2005 Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi Sayı:17, s.33-45

TESTLER GÜVENİLİR DEĞİLDİR : ÖLÇÜM GÜVENİRLİĞİNE YETERLİ DİKKAT VE GÜVENİRLİK ÇALIŞMALARI İÇİN ÖRNEKLEM BÜYÜKLÜĞÜ

Vahit BADEMCİ [1]

ÖZET

Testin güvenirliği veya test güvenilirdir ya da aracın güvenirliği diye ifade etmek doğru değildir. Güvenirlik, aracın kendisine değil, bir bellilendirme aracıyla elde edilmiş ölçümlere (veya sonuçlara) işaret eder. Böylece, testler değil, ölçümler güvenilirdir. Basit şekliyle, testler güvenilir değildir. Ölçme güvenirliğinden veya test ölçümlerinin güvenirliğinden bahsetmek, çok daha uygundur. Güvenirlik, yalnızca kullanılmış olan bellilendirme aracı tarafından değil, örneklem tarafından da etkilenmektedir. Güvenirlik, örneklemin de bir fonksiyonudur. Güvenirlik, örneklemin diğer özellikleri kadar, örneklem büyüklüğünden de etkilenmektedir. Örneklem büyüklüğü, evren güvenirliğinin daha duyar kestirimi için önemlidir. Güvenirlik çalışmaları için örneklem büyüklüğü kestirimindeki bir yaklaşım, sabit bir sayı kullanmadır. Bu makalede, ölçüm güvenirliğine yeterli dikkat edilmesi ve güvenirlik çalışmaları için gerekli olan örneklem büyüklüğünün en az 400 denek olması incelenmiş ve tavsiye edilmiştir.

Anahtar Sözcükler: Örneklem büyüklüğü, güven aralığı, güvenirlik, ölçüm güvenirliği

ABSTRACT

It is incorrect to speak of “the reliability of the test” or “ the test is reliable” or “ the reliability of the instrument”. Reliability refers to the scores (or results) obtained with an assessment instrument and not to the instrument itself. Thus, scores, not tests, are reliable. Simply, tests are not reliable. It is more appropriate to speak of the reliability of “test scores” or the “measurement”. Reliability is not only influenced by an assessment instrument used, it is influenced by the sample as well as of the instrument. Reliability is also a function of the sample. Reliability is influenced by the sample size as well as other features of the sample. The sample size is important for more precise the estimate of the population reliability. Still an approach to estimating the sample size for reliability studies is to use a fixed number. In this article, to pay sufficient attention to score reliability and, a minimum of 400 subjects of sample size needed for reliability studies is investigated and, recommended.

Key Words: Sample size, confidence interval, reliability, score reliability

---

[1]Gazi Üniversitesi, Endüstriyel Sanatlar Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Beşevler-ANKARA,bademci@gazi.edu.tr

GİRİŞ

Gronlund ve Linn (1990), güvenirliğin, aracın kendisine değil bir değerlendirme aracı ile elde edilmiş sonuçlara işaret ettiğine dikkat çekmiş ve aracın veya testin yerine, ölçmenin veya test ölçümlerinin güvenirliğinden bahsetmenin çok daha uygun olduğunu belirtmiştir. Henson ve Thompson (2002), Gronlund ve Linn'in yukarıda 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.

Güvenirlik, testin kendisinin değil elde edilmiş ölçümlerinin bir özelliğidir (Lane, White ve Henson, 2002). Kısaca, "…bir test güvenilir veya güvenilmez değildir" (Crocker ve Algina, 1986:144; Wilkinson ve APA Task Force on Statistical Inference, 1999:596). Güvenilir veya güvenilmez olan testler değil, bir test veya ölçme aracından elde edilmiş ölçme sonuçları veya ölçümlerdir.

Güvenirlik, bir bellilendirme aracı ile elde edilmiş ölçümlere ya da bellilendirme sonuçlarına işaret eder.

"Ölçme aracının/araçlarının güvenirliği" ifadesi de doğru değildir, zira Rowley'nin de (1976:53) ifade ettiği gibi "…bir aracın kendisi ne güvenilirdir ne de güvenilir değildir." Bir başka söyleyişle, bir aracın kendisi ne güvenilir, ne de güvenilmezdir. Çünkü, "teknik bakımdan, güvenirlik, aracın kendisine değil, elde edilmiş *sonuçlar*ının tutarlılığına işaret eder" (Worthen, White, Fan ve Sudweeks, 1999:95). Bir başka ifadeyle güvenirlik, aracın kendisine değil, bir bellilendirme (assessment) aracı ile elde edilmiş ölçümlere (Bademci, 2004; Linn ve Gronlund, 2000) veya bellilendirme sonuçlarına (Nitko, 2001) işaret eder. Vurgulamak gerekirse, *güvenirlik, aracın kendisinin değil elde edilmiş ölçümlerinin bir özelliğidir* ve bellilendirme sonuçlarının güvenirliği (Linn ve Gronlund, 2000) ya da ölçme aracından elde edilmiş ölçümlerin güvenirliği şeklinde ifade etmek, çok daha uygundur.

GÜVENİRLİK, TESTİN/ÖLÇME ARACININ KENDİSİNİN DEĞİL ELDE EDİLMİŞ ÖLÇÜMLERİNİN BİR ÖZELLİĞİDİR

Her ne kadar test güvenirliğinin işevuruk bir tanımı başlığı altında da olsa, güvenirliğin testin bir özelliği olmadığını yorumlayan Ebel (1972), bu tartışma konusunun ilkleri arasında sayılabilir (Bademci, 2004). Güvenirliğin eldeki veriler veya ölçümlerin bir özelliği olduğunu vurgulayan Thompson'un (1994) ise, bu konuda öncü olduğu söylenebilir. Thompson'un (1991), güvenirliğin verilerin bir özelliği olduğunu vurguladığı çalışmasından sonra, 1994 yılında *Educational and Psychological Measurement* Dergisinde "Guidelines For Authors" (1994:837-847) başlığı altında, test güvenirliği söyleminin doğru olmadığına yönelik açıklamalarını da içeren bir yazısı yayımlanmıştır; güvenirliğin testlerin değil ölçümlerin bir özelliği olduğu vurgusu ve ölçüm güvenirliği söylemi, halihazırda alanının en önde gelen dergilerinden biri olan *Educational and Psychological Measurement* Dergisinin editöryel politikalarından biri durumundadır (Thompson, 1994; Vacha-Haase, 1998; Thompson ve Vacha-Haase, 2000). Benzer düşünce, Wilkinson ve APA -American Psychological Association- Task Force on Statistical Inference'ın kelimelerinde (Wilkinson ve APA Task Force on Statistical Inference, l999:596) "…bir test güvenilir veya güvenilmez değildir. Güvenirlik, sınavı alanların belirli bir evreni için bir test üzerindeki ölçümlerin bir özelliğidir..." şeklinde

yerini almıştır. Bir diğer söyleyişle, bir test veya bir ölçme aracı güvenilir veya güvenilmez değildir, zira güvenirlik, sınavı alanların belirli bir evrenine uygulanmış bir test ya da bir ölçme aracından elde edilmiş ölçümlerin bir özelliğidir.

Ölçüm güvenirliğine yeterli dikkat gösterilmelidir

Güvenirliğin, "testlerin değil ölçümlerin bir özelliği" (Caruso, 2000) olduğu şeklinde vurgulanmakta olan bu nokta, Türkiye'de de ilk defa Bademci'nin (2004) çalışmasıyla eğitim ve bilim gündemine taşınmıştır. Türkiye'de de 1950'lerden [örneğin, 'testin güvenliği –güvenirliği-' ya da 'güvenilir bir testtir' (Weitzman ve McNamara,1953:145)] bu yana kullanılan, "testin güvenirliği" veya "test güvenilirdir" ya da "aracın güvenirliği" ve benzeri ifade biçimlerinin doğru olmadığını belirten Bademci (2004:367-372), *güvenirliğin, testin/ölçme aracının kendisinin değil elde edilmiş ölçümlerinin bir özelliği olduğunu* ve ifade edilmesi gerekenin de *ölçüm güvenirliği* olduğunu vurgulamış, doğru ifadeler için de örnekler vermiştir.

Güvenirlik, ölçümlerin bir özelliğidir

Güvenirliğin ölçümlerin bir özelliği olduğuna dair dikkat çekici bir açıklama, Crocker ve Algina'da (1986) bulunmaktadır. 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. Testi alan öğrenci grubunun bağdaşık (homojen) ya da ayrışık (heterojen) olması kadar, grubun büyüklüğü veya küçüklüğü de ölçümlerin güvenirlik kestirimini ve duyarlığını olumlu veya olumsuz etkileyebilmektedir.

ÖLÇÜM GÜVENİRLİĞİ ÖRNEKLEMDEN ÖRNEKLEME DEĞİŞİR

Dawis'in (1987:486) ifadesine göre "…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"; buradaki "bazen gözden kaçırılmış" ifadesi önemli ve dikkat çekicidir. Daha ayrışık örneklemler sıklıkla daha çok değişken ölçümlere ve bu durumda daha yüksek güvenirliğe yol açar (Thompson, 1994). Bu durumda aynı ölçme aracı [veya test], daha bağdaşık ya da daha ayrışık öğrencilerden oluşan gruplara ya da örneklemlere uygulandığında, birbirlerini onamayan ölçümler güvenirliği ortaya çıkacaktır (Thompson, 1994). Ölçüm güvenirliğinin örneklemden örnekleme değiştiği (Capraro ve Capraro, 2002) göz önünde tutulursa, örneklemin diğer özellikleri (örneğin, değişkenliği) kadar, örneklemin gözden kaçırılmış özelliklerinden biri olan *örneklem büyüklüğü*nün de, ölçüm güvenirliğinin duyarlı kestirimi konusunda göz ardı edilemeyecek faktörlerden biri olduğu söylenebilir. Ölçüm güvenirliğine yeterli dikkat gösterildiğinde, gözden kaçırılmaması gereken etmenlerden biri olarak önem kazanan güvenirlik çalışmalarındaki örneklem büyüklüğü konusuna ve bu hususun önemine de, bu çalışmayla dikkat çekilmektedir.

Güvenirlik çalışmaları için örneklem büyüklüğü kestirimindeki bir yaklaşım, sabit bir sayı kullanmadır.

Güvenirlik çalışmaları için örneklem büyüklüğü (n) kestirimindeki bir yaklaşım, sabit bir sayı kullanmadır (Streiner ve Norman, 1995). Ancak, test elkitapları ve güvenirlik çalışmaları incelendiği zaman, güvenirlik katsayısı'nın (r) duyarlı (ya da isabetli) bir kestirimi için gerekli olan örneklem büyüklüğü üzerine yazına (literature) sunulmuş çalışmaların sayısının çok az olduğu görülecektir (Charter, 1999).

Güvenirlik çalışmalarında örneklem büyüklüğünün kestirimi için sabit bir sayı kullanma yaklaşımı konusunda yayımlanmış çalışma sayısı *çok az* olduğu kadar, önerilen sayı konusunda da görüşler değişiktir. Faktör analiziyle bağıntılı çalışmasında Guilford (1954), örneklem büyüklüğünün en az 200 denek (kişi, sınavı alan) olması gerektiğini ileri sürmüştür. Tüm güvenirlik çalışmalarında örneklem büyüklüğünün 200 veya daha fazla olması gerektiğini ileri süren bir başka görüş ise, Kline'a (1986) aittir. Nunnally, önceki çalışmalarında 300'den küçük kişi (durum–örnek olay) sayısını "göreceli küçük sayı" (1967:218 ve 1978:237) olarak belirtirken, daha sonraki çalışmasında güvenirlik çalışmalarındaki denek sayısının "en az 300" olması gerektiğini ifade etmiştir (1982:1600). Segall (1994:361), doğrusal eşitlenmiş testlerin güvenirliği üzerine yaptığı çalışmasında 300 sayısını "küçük" olarak nitelendirirken, Nunnally ve Bernstein (1994) örneklem büyüklüğünün 300 veya daha fazla kişiden oluşmasını önermiştir.

ÖLÇÜM GÜVENİRLİĞİNİN KESTİRİLDİĞİ HALLERDE BÜYÜK ÖRNEKLEMLER KULLANILMALIDIR

Charter (1999) ve Charter ve Feldt (2002), Nunnally ve Bernstein'ın örneklem büyüklüğünün 300 veya daha fazla kişiden oluşması önerisinin görgül (empirical) kanıt üzerine temellenmediğini ifade etmişlerdir. Görgül kanıt üzerine temellenmiş çalışmalarında ise, Charter (1999; 2001) örneklem büyüklüğünün en az 400 denekten oluşması gerektiğini ileri sürmüştür (Charter ve Feldt ,2002). Ölçüm güvenirliğinin kestirildiği hallerde büyük örneklemler kullanılması, güvenirliğin daha duyarlı kestirimi için önemli ve yararlıdır.

Bu çalışmada, en küçük örneklem büyüklüğü 50 kişi-denek olarak alınmış ve Guilford (1954) ve Kline'ın (1986) ifade ettiği sayı olan 200 kişi-denek, Segall'ın (1994) küçük olarak da nitelediği, Nunnally (1982), Nunnally ve Bernstein'ın (1994) dikkat çektiği sayı olan 300 kişi-denek ve de Charter'ın (1999;2001) ileri sürdüğü sayı olan örneklem büyüklüğünün 400 kişi-denekten oluşması gerektiği önerileri, güven aralığı'ndan (GA) yararlanılarak incelenecektir. Charter'ın (1999) güvenirlik katsayısının duyarlığının ölçüsü gibi kullandığı güvenirlik katsayısı için güven aralığı [genişliğinin] saptanması işleminden, bu çalışmada da faydalanılmaktadır.

Güvenirlik kestirimi için çeşitli yöntemler vardır (Worthen,White,Fan ve Sudweeks, 1999; Linn ve Gronlund, 2000) ve güvenirlik kestirim yöntemlerinin bir çoğunda Pearson çarpım moment korelasyon katsayısından faydalanılmaktadır (Mehrens ve Lehmann, 1991). Pearson moment çarpım korelasyon katsayısının kullanıldığı bir güvenirlik kestirim yöntemi ise, test-tekrar test yöntemidir (Crocker ve Algina , 1986; Suen,1990).

Bir Pearson moment çarpım korelasyon katsayısı için güven aralığı saptanması, Fisher'in Z dönüşümü (Akhun,1986; Sheskin,2004) üzerine temellenebilir. Bu çalışmada örnek olarak verilen test-tekrar test yöntemiyle kestirilen güvenirlik katsayısının duyarlığının bir ölçüsü olarak kullanılan güven aralığı saptanması işleminde de, Fisher'in Z dönüşümü kullanılmıştır.

Örneklem büyüklüğü 50 olan bir örneklem üzerine temellenmiş ve Pearson çarpım moment korelasyon formülü kullanılarak hesaplanmış bir test-tekrar test güvenirliği, r = .90 olsun. Bu güvenirlik katsayısı için %95 güven aralıkları saptansın. Daha sonra da örneklem büyüklüğü 200, 300 ve 400 olarak alınıp, aynı güvenirlik katsayısı, r = .90 için, yine %95 güven aralıkları bulunsun. - Korelasyon katsayısı için güven aralığı saptanması, değişik bazı kaynaklarda da (Cox, 1987; Spiegel, 1988; Smithson, 2000) bulunmaktadır.-

Bir korelasyon katsayısı için güven aralığı saptanması, bazı çalışmalarda dört veya beş basamak halinde gösterilmiştir. (Popham ve Sirotnik, 1992; Glass ve Hopkins, 1996). Bu çalışmada da, bir Pearson çarpım moment korelasyon katsayısı için güven aralığı saptanması, aşağıda olduğu gibi "dört basamak halinde" (Glass ve Hopkins, 1996:357-358) ifade edilmiştir:

1. r 'yi Zr 'ye dönüştür (Fisher'in Z dönüşümü)
2. σz 'yi hesapla : $\sigma z = 1 / \sqrt{n-3}$ (n: denek veya sınavı alanların sayısı)
3. Zρ için GA elde et : Zr ± 1.96 σz (GA %95 için)
4. Alt ve üst sınırları tekrar r 'ye dönüştür.

Şimdi, bu çalışmada ele alınan 50, 200, 300 ve 400 kişilik dört ayrı örneklem büyüklüğü dikkate alınarak, dört basamak halinde, test- tekrar test yöntemi kullanılarak hesaplanmış güvenirlik katsayısı .90 için %95 güven aralıkları aşağıda olduğu gibi saptanılabilir.

n = 50 r = 0.90 .95 GA	n = 200 r = 0.90 .95 GA
1. Zr = .5 ln [(1+ |r|) / (1- |r|)] = .5 ln (1.9 / 0.10) = 1.4722	***1***. Zr = .5 ln [(1+ |r|) / (1- |r|)] = .5 ln (1.9 / 0.10) = 1.4722
2. $\sigma z = 1 / \sqrt{n-3} = 1 / \sqrt{50 - 3}$ = 0.1458	***2***. $\sigma z = 1 / \sqrt{n-3} = 1 / \sqrt{200 - 3}$ = 0.0712
3. Zr için GA.95 : Zr ± 1.96 σz 1.4722 ± 1.96 x = (1.1864, 1.7579) :	***3***. Zr için GA.95 : Zr ± 1.96 σz 1.4722 ± 1.96 x 0.0712 = (1.3326, 1.6117) :

4. Alt ve üst sınırları r 'ye dönüştür
r için GA.95
alt Zr = 1.1864 → alt r = .8294
üst Zr = 1.7579 → üst r = .9422

r için genişlik (range) → .9422 - .8294
= 0.1128

4. Alt ve üst sınırları r 'ye dönüştür
r için GA.95
alt Zr = 1.3326 → alt r = .8698
üst Zr = 1.6117 → üst r = .9234

r için genişlik → .9234 - .8698
= 0.0536

n = 300
r = 0.90
.95 GA

1. Zr = .5 ln [(1+ |r|) / (1- |r|)]
= .5 ln (1.9 / 0.10)
= 1.4722

2. σz = 1 / √ n-3 = 1 / √ 300 - 3
= 0.0580

3. Zr için GA.95 :
Zr ± 1.96 σz
1.4722 ± 1.96 x 0.05
= (1.3585, 1.5858) :

4. Alt ve üst sınırları r 'ye dönüştür
r için GA.95
alt Zr = 1.3585 → alt r = .8760
üst Zr = 1.5858 → üst r = .9195

r için genişlik → .9195 - .8760
= 0.0435

n = 400
r = 0.90
.95 GA

1. Zr = .5 ln [(1+ |r|) / (1- |r|)]
= .5 ln (1.9 / 0.10)
= 1.4722

2. σz = 1 / √ n-3 = 1 / √ 400 - 3
= 0.0501

3. Zr için GA.95 :
Zr ± 1.96 σz
1.4722 ± 1.96 x0.0501
= (1.3740, 1.5703) :

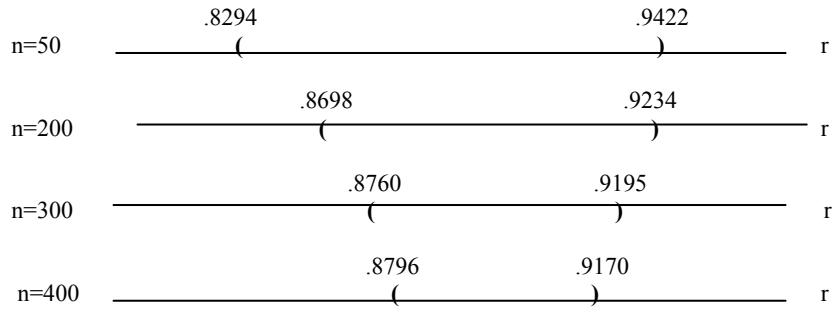
4. Alt ve üst sınırları r 'ye dönüştür
r için GA.95
alt Zr = 1.3740 → alt r = .8796
üst Zr = 1.5703 → üst r = .9170

r için genişlik →.9170 - .8796
= 0.0374

Dönüşümler için bu konuda hazırlanmış tablolardan yararlanılabilir (Shavelson,1996; Hays, 1994). Guilford ve Fruchter'da (1973) r 'den Z 'ye ve Z 'den r 'ye dönüşümler için tek tablonun kullanılabileceği ifade edilirken, Zar'da (1996) her iki dönüşüm için ayrı ayrı ve ayrıntılı tablolar bulunmaktadır. Yine bu dönüşümler için Microsoft Excel veya birer istatistik programı olan SPSS, SAS gibi programlardan yararlanılabilir. Ya da yine bu dönüşüm hesaplamaları için fonksiyonlu bir hesap makinesi kullanılabilir. Bu çalışmadaki hesaplamalar için, örnek olması amacıyla, CASIO fx-5000F marka fonksiyonlu bir hesap makinesinden faydalanılmış ve çıkan sonuçlarda yuvarlamalara gidilmemiştir.

Yukarıdaki hesaplamalarda görüleceği üzere, 50 kişilik bir örneklem üzerine temellenmiş .90 olan test-tekrar test güvenirlik katsayısının, %95 güven aralığı ile alt ve üst sınırları .8294 ve .9422 iken, 400 kişilik bir örneklem ve aynı güven aralığı ile alt ve üst sınırları, .8796 ve .9170'dir. Buradan, örneğin 50 kişilik örneklemle ortaya çıkan bu sonuca yönelik olarak, ".95 olasılıkla, evren test-tekrar test güvenirlik katsayısı .8294 ve .9422 arasındadır"

veya 400 kişilik örneklemle ortaya çıkan sonuca yönelik olarak, ".95 olasılıkla, evren test-tekrar test güvenirlik katsayısı .8796 ve .9170 arasındadır" şeklinde bir yorum yapılabilir. Benzer yorumlar, 200 ve 300 kişilik örneklemlerle ortaya çıkan sonuçlara yönelik olarak da ifade edilebilir. Değişik yorumlama ifadeleri için, Kutsal ve Muluk (1978), Akhun (1986), Glass ve Hopkins (1996) ve Russo'dan (2003) faydalanılabilir. 50, 200, 300 ve 400 kişilik örneklemlerle ilgili olarak elde edilmiş tüm bu sonuçlardan, daha büyük bir örneklem ile daha duyar bir evren güvenirlik kestirimi yapılacağı ya da evren güvenirliğinin yeterli duyarlıkta kestirimi için örneklem büyüklüğünün en az 400 kişi-denek olması önerisinin yerinde olduğu yorumları yapılabilir. Zira, 50 kişilik bir örneklem üzerine temellenmiş güvenirlik katsayısı .90 için, genişlik 0.1128 iken, 200 kişilik örneklemde bu genişlik 0.0536, 300 kişilik örneklemde 0.0435 ve 400 kişilik örneklemde 0.0374'tür; görüldüğü üzere örneklem büyüdükçe güvenirlik katsayısı .90 için, saptanmış olan güven aralığı genişliği daralmaktadır. Bu durum, Şekil 1'de de gösterilmeye çalışılmıştır.



Şekil 1. Güven aralığı . 95 ve güvenirlik katsayısı .90 için, örneklem sayısı büyüdükçe güven aralığı genişliği daralmasının şekil üzerinde gösterilmesi

Şekil 1'de de görüleceği üzere, örneklem sayısı büyüdükçe güvenirlik katsayısı .90 için güven aralığı genişliği daralmaktadır; böylelikle, daha büyük örneklem ile çok daha duyar evren test-tekrar test güvenirliği kestirilebileceği ifade edilebilir. Bu sonuçlar da, büyük örneklem üzerine temellenmiş güvenirlik kestiriminin daha duyar, daha isabetli olduğunu ya da Charter'ın (1999) görgül kanıt üzerine temellendirerek ileri sürdüğü örneklem büyüklüğünün *en az* 400 denekten oluşması gerektiği önerisinin daha doğru, daha isabetli olduğunu göstermektedir.

Güvenirlik çalışmaları için örneklem büyüklüğü, <u>en az</u> 400 kişi (denek) olmalıdır

400 kişilik örneklemdeki güvenirlik katsayısının sahip olduğu genişlik 0.0374'tür. Peki, verilen örnekte ortaya çıkan bu miktar doyurucu mudur? Elbette. Zira Stanley (1971), yüksek bir güvenirlik katsayısındaki yalnızca .01'lik bir iyileştirmenin, testin yüzde 10 veya daha fazla uzatılmasıyla elde edilmiş güvenirlikteki artışa eşdeğer olduğunu ifade etmiştir. Test geliştirmeyle uğraşanlar bu noktanın ne kadar önemli olduğunu iyi bilirler. Charter (1999), güvenirlik katsayısını .90 ve güven aralığı genişliğini 0.46 olarak verdiği bir örnekte, bu durumun test uzunluğunda %70 civarında bir azalma veya çoğalmaya eşit sayıldığına dikkat çekmiştir. Tüm bu veriler, daha büyük örneklem üzerine temellenmiş güvenirlik katsayısı kestiriminin çok daha duyarlı, isabetli olduğunu açıkça göstermektedir.

SONUÇ VE YORUM

Güvenirlik çalışmaları için örneklem büyüklüğü kestirmedeki bir yaklaşım, sabit bir sayı kullanmaktır; bu çalışmada, güvenirlik çalışmaları için 50 kişi-denek, Guilford (1954), Kline'ın (1986) en az '200' veya daha fazla kişi-denek, Nunnally (1967;1978;1982), Segal (1994) ve Nunnally ve Bernstein'ın (1994) en az '300' veya daha fazla kişi-denek ve Charter'ın (1999;2001) görgül kanıt üzerine temellendirdiği en az- '400' kişi-denek olması biçiminde önerilmiş bulunan örneklem büyüklükleri, güvenirlik katsayısının duyarlığının ölçüsü gibi kullanılan güvenirlik katsayısı için güven aralığı saptanması işleminden faydalanılarak incelenmiştir. *En az* '400' veya daha büyük kişiden-denekten oluşan örneklem üzerine temellenmiş güvenirlik kestiriminin, daha duyar, daha isabetli olduğu bu çalışmada ortaya konulmuştur.

Güvenirlik çalışmaları için büyük örneklemlerden faydalanmak en iyisidir. *En az* 400 kişiden-denekten oluşan örneklem kullanılması, *test-tekrar test, eşdeğer formlar, iki yarı, alfa katsayısı yöntemleri kullanılarak bir test ölçüm güvenirliğinin hesaplandığı* hallerde, ölçüm güvenirlik kestiriminin, daha duyar, daha isabetli olması için bir gereklilik olarak görünmektedir; bu duruma önemle dikkat edilmesi ise, bu çalışmayla tavsiye edilmektedir.

Vurgulamak gerekirse, güvenirlik, "…tasarlanmış hedef evrenden [alınmış] bir örneklem üzerinde değerlendirilmektedir…" (Dawis, 1987:486); ancak, Dawis'inde (1987:486) dikkat çektiği gibi bu nokta, "bazen gözden kaçırılmıştır." Ölçme aracı geliştirmeyle uğraşanların ya da araştırmacıların, öncelikle ölçüm güvenirliğine yeterli dikkati vermelerinin yanı sıra, güvenirlik çalışmaları için daha büyük örneklemler kullanma ve daha yüksek ölçüm güvenirliği elde etme konusunda daha duyarlı olmaları ve de büyük çaba göstermeleri gerekmektedir.

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

KAYNAKLAR

Akhun, İ.(1986).***İstatistiklerin Manidarlığı ve Örneklem***.(Geliştirilmiş İkinci Baskı). Ankara.

Bademci, V. (2004). "Testin Güvenirliği" veya "Test Güvenilirdir" Diye İfade Etmek Doğru Değildir. ***Türk Eğitim Bilimleri Dergisi***, Cilt 2, 367-373.

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, 235-254.

Charter, R. A. (1999). Sample Size Requirements for Precise Estimates of Reliability, Generalizability, and Validity Coefficients. ***Journal of Clinical and Experimental Neuropsychology***, Vol.21, 559-566.

Charter, R. A. (2001). Damn the Precision, Full Speed Ahead with the Clinical Interpretation. ***Journal of Clinical and Experimental Neuropsychology***, Vol.23, 692-694.

Charter, R.A. ve Feldt, L. S. (2002). The Importance of Reliability as It Relates True Score Confidence Intervals. ***Measurement and Evaluation in Counseling and Development***, Vol. 35, 104-112.

Cox, P. C. (1987). ***A Handbook of Introductory Statistical Methods***. New York: John Wiley and Sons.

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

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

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

Glass, G. V. and Hopkins, K. D. (1996). ***Statistical Methods in Education and Psychology***. (Third Edition). Boston: Allyn and Bacon.

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

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

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

Hays, W. L. (1994). ***Statistics***. (Fifth Edition). Fort Worth: Harcourt Brace.

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.

Kline, P. (1986). ***A Handbook of Test Construction: Introduction to Psychometric Design***. New York: Methuen.

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. and Gronlund, N. E. (2000). ***Measurement and Assessment in Teaching***.(Eight Edition). Upper Saddle River, New Jersey: Prentice-Hall.

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

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

Kutsal, A. Ve Muluk, Z. (1978). ***Uygulamalı Temel İstatistik***. (Üçüncü Baskı). Ankara: H.Ü. Fen Fakültesi Basımevi.

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

Nunnally, J. C. (1982). Reliability of Measurement. ***Encyclopedia of Educational Research***. (Fifth Edition). (Ed. H.E.Mitzel). New York: The Free Press.

Nunnally, J. C. (1978). ***Psychometric Theory***. (Second Edition). New York: McGraw-Hill.

Nunnally, J. C. (1967). ***Psychometric Theory***. New York: McGraw-Hill.

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

Popham, W. J. and Sirotnik, K. A. (1992). ***Understanding Statistics in Education***. Itasca, Illinois: F. E. Peacock.

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

Russo, R. (2003). ***Statistics for the Behavioural Sciences.An Introduction***. Hove: Psychology Press.

Segall, D.O.(1994).The Reliability of Linearly Equated Tests.***Psychometrika***,Vol.59,361-375.

Shavelson, R. J. (1996). ***Statistical Reasoning for the Behavioral Sciences***. (Third Edition). Boston: Allyn and Bacon.

Sheskin, D.J. (2004). ***Handbook of Parametric and Nonparametric Statistical Procedures***. (Third Edition). Boca Raton: Chapman and Hall/CRC.

Smithson, M. (2000). ***Statistics with Confidence***. London: Sage.

Spiegel, M. R. (1988). ***Schaum's Outline of Theory and Problems of Statistics***. (Second Edition). New York: McGraw-Hill.

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

Streiner, D. L. and Norman, G. R. (1995). ***Health Measurement Scales: A Pratical Guide to Their Development and Use***. (Second Edition). Oxford University Press.

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

Thompson, B.(1994). Guidelines for Authors. ***Educational and Psychological Measurement***, Vol. 54, 837-847.

Thompson, B. (1991). Review of Generalizability Theory: A Primer by Richard J. Shavelson and Noreen M. Webb. ***Educational and Psychological Measurement***, Vol. 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.

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

Weitzman, E. ve Mc.Namara, W. J. (1953). ***Sınıfta Test Nasıl Yapılır***. (Çeviren:V. D. Pars). İstanbul: Milli Eğitim Basımevi.

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.

Zar, J. H. (1996). ***Biostatistical Analysis***. (Third Edition). Upper Saddle River, New Jersey: Prentice-Hall.

BU MAKALENİN YAZARI VAHİT BADEMCİ'DEN BİR BİLGİ NOTU

1991 yılında sürdürdüğüm eğitimde kalite ve eğitimde kalite kontrol çemberlerinin yerleştirilmesi ve doktora tezimle ilgili çalışmalarımın devamı niteliğinde, 1995 yılında, 1995 Milliyet Örsan Öymen Birincilik Ödülü alan *Aymazlığın Sonu:Geleceği Tehlikede Bir Ulus* (Ankara: Gazi,1997) isimli çalışmamla, Türk eğitim sisteminin öğretmen yetiştiren programlarında, çıkın (portfolio), alternatif bellilendirme (assessment) gibi yeni ölçme ve bellilendirme yöntem veya tekniklerinin kullanılması gerekliliğini vurgulayarak, eğitimde ve eğitimde ölçme ve değerlendirme alanında bir *yeniden yapılanma* hareketi başlattım. Türk eğitim sisteminde, çok boyutlu zeka, eğitimde toplam kalite yönetimi, beyin temelli öğrenme gibi yeni yaklaşımların kullanılmasının yanı sıra, yeni ölçme ve bellilendirme yöntem veya yaklaşımlarının [erişim (performance) bellilendirme, çıkın, bellik (rubric)] kullanılmasının da gerekliliğini, 1999 Milliyet Sosyal Bilimler Birincilik Ödülü alan *Türkiye'deki Okullar Ne İşe Yarar?* (Ankara: Alp,İkinci Baskı,2001) isimli çalışmamda güçlü bir biçimde, bir kez daha vurguladım. Temelleri başlattığım bu yeniden yapılanma hareketi çalışmalarıma dayalı biçimde, özellikle de 1999 yılından bu yana çok daha açık ve kesin bir şekilde, güvenirliğin testlerin değil ölçümlerin bir özelliği olduğunu ve ifade edilmesi gerekenin de ölçüm güvenirliği olduğu düşüncesini Türk eğitim bilim gündemine taşıdım. Bu hususla ilgili olarak da, özellikle de 2001 ve 2002 yılı konferanslarımda ve konuşmalarımda, güvenirliğin aracın [testin] değil ölçümlerinin bir özelliği olduğu, güvenirlik çalışmaları için örneklem büyüklüğünün en az 400 olması gerekliliği, Cronbach alfa katsayısının -1 ve -1'den küçük değerler alabildiği, ölçüm güvenirlikleri için güven aralıkları gerekliliği, yazılı sorular ve güvenirlik için bellik kullanımı, geçerlik ve faktör analizi ve de istatistik ve manidarlık ile ilgili bazı yeni yaklaşımların gerekliliği, güvenirlik ve geçerliğin doğru kullanımı ve doğru yorumlar, etki genişliği/büyüklüğü gerekliliği, değer biçiciler arası (interrater) güvenirlik ve kullanım alanları, genellenirlik kuramının etkin kullanımı, uylaşma (agreement) indekslerinin kullanımı gibi, ölçmeyle ilgili ve Türk eğitim sistemi için de yeni olan pek çok konu ve yaklaşımı açık bir biçimde gündeme getirdim ve başlattığım bu yeniden yapılanma hareketindeki yeni yaklaşımların ve bu yeni hususların neredeyse tam sırasını dahi açıklayarak, 19 farklı konudan oluşacak şekilde yayınlayacağım 19 makalelik bir makaleler dizisiyle, Türk eğitim sisteminin içerisine iyice yerleşmesini sağlayacağımı da ifade ettim. Nitekim, konferanslarımda ve konuşmalarımda belirttiğim sırada olduğu şekliyle, bu makalelerimin ilki, güvenirliğin testlerin değil, ölçümlerin bir özelliği olduğunu açıkça bildiren bir biçimde, *Türk Eğitim Bilimleri Dergisi*nin 2004 yılı Yaz sayısında yayınlanmış, belirttiğim ikinci sıradaki örneklem büyüklüğü ile ilgili makalem *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*nde yayın için sıraya girmiş, üçüncü sırada belirttiğim Cronbach alfa katsayısının -1 ve -1'den küçük değerler alabileceği hususu, 2005 yılında Gazi Üniversitesi, Gazi Eğitim Fakültesi'nce düzenlenen *Eğitim Fakültelerinde Yeniden Yapılandırmanın Sonuçları ve Öğretmen Yetiştirme Sempozyumu*'nda sunduğum bir bildiriyle daha da geniş ve ayrıntılı bir biçimde, tarafımdan tekrar eğitim ve bilim gündemine getirilmiştir.

> Türk eğitim bilim dünyasındaki her yeni fikrin bir sahibi veya o fikri veya fikirleri ilk gündeme taşıyanı, bir çalışanı vardır. Yaklaşık son on beş yıldır, yeni olduğu kabul edilerek eğitim bilim alanında kullanılan, yararlanılan bu yeni fikirlerin büyük bir çoğunluğu veya büyük bir çoğunluğunun ilk gündeme taşınması bana aittir; tüm bunlar da, büyük ödüllerle onurlandırılmış kitaplarım, makalelerim, konferanslarım, konuşmalarım, çeşitli belgeler, bilgi ve kanıtlarıyla açıkça ortadadır.

İntihal (*aşırma*) bir **suçtur** ve Fikir ve Sanat Eserleri Kanunu "(FSEK) m. 71 uyarınca intihal fiilini işleyen kişilere, dört yıldan altı yıla kadar hapis ve ağır para cezası

verilecektir. FSEK md. 71'deki ağır para cezasının alt sınırı elli milyar, üst sınırı ise yüzelli milyar liradır" (www.fisaum.org.tr/metin.doc, en son 06 Şubat 2006). **Aşırma**dan kaçınmanın yollarından birisi de, "kullanılan her çeşit kaynağın belirtilmesi" gerekliliğidir (http://www.fbe.metu.edu.tr/Intihal/intihal. htm, en son 29 Ocak 2006): "Bir internet sayfası, internette bir tartışma forumu, resmi olmayan bir sunum fotokopisi, veya bir konuşmadan bile bir fikir veya bilgi alıp kullandıysanız kaynak göstermek gerekmektedir. Kaynak göstermeden başka bir yerde gördüğünüz ya da duyduğunuz bilgi veya yorumu kullanamazsınız" (http://www.fbe.metu.edu.tr/ Intihal/intihal.htm, en son 29 Ocak 2006). "Bir kaynaktan ister bir kelime ister bütün bir paragraf alın, önüne ve arkasına tırnak işareti ("....") koymalı ve bilgiyi aldığınız kaynağı hemen yanına yazmalısınız" (http://www.fbe.metu.edu.tr/Intihal/ intihal.htm, en son 29 Ocak 2006). *Çalışmalarımın, bilgilerimin ve fikirlerimin, kaynak gösterilmeden kullanılmasına karşı, bütün yasal yollarla ve daha fazla ve de yakinen takipçisi olacağımı bu Makale vesileyle açıkça ifade ederim*.

İntihal (aşırma) ile ilgili bilgi için, örneğin bkz.; http://www.metu.edu.tr/~wwwfbe/Intihal/ intihal.htm *ile* http://www.fisaum.org.tr/metin.doc