Validity and Truth

Denny Borsboom Jaap van Heerden Gideon J. Mellenbergh

Department of Psychology, University of Amsterdam
ml_borsboom.d@macmail.psy.uva.nl

Summary. This paper analyzes the semantics of test validity. First, a simple definition of validity is given: Test X is valid for the measurement of attribute Y, if and only if the proposition 'Scores on test X measure attribute Y' is true. We analyze the meaning of validity by examining the truth conditions of the proposition. These truth conditions depend on the interpretation of the term 'measures'. Three measurement systems that may provide such an interpretation are examined: Fundamental measurement theory, classical test theory, and latent variable theory. It is argued that the semantics of validity depend on the choice of measurement system. Because there is no logically or empirically compelling argument for or against any particular theory, the meaning of validity is to a certain extent indeterminate.

Key words: validity, truth, measurement, semantics, psychometrics

1 Introduction

Test validity is related to methodological aspects of test construction, psychometric properties of tests scores, philosophical perspectives on psychological constructs, as well as various legal and ethical issues surrounding test use. In view of the fact that validity is connected to so many questions from so many disciplines, it is no surprise that presently endorsed conceptualizations of validity (Messick, 1989) have come to cover virtually every aspect one could possibly imagine to be related to psychological testing.

If one insists on addressing all these matters by a single term, extending the scope of the validity concept in order to cover such diverse topics is not only understandable, but necessary. However, the problem with using validity as an umbrella term is that the meaning of the concept becomes clouded. And, in our view, the broadening of the validity concept in recent theoretical developments has not lead to a clear articulation of the semantics of validity. This is to say that a very basic question, namely 'what does it mean for a test to be valid' has received little attention, especially in comparison to epistemological ('how can we know that a test is valid?'), methodological ('how can we investigate whether a test is valid?'), and ethical ('when and how should we use test scores?') questions. The present paper therefore attempts to provide an overview of possible semantic interpretations of the validity concept. We explicitly do not intend the analysis to cover evidential or consequential issues; we are purely concerned with the meaning of validity.

Readers familiar with the literature on validity will note a feature of the above introduction that is dissonant with the current consensus on validity that has dominated the literature in the past two decades. The above introduction considers validity to be a characteristic of a test, not of a test score or test score interpretation. This is in sharp contrast to the currently endorsed position, which states that validity is always a feature of a test score interpretation (Messick, 1989). The reason for our diverging view on this matter is at the heart of the present analysis and the topic of the next section. We argue that validity may be viewed as a characteristic of a test, and that whether a test has the characteristic of validity depends on the truth of a given test score interpretation – namely, the interpretation that the test measures a certain attribute. This yields an exceedingly simple formulation of validity. Also, it allows for a standard type of inquiry into the semantics of validity by analyzing the following question: what makes the proposition 'Scores on test X measure attribute Y' true?

2 Defining validity in terms of truth

In the past century, the meaning of the term 'validity' has shifted from tests to test score interpretations. Originally, validity was conceived of as a characteristic of psychological tests. As Kelley (1927) put it, to ask the question of validity is to ask whether a test measures what it purports to measure. In later developments, validity was seen as a characteristic of test scores rather than of psychological tests. The reason for this may have been that validity was conceptualized in terms of the correlation between test scores and criterion scores (e.g., Lord & Novick, 1968), and the test scores, rather than the tests, are correlated with criterion scores. Still later, validity was deemed a characteristic of the test score interpretation rather than of tests or test scores, and this is the current position as expressed, for example, in Messick (1989). From an epistemological perspective, this shift of meaning is perfectly understandable. What is to be supported evidentially is not a test score, but a test score interpretation; and there is little room for argument on this point.

From a semantic perpective, however, the situation becomes slightly confused. To illustrate this, consider the following test score interpretation, which will be our working example in this paper:

IQ-scores measure intelligence.

Most psychologists and psychometricians will agree that, when we discuss the validity of test score interpretations, this is the type of test score interpretation we are concerned with. Now, there is something interesting about the above interpretation. Namely, it has the form of a proposition, and we may reasonably assume that all test score interpretations can be cast in this form. But what does it mean to say that a test score interpretation is valid, if not that the proposition that expresses this interpretation is true? That is, there seems little harm in a restatement of validity as

The test score interpretation 'IQ-scores measure intelligence' is valid, if and only if the proposition 'IQ-scores measure intelligence' is true.

Upon this restatement, however, the term 'valid' is superfluous, because its meaning has been reduced to the notion of truth. To assign the predicate 'valid' to the test score interpretation is to assign it the predicate 'true'. We now have two words for expressing the same idea.

We do not feel that much progress is being made by considering validity to be a characteristic of a test score interpretation. Not only does the concept end up doing nothing more than the concept of truth was already doing, it also seems more natural to consider validity to be a feature of a test: It was this conceptualization that made validity famous, so to speak, and it is also the way the great majority of test developers and users think about the issue. Further, we think that the argument that shifted the meaning of validity from tests to test score interpretations is erroneous. Namely, although the observation that one validates test score interpretations rather that test scores is correct, we see no reason why one would be forced to conclude from this that the term 'validity' can only meaningfully be applied to test score interpretations. Semantically, validity can be considered a characteristic of tests, rather than of test score interpretations, even though one can only validate interpretations. Namely, one can define the validity of a test in terms of the truth of a test score interpretation. That this can be done in a consistent manner is illustrated in the following definition of validity:

A test X is valid for the measurement of attribute Y, if and only if the proposition 'Scores on test X measure attribute Y' is true.

In terms of our working example, this becomes

An IQ-test is valid for the measurement of intelligence, if and only if the proposition 'IQ-scores measure intelligence' is true.

The two interesting features of this definition are a) that validity is explicitly considered to be a characteristic of the test in question, and b) that it is not at all applied to the test score interpretation: This role is taken over by the notion of truth. At the same time, the definition is not inconsistent with the idea that, in research, one validates interpretations of test scores rather than the test scores themselves. We conclude that it is not necessary to characterize validity as a property of test score interpretations, rather than of tests.

The semantics of validity are thus shifted back to where they originally came from (e.g., Kelly, 1927), and are brought in close agreement with the notion of validity that dominates the thinking of most psychologists. At the same time, however, the definition invites closer analysis; not of the concept of validity, which, as a predicate for tests, is defined in terms of the truth of a corresponding test score interpretation, but of the proposition 'IQ-scores measure intelligence'. That is, the semantics of validity can be clarified by looking at the conditions that would make the proposition true.

3 An investigation into the semantics of validity

One of the benefits of the definition above is that it makes explicit that an account of measurement is required. It is remarkable that influential treatises on validity, a concept deemed central to measurement, only superficially address theories of measurement, if at all. It seems to be tacitly assumed that it does not really matter whether one conceives of measurement from, say, a true score perspective (Lord & Novick, 1968), a latent variables perspective (Hambleton & Swaminathan, 1985), or a fundamental measurement theory perspective (Krantz, Luce, Suppes, & Tversky, 1971). As these theories conceive of the measurement process differently, however, we may expect that they assign different truth values to the proposition 'IQ-scores measure intelligence'. That is, it is likely that the semantics of validity are not independent of the measurement theory that gives the interpretation of the term 'measures'.

First, consider fundamental measurement theory. In this theory, measurement is a process of representing observed relations between objects (henceforth: persons) in a number system. Each person is assigned a number so that the relations between the assigned numbers are homomorphous with the observed relations between persons. This means that the observed relations are preserved in the numerical representation. The product of this process (i.e., the numerical assignment) is called a scale. Note that the scale is a product of human activity: it is therefore not necessary to assume, a priori, that scales exist independently of the act of measurement, and that they are somehow responsible for the observed relations. This is in contrast to, for example, latent variable models. Scales represent relations, they do not cause relations. Now, if observed relations can be represented in the number system (that is, if a homomorphism can be constructed), the resulting scale is an adequate representation by definition, and therefore measurement has succeeded. If the procedure fails, measurement has not taken place.

Let us consider our paradigm example, and interpret the proposition 'IQ-scores measure intelligence' from this perspective. As has been pointed out by Michell (1986), IQ-scores are not representations of observed relations because they involve the summation of item scores that do not form a unidimensional Guttman scale. This simply means that two people with different answer patterns may be assigned the same IQ-score. Thus, the assignment of IQ-scores does not produce a homomorphism between observed and numerical relations. Because IQ-scores are not the product of a measurement process, they cannot be properly considered measurements of anything. The proposition 'IQ-scores measure intelligence' is thus false. Moreover, from a fundamental measurement perspective, measurement is extremely rare in psychology (if it occurs at all), because very few psychological tests produce the type of consistencies required for representational theory to operate. Thus, according to this definition of measurement, most or all psychological tests are invalid.

Second, consider the measurent process from a classical test theory perspective. Classical test theory conceives of measurement in a statistical fashion. As Lord & Novick (1968, p. 20) put it, a test score is a measure of a theoretical construct if its expected value increases monotonically with that construct. At first sight, the theoretical construct could be taken to be the true score (Lord & Novick never explicitly say this, but suggest it in a number of places). Oddly enough, however, the true score is itself defined as the expected test score. Because true scores are identical to expected scores, and because any variable increases monotonically with itself, every test must measure its own true score perfectly. Therefore, if the true score on an IQ-test is considered to be identical to intelligence, the proposition 'IQ scores measure intelligence' is true by definition. This is because the proposition 'IQ-scores measure intelligence' is tranformed to 'the expected IQ-scores are monotonically related to the true scores on the IQ-test' which is vacuously true since the true scores are identical to the expected scores. Because the line of reasoning succeeds for every conceivable test, in this interpretation every psychological test is valid. However, it is only valid for its own true score. This is the price of operationalism: If the construct is equated with the true score, each distinct test defines a distinct construct, because it defines a distinct true score.

An alternative interpretation of classical test theory is that the observed scores do not measure the true scores (after all, it is rather odd to say that an expected value measures itself), but that the true scores measure something else, in the sense that they are themselves monotonically related to the theoretical construct in question. Viewing the issue in this way, the sentence 'IQ-scores measure intelligence' is true if the true scores on the test are monotonically related to intelligence. From a classical test theory perspective, this means that the theoretical construct cannot be conceived of as represented in the measurement model for the test in question, but must be viewed as an external variable. This prompts the conceptualization of validity as correlation with a criterion variable, which yields the concept of criterion validity.

Criterion validity has been extremely important to the theoretical development of the validity concept, for the following reason. Originally, the criterion was considered to be an observed variable, such as grades in college.

Because the validity question refers to measurement and not to prediction, and because IQ-scores do not attempt to measure college grades (which are, after all, observable) but intelligence, the criterion validity view was never an adequate conceptualization of test validity. What happened in response to this, is that the criterion variable was swept under the carpet of unobservability, and attained the status of a hypothetical entity. The definition of validity in terms of a statistical relation (i.e., the true score increases monotonically with the theoretical construct) was, however, retained. The measurability of the intended construct (intelligence) is thereby hypothesized a priori, and the validity of the measurements (IQ-scores) is conceptualized as a monotone relation of the true scores on the IQ-test with this hypothetically measurable attribute. In this view, validity is external to the measurement model, because in classical test theory a theoretical construct such as intelligence cannot be non-vacuously represented inside the measurement model. The proposition 'IQ-scores measure intelligence' thus becomes 'the true IQ-scores increase monotonically with a hypothetical criterion variable called intelligence'. Attempts to find 'perfect' measurements of intelligence that could function as a standard, analogeous to the standard meter in Paris, were, of course, fruitless. The type of thinking introduced by looking at intelligence as a criterion variable outside the measurement model is, however, still a very common way of thinking about test validity. That is, there is 'something out there', and the question of validity is how high the correlation between our test scores and that something is. This renders the semantics of validity dependent on two assumptions: 1) there really is something out there (intelligence), and 2) the test scores have a positive correlation with that something. If this is the case, then the proposition 'IQ-scores measure intelligence' is true. An interesting aspect of this view is that, because expected test scores will have monotonic relations to many attributes, any given test measures an indeterminate number of attributes. Thus, measures are not uniquely tied to a construct. If measurement is further reduced to correlation, everything measures everything else to a certain extent, and all tests must be valid.

The reason that classical test theory must consider theoretical constructs as external to the measurement model is that the syntactical machinery of the theory is not rich enough to represent constructs inside the model. As we have seen, the true score cannot perform this function without rendering a completely trivial account of measurement. Latent variable models (Hambleton & Swaminathan, 1985) do possess the required terminology. Such models relate the true scores on a number of items or tests to an underlying dimension. This dimension functions as a representative for the theoretical construct (to be distinguished from the function of fundamental measurement scales, which are representations of observed relations). The relation of measurement in latent variable models is rather similar to the statistical formulation of classical test theory; namely, it is conceived of in terms of a stochastic relation of the observed scores to the latent variable. However, these models do have the power to dispose of the problem that tests are valid for any attribute they are monotonically related to, because the dimensionality of the latent space can be specified in the model (this is not possible in classical test theory, except through the awkward requirement of strict parallellism). For example, in the unidimensional case, a latent variable model specifies that the true scores on each of a number of indicators are monotonically related to the same latent variable. Moreover, within such unidimensional models it is assumed that the indicators measure only this latent variable and nothing else. This implies that the indicators are independent, conditional on the latent variable. If, conditional on the latent variable, the indicators are still related to another variable (for example, group membership), the indicators are considered biased. Thus, in the unidimensional case, measurement can be seen as a monotonic relation of the expected scores with a latent variable, and only with this latent variable (in the sense that they do not systematically relate to another variable, given the latent variable). The proposition 'IQ-scores measure intelligence' thus becomes 'the expected IQ-scores increase monotonically with the latent variable intelligence, and, given the latent variable, with nothing else'. It follows that the semantics of unidimensional latent variable models do not allow indicators to be valid for more than one latent variable, in contrast to the classical test model.

This short and incomplete review of three important theories of psychological measurement shows that different definitions of the term 'measure' in the proposition 'IQ-scores measure intelligence' lead to different semantics for the validity concept. From a fundamental measurement perspective, measurement is a homomorphous mapping of observed relations into a number system; for true score theory, measurement is either a completely vacuous term, or it reduces to correlation with a possibly hypothetical criterion variable; for unidimensional latent variable theory, measurement is conceived of as an exclusive stochastic relation with a latent variable. It is not difficult to conceive of instances where the proposition 'scores on test X measure attribute Y' is assigned different truth values depending on the chosen definition of measurement. Thus, the semantics of validity cannot be considered apart from a definition of measurement, which is not altogether surprising since validity is supposed to be the central concept in measurement. Once a definition of measurement is chosen, the semantics of validity are fairly well fixed. However, the choice of a definition of measurement is not forced upon us by logical or philosophical arguments, and neither by empirical facts. Therefore, there is a degree of indeterminacy in the meaning of validity.

4 Discussion

In the present paper, we have provided a definition of validity in terms of the truth of a test score interpretation: test X is valid for the measurement of attribute Y if the proposition 'scores on test X measure attribute Y' is true. It seems to us that the benefits of formulating test validity in this way are considerable. First, defining validity in terms of truth has the advantage that it lines up nicely with pretheoretical intuitions concerning tests and test use; the consensus in the validity literature, holding that validity is a characteristic of test score interpretations instead of tests, seems not to have found its way into mainstream psychology, and understandably so. It conflicts with basic intuition which, we think, is close to, or even identical with, the definition of validity that we are suggesting here. Second, the present definition of validity clearly delineates between the semantics of validity ('what does it mean for a test to be valid?') and the evidential or epistemological problems in establishing validity ('how do we find out whether a test score interpretation is true?'). This produces a distinction between validity and validation, analogous to the distinction between truth and verification – an important distinction that, in our opinion, has often been blurred in the validity literature. It also provides a way of performing a semantic analysis of the validity concept, an undertaking of which we have reported in this paper. The analysis shows that the semantics of validity cannot be separated from the notion of measurement. Since multiple interpretations of measurement are available, none of which is logically or empirically forced upon us, the semantics of validity are to a certain extent indeterminate. Thus, the traditional question of validity, 'do IQ-scores really measure intelligence?', is not a purely factual question, although it is often presented in this way and the word 'really' invites this interpretation. The meaning of validity is fixed through a choice of measurement system that, depending on one's point of view, can either be described as arbitrary, normative, conventional, or pragmatic. After this choice is made, propositions such as 'IQ-scores measure intelligence' may be read as factual; but not before. Thus, it seems that our analysis underscores the importance of regarding validity as inherently involving norms and values, as Messick (1989) indicates. We add that this is much to our surprise.

References

1. Hambleton, R. K., Swaminathan, H. (1985): Item Response Theory: Principles and applications. Kluwer-Nijhoff, Boston
2. Kelley, T. L. (1927): Interpretation of educational measurements. Macmillan, New York
3. Krantz, D. H., Luce, R. D., Suppes, P., Tversky, A. (1971): Foundations of measurement, Vol. I. Academic Press, New York
4. Lord, F. M., Novick, M. R. (1968): Statistical theories of mental test scores. Addison-Wesley Publishing Company, Reading, MA
5. Messick, S. (1989): Validity. In Linn, L.R. (ed), Educational Measurement. American Council on Education and National Council on Measurement in Education, Washington, DC
6. Michell, J. (1986): Measurement scales and statistics: A clash of paragdigms. Psychological Bulletin, **100**, 398-407.