Publishing.

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

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198-211.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, Massachusetts: Harvard University Press.

Korkmaz, A. (2010). *Vahit Bademci'nin paradigma değişikliği üzerine bir araştırma: "Testler değil, ölçümler güvenilirdir"* (Yayımlanmamış yüksek lisans tezi). Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Enstitüsü, Zonguldak.

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., & Gronlund, N. E. (2000). *Measurement and assessment in education* (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. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible. *Social Indicators Research*, 45, 7-34.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

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.

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.

Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approaches* (6th ed.). Boston: Pearson Education.

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.

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.

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.

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, 9, 167-194.

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

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481.

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.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, New Jersey: Lawrence Erlbaum.

Thompson, B. (1994). *It is incorrect to say "the test is reliable": Bad language habits can contribute to incorrect or meaningless research conclusions*. (ERIC Document Reproduction Service No. ED 367 707).

Thompson, B. (Ed.) (2003). *Score reliability. Contemporary thinking on reliability issues*. Thousand Oaks, California: Sage.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60, 174-195.

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Longman.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics, Volume 26: Psychometrics* (pp. 45-79). Amsterdam: Elsevier Science B. V.