Opinion Article

Correcting Fallacies about Validity as the Most Fundamental Concept in Educational and Psychological Measurement *

Vahit BADEMCİ [1]

Abstract

Validity is the most fundamental cerebration in educational and psychological testing. That is to say, validity is a crucial concept in psychometrics, but it is still misunderstood and misused. Validity has changed in the last 100 years, in other words, evolved. Validity is the degree to which evidence and theory support the adequacy and appropriateness of the proposed interpretations and uses of the scores obtained from the test or measurement instrument applied to a particular population or sample. In short, validity is not a property of a test or measurement instrument itself, but it is a property of the proposed interpretations and uses of the scores. Thus, such statements as 'the test is valid', 'the validity of scale' or 'the scores are valid' should not be used. The most authoritative source regarding the development and evaluation of educational and psychological tests is published by name of the Standards for Educational and Psychological Testing and briefly referred to as the Standards. The view of content validity, criterion-related validity and construct validity supported in 1966 Standards was quitted in 1999 Standards.

Keywords: Validity, validation, sources of validity evidence, reliability, misconceptions in educational and psychological testing

1. INTRODUCTION

The field of educational and psychological testing is replete with fallacies, urban legends or misconceptions; reliability and validity concepts have also got one's share of these (Bademci, 2007, 2014; Goodwin & Goodwin, 1999; Phelps, 2009). However, validity is the most fundamental cerebration in educational and psychological measurement. In other words, measurement is at the core of scientific research and validity is at the heart of measurement (Bademci, 2013; Viswanathan, 2005).

Validity is the most important concept in educational and psychological testing, but it has been the most misunderstood or widely misused for a long time (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, 2014; Frisbie, 2005; Rogers, 1995). On the other hand, validity evolved and it still continues to evolve (Kane, 2001; Messick, 1989). Conceptions of validity have also changed remarkably over the past 100 years (Angoff, 1988; Kane, 2006).

1.1. Current Definitions of Validity, Validation, and Reliability

Validity and validation are two closely related but different concepts used in measurement (Kane, 2006; Newton & Shaw, 2014). Validity is the degree to which evidence and theory support the adequacy and appropriateness of the proposed interpretations and uses of the scores obtained from the test or measurement instrument applied to a particular population or sample (Bademci, 1999, 2019). Validation, on the other hand, is the process by which the evidence of the validity of score interpretations is collected (Bademci, 1999, 2017b). Besides, reliability is the reproducibility or the

Received Date: 04/07/2022 **Accepted Date:** 05/09/2022 **Publication Language:** English

* ***To cite this article:*** Bademci, V. (2022). Correcting fallacies about validity as the most fundamental concept in educational and psychological measurement. *International e-Journal of Educational Studies, 6* (12), 148-154. https://doi.org/10.31458/iejes.1140672

[1] *Assist.Prof.Dr., Gazi University, Ankara, Turkey bademci@gazi.edu.tr*

* *Corresponding Author e-mail adress: bademci@gazi.edu.tr*

consistency of the scores obtained from the test or measurement instrument applied to a particular population or sample (Bademci, 1999, 2011). It must be borne in mind that score reliability is necessary but not sufficient for score interpretation validity (Thompson, 2003).

2. MODERN VIEW ON VALIDITY AND CORRECTING FALLACIES ABOUT VALIDITY

Validity is a property of the proposed interpretations and uses of the scores; in other words, validity is not a property of a test or measurement instrument itself or of test scores (Bademci, 1999, 2017a; Cronbach, 1971; Furr & Bacharach, 2008; Kane, 2006). Therefore, the fallacious expressions such as 'validity of the test', 'the test is valid', 'the validity of scale', 'the validity of measurement instrument (or method)', 'the measurement procedure is valid', 'assessment validity', 'the validity of raters', 'the validity of exam', 'the validity of test scores', 'the scores are valid' and so on should never be used (AERA, APA, & NCME, 1985; Bademci, 2007). For example, the question "Is the test valid" is incorrect; it is appropriate to ask the question "Is it valid the interpretation of the scores from the test?"

Today, there is a broad consensus on the point that validity is related to the interpretations that have been made according to the test scores but not the tests themselves (AERA, APA, & NCME, 1999, 2014; Cizek, 2016; Cronbach, 1971; Kane, 2006; Messick, 1989). Also, at the core of this consensus, there is the underlying opinion that the interpretation of test scores is valid (Cronbach, 1971; Newton, 2012). Validity is a matter of degree; that is, validity is not a concept of all-or-none (Bademci, 1999, 2019; Kane, 2013; Nunnally, 1978). Instead, validity of the interpretation of the scores should be stated with certain degrees such as high validity, medium validity, low validity or no validity (Linn, 2010; Linn & Gronlund, 1995). That is to say, validity is not presented as a dichotomy (valid or not), because it is a continuum, one end of which is anchored by interpretations of scores that simply are not justified (Koretz, 2008). Validity is also dependent on the population or the sample like reliability; in other words, it is always specific to a particular population or sample or group (Bademci, 1999, 2011; Linn & Gronlund, 1995). It should not be neglected that "…validity information varies with the group tested…" (Linn & Gronlund, 1995, p. 77).

Validity is an evaluation argument and includes an evaluative judgement; it was founded on empirical evidence and theoretical rationales (Bademci, 1999, 2017a; Linn & Miller, 2005; Messick, 1989; Osterlind, 2006). In other saying, validity requires an evaluation of the degree to which the proposed interpretations and uses of the scores are justified by supporting evidence (Linn & Miller, 2005). Philosophical bases of the validity theory have also changed in years. The traditional psychometric viewpoint on validity which was put forward in the early twentieth century was rooted in positivism; nevertheless, the practices of contemporary validity theory and validation which point out that validity is a property of interpretations which were made from scores have been strongly influenced by constructivism (constructive realism, especially since 1980s) (Bademci, 1999, 2017a; Messick, 1989; Mislevy, 2018; Sijtsma, 2009).

3. CONTEMPORARY VALIDITY AND *1999 STANDARDS*: REJECTION OF THE HOLY TRINITY OF VALIDITY (CONTENT VALIDITY, CRITERION-RELATED VALIDITY, AND CONSTRUCT VALIDITY)

In fact, the most authoritative source regarding the development and evaluation of educational and psychological tests is published by name of the *Standards for Educational and Psychological Testing* (AERA et al., 1985, 1999, 2014; APA et al., 1966) and briefly referred to as the *Standards*. The most major change in concept of validity also occurred in *1985 Standards*; validity is a unitary concept (AERA, APA, & NCME, 1985; Algina & Penfield, 2009; Bademci, 1999, 2007; Messick,

1989). "The trinitarian doctrine" or "the holy trinity" of validity (Guion, 1980) which accepts that there are three kinds of validity such as content validity, criterion-related validity and construct validity supported in *1966 Standards* was rejected and abandoned in *1999 Standards* (APA, AERA, & NCME, 1966; AERA, APA, & NCME, 1999; Bademci, 1999, 2017b).

However, in *1999 Standards* that have represented the modern view arguing validity as a unitary concept based on various types of validity evidence, under the title of "sources of validity evidence", the types of validity evidence was presented as 1) evidence based on test content, 2) evidence based on response processes, 3) evidence based on internal structure, 4) evidence based on relations to other variables, 5) evidence based on consequences of testing [evidence for validity and consequences of testing] (AERA, APA, & NCME, 1999, 2014); the latest edition of the *Standards* was published in 2014. The types of validity evidence are encapsulated below.

3.1. Sources of Validity Evidence

Evidence based on test content "can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA, APA, & NCME, 2014, p.14). Such evidence includes "traditional content validity studies and alignment studies that require independent subject matter experts (SMEs) to review and rate test items according to their content relevance, representativeness, or alignment to curricular objectives as well as practice (job) analyses in the case of employment, licensure, or certification tests" (Sireci & Faulkner-Bond, 2015, p. 221-222).

Evidence based on response processes refers to "concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers (AERA, APA, & NCME, 2014, p.15). Validity evidence in this type include think-aloud protocols, cognitive interviews that rely on examinees' verbalizations about their own thinking processes, eye-movement patterns and timing of responses (Ercikan & Pellegrino, 2017; Urbina, 2014).

Evidence based on internal structure comes from "analyses of the relationships of responses to different items on the test. The central idea is to investigate whether the relationships among item scores or score on parts of the test are as expected from the theory of the construct" (Algina & Penfield, 2009, p.118). In other words, "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, & NCME, 2014, p.16). Approaches or methods for gathering such evidence include factor analysis, item response theory, multidimensional scaling, differential item functioning, structural equating modeling, and cluster analysis (AERA, APA, & NCME, 1999, 2014; Algina & Penfield, 2009; Osterlind, 2006). Besides, it has been suggested strategies involving generalizability theory or internal consistency methods and other indexes of score reliability as validity evidence in this type (Osterlind, 2006; Urbina, 2014). Thus, Sireci and Soto (2016) remarked "Internal structure evidence also evaluates the "strength" or "salience" of the major dimensions underlying an assessment, and this salience has a relationship to internal consistency reliability " (p.152). Urbina (2014) noted "…for example, a test is designed to assess a unidimensional construct such as spelling ability or test anxiety. For these kinds of instruments, high internal consistency coefficients, like the coefficient alpha…, support the contention of unidimensionality" (p. 185). Nevertheless, Crocker and Algina (1986) noted "…alpha should not be interpreted as a measure of the test's unidimensionality" (p. 142). Bademci (2014) also emphasized that "Unidimensionality may be examined using exploratory factor analysis or especially confirmatory factor analysis…But, Cronbach's alpha should not be used as a measure of unidimensionality [or homogeneity]…Cronbach's alpha should be used to estimate of the score reliability based on the internal consistency among the [item] scores after unidimensionality is examined" (p. 23). However, it must be borne in mind that reliability serves as an integral component to the interpretation of the scores in many validation studies (Algina & Penfield, 2009).

Evidence based on relations to other variables refers to analyses of the relationship test scores and other variables. In other words, "In many cases, the intended interpretation for a given use implies that the construct should be related to some other variables, and, as a result, analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (APA, AERA, & NCME, 2014, p.16). Such evidence can include multitrait-multimethod study, test-criterion relationships (predictive and concurrent studies), validity generalization study, contrasted groups studies (APA, AERA, & NCME, 1999, 2014; Reynolds & Livingston, 2012; Suen & Rzasa, 2004). However, Algina and Penfield (2009) noted "…validation methods making use of correlational approaches (e.g., the correlation of multiple tests and multi-trait multi-method studies) can be impacted by the reliability of the obtained test scores, and thus the proper estimation of the reliability of the scores is an important consideration in interpreting the obtained validity evidence" (p. 119).

Evidence based on consequences of testing refers to evaluation of the intended (positive and negative) and unintended (positive and negative) consequences associated with interpretations and uses of test scores (AERA, APA, & NCME, 2014; Sireci & Faulkner-Bond, 2015). Examples of evidence based on consequences of testing include increased student dropout, increased teacher stress, improved student achievement, enhanced teacher and student motivation (Linn, 2010). The standard sets which were produced in *1999 Standards* have been maintained exactly and in an enhanced way in *2014 Standards* (AERA, APA, & NCME, 1999, 2014).

4. IN LIEU OF CONCLUSION: VALIDITY IS A UNITARY CONCEPT

In contemporary validity, distinct types of validity was rejected such as content validity, criterion-related validity and construct validity. As *1999 Standards* and *2014 Standards* pointed out, validity is a unitary concept and there are various types of validity evidence as evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, evidence based on consequences of testing [evidence for validity and consequences of testing] (AERA, APA, & NCME, 1999, 2014). Contemporary validity and the sources of validity evidence was manifested in Figure 1.

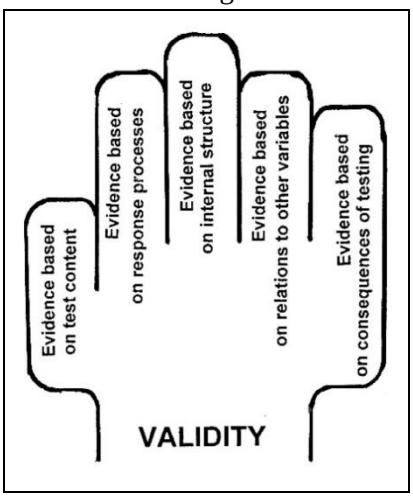

Figure 1. Validity and the sources of validity evidence

In addition, the radical changes related to validity and reliability were brought up to Turkey's agenda within the framework of a paradigm change by Bademci (1999, 2004, 2017a) 23 years ago for the first time.

Acknowledgement

Due to the scope and method of the study, ethics committee permission was not required.

3. REFERENCES

Algina, J., & Penfield, R. D. (2009). Classical test theory. In R. Millsap, & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 93-122). Los Angeles: Sage.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (APA, AERA, & NCME) (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, New Jersey: Lawrence Erlbaum.

Bademci, V. (1999). *Türkiye'de eğitim fakülteleri ve öğretmen yetiştirme: Öğretmen yetiştiren programlar nasıl olmalı?* [*Education faculties and teacher training in Turkey: How should teacher training programs be?*] Panel. Düzenleyen: ESEF İşletme Araştırma Topluluğu. Ankara: G.Ü. Mesleki Eğitim Fakültesi Konferans Salonu, 21 Mayıs 1999.

Bademci, V. (2004). Testin güvenirliği" veya "test güvenilirdir" diye ifade etmek doğru değildir [It is incorrect to express of "the reliability of the test" or " the test is reliable"]. *Türk Eğitim Bilimleri Dergisi*, *2*, 367-373.

Bademci, V. (2007). *Ölçme ve araştırma yöntembiliminde paradigma değişikliği: Testler güvenilir değildir* [*A paradigm change in measurement and research methodology: Tests are not reliable*]. Ankara: Yenyap.

Bademci, V. (2011). Türk eğitim ve biliminde bilimsel devrim: Testler ya da ölçme araçları güvenilir ve geçerli değildir [Scientific revolution in Turkish education and science: Tests or measurement instruments are not reliable and valid]. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, *16*, 116-132.

Bademci, V. (2013). Değerbiçiciler arası (interrater) ölçüm güvenirliğinin Cronbach'ın alfası ile kestirilmesi [Estimation of interrater score reliability by the Cronbach's alpha]. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, *30*, 55-62.

Bademci, V. (2014). Cronbach's alpha is not a measure of unidimensionality or homogeneity. *Journal of Computer and Educational Research*, *2*(3), 19-27.

Bademci, V. (2017a). Ölçme ve araştırma yöntembiliminde çağdaş gelişmeler ve yeni standartlar 1: Geçerlik, ölçümlerin kullanımlarının ve önerilen yorumlarının bir özelliğidir [Contemporary developments and new standards in measurement and research methodology 1: Validity is a property of the proposed interpretations and uses of scores]. *JRES*, *4*(1), 63-80.

Bademci, V. (2017b). Ölçme ve araştırma yöntembiliminde çağdaş gelişmeler ve yeni standartlar 2: Geçerlikte üçleme (kapsam, ölçüt ilişkili ve yapı geçerlikleri) öğretisinin reddi ve geçerlik kanıtının kaynakları [Contemporary developments and new standards in measurement and research methodology 2: Rejection of the trinitarian (content, criterion-related, and construct validities) doctrine in validity and sources of validity evidence]. *JRES*, *4*(1), 81-97.

Bademci, V. (2019). Geçerlik: Nedir? Ne değildir? [Validity: What is it? What is it not?] *JRES*, *6*(2), 373-385.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212-225.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 1-8). New York: Routledge.

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

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

Goodwin, L. D., & Goodwin, W. L. (1999). Measurement myths and misconceptions. *School Psychology Quarterly*, *14*(1), 408-427.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, *11*(3), 385-398.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education & Praeger.

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

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

Linn, R. L. (2010). Validity. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education, Volume 4* (pp. 181-185). Oxford: Elsevier.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.).Upper Saddle River, New Jersey: Pearson.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan Publishing Company.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement, 10*(1-2), 1-29.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. London: Sage.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, New Jersey: Pearson.

Phelps, R. P. (Ed.). (2009). *Correcting fallacies about educational and psychological testing*. Washington, DC: American Psychological Association.

Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory & methods*. Boston: Pearson.

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction*. Pasific Grove, California: Brooks/Cole.

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

Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, *39* (1), 215-252.

Sireci, S. G., & Soto, A. (2016). Validity and accountability: Test validation for 21st-century educational assessments. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 149-167). New York: Routledge.

Suen, H. K., & Rzasa, S. E. (2004). Psychometric foundations of behavioral assessment. In S.N. Haynes, & E. M. Heiby (Eds.), M. Hersen (Series Ed.), *Comprehensive handbook of psychological assessment, Volume 3* (pp. 37- 56). Hoboken, New Jersey: John Wiley & Sons.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability* (pp. 3-23). Thousand Oaks, California: Sage.

Urbina, S. (2014). *Essentials of psychological testing* (2nd ed.). Hoboken, New Jersey: Wiley.

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, California: Sage.