"TESTİN GÜVENİRLİĞİ" VEYA "TEST GÜVENİLİRDİR" DİYE İFADE ETMEK DOĞRU DEĞİLDİR

Vahit BADEMCİ*

Özet

"Testin güvenirliği" veya "Test Güvenilirdir" diye ifade etmek doğru değildir. Zira güvenirlik, testin değil, eldeki veriler veya ölçümlerin bir özelliğidir.

Anahtar sözcükler: *Ölçüm (score)** güvenirliği, güvenirlik, "test güvenirliği."*

Abstract

It is incorrect to speak of "the reliability of the test" or " the test is reliable". Because, reliability is a characteristic of scores or the data in hand, not of the test itself.

Key words: *Score reliability, reliability, "test reliability."*

"Test güvenilirdir." sözü veya "testin güvenirliği" ifadesi eğitim, psikoloji vb. alanların ölçmelerinde yaygın kullanılmaktadır. Ancak, bu iki anlatım biçimi de doğru değildir. Çünkü güvenirlik, testlerin değil ölçümlerin bir özelliğidir (Caruso, 2000).

Thompson (1994) çok az araştırmacının, güvenirliğin eldeki veriler veya ölçümlerin bir özelliği olduğunu bilinçli kabul edip, buna göre davrandıklarını belirtmiştir. Bu çalışmanın kökleri de, Bademci'nin (1999) eğitimle ilgili bazı terim ve tanımların birbirlerinin yerlerine kullanılmamasıyla ilgili yaptığı ve sürdürdüğü çalışmaya dayanmaktadır.

Güvenirlik, ölçümlerin bir özelliğidir

Her ne kadar test güvenirliğinin işe vuruk bir tanımı başlığı altında da olsa, güvenirliğin testin bir özelliği olmadığını yorumlayan Ebel (1972), bu tartışma konusunun öncüleri arasında sayılabilir. Ebel'in bu 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 olduğu söylenebilir.

Yazışma Adresi: * Yard. Doç. Dr. Vahit Bademci, Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Eğitim Bilimleri Bölümü, 06500 Beşevler/Ankara.

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." Bu açıklamaları destekleyen bir görüş de Sax' dan gelmiştir; 1980 yılı tarihli çalışmasında Sax, ölçmelerin (veriler, ölçümler ve gözlemler) güvenirliğine dair konuşmanın, testlerin (sorular, maddeler, diğer izlentiler) güvenirliğine dair konuşmadan çok daha doğru olduğunu ifade ederek, bir testin güvenirliği ima edildiğinde de daima bir testten elde edilmiş gözlemler (verilerin belirli bir grubu) veya ölçmelerin güvenirliği anlamında yorumlanması gerektiğini belirtmiştir (Sax'ın ifadelerinin bulunduğu yer: Thompson, 1991; Thompson 1992).

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, güvenirliği etkileyen bir faktördür

Belirli bir grupta testi alan öğrencilerin veya kişilerin kendileri, ölçümlerin güvenirliğini etkilemektedir. Hâl 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, 1994). Böylece daha ayrışık (heterojen) 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 uygulandığında, birbirlerini onamayan ölçümler güvenirliği ortaya çıkacaktır.

Bu durumun mantığı, Klâsik Kuram (Suen, l990), Klâsik Güvenirlik Modeli (Thorndike, l982), Klâsik Test Kuramı (Pedhazur ve Schmelkin, 1991), Klâsik Test Kuram Modeli (Lord ve Novick, l968), Klâsik Gerçek Ölçüm Modeli (Crocker ve Algina, l986), Klâsik Gerçek Ölçüm Kuramı (Allen ve Yen, l979) gibi adlandırılan ölçme kuramının bazı eşitliklerinden yararlanılarak ve bu kuramın derinliğine girilmeden açıklanabilir.

Güvenirlik kat sayısı matematiksel olarak aşağıdaki şekilde tanımlanabilir (Allen ve Yen, l979):

rxx¢ = s2T / s2x
rxx¢ = güvenirlik kat sayısı (X ve X¢ paralel ölçmeler)
s2T = gerçek ölçüm varyansı
s2x = gözlenmiş ölçüm varyansı

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

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

rxx¢ = s2T / s2x = s2x -s2E / s2x =1-s2E / s2x

s2E = hata ölçüm varyansı

rxx¢=1-s2E / s2x formülü, diğer şeyler eşit olmak üzere, daha ayrışık gruptan daha yüksek güvenirlik elde edileceğini açıklayıcı niteliktedir (Mehrens ve Lehmann, 1991; Allen ve Yen, l979). Aynı test, benzer büyüklükte biri bağdaşık diğeri ayrışık iki gruba uygulansın. Hata ölçüm varyansı eşit olma sayıtlısı 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." (Grup ayrışıklığı ve onun [test] güvenirlik üzerindeki etkisi arasındaki ilişkinin psikometrik tartışması, Gulliksen (1950) ve Magnusson' da (1967) mevcuttur.)

Güvenirlik, aracın kendisine değil bir ölçme aracı ile elde edilmiş ölçümlere işaret 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 tartışmalar özellikle Thompson (1994; 2001), Vacha-Haase (1998), Thompson ve Vacha-Haase'den (2000) gelirken, karşıt görüş sunan tartışmalar da Sawilowsky (2000), Knapp ve Sawilowsky'de (2001) ifade edilmiştir.

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 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.

Sonuç

"Test güvenilirdir." veya "testin güvenirliği" ya da "aracın güvenirliği" gibi ifade biçimleri doğru değildir. Zira güvenirlik, aracın kendisine değil, bir bellilendirme (assessment) aracı ile elde edilmiş ölçümlere işaret eder (Linn ve Gronlund, 2000). Bir diğer söyleyişle, güvenirlik testin kendisinin değil elde edilmiş ölçümlerinin bir özelliğidir (Lane, White ve Henson, 2002).

Kısacası, güvenirlik, testlerin değil, eldeki veriler veya ölçümlerin bir özelliği (Thompson, 1994), ölçümlerin bir fonksiyonudur (Capraro, Capraro ve Henson, 2001).

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. Hâl 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. Çünkü güvenirlik, testlerin değil ölçümlerin bir özelliğidir ve ifade edilmesi gereken de ölçüm güvenirliğidir.

Ana metne ek; doğru ifadelerin kullanılması için olası örnekler

Doğru ifadeler için , Thompson (1994), Cecil ve Stanley (1997), Arnau, Thompson ve Rosen (1999), Shiarella, McCarthy ve Tucker (2000), Yin ve Fan (2000), Steed (2001), Capraro ve Capraro (2002), Kieffer ve Reese (2002), Turgeon ve Chartrand'dan (2003) faydalanılmış ve bazı ifade örnekleri de uyarlanmış bir şekilde aşağıda sunulmuştur:

"X testi ölçümleri için test-tekrar test güvenirliği .91'dir."

"X ölçeğinden elde edilen ölçümler için Cronbach a .92'dir."

"X ölçme aracından elde edilen ölçümler için alfa güvenirliği .76'dır."

"Bu çalışmadaki ölçümlerin güvenirliği iki yarı yöntemiyle kestirilmiştir ve r = 0.85'tir."

"X testi ölçümlerinin güvenirliği düşüktür."

"X testine dair ölçümlerin iç tutarlılığı Cronbach a= .95'tir."

"Eldeki mevcut veriler için iki yarı güvenirliği .82 olarak hesaplanmıştır."

"X testi ölçümlerinin test-tekrar test güvenirliği .58'dir."

"Bu çalışmadaki ölçümlerin test-tekrar test güvenirliği .8l'dir."

"… testi ölçüm güvenirliği kestiriminde test-tekrar test yöntemi kullanılmıştır."

"X testi ölçümleri için iç tutarlılık güvenirlik kestiriminde KR- 20 formülü kullanılmıştır."

"Bu çalışmadaki ölçümlerin iç tutarlılığı Cronbach a ile hesaplanmıştır ve a = .87'dir."

"…ölçümlerinin iç tutarlılığı tatmin edicidir (Cronbach _= .86'dır)."

"X testi ölçümlerinin güvenirliği iki yarı yöntemi kullanılarak hesaplanmıştır."

Hiç şüphesiz ki teknik olarak doğru dilin kullanılması, en iyi uygulamalarla pekişecektir. Ancak, yıllardır süregelmiş genel bir hatadan vazgeçilerek doğru ifadelerin kullanılmasında, araştırmacılara, yazarlara, bilim adamlarına, bilimsel dergilerin editörlerine ve yazı işleri müdürlerine önemli bir sorumluluk ve görev düşmektedir.

**Ölçüm: (score)

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

Ölçümleme: (scoring)

Bellilendirme: (assessment)

***Metin içindeki […] arasındaki ifadeler Vahit Bademci tarafından eklenmiştir.

Kaynaklar

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

Arnau, R. C.; Thompson, B. ve Rosen, D. H. (1999). Alternative Measures of Jungian Personality Construct. Measurement and Evaluation in Counseling and Development, Vol. 32, 90-104.

Bademci, V. (1999). Hedefin Davranışlara Çevrilmesi, Davranışlardan Seçmeli Test Maddeleri Yazılması. (Geliştirilmiş Üçüncü Baskı). Ankara: Gazi Kitabevi.

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.

Cecil, H. ve Stanley, M. A. (1997). Reliability and Validity of Adolescents' Scores on the Body Esteem Scale. Educational and Psychological Measurement, Vol. 57, 340-356.

Crocker, L. ve 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.

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

Guilford, J. P. ve 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 and Sons.

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

Knapp, T. R. ve Sawilowsky, S. S. (2001). Constructive Criticisms of Methodological and Editorial Practices. The Journal of Experimental Education, Vol. 70, 65-79.

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 Gronlund, N.E. (2000). Measurement and Assessment in Teaching. Eigth 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.

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

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

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

Pedhazur, E. J. ve Schmelkin, L. P. (1991). Measurement, Design And Analysis: An Integrated Approach. Hillsdale, New Jersey: Lawrence Erlbaum.

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.

Shiarella, A. H.; McCarthy, A. ve Tucker, M. L. (2000). Development and Construct Validity of Scores on the Community Service Attitudes Scale. Educational and Psychological Measurement, Vol. 60, 286-300.

Steed, L. (2001). Further Validity and Reliability Evidence for Beck Hopelessness Scale Scores in A Nonclinical Sample. Educational and Psychological Measurement, Vol. 61, 303-316.

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

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. (1994). Guidelines for Authors. Educational and Psychological Measurement, Vol. 54, 834-47.

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 Genaralizability 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.

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

Turgeon, L. ve Chartrand, É. (2003). Psychometric Properties of the French Canadian Version of the State-Trait Anxiety Inventory for Children. Educational and Psychological Measurement, Vol. 63, 174-185.

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

Wilkinson, L. ve APA Task Force on Statistical Inference. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist, Vol. 54, 594-604.

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

Summary

IT IS INCORRECT TO SPEAK OF *"THE RELIABILITY OF THE TEST "* OR *"THE TEST IS RELIABLE"*

Vahit BADEMCİ*

It is incorrect to say " the test is reliable" or "the reliability of the test." Because, reliability is an attribute of obtained scores, not of a test.

Reliability is a function of sample. Because, reliability should be evaluated on selecting a sample from planned population. Reliability fluctuates from sample to sample, and so more heterogeneous samples often lead to more variable scores and thus to higher reliability. Therefore, same test or instrument, when administered to more heterogeneous or to more homogeneous the groups, will yield scores with different reliability. In that case, it is more appropriate to speak of the reliability of "test scores" or the "measurement" .

In a few words speaking, a test is not reliable or unreliable and the data in hand or the test scores are reliable or unreliable.

Address for Correspondence. * Yard. Doç. Dr. Vahit Bademci, Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Eğitim Bilimleri Bölümü, 06500 Beşevler/Ankara.