Eğitim Fakültesi Dergisi Cilt: 7 Sayı:12 Güz 2006

s. 3- 26

GÜVENİRLİĞİ DOĞRU ANLAMAK VE BAZI KLİŞELERİ YIKMAK: BİLİNENLERİN AKSİNE, CRONBACH'IN ALFA KATSAYISI, NEGATİF VE "–1" DEN KÜÇÜK OLABİLİR

Vahit BADEMCİ *

Özet

Güvenirlik, aracın kendisine değil, bir ölçme aracıyla elde edilmiş ölçümlere (veya sonuçlara) işaret eder. Güvenirlik testin kendisinin değil, eldeki veriler veya ölçümlerin bir özelliğidir. Böylelikle, "testin güvenirliği" veya "aracın güvenirliği" ya da "test güvenilirdir" diye ifade etmek doğru değildir. Zira ölçümler güvenilir veya güvenilir değildir. Basit şekliyle, testler güvenilir değildir. Cronbach'ın (1951) alfa katsayısı yöntemi, ölçüm güvenirliği kestirimi için kullanılmaktadır. Alfa katsayısı değişik faktörler tarafından etkilenmektedir. Örneğin, toplam test ölçüm varyansı, alfa katsayısı üzerinde en büyük etkiye sahiptir. Böylece, toplam test ölçümlerinin varyansı, madde varyansları toplamından daha küçükse, sonrasında Cronbach'ın alfa katsayısı negatif olabilecektir. Matematiksel olarak, Cronbach alfa, negatif ve "–1"'den küçük olabilir.

Anahtar Kelimeler: *Ölçüm (score)*** *güvenirliği, güvenirlik, Cronbach'ın alfa katsayısı.*

* *Gazi Üniversitesi, Endüstriyel Sanatlar Eğitim Fakültesi, Eğitim Bilimleri Bölümü,* bademci@gazi.edu.tr

** Ölçüm; (score)

M. Fuat Turgut, ölçme işlemleri sonunda elde edilen sayılara ölçüm denilmesini önermektedir (Bademci,1999:7–8).

Ölçümleme; (scoring)

Bellilendirme; (assessment)

TO UNDERSTAND RELIABILITY PROPERLY AND TO OVERTHROW SOME CLICHÉS: CONTRARY TO THE KNOWN FACTS, CRONBACH'S COEFFICIENT ALPHA CAN BE NEGATIVE AND SMALLER THAN "–1"***

Abstract

Reliability refers to the scores (or results) obtained with an measurement instrument and not to the instrument itself. Reliability is a characteristic of scores or the data in hand, not of the test itself. Thus, it is incorrect to speak of "the reliability of the test" or "the reliability of the instrument" or "the test is reliable." Because, it is the scores that are reliable or unreliable. Simply, tests are not reliable.

Cronbach's (1951) coefficient alpha method is used for estimating score reliability. Coefficient alpha is affected by various factors. For example, total test score variance has the biggest effect on coefficient alpha. Thus, if the variance of total test scores is less than the the sum of the item variances, then Cronbach's coefficient alpha will be negative. Mathematically, Cronbach's alpha can be negative and smaller than" –1"

Key Words: *Score reliability, reliability, Cronbach's coefficient alpha.*

GİRİŞ

*** Cronbach alfa katsayısının negatif ve "–1"den küçük değerler alabileceği hususu, Bademci (2001b; 2002) tarafından çeşitli konferanslarda dile getirilmiştir. Cronbach alfa katsayısı ile ilgili bu husus, 2005 yılında düzenlenen bir sempozyumda "*Araştırmalarda Ölçme İle İlgili Bazı Büyük Hataları Düzeltmek ve Bir Reformu Başlatmak: Güvenirlik, Testlerin Bir Özelliği Değildir*" başlıklı bir bildiri içinde, bazı örnekleri ve ana hatlarıyla sempozyumdaki izleyicilere tekrar aktarılmıştır (Bademci, 2005a). Sempozyumu izleyen kimi bilim adamları, sunulan bildirinin ve bildirinin içinde aktarılan Cronbach alfa katsayısı ile ilgili hususların yeni ve önemli olduğunu ve de Cronbach alfa katsayısı ile ilgili bu hususun genişletilerek ayrı bir çalışma biçiminde yeniden sunulmasını da beklediklerini ifade etmişlerdir. Bu görüşler paralelinde, Cronbach alfa katsayısı ile ilgili bu husus, bu makalede oldukça genişletilerek yeniden hazırlanmıştır. Ancak, ilgili konuyla bir bütünlük sağlanması ve bir değerler dizisi değişikliği olarak Türk eğitim bilim topluluğunun gündemine Bademci (2001b; 2002; 2004; 2005a; 2005b; 2005c; 2006) tarafından taşınılan, "güvenirliğin, testlerin bir özelliği olmadığı" hususunun daha iyi vurgulanması gibi bazı zaruretlerden dolayı, "*Araştırmalarda Ölçme İle İlgili Bazı Büyük Hataları Düzeltmek ve Eğitimde Yeniden Yapılanmayı Sürdürmek: Güvenirlik, Testlerin Bir Özelliği Değildir*" (Bademci, 2005d) başlıklı ve ilgili ve de bağlantılı bir diğer çalışmadaki bazı ifadelerden, aynen veya değiştirilmiş biçimde, bu çalışmada da yararlanılmıştır.

Bu makale, Cronbach alfa katsayısının -1 ve -1'den küçük değerler alabileceği hususunu, ölçüm güvenirliğiyle etkileşimli ve bağlantılı bir biçimde ortaya koymaya çalışmakta ve bu konu üzerindeki tartışmayı da sonlandırmayı amaçlamaktadır

Test ölçümlerinin güvenirliğini kestirmenin yöntemlerinden birisi de, Cronbach'ın (1951) alfa katsayısı yöntemidir (Linn ve Miller, 2005; Mehrens ve Lehmann, 1991). Alfa katsayısı, aşağıdaki formül yoluyla hesaplanmaktadır (Crocker ve Algina, 1986; Reinhardt, 1996; Worthen, White, Fan ve Sudweeks, 1999).

$$[\text{Alfa}]\ \alpha = k / (k-1)*[1- (\Sigma\sigma^2_i / \sigma^2_T)]$$

k = test üzerindeki madde sayısı
σ^2_i = i madde ölçüm varyansı [ya da bir madde üzerindeki bir grup bireyden elde edilen ölçümlerin varyansı]
$\Sigma\sigma^2_i$ = i madde ölçüm varyanslarının toplamı
σ^2_T = toplam test ölçümlerinin varyansı

Bilinenlerin Aksine, Cronbach Alfa, Negatif ve -1'den Küçük Değerler Alabilir

Tablo 1'de ayrışık ve bağdaşık iki ayrı örneklem için denencel veriler sunulmuştur. Buna göre, 10 kişiden oluşan ayrışık ve bağdaşık örneklemlere, 7 maddelik bir test uygulanmış ve maddeler doğru (1) ve yanlış (0) olarak ölçümlenmiştir.

Alfa katsayısı, "…madde varyansları toplamı, madde güçlüğü ve toplam test ölçüm varyansı tarafından etkilenmektedir" (Helms, 1999: 10). "Reinhardt (1996), …*toplam test ölçüm varyansının alfa katsayısı üzerinde en büyük etkiye sahip olduğunu göstermiştir*"(Bulunduğu yer, Helms, 1999: 10). [Cümle, Bademci tarafından italik yazdırılmıştır.][****] Daha küçük toplam test ölçüm varyansı, daha küçük alfa katsayısına, daha büyük toplam test ölçüm varyansı ise, daha büyük alfa katsayısına vesile olmaktadır (Arnold, 1996; Helms, 1999). Zira klasik test kuramı güvenirlik kestirimleri toplam test ölçüm varyansı tarafından (Capraro, Capraro ve Henson, 2001), toplam test ölçüm varyansı da, sınavı alan grubun ne derece bağdaşık ya da ayrışık olmasından çokça etkilenmektedir (Helms, 1999). Eğer bir test, bağdaşık [homojen] bir gruba verilirse, toplam test ölçümü içindeki değişkenlik azalacak, dolayısıyla alfa katsayısı küçülecek, aynı test daha ayrışık [heterojen] bir gruba verilirse toplam test ölçümü içindeki değişkenlik artacak, dolayısıyla alfa katsayısı da büyüyecektir (Arnold, 1996; Helms,

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

1999). Kuramsal olarak da beklendiği üzere, bağdaşık örnekleme [σ^2= 0.25; α= -7.00] göre ayrışık örneklem [σ^2 =3.36; α= .8333], daha değişken test ölçümlerine, daha büyük toplam test ölçümleri varyansına ve daha büyük alfa katsayısına yol açmıştır.

Tablo 1. Ayrışık ve Bağdaşık Örneklemler İçin Denencel Veriler*

Kişiler	Maddeler							Toplam Ölçüm
	1	2	3	4	5	6	7	
				Ayrışık Örneklem				
A	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0
F	0	0	0	1	1	1	1	4
G	0	0	0	0	1	1	1	3
H	0	0	0	1	1	1	1	4
I	0	0	0	0	1	1	1	3
İ	0	0	0	1	1	1	1	4
Madde σ^2	0	0	0	.21	.25	.25	.25	
Toplam ölçümlerin σ^2								3.36
α = k / (k–1)*[1- ($\Sigma\sigma^2_i$ / σ^2_T)] = (7/6) x [1- (0.96/ 3.36)] = **.8333**								
				Bağdaşık Örneklem				
A	1	0	1	0	1	0	1	4
B	0	1	0	1	0	1	0	3
C	1	0	1	0	1	0	1	4
D	0	1	0	1	0	1	0	3
E	1	0	1	0	1	0	1	4
F	0	1	0	1	0	1	0	3
G	1	0	1	0	1	0	1	4
H	0	1	0	1	0	1	0	3
I	1	0	1	0	1	0	1	4
İ	0	1	0	1	0	1	0	3
Madde σ^2	.25	.25	.25	.25	.25	.25	.25	
Toplam ölçümlerin σ^2								0.25
α = k / (k-1)*[1- ($\Sigma\sigma^2_i$/σ^2_T)] = (7/6) x [1- (1.75/ 0.25)] = **-7.0000**								

* Tablo, Dawson (1997), Henson (2000b) ve Reinhardt'tan (1996) uyarlanmıştır.

Tablo 1'e bağlı olarak verilen denencel örnekten de görüleceği üzere, güvenirlik (alfa katsayısı) örneklem özelliklerinden güçlü şekilde etkilenmektedir. Bu durumda aynı ölçme aracı [veya test], daha ayrışık ya da daha bağdaşık öğrencilerden oluşan gruplara ya da örneklemlere uygulandığında, birbirlerini onamayan ölçümler güvenirliği ortaya çıkacaktır (Thompson, 1994a). Tüm bu bulgular, güvenirliğin, örnekleme, dolayısıyla örneklemden elde edilen ölçümlere bağlı olduğunu göstermektedir (Guthrie, 2000). Buradan, güvenirliğin testlerin değil, eldeki verilerin veya ölçümlerin bir özelliği olduğu (Thompson, 1994a; Thompson, 1999) ifade edilebilir. O halde güvenirliği, ölçümlerin değil de, testlerin veya ölçme araçlarının bir özelliği olarak kabul etmek, araştırmalarda bu "bilinçsiz paradigmatik inanç" (Cousin ve Henson, 2000: 6) doğrultusunda hareket etmek ve "bir ölçme aracına işaret ettiği zaman kullanılan 'test güvenilirdir' veya 'testin güvenirliği' ifadelerini" (Guthrie, 2000) kullanmak, *doğru değildir*.

Türkçe literatürde, *güvenirlik katsayısı ile ilgili olarak,* Baykul'da (2000: 143) "$\rho(X,X')$ korelâsyonuna güvenirlik katsayısı adı verilir... Korelasyon katsayısı [-1,1] aralığında değerler almasına rağmen, $\rho(X,X') = \rho^2(X,T)$ eşitliğinden dolayı, güvenirlik katsayısı negatif değerler alamaz ve [0,1] aralığında değişir... Pratikte $\rho(X,X')$ korelasyonunu [güvenirlik katsayısını]**** hesaplamaya yarayan yöntemler geliştirilmiştir. Bunlar, test-tekrar test, paralel formlar, eşdeğer yarılar, maddelerin testin bütünüyle korelasyonu [α, KR-20, KR-21]**** ve varyans analizine dayalı yöntemlerdir" ifadesi mevcuttur. Yine bir başka çalışmasında Baykul (2001: 16), "Ölçme araçlarının veya ölçme sonuçlarının güvenirlikleri hesaplanabilir, elde edilen sayıya güvenirlik kat sayısı denir. Güvenirlik kat sayısı bir korelasyon kat sayısı olup (0, +1) aralığında değerler alır" ifadesini kullanmıştır. Aynı konu ile ilgili olarak Turgut (1993:33), "Korelasyon katsayısı –1,00 ile +1,00 arasında değişmekle birlikte, güvenirlik katsayıları hemen her zaman 0,00 ile +1,00 arasında değişir" ifadesini [benzer ifade, Tekin'de de (1982:58) bulunmaktadır], Özçelik (1989:113) "Güvenirlik tahmininde izlenen yöntem [test-tekrar-test, iki yarım yolu, eşdeğer (paralel) formlar, KR-20]**** ne olursa olsun, güvenirlik tahmini sonucunda 0,00 ile 1,00 arasında bir korelasyon elde edilir" ifadesini, Öncü (1997: 255), "Bu sebeple güvenirlik katsayısı 0 ile +1 arasında değişmektedir. Ancak, 1,00 ve 0,00 gibi uç değerlere pek rastlanmaz" ifadesini, Akdeniz, Aydemir, Akdeniz, Gülseren ve Kültür (1999: 106) "Cronbach alfası 0 ile 1 arasında değişen değerlere sahip bir korelasyon katsayısıdır" ifadesini, Gelbal (1999: 237 ve 2002: 108) "İki değişken arasındaki korelasyon katsayısı –1.00 ile +1.00 arasında

değişir, ancak bir testin güvenirliği hemen her zaman 0,00 ile +1,00 arasında değişir" ifadesini, Anıl ve arkadaşları (2003: 86–87), "Bu sayısal değer genellikle bir korelasyon katsayısıyla ifade edilmesine karşın, güvenirlik katsayısı daima 0 ile +1 arasında değerler almaktadır" ve "KR–20 ve KR–21 güvenirlik katsayıları testi oluşturan maddelerin iç tutarlılığının bir ölçüsüdür ve bu güvenirlik katsayıları 0.00 ve 1.00 arasında değişen değerler alır" ifadelerini, Erkuş (2003: 36) "Güvenirlik mutlaka görgül yollarla saptanır ve sayısal bir değerle ifade edilir. Bu sayısal değer genellikle bir korelasyon katsayısıyla ifade edilmesine karşın, güvenirlik katsayısı daima 0-1 arasında bir değer alır" ifadesini, Özdamar (2004: 623) her ne kadar "Sorular arasında negatif korelasyon varsa Alfa katsayısı da negatif çıkar" ifadesi kullanmışsa da, yine aynı sayfada "Cronbach alfa katsayısı, 0 ile 1 arasında değişim gösterir" ifadesini, Tan ve Erdoğan (2004:176 ve 183) "Güvenirlik katsayıları genellikle 0 ile 1 arasında bir değer alır" ve "Alpha güvenirlik katsayısı da tıpkı KR_{20} güvenirlik katsayısı gibi bir iç-tutarlılık katsayısıdır ve 0 ile 1.00 arasında değer almaktadır" ifadelerini, Aygin ve Eti Aslan (2005:396) dört farklı kaynağı da referans göstererek "Cronbach Alfa katsayısı 0.0 ile 1.00 arasında değişim gösterir ve 1'e ne kadar yakınsa o kadar güvenilir olduğu düşünülür" ifadesini kullanmışlardır. Ancak bilinenlerin aksine, son yıllardaki yapılan çalışmalarda (Arnold, 1996; Bademci, 2001b; Bademci, 2002; Bademci, 2005a; Cousin ve Henson, 2000; Dawson, 1997; Henson, 2000a; Henson, 2000b; Henson, 2001; Reinhardt, 1996) ölçüm güvenirliğinin kestiriminde kullanılan ***a katsayısının***, -ki alfa, [iç tutarlılık] güvenirlik kestirimlerinde kullanılan bir güvenirlik katsayısı olarak ifade edilmektedir- [*tüm maddeler 0,1 ile ölçümlendiğinde* de, kısaca, $\sigma^2_i = p_i\ q_i$, *a= KR-20*] ***negatif ve de -1'den küçük değerler alabileceği matematiksel olarak ortaya konulmuştur***. Denencel veriler kullanılan yukarıdaki tabloda da, bu durum [**α = -7.00**] görülmektedir. Tablo 1'deki denencel verilerden de görüleceği üzere, *alfa* (α) [veya KR 20] iç tutarlılık *güvenirlik katsayıları*nın, [pratikte kabul edilemez gibi görünse de] *matematiksel olarak*, *negatif ve –1'den daha küçük değerler alabileceği* ifade edilebilir. Bu sebeple, güvenirlik katsayıları ile ilgili yapılan bir açıklamanın ardından, "güvenirlik katsayılarının 0 ile 1 arasında değiştiği" şeklinde *genel bir yorum* yapılması yerine, [en azından Cronbach α veya KR–20 göz önünde tutularak] güvenirlik katsayıları ile ilgili *daha dikkatli ve özenli yorumlar* yapılması gereklidir; Türkçe literatüre 'kalıplaşmış biçimde' yerleşmiş olan, güvenirlik katsayıları ile ilgili 'güvenirlik katsayıları negatif değerler alamaz ya da [hemen her zaman, daima, genellikle] 0,00 ile +1,00 arasında değişir' gibi *genel ve kesin ifadeler*in de [yine, en azından Cronbach α veya KR–20 yönünden] düzeltilmesi, yerinde ve doğru olacaktır. [Cronbach'ın alfa katsayısıyla ve ölçüm güvenirliğini etkileyen bazı faktörlerle ilgili olarak Arnold (1996), Cousin ve Henson (2000), Dawson (1997), Henson (2000a),

Henson (2001) ve Reinhardt (1996) tarafından yapılan çalışmalar, etkili çalışmalar olarak göze çarpmaktadır ve incelenmesi amacıyla okuyucuya tavsiye edilir.]

"Test Güvenilirdir" veya "Testin Güvenirliği" Diye İfade Etmek Doğru Değildir

"Test ölçümlerinin güvenirliği" gibi uzun ama doğru bir ifadeyi, "testin güvenirliği" biçiminde kısaltarak konuşma tarzının (Thompson, 1999; Thompson ve Vacha-Haase, 2000), güvenirliğin doğası hakkında yanlış anlamalara yol açtığı ileri sürülmüştür (Baugh, 2002). Pedhazur ve Schmelkin'e (1991:82) göre, bir ölçünün güvenirliğine dair ifadeler, uygun değildir ve olanak dahilinde [uygun şart veya durum sağlandığında] yanlış yola sevkeder. "Testin güvenirliği" biçiminde kısaltarak ifade etme yolu doğal olarak şüpheli görünmemektedir, ancak sonrasında, "testin güvenirliği" şeklindeki kısaltılan ifadeye, bilinçsizce, kelime kelimesine [kısaltılan] aslına uygun anlam yüklenir ve bu da doğru değildir (Thompson, 1994a; Thompson, 1999).

Thompson (1994a), çok az araştırmacının, güvenirliğin eldeki veriler veya ölçümlerin bir özelliği olduğunu bilinçli kabul edip, buna göre davrandıklarını belirtmiştir. Ancak, pek çok insan da ölçüm güvenirliğini [hâlâ] anlamamıştır (Mittag ve Thompson, 2000; Vacha-Haase, Kogan ve Thompson, 2000).

Ölçüm Güvenirliğinin Özü

Her ne kadar test güvenirliğinin işevuruk bir tanımı başlığı altında da olsa, güvenirliğin testin bir özelliği olmadığını yorumlayan Ebel (1972) ve de Gronlund (1965), bu tartışma konusunun ilkleri arasında sayılabilir. Ebel'in (1972) [buradaki] yorumuna benzer bir vurgu, ölçme aracından ziyade ölçmelerin güvenirlik özelliğine sahip olduğu şeklinde, Guilford ve Fruchter'dan (1973) gelmiştir. Bu konuda başı çeken kişinin ise, Rowley (1976) olduğu söylenebilir. Güvenirliğin bir aracın (örneğin, test) değil ölçmenin bir özelliği olduğunu açıklamaya çalışan Rowley (1976: 53), bu konuda net bir ifade kullanmıştır; "…bir aracın kendisi ne güvenilirdir, ne de güvenilir değildir."

Güvenirlik, Aracın Kendisine Değil Bir Ölçme Aracı ile Elde Edilmiş Ölçümlere İşaret Eder

Buraya kadar yapılan açıklamalara ilave edilebilecek aydınlatıcı ifadeler, Gronlund ve Linn'in l990 tarihli çalışmasında mevcuttur. Gronlund ve Linn, (1990: 78) güvenirliğin, aracın kendisine değil bir değerlendirme aracı ile elde edilmiş ölçümlere işaret ettiğine dikkat çekerek, aracın veya testin yerine, ölçmenin veya test ölçümlerinin güvenirliğinden bahsetmenin çok daha uygun olduğunu belirtmiştir. Bu görüşü destekleyici ve açımlayıcı bazı tartışmalar, Bademci (2001a; 2001b; 2002; 2004; 2005a; 2005b; 2005c), Ebel ve Frisbie (1991), Thompson (1994a; 2001), Thompson ve Vacha-Haase (2000) ve Vacha-Haase'de (1998) vardır.

Henson ve Thompson (2002), Gronlund ve Linn'in (1990) yukarıda da belirtilen görüşünün, American Educational Research Association, American Psychological Association ve National Council on Measurement in Education (AERA/APA/NCME) test etme standartlarının, Standart 2.l ve 2.2'sine yansımış olduğunu ifade etmiştir. Benzer düşünce APA Task Force on Statistical Inference (Wilkinson ve APA Task Force on Statistical Inference, l999) tarafından, "güvenirlik, sınavı alanların belirli bir evreni için bir test üzerindeki ölçümlerin bir özelliğidir" şeklinde ifade edilmiştir.

[Test ölçümlerinin güvenirliği ve de güvenirliğin testin kendisinin değil, test ölçümlerinin [belirli] bir grubunun bir özelliği olduğunu vurgulama hususunda, Gronlund ve Linn (1990) ve Ebel ve Frisbie'nin (1991) derli toplu çalışmaları önemli bir yer tutsa da, unutulmaması ve gözden kaçırılmaması gereken bir nokta, güvenirliğin eldeki veriler veya ölçümlerin bir özelliği olduğunu ısrarla vurgulayan, etkin bir şekilde tartışmalara katılan ve konunun kristalleşmesine önemli katkılarda bulunan -ve alanın en önde gelen süreli yayınlarından olan *Educational and Psychological Measurement* dergisinin 1995–2003 yılları arasında editörlüğünü de yapan- Thompson'un (1994a; 1994b), ölçme ve ölçmeyi yakından ilgilendiren değişik konularda, yurt dışındaki köklü reform hareketinin -özellikle 1994 yılından bu yana- başlatıcısı, öncüsü olduğu ve de düşünceleriyle, reform hareketi üzerinde de oldukça etkili bir bilim adamı durumunda bulunduğu söylenebilir.]

Güvenirlik, Ölçümlerin Bir Özelliğidir

Güvenirliğin ölçümlerin bir özelliği olduğuna dair belki de en çarpıcı açıklama Crocker ve Algina'dan (1986) gelmiştir. Crocker ve Algina (1986) güvenirlik katsayısını etkileyen faktörlerden grup bağdaşıklığını (homojenliğini) tartışmış ve denencel bir örnek vererek, güvenirliği, sınavı alanların belirli bir grubu için bir test üzerindeki ölçümlerin bir özelliği şeklinde ifade etmiştir. Bir başka söyleyişle güvenirlik, sınava giren belirli bir gruba uygulanmış bir testten elde edilmiş ölçümlerin bir özelliğidir. Yani güvenirlik, test sonuçlarının bir özelliğidir (Livingston, 1988). Livingston (1988), test sonuçlarının güvenirliğinin ise, testi alan öğrencilerin grubuna bağlı olacağına dikkat çekmiştir.

Grup Bağdaşıklığı veya Ayrışıklığı, Güvenirliği Etkileyen Bir Faktördür

Belirli bir grupta testi alan öğrencilerin veya kişilerin kendileri, ölçümlerin güvenirliğini etkilemektedir. Hal böyle iken, testin bir gruba uygulandığını dikkate almaksızın testin güvenirliğinden bahsetmek anlamsız olacaktır. Güvenirlik, varyans tarafından yönlendirilmektedir; daha büyük ölçümler varyansı, daha büyük ölçümler güvenirliğine olanak tanır (Thompson, 1994a). Böylece daha ayrışık (heterojen) örneklemler sıklıkla daha çok değişken ölçümlere ve bu durumda daha yüksek güvenirliğe yol açar (Thompson, 1994a). Bu durumun mantığı, Klasik Kuram (Suen, l990), Klasik Güvenirlik Modeli (Thorndike, l982), Klasik Test Kuramı (Pedhazur ve Schmelkin, 1991), Klasik Test Kuram Modeli (Lord ve Novick, l968), Klasik Gerçek Ölçüm Modeli (Crocker ve Algina,l986), Klasik Gerçek Ölçüm Kuramı (Allen ve Yen, l979) gibi adlandırılan ölçme kuramının bazı eşitliklerinden yararlanılarak ve bu kuramın derinliğine girilmeden açıklanabilir.

Güvenirlik katsayısı matematiksel olarak aşağıdaki şekilde tanımlanabilir (Allen ve Yen, l979).

$\rho_{XX'} = \sigma^2_T / \sigma^2_X$
$\rho_{XX'}$ = güvenirlik katsayısı (X ve X' paralel ölçmeler)
σ^2_T = gerçek ölçüm varyansı
σ^2_X = gözlenmiş ölçüm varyansı

Güvenirlik katsayısı, gerçek ölçüm varyansının gözlenmiş (toplam) ölçüm varyansına oranıdır (Allen ve Yen, l979; Nunnally ve Bernstein, l994).

Güvenirlik katsayısıyla ilgili yukarıda verilen formül, öteki biçimde de yazılabilir (Allen ve Yen, l979; Pedhazur ve Schmelkin, 199l).

$$\rho_{XX'} = \sigma^2_T / \sigma^2x = \sigma^2x - \sigma^2_E / \sigma^2x = 1- \sigma^2_E / \sigma^2x$$

σ^2_E = hata ölçüm varyansı

$\rho_{XX'} = 1- \sigma^2_E / \sigma^2x$ formülü, diğer şeyler eşit olmak üzere, daha ayrışık gruptan daha yüksek güvenirlik elde edileceğini açıklayıcı niteliktedir (Allen ve Yen, l979; Mehrens ve Lehmann, 1991). Aynı test, benzer büyüklükte biri bağdaşık diğeri ayrışık iki gruba uygulansın. Hata ölçüm varyansı eşit olma sayıltısı altında, ayrışık gruptaki gözlenmiş ölçüm varyansı, bağdaşık gruptaki gözlenmiş ölçüm varyansından daha büyük olacaktır. Çünkü ayrışık gruptaki ölçümler, bağdaşık gruptaki ölçümlere göre çok daha değişken olacaktır. Böylelikle ayrışık gruptaki gözlenmiş ölçüm varyansı büyüyeceğinden, güvenirlikte buna bağlı olarak artacaktır. Bu durumu, belki de en iyi Dawis (1987: 486) açıklamıştır; "çünkü güvenirlik, aracın olduğu kadar, örneklemin de bir fonksiyonudur. [Zira] güvenirlik, tasarlanmış hedef evrenden [alınmış] bir örneklem üzerinde değerlendirilmektedir", ancak Dawis'inde (1987: 486) ifade ettiği gibi bu nokta, "bazen gözden kaçırılmıştır."

Ölçüm Güvenirliği, Örneklemden Örnekleme Değişir

Örneklem özellikleri ölçüm güvenirliğini etkileyebilmekte (Henson, Kogan ve Vacha-Haase, 2001), bir testin veya ölçme aracının uygulandığı örneklemin bağdaşık ya da ayrışık 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. Hal böyle iken "test güvenilirdir" ya da "testin güvenirliği" demek ve güvenirliği, testin veya aracın bir özelliği gibi ima veya ifade etmek uygun değildir, *doğru değildir*.

Thompson (1992; 1994a; 1999; Thompson ve Snyder, 1998; Thompson ve Vacha-Haase 2000), "test ölçümlerinin güvenirliği" yerine, testin güvenirliği biçiminde kısaltarak konuşmayı [ve de yazmayı], "dikkatsiz (sloopy)" [sloppy; düzensiz-disiplinsiz-dikkatsiz-özensiz-şapsal anlamları da var, (Töreci, 2005)] konuşma olarak nitelendirmektedir. "Dikkatsiz" konuşma ise, bazen, "dikkatsiz" düşünmeye, "dikkatsiz" uygulamaya ve daha fazla zararlı bir sonuca kılavuzluk etmektedir (Thompson,1992; Thompson, 1994a; Vacha-Haase, 1998).

"Testler Güvenilirdir" Şeklindeki Hatalı Söylemin Bazı Sonuçları

Ölçme, sosyodavranışsal araştırmanın "Aşil topuğu"dur (Pedhazur ve Schmelkin, 1991: 2) yani, zayıf noktasıdır (Bademci, 2005d). Bu duruma, öncelikle lisansüstü programlar olmak üzere, yüksek lisans ve özellikle doktora programlarındaki ölçmeyle ilişkili konuların giderek azalmasının ve de ölçme konularıyla ilgili zayıf ve kalitesiz eğitim verilmesinin sebep olduğu söylenebilir. Bu hususla ilgili olarak, Pedhazur ve Schmelkin (1991), özellikle doktora programlarında, araştırma deseni ve istatistiklere bir nebze de olsa gereksinim gösterilirken, ölçme üzerindeki vurgunun azaldığını, böylelikle de, pek çok öğrencinin, ölçülerin kullanılması ve geliştirilmesi için zorunlu olan özel yeterliklerin hiçbirini elde edemediğini, bunun sonucunda da, pek çok araştırmada kullanılmış olan ölçülerin özelliklerine hiç dikkat edilmediğini veya çok az dikkat edildiğini, ortaya çıkan bu durumun ise, beklenmedik bir olay olmadığını ileri sürmüştür (Henson,2000a; Henson, 2001; Pedhazur ve Schmelkin, 1991; Thompson, 2001). Bu izlenim, Aiken ve arkadaşları (1990) tarafından yapılan bir çalışmada, doktora eğitim programları içindeki ölçmeyle ilişkili konuların esasen azaldığı şeklinde teyit edilmiş, bir başka ifadeyle doktora eğitim programları içindeki bu ölçme boşluğu, Aiken ve arkadaşlarınca da (1990) doğrulanmıştır (Aiken ve arkadaşları, 1990; Henson, 2001; Thompson, 1999). Eğitimin bu yetersizliği, yayımlanmış araştırmaların kalitesi üzerinde de kendisini göstermiştir. *American Educational Research Journal (AERJ)* ile ilgili yaptığı ve 1980 yılında yayımlanmış çalışmasında Wilson, *AERJ* makalelerinin yalnızca % 37'since analiz edilmiş veriler için güvenirlik katsayılarının açıkça rapor edildiğini, diğer % 18'inin daha önceki araştırmaya referans verme suretiyle güvenirliği dolaylı olarak rapor ettiğini, yayımlanmış araştırmanın yaklaşık yarısında ise, [mazeret kabul edilemez biçimde] güvenirliğin rapor edilmediğini ifade etmiştir (Bulunduğu yer, Thompson,1994a; Thompson, 1994b; Vacha-Haase,1998).

Meier ve Davis (1990) ise, *Journal of Counseling Psychology (JCP)* 'nin 1967, 1977 ve 1987 ciltleri üzerine yaptıkları çalışmalarında, üç *JCP* cildi içinde, tanımlanmış ölçeklerle ilgili olarak, 1967 cildi içindeki ölçeklerin %95'inin, 1977 cildi içindeki ölçeklerin %85'inin ve 1987 cildi içindeki ölçeklerin %60'ının psikometrik özelliklerinin [veriler güvenirlik kestirimlerine yöneliktir] raporlara eklenmediğini [ya da aktarılmadığını] ifade etmişlerdir (Meier ve Davis, 1990; Thompson,1994a; Vacha-Haase,1998). Vacha-Haase, Ness, Nilsson ve Reetz (1999) yaptıkları bir araştırmada, *Journal of Counseling Psychology (JCP)*, *Psychology & Aging (P&A)* ve *Professional Psychology: Research and Practice (PP)* adlı üç derginin, 1990'dan, 1997'ye yayımlanmış makalelerini incelenmişler ve toplam 839 makalenin %36.4'ünde makale yazarlarınca güvenirliğin ifade edilmediğini, bu makalelerde ölçüm güvenirliğinden ziyade, hatalı biçimde test güvenirliği üzerinde odaklanıldığını belirtmişlerdir. Bu çalışmalar ve benzerlerindeki zayıf güvenirlik rapor etme uygulamaları ve bunlarla ilgili ortaya konulmuş olan bulgular (Meier ve Davis 1990; Thompson, 1994b; Vacha-Haase,1998; Vacha-Haase, Kogan ve Thompson,2000; Whittington, 1998), bir başka söyleyişle ölçümlerin psikometrik özelliklerini rapor etmedeki bazı problemler, güvenirliğin ölçümlerin bir özelliği olduğunun karşıtı düşüncede olan ve güvenirliğin testin bir özelliği olduğuna inanan araştırmacılara değin izler olabilir (Whittington, 1998). Bu bulgular ise, güvenirliğin kesin görünüşünün iyi anlaşılamamış olabileceğini ortaya koymaktadır (Shields ve Caruso, 2004).

Güvenirliğin özünün yeterince anlaşılamamış olabileceğinin, bir başka ifadeyle, güvenirliği ölçümlerin değil de, hatalı biçimde testlerin bir özelliği gibi kabul etmenin uzantılarının, güvenirlik katsayılarının hatalı biçimde yorumlanması şeklinde de ortaya çıktığı söylenebilir. Onwuegbuzie ve Daniel (2000), bazı araştırmacıların .70 olarak verilen [X testinden elde edilen] bir ölçümlerin güvenirlik katsayısını, aracın kendisi %70 güvenilir şeklinde ve doğru olmayan biçimde yorumladıklarını ifade etmişlerdir. Güvenirlik katsayısıyla ilgili buna benzer hatalı yorumlar yurt içindeki çalışmalarda da görülmektedir; örneğin, Büyüköztürk (2004:164 ve 2005:170) güvenirlik katsayısı ile ilgili yaptığı yorumunda, "Güvenirlik katsayısı .80 olan bir test için bireyler arası gözlenen test puanlarındaki farkların %80 oranında gerçek farkları, %20 oranında ise hatayı yansıttığı söylenebilir" şeklinde bir ifade kullanmıştır ve Büyüköztürk'ün (2004:164 ve 2005:170) güvenirlik katsayısı ile ilgili yaptığı bu yorum yanlıştır. Çok daha doğru yorumlar Crocker ve Algina (1986), Onwuegbuzie ve Daniel (2000) ve Bademci'de (2005a) vardır. Klasik test kuramında, güvenirlik katsayısı [$\rho_{XX'}$], matematiksel olarak, gerçek ölçüm varyansının gözlenmiş ölçüm varyansına

oranı [$\rho_{XX'}=\sigma^2_T/\sigma^2x$] biçiminde tanımlanmıştır (Allen ve Yen, 1979; Crocker ve Algina, 1986; Helmstadter, 1964; McDonald, 1999; Nunnally ve Bernstein,1994; Thompson ve Vacha-Haase, 2000; Bademci, 2005a). Örneğin, X testi ölçümlerinin test-tekrar test güvenirliği .90 olsun. Bu aşamadan sonra, güvenirlik katsayısı ile ilgili yapılan yorumlar, uygun ve doğru olmalıdır. Birinci yorum, sınavı alan bu grup için, gözlenmiş ölçüm varyansının %90'ı, gerçek ölçüm varyansına atfedilebilir [ya da dayandırılabilir] veya toplam ölçüm varyansının en azından %90'ı, gerçek ölçüm varyansı nedeniyledir şeklinde yapılabilir. İkincisi, ikinci test üzerindeki gözlenmiş ölçüm varyansının ($.90^2$) veya [$\rho_{xx'}{}^2$ =] % 81'i, birinci test üzerindeki gözlenmiş ölçümlerin varyansından yordanmış olabilir biçiminde söylenebilir. Sonuncu [ve üçüncü] yorum ise, sınavı alanlara dair, gerçek ölçümler ve gözlenmiş ölçümler arasındaki korelasyon [ρ_{XT}], $\sqrt{.90}$ [veya .9486] ya da .95'dir şeklinde yapılabilir (Bademci,2005a; Crocker ve Algina,1986; Onwuegbuzie ve Daniel,2000; Thompson ve Vacha-Haase,2000). [Sonuncu yorum için gerekli bir not: Güvenirlik katsayısı, gözlenmiş ölçümler ve gerçek ölçümler arasındaki korelasyonun karesidir, $\rho_{XX'} = \sigma^2_T / \sigma^2x = \rho^2_{XT}$, (Lord ve Novick, 1968: 61; McDonald, 1999: 66)].

Güvenirliğin iyi anlaşılamadığının bir diğer örneği ise, kendi örnekleminden elde ettiği veriler için, bir başkasının ölçümlerinden hesaplanmış olan bir güvenirlik katsayısını yorumlama ve de rapor etme şeklinde, dikkatsizce yapılmış ve son derece hatalı olan bir uygulamadır. *Ölçüm güvenirliğini, örneklemdeki deneklerin ya da kişilerin kendileri etkilemektedir* (Arnold, 1996; Thompson, 1994a). Testi [veya ölçme aracını] uygulamaksızın testin güvenirliğinden bahsetmek veya bir başkasının hesapladığı bir teste [ya da ölçme aracına] ait ölçümlerin güvenirlik katsayısını, hatalı bir kabul ile, testin [ya da ölçme aracının] kendi özelliği gibi kabul edip ve yine yanlış bir biçimde ve bilinçsizce bir kabul ile, ölçümlerin ve onlardan elde edilen güvenirlik katsayılarının değişmeyeceğini kabul ederek (Reinhardt, 1996; Vacha-Haase, Ness, Nilsson ve Reetz, 1999) *uygulama yapmaksızın* bir test hakkında 'test güvenilirdir' ya da 'güvenilir testler' şeklinde ifadeler kullanmak, Thompson'un (1994a: 839) ifadesiyle bir "oxymoron" [oxymoron; yan yana kullanılması imkansız ve kullanıldığında da saçma olan iki kelime, (Turgut; 2002)] olmaktadır. Zira güvenirlik, yalnızca testin kendisinin bir fonksiyonu değildir (Reinhardt, 1996), örneklemin [özelliklerinin] de bir fonksiyonudur (Dawis, 1987; Henson,2000b); bir başka ifadeyle güvenirlik, en azından test ve testi alanların her ikisinin de bir fonksiyonudur (Arnold, 1996). O halde bir başkasının kendi örneklemine uyguladığı bir ölçme aracından elde edilen ölçümlere ait bir güvenirlik katsayısını, bir diğerinin [ya da aynı kişinin], o

güvenirlik katsayısını, aynen kendi yeni çalışmasında kullanması, yani daha önce bir testten elde edilen ölçümlere yönelik olarak hesaplanmış güvenirlik katsayısını, bir başkasının o güvenirlik katsayısını o testin bir özelliği gibi kabul ederek, [hiç hesap yapmaksızın] aynen kendi araştırmasında kullanması, çok ciddi bir ölçme ve yöntembilim hatasıdır. Zira iyice bilinmelidir ki, güvenirlik, bir testin değil, sınava giren belirli bir gruba uygulanmış o testten elde edilmiş ölçümlerin bir özelliğidir (Bademci, 2004; Bademci, 2005a) ve ölçüm güvenirliği de, örneklemden örnekleme değişmektedir (Capraro ve Capraro, 2002).

Yurt dışındaki yayımlanmış çalışmalarda olduğu kadar (Meier ve Davis, 1990; Thompson, 1994a; Thompson ve Snyder,1998; Whittington, 1998), yurt içindeki yayınlanmış çalışmalarda da (bu konudaki geniş bilgi için; Bademci, 2005a; Bademci, 2005d), bir başkasının kendi örneklemine uyguladığı bir ölçme aracından elde edilen ölçümlere ait bir güvenirlik katsayısını, bir diğerinin [ya da aynı kişinin] aynen kendi örnekleminde [hatalı biçimde] kullandığı, yani daha önce bir testten elde edilen ölçümlere yönelik olarak hesaplanmış güvenirlik katsayısını, bir başkasının o güvenirlik katsayısını o testin [ya da ölçme aracının] bir özelliği gibi kabul ederek, aynen yeni araştırmasında [doğru olmayan biçimde] kullandığı görülmektedir. Daha önceden hesaplanmış bir ölçüm güvenirlik katsayısını, hatalı olarak, [hiç hesap yapmaksızın] kendi araştırmalarında [aynen] kullanmayla ilgili tipik bir örnek, Gürsoy, Aral, Bütün Ayhan ve Aydoğan'ın (2004) yayımlanmış makalesinde görülebilir. Gürsoy, Aral, Bütün Ayhan ve Aydoğan (2004: 64), araştırmalarında, Flanders, Anderson ve Amidon tarafından geliştirilmiş ve Türk çocuklarına adaptasyonu ile 'geçerlilik' ve 'güvenilirlik' çalışması Uluğtekin (1976) tarafından yapılmış olan "Bağımlılık Eğilimi Ölçeği"ni kullanmış olduklarını ifade etmişler ve "Bağımlılık Eğilimi Ölçeği"nin güvenirlik katsayısı'nın r=.68 olduğunu da çalışmalarında rapor etmişlerdir. Gürsoy, Aral, Bütün Ayhan ve Aydoğan (2004:64) tarafından atıfta bulunulan Uluğtekin'in (1976: 124) doktora tezi incelendiğinde ise, "Bağımlılık eğilimi ölçeğinin yazarları tarafından Hoyt'un varyans analizi tekniğine göre hesaplanan güvenirlik katsayısı r= .68'dir" şeklinde yine ve bir başkasının çalışmasına daha atıf yapıldığı görülebilmektedir; bu atıf ise, doktora tezini hazırlamış olan Uluğtekin (1976:124) tarafından yapılmıştır ve de "Flanders, Anderson, Amidon, Ön. Ver., s.583" şeklindedir. Uluğtekin (1976) tarafından yapılmış bu atıf doğrultusunda, Flanders, Anderson ve Amidon'un (1961: 583) çalışmaları incelendiğinde ise, "... the reliability coefficient is .68..." ifadesi, ilgili sayfada [sayfa 583] görülebilmektedir. Oradan oraya atıf yapılarak aktarıldığı görülen ve Flanders, Anderson ve Amidon (1961:583) tarafından 43 yıl önce

hesaplanmış ölçüm güvenirlik katsayısını [.68], ciddi bir yöntembilim hatası yaptıkları da söylenebilerek, Gürsoy, Aral, Bütün Ayhan ve Aydoğan'ın (2004:64) aynen ve sorgulamadan kendi çalışmalarında kullandıkları ve Flanders, Anderson ve Amidon (1961:583) tarafından 43 yıl önce hesaplanmış [Bağımlılık Eğilimi Ölçeği'ne ilişkin] ölçüm güvenirlik katsayısıyla, hatalı biçimde, kendi çalışmalarında elde edilen ölçümlerle ilgili çeşitli [istatistiklerle] yorumlara gittikleri de ifade edilebilir.

SONUÇ

Nunnally (1982), bilimsel topluluk içinde ölçmenin oynadığı [kritik] rolü [mükemmel biçimde] teşhis etmiştir (Vacha-Haase, Ness, Nilsson ve Reetz, 1999). Nunnally'ye (1982) göre, bilim tekrar edilir [edilebilir] deneylerle ilgilenmektedir. Klasik ölçme kuramı, esasen bir '*büyük örneklem*' kuramıdır (Nunnally ve Bernstein, 1994) ve güvenirlik, "test ölçümlerinin [istendik] tutarlılığı veya tekrarlanabilirliği" şeklinde tanımlanabilir (Crocker ve Algina, 1986:105). O halde, bilim tekrar edilir [edilebilir] deneylerle ilgileniyorsa, güvenirlik ise,"test ölçümlerinin [istendik] tutarlılığı veya tekrarlanabilirliği" şeklinde tanımlanabiliyorsa ve ölçümlerin güvenirliğinin örneklemden örnekleme değiştiği ifade ediliyor ve biliniyorsa, bir ölçme aracından elde edilmiş ölçümleri ve o ölçümlerden hesaplanmış güvenirlik katsayısını *değişmez* kabul edip "test güvenilirdir" demek veya "testin güvenirliği şudur" şeklinde ifade etmek veya daha önceden yapılmış araştırmalardaki güvenirlik katsayılarını kendi çalışmalarındaki verilerde kullanmak ve de rapor etmek, *doğru değildir*. Çünkü "bir test güvenilir veya güvenilmez değildir" (Crocker ve Algina, 1986: 144), güvenilir veya güvenilmez olan ölçümlerdir (Kieffer, 1999) ve güvenirlik, testlerin değil, eldeki veriler veya ölçümlerin bir özelliğidir (Thompson, 1994a; Thompson, 1999). Bir diğer söyleyişle, güvenirlik, testin kendisinin *değil*, elde edilmiş ölçümlerinin bir özelliğidir (Lane, White ve Henson, 2002).

Güvenirlik, aracın kendisine değil, bir bellilendirme (assessment) aracı ile elde edilmiş ölçümlere işaret eder (Linn ve Gronlund, 2000; Linn ve Miller, 2005). Böylelikle, bir ölçme aracına [testin kendisine] işaret ettiği zaman kullanılan "test güvenilirdir" veya "testin güvenirliği" ifadelerini (Guthrie, 2000) kullanmak da, doğru değildir, uygun değildir (Thompson, 1994b; Thompson ve Vacha-Haase, 2000). Bu bakış açısıyla, gözden kaçırılmaması gereken önemli bir nokta, *güvenilir ölçümler* ile *güvenilir testler* terimlerinin *eşanlamlılıktan uzak olduğu*dur (Vacha-Haase, Kogan, Tani ve

Woodall, 2001). [Türkiye'deki kimi öğretim elemanlarının, ölçme ile bağlantılı bu ve benzeri birtakım terimleri birbirlerinden ayırt edemedikleri ve de bazı *çok ciddî* bilimsel hatalar yaptıkları da görülmektedir (Bademci, 2005c).]

Yurt dışındaki çalışmalarda da yaygın kullanılan "test güvenilirdir" veya "testin güvenirliği" ya da "aracın güvenirliği" gibi hatalı şekilde kısaltarak ifade etme biçimleri, Türk eğitim ve bilim topluluğunda da 1940'lardan bu yana kullanılmaktadır ve Bademci (2001a; 2001b; 2002; 2004; 2005a; 2005b; 2005c) yaklaşık 60 yılı aşkın bir süredir süregelen bu doğru olmayan kullanım biçimine ve güvenirliğin hatalı yorumlanış ve uygulama şekillerine karşı çıkmıştır. Bademci (2001a; 2001b; 2002; 2004; 2005a; 2005b; 2005c; 2006), güvenirliği, testin ya da ölçme aracının bir özelliği gibi kabul eden [yerleşik] düşünme tarzına [paradigmaya] da karşı çıkmış, güvenirliğin, testlerin ya da kullanılan ölçme araçlarının değil, ilgili araçlar veya testlerden elde edilen ölçümlerin ya da ölçme sonuçlarının bir özelliği olduğunun altını çizmiş ve doğru ifade biçiminin de, testin [ya da testlerin] güvenirliği biçiminde değil, test ölçümlerinin güvenirliği [veya ölçüm güvenirliği] şeklinde olması gerektiğini ifade etmiş ve de bir paradigma değişikliği gerekliliğini vurgulayarak, bir *yeni* paradigmayı [adayını] da yine bilimsel kanıtlarıyla Türk eğitim ve bilim topluluğunun gündemine taşımıştır.

Güvenirlik, testin kendisinin değil, ölçümlerin bir fonksiyonudur (Capraro, Capraro ve Henson, 2001). Bir başka ifadeyle güvenirlik, en azından test ve testi alanların her ikisinin de bir fonksiyonudur (Arnold, 1996). Örneklem özellikleri ise, ölçümleri ve güvenirliği etkileyebilmektedir (Capraro, Capraro ve Henson, 2001). Klasik test kuram güvenirlik kestirimleri toplam test ölçüm varyansı tarafından (Capraro, Capraro ve Henson, 2001), toplam test ölçüm varyansı da, sınavı alan grubun ne derece bağdaşık ya da ayrışık olmasından çokça etkilenmektedir (Helms,1999). Buradan, güvenirliğin, testin kendisinin değil, örneklemin özelliklerinin bir fonksiyonu olduğu da söylenebilir (Capraro, Capraro ve Henson, 2001). Ölçüm güvenirliği ise, örneklemden örnekleme değişir (Buhi, 2005; Capraro ve Capraro, 2002). Örneğin, aynı ölçek [test veya ölçme aracı], 100 farklı örnekleme uygulansa, 100 farklı güvenirlik katsayısı ortaya çıkabilir (Buhi, 2005). Hal böyle iken, güvenirliği testin bir özelliği gibi kabul etmek, aynı hatalı düşüncenin uzantısı olarak o test ya da ölçme aracından elde edilen ölçümleri ve hesaplanmış güvenirlik katsayısını [katsayılarını] da değişmez gibi kabul etmek, dolayısıyla aynı ölçme aracının kullanıldığı önceki çalışmalardaki hesaplanmış ve rapor edilmiş güvenirlik katsayılarını, hesaplama yapmaksızın kendi çalışmalarında aynen kullanmak, bir ölçme

aracına işaret eden "testin güvenirliği" veya "test güvenilirdir" ya da "ölçme aracının güvenirliği" [ve benzerî] ifadelerini kullanmak, *doğru değildir*.

KAYNAKÇA

Aiken, L. S. ve Arkadaşları (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, Vol. 45, 721–734.

Akdeniz, C., Aydemir, Ö., Akdeniz, F., Gülseren, Ş. ve Kültür, S. (1999). Sağlık düzeyi ölçeği'nin Türkçe'ye uyarlanması ve güvenilirliği. *Klinik Psikofarmakoloji Bülteni*, Cilt 9 (2), 104–108.

Allen, M. J. ve Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, California: Brooks/Cole

Anıl, D. ve Arkadaşları. (2003). *Öğretmen adayları için tamamı konu anlatımlı KPSS*. Ankara: Çağdaş Öğretmen Yayınları

Arnold, M. E. (1996). *Influences on and limitations of classical test theory reliability estimates*. (ERIC Document Reproduction Service No. ED 395 950).

Aygin, D. ve Eti Aslan, F. (2005). Kadın cinsel işlev ölçeği'nin Türkçeye uyarlanması. *Türkiye Klinikleri Tıp Bilimleri Dergisi*, Cilt 25 (3), 393–399.

Bademci, V. (2006). *Paradigma değişikliği*: *Testler güvenilir değildir*. Düzenleyen: G.Ü. 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 ilgili bazı büyük hataları düzeltmek ve bir reformu başlatmak*: *Güvenirlik, testlerin bir özelliği değildir*. Eğitim Fakültelerinde Yeniden Yapılandırmanın Sonuçları ve Öğretmen Yetiştirme Sempozyumunda Sunulan Bildiri. Ankara: Gazi Üniversitesi, Gazi Eğitim Fakültesi, 22–24 Eylül.

Bademci, V. (2005b). Testler güvenilir değildir: Ölçüm güvenirliğine yeterli dikkat ve güvenirlik çalışmaları için örneklem büyüklüğü. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, Sayı 17, 33–45.

Bademci, V. (2005c). Hakemlerin değerlendirmelerindeki hatalar üzerine: Fisher'in Z dönüşümü ve güvenirlik çalışmaları için örneklem büyüklüğü. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, Sayı 17, 46–75.

Bademci, V. (2005d). Araştırmalarda ölçme ile ilgili bazı büyük hataları düzeltmek ve eğitimde yeniden yapılanmayı sürdürmek: Güvenirlik, testlerin bir özelliği değildir. Yayına Hazırlanmış Makale. (28 sayfa).

Bademci, V. (2004). Testin güvenirliği veya test güvenilirdir diye ifade etmek doğru değildir. Türk Eğitim Bilimleri Dergisi, Cilt 2 (3), 367–372.

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

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

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

Bademci, V. (1999). *Hedefin davranışlara çevrilmesi, Davranışlardan seçmeli test maddeleri yazılması*. (Geliştirilmiş Üçüncü Baskı). Ankara: Gazi Kitabevi.

Baugh, F. (2002).Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, Vol. 62, 254-263.

Baykul, Y. (2001). *İlköğretimde ölçme ve değerlendirme*. Ankara: T.C. MEB Projeler Koordinasyon Merkezi Başkanlığı.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM.

Buhi, E. R. (2005). Reliability reporting practices in rape myth research. *Journal of School Health*, Vol. 75 (2), 63-66.

Büyüköztürk, Ş. (2004). *Sosyal bilimler için veri analizi el kitabı*. (Dördüncü Baskı). Ankara: PegemA.

Büyüköztürk, Ş. (2005). *Sosyal bilimler için veri analizi el kitabı*. (Gözden Geçirilmiş 5. Baskı). Ankara: PegemA.

Capraro, M. M., Capraro, R. M. ve Henson, R. K. (2001). Measurement error of scores on the Mathematics anxiety rating scale across studies. *Educational and Psychological Measurement*, Vol. 61, 373-386.

Capraro, R. M. ve 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.

Caruso, J. C. (2000). Reliability generalization of the neo personality scales. *Educational and Psychological Measurement*, Vol.60, 236-254.

Cousin, S. L. ve Henson, R. K. (2000). *What is reliability generalization, and why is it important*? (ERIC Document Reproduction Service No. ED 445 077).

Crocker, L. ve Algina, J. (1986). *Introduction to classical and modern test theory*. fort worth: Holt, Rinehart and Winston.

Cronbach, L. J.(1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, Vol. 16, 297-334.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology,* Vol. 34, 481-489.

Dawson, T. E. (1997). *Basic concepts in classical test theory: Relating variance partitioning in substantive analyses to the same process in measurement analyses*. (ERIC Document Reproduction Service No. ED 406 443).

Ebel, R. L. (1972). *Essential of educational measurement*. (Second Edition). Englewood Cliffs, New Jersey: Prentice- Hall, Inc.

Ebel, R. L. ve Frisbie, D. A. (1991). *Essentials of educational measurement.* (Fifth Edition). Englewood Cliffs, New Jersey: Prentice Hall.

Erkuş, A. (2003). *Psikometri üzerine yazılar*. (1. Basım). Ankara: Türk Psikologlar Derneği Yayınları, 24.

Flanders, N. A., Anderson, J. P. ve Amidon, E. J. (1961). Measuring dependence proneness in the classroom. *Educational and Psychological Measurement*, Vol. 21, 575-587.

Gelbal, S. (2002). Ölçme ve değerlendirme. *Öğretmen adayları için konu anlatımlı KMS*. (Beşinci Baskı). Ö. Demirel ve diğerleri. Ankara: Pegem A.

Gelbal, S. (1999). Öğrenci başarısının ölçülmesinde ölçme ve değerlendirme teknikleri. *Cumhuriyet Döneminde Eğitim II*. Ankara: T.C. MEB Talim ve Terbiye Kurulu Başkanlığı.

Gronlund, N. E. ve Linn, R. L. (1990). *Measurement and evaluation in teaching*. (Sixth Edition). New York: Macmillan.

Gronlund, N. E. (1965). *Measurement and evaluation in teaching*. New York: Macmillan.

Guilford, J. P. ve Fruchter, B. (1973). *Fundamental statistics in psychology and education*. (Fifth Edition). New York: McGraw-Hill.

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).

Gürsoy, F., Aral, N., Bütün Ayhan, A. ve Aydoğan, Y. (2004). Annesi çalışan ve çalışmayan çocukların bağımlılık eğilimlerinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, Sayı 26, 62–71.

Helms, L. S. (1999). *Basic concepts in classical test theory: Tests aren't reliable, the nature of alpha, and reliability generalization as meta-analytic method*. (ERIC Document Reproduction Service No. ED 427 083).

Helmstadter, G. C. (1964). *Principles of psychological measurement*. New York: Appleton-Century-Crofts.

Henson, R. K. (2000a). *A Primer on Coefficient Alpha*. (ERIC Document Reproduction Service No. ED 447 210).

Henson, R. K. (2000b). *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).

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, Vol. 34, 177-189.

Henson, R. K. ve Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, Vol. 35, 113-126.

Henson, R. K., Kogan, L. R. ve Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement*, Vol.61, 404-420.

Kieffer, K. M. (1999). Why reliability theory is essential and classical test theory is often inadequate. *Advances in Social Science Methodology, Volume 5*. Ed. B. Thompson. Stamford, Connecticut: JAI.

Lane, G. G., White, A. E. ve Henson, R. K. (2002). Expanding reliability generalization methods with kr-21 estimates: An RG study of coopersmith self-esteem inventory. *Educational and Psychological Measurement*, Vol.62, 685-711.

Linn, R. L. ve Miller, M.D. (2005). *Measurement and assessment in teaching*. (Ninth Edition). Upper Saddle River, New Jersey: Pearson.

Linn, R. L. ve Gronlund, N. E. (2000). *Measurement and assessment in teaching*. (Eighth Edition). Upper Saddle River, New Jersey: Merrill.

Livingston, S. A. (1988). Reliability of test results. *Educational Research, Methodology, And Measurement: An International Handbook*. Ed. John P.Keeves. Oxford: Pergamon.

Lord, F. M. ve Novick, M. R. (1968). *Statistical theories of mental test scores*. reading, Massachusetts:Addison-Wesley.

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

Mehrens, W. A. ve Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. (Fourth Edition). Fort Worth: Harcourt Brace.

Meier, S. T. ve Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology*, Vol. 37, 113-115.

Mittag, K. C. ve Thompson, B.(2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, May, 14-20.

Nunnally, J. C. (1982). Reliability of measurement. *Encylopedia of educational research*. (Fifth Edition). Ed. H.E.Mitzel. New York: The Free Press.

Nunnally, J. C. ve Bernstein, I. H. (1994). *Psychometric theory*. (Third Edition). New York: McGraw-Hill.

Onwuegbuzie, A. J. ve Daniel, L. G. (2000). *Reliability generalization: The importance of considering sample specificity, confidence interval, and subgroup differences*. (ERIC Document Reproduction Service No. ED 448 204).

Öncü, H. (1997). Eğitimde ölçme ve değerlendirme. *Eğitim bilime giriş*. Ed. L. Küçükahmet. Ankara: Gazi Kitabevi.

Özçelik, D. A. (1989). *Test hazırlama kılavuzu*.(Genişletilmiş İkinci Baskı). Ankara: ÖSYM Eğitim Yayınları,5.

Özdamar, K. (2004). *Paket programlar ile istatiksel veri analizi-1*. (Genişletilmiş 5. Baskı). Eskişehir : Kaan.

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

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini monte carlo study. *Advances in Social Science Methodology, Volume 4*. Ed. B. Thompson. Greenwich, Connecticut: JAI.

Rowley, G. R. (1976). The reliability of observational measures. *American Educational Research Journal*, Vol.13, 51-59.

Shields, A. L. ve Caruso, J. C. (2004). A reliability induction and reliability generalization study of the cage questionnaire. *Educational and Psychological Measurement*, Vol. 64, 254-270.

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

Tan, Ş. ve Erdoğan, A. (2004). *Öğretimi planlama ve değerlendirme*. (Genişletilmiş 5.Baskı). Ankara: PegemA.

Tekin, H. (1982). *Eğitimde ölçme ve değerlendirme*. (Üçüncü Baskı). Ankara: Daily News Web Ofset Tesisleri.

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. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other Faux Pas. *Advances in Social Science Methodology, Volume 5*. Ed. B. Thompson. Stamford, Connecticut: JAI.

Thompson, B. (1994a). Guidelines for authors. *Educational and Psychological Measurement*, Vol.54, 834-47.

Thompson, B. (1994b). *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. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Measurement*, Vol. 70,434-438.

Thompson, B. (1991). Review of generalizability theory: A Primer by Richard J. Shavelson and Noreen M. Webb. *Educational and Psychological Measurement*, 51, 1069-1075.

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

Thompson, B. ve Snyder, P. A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling & Development research articles. *Journal of Counseling and Development*, Vol. 76, 436-441.

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

Töreci, K. (2005). Yayın etiği. *http://www.endokrin.com* (en son 25.10.2005'de ulaşılmıştır).

Turgut, M. F. (1993). *Eğitimde ölçme ve değerlendirme metotları*. (Dokuzuncu Baskı). Ankara: Saydam Matbaacılık.

Turgut, S. (2002). Akıl defterimden notlar. *http://www.hürriyetim.com.tr* (en son 25.10.2005'de ulaşılmıştır.)

Uluğtekin, S. (1976). *Çocuk yetiştirme açısından anababa çocuk ilişkileri. Anababa davranışlarıyla çocuğun saldırganlık ve bağımlılık eğilimi arasındaki ilişkilerin araştırılması*. Yayımlanmamış Doktora Tezi. Ankara: A.Ü.Eğitim Fakültesi.

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. ve Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of mmpı clinical scales scores. *Educational and Psychological Measurements*, Vol. 61, 45–59.

Vacha-Haase, T., Kogan, L. R. ve Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, Vol. 60, 509-522.

Vacha-Haase, T., Ness, C., Nilsson, J. ve Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education*, Vol. 67 (4), 335-341.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, Vol. 58, 21-37.

Wilkinson, L. ve APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, Vol. 54, 594-604.

Worthen, B. R., White, K. R., Fan, X ve Sudweeks, R. R. (1999). *Measurement and assessment in schools*. (Second Edition). New York: Addison Wesley Longman.

Alınış Tarihi: Temmuz 2006
Hakemlerden Dönüş: Ekim 2006