TÜRK EĞİTİM VE BİLİMİNDE BİLİMSEL DEVRİM: TESTLER YA DA ÖLÇME ARAÇLARI GÜVENİLİR VE GEÇERLİ *DEĞİLDİR*

Scientific Revolution in Turkish Education and Science: Tests or Measurement Instruments *are not* Reliable and Valid

Vahit BADEMCİ [1]

Özet

Güvenirlik ve geçerlik çok sık yanlış anlaşılmıştır. Testler ya da ölçme araçları güvenilir ve geçerli değildir. Çünkü, güvenirlik, ölçümlerin bir özelliği; geçerlik ise, ölçümlerin kullanımlarının ve yorumlarının bir özelliğidir. Güvenirlik ve geçerlik evren ya da örneklem veya grup bağımlı kavramlardır. Güvenirlik katsayıları gibi geçerlik katsayıları da, evrenden evrene, örneklemden örnekleme değişir. "Testin güvenirliği", "ölçeğin geçerliği", "bellilendirmenin geçerliği" veya "ölçme aracı güvenilirdir" gibi ifadelerin kullanılması doğru değildir. Böylelikle, "test güvenirliği" yerine, "ölçüm güvenirliği" kavramının kullanılması çok daha uygundur. Geçerlik iddiaları ise, test ölçümlerinin belirli kullanımları ve yorumlarına ilişkin yapılmalıdır.

***Anahtar kelimeler**: Ölçüm güvenirliği, geçerlik, paradigma, bilimsel devrim, Türk eğitim ve bilimi*

Abstract

Reliability and validity are very often misunderstood. Tests or measurement instruments are not reliable and valid. Because, reliability is a characteristic of scores, as for validity is a property of interpretations and uses of scores. Reliability and validity are population or sample or group dependent concepts. As for reliability coefficients, validity coefficients fluctuate from population to population, from sample to sample as well. It is not correct to use the statements such as "the reliability of the test", "the validity of the scale", "the validity of assessment" or "measurement instrument is reliable". Thus, it is more appropriate to use the term "score reliability" instead of "test reliability". As to validity claims should be made in relation to specific uses and interpretations of test scores.

***Keywords**: Score reliability, validity, paradigm, scientific revolution, Turkish education and science*

1. PARADİGMA DEĞİŞİKLİĞİ YA DA BİLİMSEL DEVRİM ÜZERİNE

Paradigmalar, kuramlar değil, düşünme tarzları veya araştırma için örnekler ya da modellerdir (Gage, 1963). "Ispanağın bol miktarda demir içerdiği" inancı ya da kabulü, Norton'un (2001) ifadesiyle bir "yerleşik düşünce", paradigma kelimesini ilk defa kullanan Kuhn'un (1995) ifadesiyle ise, bir "paradigma"dır. "Yerleşik düşünce"lerin en büyük özelliği, yanlış ya

[1] Yrd.Doç.Dr.; Gazi Üniversitesi, Endüstriyel Sanatlar Eğitim Fakültesi, 06830 Gölbaşı - Ankara, vahitbademci@yahoo.com

da banal olmaları değil, üzerinde hiç düşünülmeden kabul edilmeleridir (Norton, 2001). Örneğin, birkaç temel ilkeye dayanan Newton [1642-1727; *Principia*, 1686]* yasaları, evrenin bütün geçmişini ve geleceğini açıklamaya yeterli görünüyordu ve fizikçiler sonraki 250 yıl Newton sistemini geliştirmekle uğraştılar ve de Newton yasalarının aslında yanlış olabileceği kimsenin –en azından Einstein'a kadar- aklına gelmedi (Bernstein, 2006: 41-42). Kuhn'un (1995) açıklamalarına göre, -en azından Einstein'a kadar- 200 yılı aşmış bir süre doğruluğu sorgulanmamış olan Newton mekaniği *eski paradigma; onunla bağdaşmayan ve çığır açıcı* Einstein'ın Özel İzafiyeti [1905] ise, *yeni paradigma* olarak adlandırılır.

1.1. Türk Eğitim ve Biliminde "Vahit Bademci'nin Paradigma Değişikliği ya da Vahit Bademci Markası: Testler veya Ölçekler Güvenilir ve Geçerli Değildir"

(Gazi Haber, 2010: 48; Korkmaz, 2010: 21).

1940'lardan 2000'lerin başına kadar Türk eğitim ve bilim dünyasında egemenliğini sürdürmüş olan "testler güvenilir ve geçerlidir" şeklindeki yerleşik düşünme tarzı *eski paradigma;* 60 yılı aşkın bir süre sonra Bademci'nin (2001a; 2001b; 2002; 2004; 2005a; 2005b; 2005c; 2006a; 2006b; 2006c; 2007; 2008; 2010) ortaya koyup, ispatladığı "testler ya da ölçme araçları güvenilir ve geçerli değildir" ya da "güvenirlik ve geçerlik, ölçümlerin fonksiyonlarıdır" şekillerindeki çağdaş düşünme tarzı ise, *yeni paradigma* olarak adlandırılır. Yeni paradigma ile bütün olgular da yeni bir anlam kazanmaktadır (Topdemir, 2002). Einstein'cı kavramların değindikleri fiziksel olgular, aynı isimleri taşıyan Newton'cu kavramların çağrıştırdığı olgularla *özdeş değildir*; Newton'cu *kitle*, değişmez korunur; Einstein'cı *kitle* ise, her zaman enerjiye dönüştürülebilir (Kuhn, 1995). Bademci'nin (2001a; 2001b; 2002; 2004; 2007; 2008; 2010) gerçekleştirdiği *yeni paradigma*daki güvenirlik ve geçerlik kavramlarının çağrıştırdıkları da, *eski ya da çağdışı* paradigmadaki ile *özdeş* ya da *aynı değildir.*

1.2. Paradigma Değişikliği ya da Bilimsel Devrim Nadiren Ortaya Çıkar

Her paradigma, paradigmayı tanımlayıp çerçevesini çizecek bir büyük eser üretir; paradigma değişikliği ya da yeni paradigmaya geçiş bilimsel bir devrimdir ve bilimsel ilerleme, devrimsel bir süreçtir ve de *bilimsel devrimler, nadiren ortaya çıkan olağan dışı bilimsel süreçlerdir* (Bademci, 2007; 2010; Kuhn, 1995; Serdar, 2001; Topdemir, 2002). "Paradigmadaki bir değişim, araştırmanın temel kavramlarını değiştirir ve eskilerine hiç mi hiç uymayan yeni kanıt standartları, yeni araştırma teknikleri ve yeni kuram düzlemlerinin önünü açar" (Serdar, 2001: 37). Paradigmanın değişmesiyle birlikte dünya görüşü de değişmektedir; bundan dolayı paradigma değişikliği bilim adamlarının bağlanmış oldukları dünyayı farklı şekilde görmelerine neden olur; öyle ki, *bilim adamının dünyasında ördek sayılan nesne, tavşan olmuştur; bu nedenle devrim dönemlerinde, yani olağan bilimsel gelenek*

değiştiği zamanlar, bilim adamı çevresini algılamayı yeniden öğrenmek zorundadır (Kuhn, 1995; Topdemir, 2002).

Bir paradigmadan, diğerine veya yenisine geçmek ya da bağlılık değiştirmek, zor olmayacak bir dönüş deneyimidir (Kuhn, 1995). Yeni bir paradigma ya da paradigma değişikliği, *işlerinin ehli* bazı bilim adamlarınca hızlıca kabul görmekte ve savunulmaktadır. Ancak bazı bilim adamları da, özellikle de [göreli] daha yaşlı ve deneyimli olanlar, paradigma değişikliğine karşı direnç göstermektedirler (Kuhn, 1995). Paradigma değişikliğine karşı direnen bazı kişilerin, yeni paradigmanın fazlaca bir üstünlük sağlamadığını iddia etmeleri ise, alışılmış bir olaydır (Kuhn, 1995). Hiç şüphe yok ki, " eğer yeni bir paradigma adayı [ya da paradigma değişikliği], daha başlangıçta yalnızca göreli problem çözme yeteneğini ölçen kalın kafalı kişilerce yargılanacak olursa, bilimlerin geçirdiği büyük devrimlerin sayısı oldukça azalırdı" (Kuhn, 1995: 165). Eğer, reddedilen [eski] paradigmanın yerini, yeni paradigmanın alması eşzamanlı değilse, reddedilen paradigma değil, bilim olmaktadır; bilimi reddetmek ise, paradigmanın değil, bilim adamının işidir; böyle bir bilim adamı da kendi beceriksizliğinin suçunu aletlerinde arayan marangoza benzetilebilir ve 'kötü marangoz aletini suçlar' (Kuhn, 1995). Bademci'nin (2007) 60 yılı aşkın bir süre sonra, ölçme ve araştırma yöntembiliminde Türk eğitim ve bilim topluluğuna yönelik ortaya koyduğu *yeni paradigma* etrafındaki görüşleri, yaklaşımları ve çalışmalarında meydana vurduğu bazı bilimsel kanıtlamaları ise, *yeni paradigma*nın bilimsel doğruluğuna, güncelliğine, etkililiğine ve verimliliğine vurgu yapan ve bilimdeki çağdaş gelişmelerin, düşüncelerin ve yenileşmelerin yanında olan araştırmacıların bilimsel çalışmalarındaki yerini almaya başlamıştır (örneğin, bakınız, Beycioğlu, 2007; Cebeci, 2006; Hotaman & Yüksel-Şahin, 2010; Kartal, 2009; Kartal & Pekkanlı, 2011; Korkmaz, 2010; Özsoy, Keleş & Uzun, 2009; Sayın, 2008; Sayın, 2010; Sever, 2008).

2. KLASİK GERÇEK ÖLÇÜM KURAMINA KISA BİR GİRİŞ

Klasik Gerçek Ölçüm Kuramı (Allen & Yen, 1979), ölçme hatasının bir kuramıdır (van der Linden, 2005). *Test ölçümleri hakkında bir kuram* olarak da ifade edilebilen (Hambleton & Jones, 1993) Klasik Gerçek Ölçüm Kuramı ya da diğer bazı adlandırılmalarıyla Klasik Test Kuramı (Pedhazur & Schmelkin, 1991) veya Klasik Gerçek Ölçüm Modeli (Crocker & Algina, 1986), bir sayıltı ya da sayıltıların ilki üzerine temellenmiştir; bir *gözlenmiş ölçüm* X, *gerçek ölçüm* T ve *ölçme hatası* ya da *hata ölçümü*nün E toplamıdır (Allen & Yen, 1979; Feldt & Brennan, 1989; Hopkins, 1998) ve

$$X = T + E \qquad (1)$$

şeklinde de gösterilebilir (Algina, 1992; Mehrens & Lehmann, 1991). Bir başka söyleyişle, $X = T + E$, Klasik Gerçek Ölçüm Kuramının temel eşitliğidir (Crocker & Algina, 1986; Traub, 1994).

Gözlenmiş ölçüm X, "elde edilmiş ölçüm" veya "ölçme" ya da "ham ölçüm" veya "test ölçümü" biçiminde de adlandırılmaktadır (Gronlund, 1998; Guilford, 1954; Guilford & Fruchter, 1978; Gulliksen, 1950; Hambleton & Jones, 1993). Ölçme hatası E ise, "hata ölçümü" veya "random hata" olarak da isimlendirilmektedir (Kieffer, 1999; Magnusson, 1967).

Güvenirlik, gerçek ölçüm ve gözlenmiş ölçüm arasındaki ilişkinin gücüdür; bu, gerçek ölçüm T ve gözlenmiş ölçüm X arasında Pearson'ın korelasyonu olarak ifade edilebilmekte ve ρ_{XT} şeklinde de gösterilebilmektedir; bu korelasyon, *güvenirlik indeksi* gibi adlandırılmaktadır; bir başka ifadeyle, bir test üzerindeki gerçek ve gözlenmiş ölçümler arasındaki ilişkinin derecesini ifade eden korelasyon katsayısı, *güvenirlik indeksi* olarak bilinmektedir; bu güvenirlik indeksinin karesi ise, *güvenirlik katsayısı* gibi isimlendirilmektedir ve ρ_{XT}^2 biçiminde ifade edilebilmektedir; ρ_{XT}, gözlenmiş ölçümlerden veya verilerden doğrudan kestirilemez, ρ_{XT}^2'yi ise, [belirli sayıltılar altında] kestirmek mümkündür (Algina, 1992; Crocker & Algina, 1986; Lord & Novick, 1968; Suen, 1990). Gözlenmiş ölçümler ve gerçek ölçümler arasındaki korelasyonun karesi ρ_{XT}^2 [ya da güvenirlik katsayısı] , gerçek ölçüm varyansının σ_T^2, gözlenmiş ölçüm varyansının σ_X^2 oranına eşittir ve

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \qquad (2)$$

şeklinde gösterilebilir (Algina, 1992; Lord & Novick, 1968).

Klasik Gerçek Ölçüm Kuramının sayıltıları (Allen & Yen, 1979; Thorndike, 1982) altında, gözlenmiş ölçüm varyansı σ_X^2, iki bileşene bölünmüş biçimde veya bir başka ifadeyle, iki bileşenin toplamı şeklinde yazılabilir (Algina, 1992; Kane, 1996).

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \qquad (3)$$

3 numaralı eşitlik, kişilerin evrenindeki gözlenmiş ölçüm varyansının σ_X^2 , gerçek ölçüm varyansı σ_T^2 ve hata ölçüm varyansının σ_E^2 toplamına eşit olduğunu ifade eder (Lord & Novick, 1968).

Eşitlik 2 ve 3'den yararlanılarak ve belirli sayıltılar altında olmak üzere, bir *test ölçüm güvenirliği* ya da kısaca *güvenirlik*

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \qquad (4)$$

biçiminde (de Gruijter & van der Kamp, 2008; Kane, 1996; Lord & Novick, 1968) veya öteki şekilde,

$$\rho_{XT}^2 = 1 - \frac{\sigma_E^2}{\sigma_X^2} = \rho_{XX'} \qquad (5)$$

$\rho_{XX'}$ = güvenirlik katsayısı (X ve X' paralel ölçmeler ya da paralel testler üzerindeki ölçümler)

gibi de (Allen & Yen, 1979; Lord & Novick, 1968; Pedhazur & Schmelkin, 1991; Stanley, 1971) ifade edilebilmektedir.

Klasik Gerçek Ölçüm Kuramının bir temel tanımı olarak (Feldt & Brennan, 1989) Eşitlik 4, güvenirliğin ya da bir diğer söyleyişle test ölçüm güvenirliği büyüklüğünün, evrene bağımlı olduğunu açıkça göstermektedir (de Gruijter & van der Kamp, 2008; Mellenbergh, 1996).

Yine, Eşitlik 5 de, diğer şeyler eşit olmak üzere, *daha ayrışık* [heterojen] evren veya örneklem ya da gruptan, daha yüksek güvenirlik elde edileceğini açıklayıcı niteliktedir; bir başka söyleyişle Eşitlik 5, güvenirlik ya da güvenirlik katsayısı büyüklüğünün, [diğer şeyler eşit olmak üzere] doğrudan evren ya da örneklem veya grubun *ayrışıklığı* üzerine bağlı olacağını göstermektedir (Allen & Yen, 1979; Bademci, 2001a; 2004; 2007; 2010; Guilford, 1954; Mehrens & Lehmann, 1991).

3. GÜVENİRLİK, EVREN YA DA ÖRNEKLEM BAĞIMLIDIR: TESTLER VEYA ÖLÇEKLER YA DA ÖLÇME ARAÇLARI GÜVENİLİR *DEĞİLDİR*

Yukarıda yapılan bu açıklamalar, Klasik Gerçek Ölçüm Kuramındaki yıllardır unutulmuş ya da gözden kaçırılmış bir gerçeğin altını önemle çizer veya üzerine basa basa tekrar hatırlatır: ***Güvenirlik, evren ya da örneklem bağımlıdır*** (Bademci, 2001a; 2001b; 2004; 2007; 2010; Borsboom, Romeijn & Wicherts, 2008; de Gruijter & van der Kamp, 2008; Mellenbergh, 1996; Mellenbergh, 1999; Rouse, 2007; Tyson, Dulmus & Wodarski, 2002).

3.1. Aynı Ölçek veya Test, 100 Farklı Örnekleme Uygulansa, 100 Farklı Güvenirlik Katsayısı Ortaya Çıkabilir: Ölçeğin ya da Testin Kendisi, Güvenilir *Değildir*

Bu aydınlatıcı bilgilerin ışığında ve tam da bu noktada, bir olguyu vurgulamakta da fayda bulunmaktadır; güvenirlik katsayıları, neredeyse her zaman, tüm evrenlerden değil, kişilerin örneklemlerinden alınmış ölçmelerden hesaplanmaktadır (Traub, 1994). Dolayısıyla,

> "örneklem özellikleri ölçüm güvenirliğini etkileyebilmekte (Henson, Kogan ve Vacha-Haase, 2001), bir testin veya ölçme aracının uygulandığı örneklemin bağdaşık [homojen] ya da ayrışık [heterojen] olması, ölçüm güvenirliğinin azalmasına veya artmasına neden olmaktadır. Bir başka ifadeyle ölçüm güvenirliği, örneklemden örnekleme değişmektedir (Capraro ve Capraro, 2002). Aynı test, bağdaşık veya ayrışık örneklemlere uygulandığı zaman güvenirliğe ilişkin farklı sonuçlar doğurabilecektir... Örneğin, *aynı ölçek [test veya ölçme aracı], 100 farklı örnekleme uygulansa, 100 farklı güvenirlik katsayısı ortaya çıkabilir* (Buhi, 2005). Hâl böyle iken, "test/araç/ölçek güvenilirdir" ya da "testin/aracın/ölçeğin güvenirliği" demek ve güvenirliği, testin veya aracın ya da ölçeğin bir özelliği gibi ima veya ifade etmek *uygun değildir*, *doğru değildir*" (Bademci, 2007: 95 ve 206).

Tekrar ve kısaca ifade etmek gerekirse, güvenirlik, evren ya da örneklem bağımlıdır ve sınavı alanların belirli evreninde [veya örnekleminde] gerçek ölçümler ve gözlenmiş ölçümler arasındaki korelasyonunun karesi gibi tanımlanmıştır (de Gruijter & van der Kamp, 2008; Lord & Novick, 1968; Mellenbergh, 1999).

3.2. *Güvenirlik*, Sınava Giren Belirli Bir Gruba Uygulanmış Bir Testten Elde Edilmiş *Ölçümlerin Bir Özelliğidir*

Güvenirlik sıklıkla yanlış anlaşılmıştır (Aycock, 1993; Bademci, 2007; Capraro & Capraro, 2002). Şu çok açıktır ki, güvenirlik, [tek başına] testin kendisinin değil, [daha çok] örneklemin özelliklerinin [de] bir fonksiyonudur ya da öteki söyleyişle, güvenirlik, ölçümlerin elde edildiği örneklemin özelliklerine [doğrudan] bağımlıdır; bir diğer net anlatımla, güvenirlik, ölçme duyarlığının evren ya da örneklem bağımlı bir kavramıdır; yapılan tüm bu açıklamaların doğrultusunda, *güvenirlik*, sınavı alanların belirli bir [evreni ya da örneklemi veya] grubu için bir test üzerindeki ölçümlerin bir özelliği şeklinde ya da bir başka ifadeyle, sınava giren belirli bir gruba uygulanmış bir testten elde edilmiş *ölçümlerin bir özelliği* biçiminde de ifade edilebilir (Bademci, 2001a; 2004; 2007; 2010; Crocker & Algina, 1986; Frisbie, 2005; Henson, 2000; Mellenbergh & van den Brink, 1998).

Bir diğer ve açık söyleyişle, güvenirlik, testin kendisinin değil, elde edilmiş ölçümlerinin bir özelliğidir; o halde, *bir test ya da ölçme aracının kendisi ne güvenilir, ne de güvenilmezdir* (Bademci, 2001a; 2004; 2007; 2010; Ebel & Frisbie, 1991; Crocker & Algina, 1986; Rouse, 2007; Rowley, 1976; Traub & Rowley, 1991). *Güvenilir ya da güvenilmez olan*, testler veya ölçekler ya da ölçme araçları *değil*, onlardan elde edilmiş olan *ölçümlerdir*; bir başka söyleyişle, güvenirlik özelliğine ölçümler sahiptir, testin veya ölçeğin ya da ölçme aracının kendisi *değil* (Bademci, 2007; Traub & Rowley, 1991; Thompson, 2003). Kısaca, *testler değil, ölçümler güvenilirdir* (Kieffer & Reese, 2002; Vacha-Haase, 1998).

Böylelikle, "test güvenilirdir" veya "ölçeğin güvenirliği" ya da "ölçme aracı güvenilirdir" ve benzeri ifadeler kullanmak, *doğru değildir*, *uygun değildir* (Bademci, 2001a; 2007; 2010; Buhi, 2005; Kieffer, 1999; Thompson, 2001; Thompson, 2003); çünkü bu tür ifadeler, güvenirliğin, testin veya ölçme aracının ya da ölçeğin bir özelliği olduğuna işaret eder veya atıfta bulunur (Bademci, 2001a; 2007; 2010; Guthrie, 2000; Ragan & Kang, 2005; Sawilowsky, 2000; Thompson & Vacha-Haase, 2000; Victorson, Barocas, Song & Cella, 2008).

Güvenirlik, ölçümlerin bir özelliğidir; dolayısıyla, güvenirliğin, ölçümlerin bir özelliği olduğuna işaret eden "*ölçüm güvenirliği*" ya da "*test ölçüm güvenirliği*" ve benzeri ifadeler kullanmak daha *doğrudur* (Buhi, 2005; Miller, Shields, Campfield, Wallace & Weiss, 2007; Thompson, 2003; Vassar & Hale, 2009; Wasserman & Bracken, 2003); olası doğru ifade örnekleri ise, Bademci'nin (2001a; 2004; 2007; 2010) bazı çalışmalarında da bulunmaktadır.

Tüm bu gerekçelerin ışığında, *güvenilir ölçümler* ve *güvenilir testler* kavramlarının *eş anlamlılıktan uzak* olduğu ise, asla gözden kaçırılmaması gereken çok önemli bir noktadır (Bademci, 2001a; 2007; 2010; Vacha-Haase, Kogan, Tani & Woodall, 2001). Bir başka ifadeyle, *test güvenirliği* ve *test ölçüm güvenirliği* kavramları arasında farklılık vardır ve bu *farklılık* yüzeysel olmayıp, *önemlidir* (Bademci, 2001a; 2007; 2010; Yin & Fan, 2000).

3.3. Korkmaz'ın (2010) Çalışması: Güvenirlikle İlgili Olarak, 2000-2009 Yılları Arasında Yapılmış Yüksek Lisans ve Doktora Tezlerinin %79'unda *Paradigmatik Kavram Yanılgısı* Bulunmaktadır

Zonguldak Karaelmas Üniversitesi'nde, Yrd. Doç. Dr. Saime Sayın'ın tez danışmanlığında yapılmış olan Ahu Korkmaz'ın (2010) yüksek lisans tezi, konu ve tespitleri yönünden Türk eğitim ve biliminde bir *ilk*i oluşturmakta ve çok önemli katkılar sunmaktadır. Korkmaz'ın (2010) bu tezi, -özellikle eğitim bilimleri olmak üzere- Türkiye'deki mevcut yüksek lisans ve doktora eğitiminin niteliğinin acilen sorgulanması ve tartışılması gerektiğini güçlü *bilimsel kanıtlarıyla* ve apaçık ortaya koymuştur. Ortaya koyduğu pek çarpıcı ve önemli bilimsel bulgularından dolayı, Korkmaz'ın (2010) tezinin Türkiye'deki tüm üniversitelerin ilgili birimlerince ve üniversitelerle bağlantılı tüm birimlerce de mutlaka ve dikkatle okunması ve bilgilenilmesi gerektiği gün gibi aşikardır. Eğitim bilimleri ile ilgili olarak, Korkmaz'ın (2010) tezinin 80. sayfasında mevcut olan ve aşağıya olduğu gibi aktarılmış bulunulan yalnızca bir bulgu dahi, bahsedilen durumun ciddiyetini anlatmaya yeter görülmektedir:

> "Testin güvenirliği" [/"ölçeğin güvenirliği"/ "aracın güvenirliği"] ifadesi, araştırma kapsamına alınmış ve 2000-2009 yılları arasında yapılmış olan; Ankara Üniversitesi'nde yapılmış tezlerin %85'inde, Gazi Üniversitesi'nde yapılmış olan tezlerin %81'nde ve Hacettepe Üniversitesi'nde yapılmış tezlerin %72'sinde olmak üzere, yani üç

> üniversitede yapılmış toplam yüksek lisans ve doktora tezlerinin %79'unda kullanılmıştır. Bir başka ifadeyle, *paradigmatik kavram yanılgısı* (Thompson ve Vacha-Haase, 2000) olarak vurgulanan bu durum, incelenen tezlerin büyük çoğunluğunda [444 tezin 349'unda, yani %79'unda] görülmektedir. American Educational Research Association, American Psychological Association ve National Council on Measurement in Education tarafından 1999 yılında yayınlanmış olan "otoriter" *Eğitimsel ve Psikolojik Test Etme/Test Yapma Standartları*'nda da (EPTS) (AERA, APA ve NCME, 1999) "testin güvenirliği" ifadesi kullanmanın "**kabul edilemez**" olduğu açıkça ifade edilmesine ve aradan 10 yıl geçmesine rağmen, üç büyük üniversiteden araştırma kapsamına alınmış yüksek lisans ve doktora tezlerinin, %79'unda "testin güvenirliği" [/"ölçeğin güvenirliği"/ "aracın güvenirliği"] ifadesinin görülmüş olması, Türk eğitim ve biliminin literatürü takip etme ve lisansüstü eğitiminin kalitesi yönünden ciddi sıkıntılarının olduğunun açık bir göstergesi olarak kabul edilebilir" (Korkmaz, 2010: 80). [Yazarından izin alınmıştır.]

4. GÜVENİRLİK GİBİ GEÇERLİK DE, ÖLÇME ARACININ KENDİSİNE DEĞİL, *ÖLÇÜMLER*E İŞARET EDER

Geçerlik de, bir testin ya da ölçme aracının doğasında olan bir özelliği değildir; ***güvenirlik katsayıları gibi***, ***geçerlik katsayıları da***, ***evren ya da örneklem veya grup bağımlıdır*** ve dolayısıyla, evrenden evrene, örneklemden örnekleme değişir ya da değişecektir; güvenirlik katsayılarında olduğu gibi, diğer şeyler eşit olmak üzere, daha ayrışık örneklem ya da gruptan daha yüksek geçerlik katsayısı elde edilir ya da edilebilecektir (Allen & Yen, 1979; Anastasi & Urbina, 1997; Bademci, 2001a; 2001b; Chartrand & Walsh, 2001; Gray, 1997; Hambleton, Swaminathan & Rogers, 1991; Kubiszyn & Borich, 1993; Le & Klein, 2002; McHorney, 1999; Mehrens & Lehmann, 1991; Streiner & Norman, 1995; Victorson, Barocas, Song & Cella, 2008).

Çağcıl psikometri kuramcılarının ya da başlıca geçerlik kuramcılarının belki de en etkilisi (Superfine, 2004; Sireci, 2005) olarak ifade edilen Messick (1989), testlerin ya da ölçme araçlarının güvenirliklerinin ve geçerliklerinin *olmadığını* vurgulamıştır. Bir diğer söyleyişle, güvenirlik ve geçerlik, testlerin ya da ölçme araçlarının *değil*, ölçümlerin fonksiyonlarıdır (Bademci, 2007; 2010; Mji & Onwuegbuzie, 2004). Güvenirlik gibi geçerlik de, bellilendirme [assessment; Bademci, 2000; 2007] ya da ölçme aracının kendisine değil, bellilendirme ya da ölçme *sonuçları*na veya *ölçümler*e işaret eder (Bademci, 2001a; 2007; 2010; Brookhart & Nitko, 2008; Nitko, 2001).

Kısaca, geçerlik, testlerden veya ölçme araçlarından elde edilen ölçümlerin yorumlarının ve kullanımlarının niteliğinin değerlendirilmesindeki en temel ve en önemli faktördür (Bademci, 2001; 2002; Linn, 1995; Linn, 2002; Linn & Gronlund, 2000).

5. GÜVENİRLİK, ÖLÇÜMLERİN BİR ÖZELLİĞİ; GEÇERLİK İSE, ÖLÇÜMLERİN YORUMLARININ VE KULLANIMLARININ BİR ÖZELLİĞİDİR

Görüldüğü üzere ve çok açıktır ki, güvenirlik ve geçerlik, testlerin ya da ölçeklerin veya ölçme araçlarının özellikleri *değildir* (Murphy & Davisshofer, 2001; Barnes, Harp & Jung, 2002; Worthen, White, Fan & Sudweeks, 1999); güvenirlik ve geçerlik, "*bellilendirmenin* de *bir özelliği değildir*" (Bademci, 2001a; 2007; 2010; Frisbie, 2005; Messick,1995); zira, güvenirlik, ölçümlerin bir özelliği; geçerlik ise, ölçümlerin yorumlarının ve kullanımlarının bir özelliğidir (Bademci, 2007; 2010; Kane, 2006b; Linn, 2002; Thompson, 2003). Dolasıyla, "test geçerlidir", "test güvenilirdir", "bu deneyin geçerliği", "testin geçerliği", "ölçeğin güvenirliği", "bellilendirmenin geçerliği", "ölçme aracının [ya da yönteminin] geçerliği", "ölçme prosedürü geçerlidir" ve benzeri ifadeler kullanmak, *uygun değildir*, *doğru değildir*; bunların yerlerine, "test ölçümlerinin güvenirliği", "ölçümlerden yapılmış kullanım ve yorumun geçerliği", "ABC ölçek ölçüm yorum geçerliği", "ölçüm güvenirliği", "ABC testinden elde edilen ölçümlerden yapılmış bir yorumun geçerliği", "ABC testi ölçümlerinin test-tekrar test güvenirliği" ve benzeri ifadeler kullanmak ise, *daha uygun* ve *doğrudur* (AERA, APA & NCME, 1999; Bademci, 2007; Brookhart & Nitko, 2008; Linn & Miller, 2005; McMillan, 2007; Nilsson, Schmidt & Meek, 2002; Reynolds, Livingston & Willson, 2009; Thompson, 2003)

5.1. Geçerlik Üzerinde Kane'in (1990; 1992; 2006a) Bakış Açısı: Tartışma Temelli Geçerleme (Argument Based Validation) ve Bir Kısa Giriş

Yaklaşık son 70 yılda geçerlik kuramı, geçerlik kavramları ve içerdikleri evrim geçirmiştir (Anastasi, 1992; Bademci, 2007; Kane, 2006a; Langenfeld & Crocker, 1994). Bu süreci, "otorite" olarak kabul edilen *Educational Measurement*'ın dört ayrı baskısında ve geçerlikle ilgili dört ayrı bölümünde görebilmek mümkündür; Cureton (1951) geçerliği, test ve ölçüt ölçümleri arasındaki korelasyon terimleri içinde tanımlarken, Cronbach'da (1971) yapı geçerliği merkezde olmuş ve geçerleme (validation) ve çıkarımlara dikkat çekilmiş, Messick (1989) ise, geçerliğe bir bütünleştirilmiş değerlendirme gibi vurgu yapmıştır. Editörlüğünü Brennan'ın (2006a) yaptığı *Educational Measurement*'ın dördüncü baskısında ise, geçerlikle ilgili bölüm Kane (2006a) tarafından ve "validation" ["geçerleme"] başlığı altında yazılmıştır; Kane'de (2006a), tartışma temelli yaklaşım ile, geçerleme için bir genel çerçeve sağlamıştır (Brennan, 2006a; 2006b; Kane, 2004; Kane, 2008). Kane'e (2006a; 2008) göre, geçerleme, ölçmelerin veya ölçümlerin kullanımları ve önerilmiş yorumlarının *değerlendirme*sini içerir. Yine, Kane'e (2006a) göre, geçerlenmiş olan, test veya test ölçümleri *değildir*; geçerlenmiş olan, test sonuçları ya da *test ölçümleri üzerine temellendirilmiş kararlar ve iddialardır*. Kane (1990; 1992) tarafından önerilmiş olan geçerliğe tartışma temelli yaklaşımın, American Educational Research Association (AERA),

American Psychological Association (APA) ve National Council on Measurement in Education (NCME) tarafından yayınlanmış olan "otorite" *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999) ile kabul edilmiş olduğu da görülmektedir (Sireci ve Parker, 2006). *İlk olarak* Bademci (2001; 2002; 2010) tarafından Türk eğitim ve bilim gündemine taşınan Kane'in (1990; 1992; 2006a) geçerliğe tartışma temelli yaklaşımının temelinde, Cronbach'ın (1988) "tartışma gibi geçerlik" önerileri yatmaktadır; Toulmin (1964; 2003; Toulmin, Rieke & Janik, 1984) ve House (1977) ve Cronbach'ın (1982) eserlerinin de, yine bu yaklaşıma katkı sağladığı, asla gözden kaçırılmamalıdır. Kane'in (1992; 2001; 2006a) tartışma gibi geçerlik sunumunda, belirli bir vurgu ile *genellenirlik kuramı* (generalizability theory; Brennan, 2001) [veya güvenirlik (Feldt & Brennan, 1989)] izlerini de görmek mümkündür. Tüm bu katkıların ışığında, Kane'in (1990; 1992; 2001; 2006a; 2008), geçerlik üzerindeki bakış açısının çoğu durumunun, *program değerlendirme* (bakınız, Cronbach, 1982) içindeki ya da çerçevesindeki kavramları çağrıştırdığı da, eklenerek ifade edilebilir (Brennan, 2006a; 2006b). Özetle, geçerlemeye [ya da geçerliğe] tartışma temelli yaklaşım (Kane, 1990; 1992; 2004; 2006a), test ölçümlerinin kullanımları ve önerilmiş yorumlarının geçerliğini değerlendirmek için bir *yöntembilim* [methodology] sağlar ve [yine, *ilk olarak* Bademci (2001; 2002; 2010) tarafından Türk eğitim ve bilim ortamına ayrıntılı olarak taşınan] Messick'in (1989) modeli ile de büyük ölçüde tutarlıdır (Kane, 2004; Brennan, 2006a; 2006b).

6. SONUÇ YA DA GÜVENİRLİK VE GEÇERLİK İLE İLGİLİ ÇAĞDAŞ TANIMLAMALAR

Türk eğitim ve bilimine yönelik olarak, ölçme ve araştırma yöntembiliminde, Bademci (2001a; 2001b; 2002; 2004; 2005a; 2005b; 2005c; 2006a; 2006b; 2006c; 2007; 2008; 2010) tarafından yaklaşık 60 yıl sonra ortaya konulan *yeni paradigma* doğrultusunda, güvenirlik ve geçerlikle ilgili, [burada] dört ayrı *yeni* tanımlama yapılmıştır; [çağdaş ya da güncel] bu tanımlamalar,

> ***güvenirlik**, bir test ya da ölçme aracından elde edilmiş ölçümlerin tutarlılığı veya tekrarlanabilirliği*; ***geçerlik*** ise, *bir test ya da ölçme aracından elde edilmiş ölçümlerden yapılmış belirli yorumların ve kullanımların uygunluğu ve yeterliği* şeklinde, ya da
>
> ***güvenirlik**, belirli bir evrene veya örnekleme uygulanmış bir test ya da ölçme aracından elde edilmiş ölçümlerin tutarlılığı veya tekrarlanabilirliği*; ***geçerlik*** ise, *belirli bir evrene veya örnekleme uygulanmış bir test ya da ölçme aracından elde edilmiş ölçümlerden yapılmış belirli yorumların ve kullanımların uygunluğu ve yeterliği* biçiminde,
>
> veya en genel haliyle, *bir test ya da ölçme aracından elde edilmiş ölçümlerin tutarlılığı veya tekrarlanabilirliğine* ***güvenirlik*** denir; *bir test ya da ölçme*

aracından elde edilmiş ölçümlerin kullanımları ve önerilmiş yorumlarının bir değerlendirmesine ***geçerlik*** denir, şeklinde

ya da en kısa durumuyla, *ölçümlerin tutarlılığı veya tekrarlanabilirliğine* ***güvenirlik***; *ölçümlerin kullanımları ve önerilmiş yorumlarının değerlendirmesine* ***geçerlik*** denir, biçiminde ifade edilebilir:

Kısaca, ***güvenirlik****, ölçümlerin tutarlılığı veya tekrarlanabilirliği* ile, ***geçerlik*** ise, *ölçümlerin yorumu ve kullanımı* ile ilgilidir (Bademci, 2001a; 2007; 2010; Crocker & Algina, 1986; Gronlund, 1998; Gronlund & Waugh, 2009; Kane, 2006a; Linn & Miller, 2005; Brookhart & Nitko, 2008).

---

* Metin içindeki [...] arasındaki ifadeler yazar tarafından eklenmiştir.

KAYNAKLAR

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

Algina, J. (1992). Reliability of Measurement. In Alkin, M. C. (Ed.), *Encyclopedia of Educational Research, Vol. 3*. (Sixth Edition). New York: Macmillan.

Allen, M. J. & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, California: Brooks/Cole.

Anastasi, A. & Urbina, S. (1997). *Psychological Testing*. (Seventh Edition). Upper Saddle River, New Jersey: Prentice-Hall.

Anastasi, A. (1992). What Counselors Should Know About the Use and Interpretation of Psychological Tests. *Journal of Counseling and Development*, Vol. 70, 610-615.

Aycock, T. (1993). *It is Incorrect to Say "the Test is Reliable": A Review of the Literature and Implications for Research Practice*. (ERIC Document Reproduction Service No. ED 355 275).

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. [Konferansla ilgili haber için; *Gazi Haber*, Nisan 2010, Sayı 104, Sayfa 48-49.]

Bademci, V. (2008). Araştırmalarda Ölçme ile İlgili Bazı Büyük Hataları Düzeltmek ve Eğitimde Yeniden Yapılanmayı Sürdürmek: Güvenirlik, Testlerin Bir Özelliği Değildir. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi,* Sayı 22, 50-69. (http://www.esef.gazi.edu.tr/html/yayinlar/22_pdf/22_5.pdf)

Bademci, V. (2007). *Ölçme ve Araştırma Yöntembiliminde Paradigma Değişikliği: Testler Güvenilir Değildir*. Ankara: Yenyap Yayınları.

Bademci, V. (2006a). Güvenirliği Doğru Anlamak ve Bazı Klişeleri Yıkmak: Bilinenlerin Aksine, Cronbach'ın Alfa Katsayısı, Negatif ve –1'den Küçük Olabilir. *İnönü Üniversitesi Eğitim Fakültesi Dergisi,* Cilt 7, Sayı 12, 3-26. (http://web.inonu.edu.tr/~efdergi/arsiv/bademci.pdf)

Bademci, V. (2006b). Tartışmayı Sonlandırmak: Cronbach'ın Alfa Katsayısı, İki Değerli [0,1] Ölçümlenmiş Maddeler ile Kullanılabilir. *Kazım Karabekir Eğitim Fakültesi Dergisi*, Sayı 13, 438-446. (http://e-dergi.atauni.edu.tr/index.php/kkefd/article/viewFile/4116/3940)

Bademci, V. (2006c). *Paradigma Değişikliği: Testler Güvenilir Değildir*. Konferans. Düzenleyen: Gazi Üniversitesi, Endüstriyel Sanatlar Eğitim Fakültesi Dekanlığı.

Ankara: G.Ü. Mesleki Eğitim Fakültesi Konferans Salonu, 28 Nisan. [Konferansla ilgili haber için; *Gazi Haber*, Nisan 2006, Sayı 66, Sayfa 64.]

Bademci, V. (2005a). *Araştırmalarda Ölçme ile İlgili Bazı Büyük Hataları Düzeltmek ve Bir Reformu Başlatmak: Güvenirlik, Testlerin Bir Özelliği Değildir*. Bildiri. Eğitim Fakültelerinde Yeniden Yapılandırmanın Sonuçları ve Öğretmen Yetiştirme Sempozyumu. Ankara: Gazi Üniversitesi, Gazi Eğitim Fakültesi, 22-23-24 Eylül.

Bademci, V. (2005b). Testler Güvenilir Değildir: Ölçüm Güvenirliğine Yeterli Dikkat ve Güvenirlik Çalışmaları İçin Örneklem Büyüklüğü. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi,* Sayı 17, 33-45. (http://www.esef.gazi.edu.tr/html/yayinlar/17_pdf/17_c.pdf)

Bademci, V. (2005c). Hakemlerin Değerlendirmelerindeki Hatalar Üzerine: Fisher'in Z Dönüşümü ve Güvenirlik Çalışmaları İçin Örneklem Büyüklüğü. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi,* Sayı 17, 46-75. (http://www.esef.gazi.edu.tr/html/yayinlar/17_pdf/17_d.pdf)

Bademci, V. (2004). Testin Güvenirliği" veya "Test Güvenilirdir" Diye İfade Etmek Doğru Değildir. *Türk Eğitim Bilimleri Dergisi,* Cilt 2, 367–373. (http://www.tebd.gazi.edu.tr/c2s3.html) (http://www.tebd.gazi.edu.tr/arsiv/2004_cilt2/sayi_3/367-373.pdf)

Bademci, V. (2002). *Türkiye'deki Okullar Ne İşe 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. (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 İşe Yarar?* Konferans. Düzenleyen: Ankara Türk Telekom Anadolu Teknik L. Ankara: Başkent Öğretmenevi Konferans Salonu, 9 Aralık 2001.

Bademci, V. (2000). *Türkiye'deki Okullar Ne İşe Yarar?* (Birinci Basım). Ankara: Başkent Basım Yayın Dağıtım.

Barnes, L. L. B., Harp, D. & Jung, W. S. (2002). Reliability Generalization of Scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, Vol. 62, 603-618.

Bernstein, J. (2006). *Albert Einstein. Fiziğin Sınırları*. (Çev.: Uzunefe Yazgan, Y.). (Birinci Basım). Ankara: TÜBİTAK.

Beycioğlu, K. (2007). Alfa Güvenirliği ve Eğitim Araştırmaları. *Çağdaş Eğitim*, 347, 37-42.

Borsboom, D., Romeijn, J-W. & Wicherts, J. M. (2008). Measurement Invariance Versus Selection Invariance: Is Fair Selection Possible? *Psychological Methods*, Vol. 13(2), 75-98.

Brennan, R. L. (Ed.) (2006a). *Educational Measurement*. (Fourth Edition). Westport, CT: American Council on Education & Praeger.

Brennan, R. L. (2006b). Perspectives on the Evolution and Future of Educational Measurement. In Brennan, R. L. (Ed.), *Educational Measurement*. (Fourth Edition). Westport, CT: American Council on Education & Praeger.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Brookhart, S. M. & Nitko, A. J. (2008). *Assessment and Grading in Classrooms*. Upper Saddle River, New Jersey: Pearson/Prentice Hall.

Buhi, E. R. (2005). Reliability Reporting Practices in Rape Myth Research. *Journal of School Health*, Vol. 75, 63- 66.

Capraro, R. M. & Capraro, M. M. (2002). Myers-Briggs Type Indicator Score Reliability Across Studies: A Meta-Analytic Reliability Generalization Study. *Educational and Psychological Measurement*, Vol. 62, 590-602.

Cebeci, S. (2006). "The Examination of Guidance and Research Centers' Administrators' Conflict Management Strategies with the Perceptions of Self and Teachers".

Unpublished Master's Thesis. Ankara: Middle East Technical University, The Graduate School of Social Sciences.

Chartrand, J. M. & Walsh, W. B. (2001). Career Assessment: Changes and Trends. In Leong, F. T. L. & Barak, A. (Eds.), *Contemporary Models in Vocational Psychology. A Volume in Honor of Samuel H. Osipow*. Mahwah, New Jersey: Lawrence Erlbaum.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Fort Worth: Holt, Rinehart and Winston.

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

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Cronbach, L. J. (1971). Test Validation. In Thorndike, R. L. (Ed.), *Educational Measurement*. (Second Edition). Washington, D. C.: American Council on Education.

Cureton, E. E. (1951). Validity. In Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D. C.: American Council on Education.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. (Fifth Edition). Englewood Cliffs, New Jersey: Prentice Hall.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In Linn, R. L. (Ed.), *Educational Measurement*. (Third Edition). New York: American Council on Education & Macmillan.

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

Gazi Haber (2010). Türk Eğitim ve Biliminde Paradigma Değişikliği: Testler veya Ölçekler Güvenilir ve Geçerli Değildir. Nisan 2010, Sayı 104, 48-49.

Gage, N. L. (1963). Paradigms for Research on Teaching. In Gage, N. L. (Ed.), *Handbook of Research on Teaching*. Chicago: Rand MçNally & Company.

Gray, B. T. (1997). *Controversies Regarding the Nature of Score Validity: Still Crazy After All These Years*. (ERIC Document Reproduction Service No. ED 407 414).

Gronlund, N. E. (1998). *Assessment of Student Achievement*. (Sixth Edition). Boston: Allyn & Bacon.

Gronlund, N. E. & Waugh, C. K. (2009). *Assessment of Student Achievement*. (Ninth Edition). Upper Saddle River, New Jersey: Pearson.

de Gruijter, D. N. M. & van der Kamp, L. J. T. (2008). *Statistical Test Theory for the Behavioral Sciences*. Boca Raton, FL: Chapman & Hall / CRC

Guilford, J. P. (1954). *Psychometric Methods*. (Second Edition). New York: McGraw-Hill.

Guilford, J. P. & Fruchter, B. (1973). *Fundamental Statistics in Psychology and Education* (Fifth Edition). New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley & Sons.

Guthrie, A. C. (2000). *A Review of Coefficient Alpha and Some Basic Tenets of Classical Measurement Theory*. (ERIC Document Reproduction Service No. ED 438 307).

Hambleton, R. K. & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, Vol. 12 (3), 38-47.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.

Henson, R. K. (2000). *Sacrificing Reliability and Exalting Sampling Error at the Altar of Parsimony: Some Cautions Concerning Short-Form Test Development*. (ERIC Document Reproduction Service No. ED 447 211).

Hopkins, K. D. (1998). *Educational and Psychological Measurement and Evaluation*. (Eight Edition). Boston: Allyn & Bacon.

Hotaman, D. & Yüksel-Şahin, F. (2010). The Effect of Instructors' Enthusiasm on University Students' Level of Achievement. *Education and Science* [*Eğitim ve Bilim*], Vol. 35(155), 89-103.

House, E. R. (1977). *The Logic of Evaluative Argument*. CSE Monograph Series in Evaluation, No. 7. Los Angeles: Center for the Study of Evaluation.

Kane, M. T. (2008). Terminology, Emphasis, and Utility in Validation. *Educational Researcher*, Vol. 37(2), 76-82.

Kane, M. T. (2006a). Validation. In Brennan, R. L. (Ed.), *Educational Measurement*. (Fourth Edition). Westport, CT: American Council on Education & Praeger.

Kane, M. (2006b). Content-Related Validity Evidence in Test Development. In Downing S. M. & Haladyna, T. M. (Eds.), *Handbook of Test Development*. Mahwah, New Jersey: Lawrence Erlbaum.

Kane, M. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement*, Vol. 2(3), 135-170.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, Vol. 38, 319-342.

Kane, M. (1996). The Precision of Measurements. *Applied Measurement in Education*, Vol. 9(4), 355-379.

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin*, Vol. 112(3), 527-535.

Kane, M. T. (1990). *An Argument-based Approach to Validation*. ACT Research Report Series, 90-13. Iowa City, Iowa: ACT.

Kartal, H. (2009). Öğretmen Adaylarının Uygulama Okullarındaki Zorbalıkla İlgili Değerlendirmeleri. *GÜ, Gazi Eğitim Fakültesi Dergisi*, Cilt 29(1), 141-172.

Kartal, E. & Pekkanlı, İ. (2011). Yabancı Dil Öğretmen Adaylarının Anadil ve Yabancı Dilde İnternet Üzerinden Okuma Alanları ve Sıklıkları. *International Journal of Human Sciences*, Vol. 8(1), 1316-1326.

Kieffer, K. M. & Reese, R. J. (2002). A Reliability Generalization Study of the Geriatric Depression Scale. *Educational and Psychological Measurement*, Vol. 62, 969-994.

Kieffer, K. M. (1999). Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate. In Thompson, B. (Ed.), *Advances in Social Science Methodology, Volume 5*. Stamford, Connecticut: JAI.

Korkmaz, A. (2010). "Vahit Bademci'nin Paradigma Değişikliği Üzerine Bir Araştırma: "Testler Değil, Ölçümler Güvenilirdir" ". Yayımlanmamış Yüksek Lisans Tezi. Zonguldak: Zonguldak Karaelmas Üniversitesi, Sosyal Bilimler Enstitüsü.

Kubiszyn, T. & Borich, G. (1993). *Educational Testing and Measurement*. Fourth Edition. New York: HarperCollins College Publishers.

Kuhn, T. S. (1995). *Bilimsel Devrimlerin Yapısı*. (Çev.: Kuyaş, N.). (Dördüncü Baskı). İstanbul: Alan Yayıncılık.

Langenfeld, T. E. & Crocker, L. M. (1994). The Evolution of Validity Theory: Public School Testing, the Courts, and Incompatible Interpretations. *Educational Assessment*, Vol. 2(2), 149-165.

Le, V-N. & Klein, S. P. (2002). Technical Criteria for Evaluating Tests. In Hamilton, L. S., Stecher, B. M. & Klein, S. P. (Eds.), *Making Sense of Test-Based Accountability in Education*. Santa Monica, CA: RAND.

Linn, R. L. (2002). Validation of the Uses and Interpretations of Results of State Assessment and Accountability Systems. In Tindal, G. & Haladyna, T. M. (Eds.), *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Impletation*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Linn, R. L. (1995). *Assessment-Based Reform: Challanges to Educational Measurement*. Princeton, New Jersey: Educational Testing Service.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and Assessment in Teaching*. (Eighth Edition). Upper Saddle River, New Jersey: Pearson.

Linn, R. L. & Miller, M. D. (2005). *Measurement and Assessment in Teaching*. (Ninth Edition). Upper Saddle River, New Jersey: Merrill.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.

Magnusson, D. (1967). *Test Theory*. Massachusetts: Addison-Wesley.

McHorney, C. A. (1999). Health Status Assessment Methods for Adults: Accomplishment and Future Challanges. *Annual Review of Public Health*, Vol. 20, 309-335.

McMillan, J. H. (2007). *Classroom Assessment. Principles and Practice for Effective Instruction*. (Fourth Edition). Boston: Allyn and Bacon.

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology*. (Fourth Edition). Fort Worth: Harcourt Brace.

Mellenbergh, G. J. (1999). A Note on Simple Gain Score Precision. *Applied Psychological Measurement*, Vol. 23, 87-89.

Mellenbergh, G. J. (1996). Measurement Precision in Test Score and Item Response Models. *Psychological Methods*, Vol. 1(3), 293-299.

Mellenberg, G. J. & van den Brink, W. (1998). The Measurement of Individual Change. *Psychological Methods*, Vol. 3(4), 470-485.

Messick, S. (1995). Validity of Psychological Assessment. Validation of Inferences From Person's Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, Vol. 50, 741-749.

Messick, S. (1989). Validity. In Linn, R. L. (Ed.), *Educational Measurement*. (Third Edition). New York: American Council on Education & Macmillan.

Miller, C. S., Shields, A. L., Campfield, D., Wallace, K. A. & Weiss, R. D. (2007). Substance Use Scales of the Minnesota Multiphasic Personality Inventory. An Exploration of Score Reliability Via Meta-Analysis. *Educational and Psychological Measurement*, Vol. 67, 1052-1065.

Mji, A. & Onwuegbuzie, A. J. (2004). Evidence of Score Reliability and Validity of the Statistical Anxiety Rating Scale Among Technikon Students in South Africa. *Measurement and Evaluation in Counseling and Development*, Vol. 36, 238-251.

Murhpy, K. R. & Davidshofer, C. O. (2001). *Psychological Testing. Principles and Applications.* (Fifth Edition). Upper Saddle River, New Jersey: Prentice Hall.

Nilsson, J. E., Schmidt, C. K. & Meek, W. D. (2002). Reliability Generalization: An Examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement*, Vol. 62, 647-658.

Nitko, A. J. (2001). *Educational Assessment of Students*. (Third Edition). Upper Saddle River, New Jersey: Merrill/ Prentice-Hall.

Norton, D. (2001). Giriş. "Yerleşik Düşünceler: Verip Veriştirmek". Bouvet, J-F. (Haz.), *Ispanaktaki Demir ve Diğer Yerleşik Düşünceler Üzerine*. (Çev.; Atuk, E.). İstanbul: YKY.

Özsoy, S., Keleş, Ö. & Uzun, N. (2009). *Fen Bilgisi Eğitimi Alanında Hazırlanan Yüksek Lisans Tezlerindeki Yöntem ve İstatistiksel Analiz Hataları*. 1. Uluslararası Türkiye Eğitim Araştırmaları Kongresi. Çanakkale: Çanakkale Onsekiz Mart Üniversitesi, 1-3 Mayıs. (http://oc.eab.org.tr/egtconf/pdfkitap/pdf/238.pdf) 11 Kasım 2009'da alınmıştır.

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. (2009). *Measurement and Assessment in Education*. (Second Edition). Upper Saddle River, New Jersey: Pearson.

Rouse, S. V. (2007). Using Reliability Generalization Methods to Explore Measurement Error: An Illustration Using the MMPI-2 PSY-5 Scales. *Journal of Personality Assessment*, Vol. 88(3), 264-275.

Ragan, B. G. & Kang, M. (2005). Reliability: Current Issues and Concerns. *Athletic Therapy Today*, Vol. 10(6), 30-33.

Rowley, G. R. (1976). The Reliability of Observational Measures. *American Educational Research Journal*, Vol. 13, 51-59.

Sawilowsky, S. S. (2000). Psychometrics Versus Datametrics: Comment on Vacha-Haase's "Reliability Generalization" Method and Some EPM Editorial Policies. *Educational and Psychological Measurement*, Vol. 60, 157-173.

Sayın, S. (2010). Bilimsel Araştırmalarda Yapılan İstatistiksel ve Yöntembilimsel Hatalar-II: Grafik, Tablo ve Gösterim Hataları. *Türk Eğitim Bilimleri Dergisi*, Cilt 8(1), 117-143.

Sayın, S. (2008). Bilimsel Araştırmalarda Yapılan Bazı İstatistiksel ve Yöntembilimsel Hatalar-III: Güvenirlik Kestirimlerine Yönelik Hatalar. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, Sayı 15, 53-69.

Sever, E. (2008). "Öğrenme Stilleri: İlköğretim 6-8. Sınıf Öğrencilerine Yönelik Bir Ölçek Geliştirme Çalışması". Yayımlanmamış Yüksek Lisans Tezi. Aydın: Adnan Menderes Üniversitesi, Sosyal Bilimler Enstitüsü.

Serdar, Z. (2001). *Thomas Kuhn ve Bilim Savaşları*. (Çev.: Kılıç, E.). İstanbul: Everest.

Sireci, S. G. (2005). Unlabeling the Disabled: A Perspective on Flagging Scores From Accommodated Test Administrations. *Educational Researcher*, Vol. 34(1), 3-12.

Sireci, S. G. & Parker, P. (2006). Validity on Trial: Psychometric and Legal Conceptualizations of Validity. *Educational Measurement: Issues and Practice*, Vol. 25(3), 27-34.

Stanley, J. C. (1971). Reliability. In Thorndike, R. L. (Ed.), *Educational Measurement*. (Second Edition). Washington, D.C.: American Council on Education.

Streiner, D. L. & Norman, G. R. (1995). *Health Measurement Scales*. (Second Edition). Oxford: Oxford University Pres.

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

Superfine, B. M. (2004). At the Intersection of Law and Psychometrics: Explaining the Validity Clause of No Child Left Behind. *Journal of Law & Education*, Vol. 33(4), 475-513.

Thompson, B. (Ed.) (2003). *Score Reliability. Contemporary Thinking on Reliability Issues*. Thousand Oaks, California: Sage.

Thompson, B. (2001). Significance, Effect Sizes, Stepwise Methods and Other Issues: Strong Arguments Move the Field. *The Journal of Experimental Education*, Vol. 70, 80-93.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is Datametrics: The Test is Not Reliable. *Educational and Psychological Measurement*, Vol. 60, 174-195.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton Mifflin.

Topdemir, H. G. (2002). Kuhn ve Bilimsel Devrimlerin Yapısı Üzerine Bir Değerlendirme. *Felsefe Dünyası*, Sayı 36, 45-62.

Toulmin, S. E. (2003). *The Uses of Argument*. (Updated Edition). New York: Cambridge.

Toulmin, S. E. (1964). *The Uses of Argument*. London: Cambridge.

Toulmin, S., Rieke, R. & Janik, A. (1984). *An Introduction to Reasoning*. New York: Macmillan.

Traub, R. E. (1994). *Reliability for the Social Sciences. Theory and Applications*. Thousand Oaks: Sage.

Traub, R. R. & Rowley, G. L. (1991). Understanding Reliability. *Educational Measurement:Issues and Practice*, Vol. 10(1), 37-45.

Tyson, E. H., Dulmus, C. N. & Wodarski, J. S. (2002). Assessing Violent Behavior. In Rapp-Paglicci, Roberts, A. R. & Wodarski, J. S. (Eds.). *Handbook of Violence*. New York: John Wiley & Sons.

Vacha-Haase, T. (1998). Reliability Generalization: Exploring Variance in Measurement Error Affecting Score Reliability Across Studies. *Educational and Psychological Measurement*, Vol. 58, 6-20.

Vacha-Haase, T., Kogan L. R., Tani, C. R. & Woodal, R. A. (2001). Reliability Generalization: Exploring Variation of Reliability Coefficients of MMPI Clinical Scales Scores. *Educational and Psychological Measurement*, Vol. 61, 45-59.

van der Linden, W. J. (2005). Classical Test Theory. In Kempf-Leonard, K. (Ed.), *Encyclopedia of Social Measurement*. Oxford: Elsevier.

Vassar, M. & Hale, W. (2009). Reliability Reporting Across Studies Using the Buss Durkee Hostility Inventory. *Journal of Interpersonal Violence*, Vol. 24, 20-37.

Victorson, D., Barocas, J., Song, J. & Cella, D. (2008). Reliability Across Studies From the Functional Assessment of Cancer Theraphy-General (FACT-G) and Its Subscales: A Reliability Generalization. *Quality of Life Research*, Vol. 17, 1137-1146.

Wasserman, J. D. & Bracken, B. A. (2003). Psychometric Characteristics of Assessment Procedures. In Weiner, I. B., Graham, J. R. & Naglieri, J. A. (Eds.), *Handbook of Psychology*. Hoboken, New Jersey: John Wiley & Sons.

Witta, E. L. & Daniel, L. G. (1998). *The Reliability and Validity of Test Scores: Are Editorial Policy Changes Reflected in Journal Articles?* (ERIC Document Reproduction Service No. ED 422 366).

Worthen, B. R., White, K. R., Fan, X. & Sudweeks, R. R. (1999). *Measurement and Assessment in Schools*. (Second Edition). New York: Longman.

Yin, P. & Fan, X. (2000). Assessing the Reliability of Beck Depression Inventory Scores: Reliability Generalization Across Studies. *Educational and Psychological Measurement*, Vol. 60, 201-223.