Psychological Review
2004, Vol. 111, No. 4, 1061–1071

0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.4.1061

The Concept of Validity

Denny Borsboom and Gideon J. Mellenbergh
University of Amsterdam

Jaap van Heerden
Maastricht University

This article advances a simple conception of test validity: A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes. This conception is shown to diverge from current validity theory in several respects. In particular, the emphasis in the proposed conception is on ontology, reference, and causality, whereas current validity theory focuses on epistemology, meaning, and correlation. It is argued that the proposed conception is not only simpler but also theoretically superior to the position taken in the existing literature. Further, it has clear theoretical and practical implications for validation research. Most important, validation research must not be directed at the relation between the measured attribute and other attributes but at the processes that convey the effect of the measured attribute on the test scores.

We start this article with a request to the reader. Please take a slip of paper and write down your definition of the term *construct validity*. Now, take the classic article of Cronbach and Meehl (1955), who invented the concept, and a more recent authoritative article on validity, for instance that of Messick (1989), and check whether you recognize your definition in these works. You are likely to fail. The odds are that you have written down something like "construct validity is about the question of whether a test measures what it should measure." If you have read the articles in question carefully, you have realized that they do not conceptualize validity like you do. They are not about a property of tests but about a property of test score interpretations. They are not about the simple, factual question of whether a test measures an attribute but about the complex question of whether test score interpretations are consistent with a nomological network involving theoretical and observational terms (Cronbach & Meehl, 1955) or with an even more complicated system of theoretical rationales, empirical data, and social consequences of testing (Messick, 1989).

This illustrates a remarkable feature of the validity literature of the past 50 years. The concept that validity theorists are concerned with seems strangely divorced from the concept that working researchers have in mind when posing the question of validity. This is because in the past century, the question of validity has evolved from the question of whether one measures what one intends to measure (Cattell, 1946; Kelley, 1927), to the question of whether the empirical relations between test scores match theoretical relations in a nomological network (Cronbach & Meehl, 1955), and finally, to the question of whether interpretations and actions based on test scores are justified—not only in the light of scientific evidence but with respect to social and ethical consequences of test use (Messick, 1989). Thus, validity theory has gradually come to treat every important test-related issue as relevant to the validity concept and aims to integrate all these issues under a single header. In doing so, however, the theory fails to serve either the theoretically oriented psychologist or the practically inclined tester: The theoretically oriented are likely to get lost in the intricate subtleties of validity theory, whereas the practically oriented are unlikely to derive a workable conceptual scheme with practical implications from it. A theory of validity that leaves one with the feeling that every single concern about psychological testing is relevant, important, and should be addressed in psychological testing cannot offer a sense of direction to the working researcher.

The objective of this article is to show that validity theory can do better. We aim to analyze the considerations that have led to the present state of affairs, to show that many of these are irrelevant, and to offer a simple, clear, and workable alternative. It is our intent to convince the reader that most of the validity literature either fails to articulate the validity problem clearly or misses the point entirely. Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what it purports to measure.

The argument to be presented is exceedingly simple; so simple, in fact, that it articulates an account of validity that may seem almost trivial. It is as follows. If something does not exist, then one cannot measure it. If it exists but does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether. Thus, a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. The general idea is based on the causal theory of measurement (e.g.,

Denny Borsboom and Gideon J. Mellenbergh, Department of Psychology, University of Amsterdam, Amsterdam; Jaap van Heerden, Department of Psychology, Maastricht University, Maastricht, the Netherlands.

We thank Ingmar Visser, Conor Dolan, Martijn Meeter, Keith Markus, and Brenda Jansen for their comments on earlier versions of this article. Niels Smits has brought some important differences between measurement and prediction to our attention.

Correspondence concerning this article should be addressed to Denny Borsboom, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands. E-mail: dborsboom@uva.nl

Trout, 1999). When considered within a latent variable framework, the position is akin to the one taken by Bollen (1989). Also, several theorists in the previous century have hinted at similar conceptions (e.g., Cattell, 1946; Loevinger, 1957). Finally, it is likely that most researchers think of construct validity in causal terms, so that one could consider the proposed conception to be a kind of underground interpretation of construct validity.

Nevertheless, in the consensus on validity that has emerged in the past 2 decades, it is difficult to find an explicit formulation resembling the above. In the writings of leading theorists (i.e., Cronbach, 1988; Kane, 2001; Messick, 1981, 1989, 1998; Shepard, 1993), one will not find much that sustains it; rather, one is likely to find this type of idea in a discussion of historical conceptions of validity (Kane, 2001, pp. 319–323). More important, even though many working researchers may be inclined to agree with the proposed conception if pressed for a definition of validity, its consequences are not followed through in mainstream validation research. These consequences are not trivial. That the crucial ingredient of validity involves the causal effect of an attribute on the test scores implies that the locus of evidence for validity lies in the processes that convey this effect. This means that tables of correlations between test scores and other measures cannot provide more than circumstantial evidence for validity. What needs to be tested is not a theory about the relation between the attribute measured and other attributes but a theory of response behavior. Somewhere in the chain of events that occurs between item administration and item response, the measured attribute must play a causal role in determining what value the measurements outcomes will take; otherwise, the test cannot be valid for measuring the attribute. It is important to note that this implies that the problem of validity cannot be solved by psychometric techniques or models alone. On the contrary, it must be addressed by substantive theory. Validity is the one problem in testing that psychology cannot contract out to methodology.

In the course of developing the conception of validity as put forward above, we aim to do two things. First, we aim to offer simple, yet adequate, semantics for the validity concept. This is done through a juxtaposition of the proposed validity concept with existing theory in three domains: ontology versus epistemology, reference versus meaning, and causality versus correlation. Second, we aim to indicate future directions in research that may demystify, pinpoint, and solve the widespread validity problem in psychology. Here, the benefits of a trimmed-down validity concept are illustrated through a discussion of some clear theoretical and practical implications that may improve both the theory and practice of psychological measurement.

Ontology Versus Epistemology

If the crucial issue in validity concerns the existence of an attribute that causally influences the outcome of the measurement procedure, then the central claim is ontological and not epistemological. This is to say that one is claiming something about which things inhabit reality and how they relate to each other. Thus, the realm of ontology includes both the existence of phenomena and their causal influences. These constituents are fleshed out in the following sections, which address reference and causality in turn.

The truth of ontological claims is distinct from the ability to find out about reality, which is the central issue in epistemology. Measurement is the prototypical epistemological activity in science, and it is therefore easy to make the mistake that one is primarily claiming something on this front. This is because if the ontological claim holds, then the measurement procedure can be used to find out about the attributes to which it refers. Put more simply, if differences in intelligence cause differences in IQ scores, then the IQ score differences can be used to find out about the intelligence differences. Thus, in this very special case, the truth of the ontological claim guarantees the epistemological access. Note that the truth of the ontological claim is logically prior to the process of measurement itself because it is a necessary condition for measurement to be possible. Nevertheless, the ontological claim that underlies the measurement procedure is itself conceptually distinct from the measurement process. One can see this by considering the following analogy. It may be a necessary condition for entering a room that the door leading to that room is not locked. Still, that the door is not locked is conceptually distinct from walking through it to enter the room.

However, in the case of measurement, it would seem that to talk about the ontology is to talk about the epistemology, and there surely is a sense in which this is correct. Now, it is a small step to conclude that, instead of laying down the ontological claims involved in measurement, which make so abundantly clear the strong assumptions one is making about psychological attributes (Borsboom, Mellenbergh, & Van Heerden, 2003; Kagan, 1988; Michell, 1999), the discussion could just as well be limited to the epistemological side of the endeavor, which is respectable and familiar, without addressing the ontological issues. It is another small step to conclude that the essential question of validity is about particular aspects of this epistemological process called measurement. The final step, leading to some very dark philosophical dungeons from which escape is impossible, is to start talking about some presumed universal characteristics of this epistemological process (usually derived from a few paradigm cases like length or temperature measurement) that, if present, would allow one to somehow be rationally justified in concluding that the ontological claims are true.

This will not work. The family of procedures that scientists—as opposed to philosophers—regard as instances of measurement is diverse and incoherent and has few universal characteristics. Length and temperature, blood pressure and brain size, and pathology and intelligence all could be said to involve measurement, but the associated measurement practices are based on vastly different lines of reasoning and use vastly different methodologies. So, now one gets into trouble. What could it be that successful measurement procedures have in common? Is it the way the test looks? Is it representative sampling from a universe of behaviors? Is it the line of reasoning on which the test is constructed? Is it the correlation between a test and some external variable called the *criterion*? Is it the (presumed) fact that the test figures in a nomological network of constructs? Is it just that one can do something useful with regard to some purpose that is presumably different from measuring the hypothesized attribute? Or, are we on the wrong track here because what is important is not a characteristic of tests or test scores but of test score interpretations—which are, again, presumably different from the obvious ones like *IQ scores measure intelligence*?

This line of reasoning quickly gets us nowhere. The reason is that there are no universal characteristics of measurement except

the ontological claim involved. The only thing that all measurement procedures have in common is the either implicit or explicit assumption that there is an attribute out there that, somewhere in the long and complicated chain of events leading up to the measurement outcome, is playing a causal role in determining what values the measurements will take. This is not some complicated and obscure conception but a very simple idea. If one, however, fails to take it into account, one ends up with an exceedingly complex construction of superficial epistemological characteristics that are irrelevant to the validity issue. And because the measurement processes and models are diverse and complicated, one is likely to buy into the mistaken idea that the concept of validity must also be complicated. So, now one gets a multiplication of terms. For the human condition is such that someone will inevitably distinguish between *kinds of validity* and *degrees of validity*, and so, theorists are bound to come up with a hundred or so *validities*, which all come in *degrees*, until someone stands up because this is clearly ridiculous and claims that "all validation is one" (Cronbach, 1980, p. 99) so that all kinds of validity can be integrated and subsumed under one giant umbrella (Messick, 1989). And because one is now thoroughly convinced that validity concerns characteristics of an epistemological process, rather than an ontological claim, one will reach the conclusion that all this time we were really just talking about the one grand epistemological process—scientific research (Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989). However, given that every attempt at drawing a line between scientific and unscientific research either fails or duplicates the distinction between good and bad research, the exciting fact discovered here is that validation research is research. In other words, nothing has been discovered at all. And the reason for this is that there was nothing to be discovered in the first place.

When claiming that a test is valid, one is taking the ontological position that the attribute being measured exists and affects the outcome of the measurement procedure. This is probably one of the more serious scientific claims one can make, and it is often difficult to prove or refute. This, however, does not mean that the validity concept itself is complicated. Every test constructor in every scientific discipline has the stated line of reasoning in mind when he or she is constructing, administering, or interpreting a test. It is the only aspect that measurement procedures have in common. If one is going to search for homogeneity in the superficial characteristics of these procedures, one is not going to find any, and one is likely to build ever more complicated systems covering different aspects of validity. These systems, however, do not cover different aspects of validity but describe different research procedures for validation. So, *asking people what they think about the test* becomes *face validity*; *checking whether we can predict some interesting things with it* becomes *predictive validity*; *investigating whether the data fit our theory about the attribute* becomes *construct validity*; and so on.

The union of all possible test-related activities of this kind is not validity but validation. These terms are sometimes used interchangeably in the literature, but they are not the same. This is clear because validity is a property, whereas validation is an activity. In particular, validation is the kind of activity researchers undertake to find out whether a test has the property of validity. Validity is a concept like truth: It represents an ideal or desirable situation. Validation is more like theory testing: the muddling around in the data to find out which way to go. Validity is about ontology; validation is about epistemology. The two should not be confused. Now, most of the validity literature has not dealt with the problem of validity but with the problem of validation. Although there is nothing wrong with describing, classifying, and evaluating validation strategies, such activities are not likely to elucidate the concept of validity itself. In fact, if one concentrates on the epistemological problems long enough, one will move away from the validity concept rather than toward it. Consider, for example, Messick's (1989) widely cited definition of validity: "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13, italics in the original). No view could be farther apart from the one advanced here. Validity is not a judgment at all. It is the property being judged.

Reference Versus Meaning

That the position taken here is so at variance with the existing conception in the literature is largely because in defining validity, we have reversed the order of reasoning. Instead of focusing on accepted epistemological processes and trying to fit in existing test practices, we start with the ontological claim and derive the adequacy of epistemological practices only in virtue of its truth. This means that the central point in validity is one of reference: The attribute to which the psychologist refers must exist in reality; otherwise, the test cannot possibly be valid for measuring that attribute. This does not imply that the attribute cannot change over time or that that psychological attributes are unchanging essences (cf. Kagan, 1988). It does imply that to construe theoretical terms as referential requires a realist position about the phenomena to which such terms refer. Thus, measurement is considered to involve realism about the measured attribute. This is because we cannot see how the sentences *Test X measures the attitude toward nuclear energy* and *Attitudes do not exist* can both be true. If you agree with us in this, then you are in disagreement with some very powerful philosophical movements that have shaped validity theory to a large extent.

The strongest of these movements was logical positivism. Philosophers and scientists endorsing this theory saw it as their mission to exorcise all reference of theoretical terms (like *attitude*) because such reference introduces metaphysics, which the logical positivists thought was bad. They therefore constructed theoretical terms as nonreferential. This led them to focus on the meaning of theoretical terms. Meaning and reference are easily confused but are very different concepts. As a classic example (Frege, 1892/1952), *the morning star* and *the evening star* have different meanings (namely, the last star still to be seen at morning and the first star to be seen at evening) but refer to the same thing (namely, the planet Venus). Because the positivists had a slightly phobic attitude toward metaphysics, they wanted to explain the use of theoretical terms like *attitude* without letting these terms refer to reality.

This was an interesting endeavor, but it failed (see Suppe, 1977, for a good overview). However, one of the relics of the approach has plagued validity theory to this day. This is the nomological network. A nomological network is a kind of system of laws relating the theoretical terms to each other and to the observations.

For the positivists, this network served to create meaning without reference for the theoretical terms. The idea is that the meaning of a theoretical term is solely determined by the place of that term in the nomological network: The meaning of the term *energy* is fixed by the network and by nothing else—certainly not by a reference to actual energy. Thus, in this view, one can have meaning without reference and can invoke theoretical terms without automatically engaging in ontological claims, which always introduce a lot of metaphysics.

This idea was used by Cronbach and Meehl (1955) to put forward their idea of construct validity. Construct validity theory depends crucially on the availability of a nomological network. This network does double duty. First, it is said to give an implicit definition of psychological constructs, in the same way that the positivists attempted to define theoretical terms. Second, it serves to generate the idea of construct validity itself, which consists in the agreement between the nomological network and empirical data. That is, a test can be considered valid for a construct if the empirical relations between test scores match the theoretical relations between constructs. One can imagine this as two path models, one hovering over the other. One model stands for theoretical relations and the other for empirical relations. If the models match, then there is construct validity for test score interpretations in terms of the nomological network. For instance, suppose the nomological network says that the construct *intelligence* is positively related to the construct *general knowledge* and negatively to the construct *criminal behavior*. Further, suppose that one observes a correlation of .5 between an IQ test and a test for general knowledge and a correlation of −.4 between the IQ test and the number of months spent in prison. There is thus a match between empirical and theoretical relations. In construct validity theory, it is this match that constitutes and defines the validity concept.

Note that neither the idea of implicit definition of constructs nor the idea of construct validity itself can be formulated in the absence of a theory that relates the construct to other constructs. This is nicely illustrated by attempts to quantify construct validity in terms of the "distance" between theoretical and empirical relations (Westen & Rosenthal, 2003). Just like construct validity itself, such attempts do not get off the ground without some kind of nomological network. Also note that to define construct validity, one needs no reference to the existence of theoretical entities, and their causal impact on the measurement outcomes is not even a topic of discussion. Read Cronbach and Meehl (1955) to see how carefully they avoided this issue. As an illustration of the ambiguity of Cronbach and Meehl's article, one may compare Bechtold (1959) and Loevinger (1957), who both discussed construct validity but were talking about two completely different interpretations of the concept—one positivist and the other realist. In principle, however, within the construct validity perspective, there is no friction between *Test X measures the attitude toward nuclear energy* and *Attitudes do not exist*. As long as the empirically observed relations, between Test X and other tests, match the theoretical relations in the nomological network, all is fine.

There are, however, serious problems with this view. First, it gets the scientific process backward. One does not start with drawing a network of relations, subsequently encountering a blank spot in the middle, and then concluding that this must be an implicitly defined construct, the character of which can be read off from the theoretical relations one has drawn. Perhaps this happens in subatomic physics, in which the existence of hitherto unknown particles is sometimes derived purely from the fact that there turns out to be a blank spot in the theory that must be filled by an entity. But, to suppose that this is regular practice in psychology, or in science in general, would be absurd. It is important to think about the way an attribute relates to other attributes because this is a crucial step in theory formation. And in the process of spelling out these relations, one develops something that bears at least a superficial resemblance to a nomological network. But, it is farfetched to presume that such a network implicitly defines the attributes in question. One could just as well say that the nomological network is implicitly defined by the construct, that they define each other, or that nothing deserving the name *definition* is in play at all. It is even more contrived to presume that the validity of a measurement procedure derives, in any sense, from the relation between the measured attribute and other attributes. Length is not implicitly defined in terms of its relation with weight, and much less is the validity of a meter stick. There is no reason to suppose that the situation would be different with psychological attributes.

A second problem is that, even if one were to consider this view seriously, there are few, if any, nomological networks in psychology that are sufficiently detailed to do the job of fixing the meaning of theoretical terms. To fix this meaning requires a very restrictive nomological network. The reason is that if one wants to evade realist metaphysics, one cannot say "intelligence is a real attribute with causal impact on our measurements" but only that "intelligence is a theoretical term that is defined in terms of the relations it has to other constructs in the nomological network." Now, it is crucial for the ideas formulated in Cronbach and Meehl (1955) that the latter type of descriptive, nonreferential account is possible because otherwise one is forced to invoke a referent for a theoretical term like *intelligence*, which brings in the very metaphysics to be avoided through the back door—introducing such metaphysics changes the rules of the game considerably.

In some highly developed theories, like the ones in physics, one could at least begin to consider this account because they are restrictive enough to single out one particular theoretical term, which is the only one that has all the right relations. In psychology, such an account does not work because there is a general lack of detailed theory. We do have loosely specified ideas on how largely undefined attributes relate to each other under limited sets of circumstances, but this is not enough. The typical network in psychology is in terms of *higher* and *lower* correlations between attributes. Such a loose network is unrestrictive and can be satisfied by an indefinite number of attributes besides the intended one. That this is not just an academic point but a decisive argument against using a descriptive, nonreferential account can be immediately seen by considering the intelligence example discussed above. One does not get anywhere by saying that "intelligence is whatever is positively related to general knowledge and negatively to criminal behavior" because there are too many theoretical terms that will satisfy this description and many of them will evidently not be the same as intelligence. Few, if any, theoretical terms in psychology can be unambiguously identified in this way. Thus, this theory will not be able to single out theoretical terms by merely describing where they stand in a nomological network. Cronbach and Meehl (1955) did discuss the problem that nomological networks are incomplete and vague in psychology, but they

did not mention the most important implication of that problem: It is fatal to any positivist reading of their account because it shows that reference, and the accompanying realist metaphysics of measurement, cannot be avoided.

In this context, it has been noted by validity theorists (Kane, 2001; Shepard, 1997) that requiring the existence of a nomological network is unrealistic in psychology, which is correct. However, if one removes the nomological network from construct validity theory, one is left with very little indeed. In fact, dropping the nomological network leaves one without the heavily needed theory of meaning, and one is likely to be forced to introduce reference again, that is, to interpret the theoretical terms as referring to things out there in the world. We think that this is a plausible move, as will be evident, but the consequence is that the main idea of construct validity, as put forward by Cronbach and Meehl (1955), loses its bite. That is, if one reintroduces reference, then it is difficult to maintain that what constitutes validity is a match between empirical relations and theoretical relations. For this match is now rendered a helpful epistemological criterion, which may be given a signaling function but not much more. Thus, if there is a grave discrepancy between the theoretical and empirical relations, one knows that something is wrong somewhere; but, this can hardly be considered news. If the theoretical and empirical relations match, this match does nothing more than corroborate the theory. The match is no longer constitutive of validity, however, because the reintroduction of the realist metaphysics forces one to shift back to reference as the primary defining feature of validity.

The emphasis that is placed on the importance of ruling out alternative rival hypotheses for corroborating data (Cronbach & Meehl, 1955; Messick, 1989) partly acknowledges this. Consider, as an example, the validity of IQ tests as measures of general intelligence. The theory of general intelligence predicts positive correlations between different types of IQ tests. The observation that such positive correlations indeed obtain can thus be taken as corroborating data for the validity of these tests as measures of general intelligence. Now, according to validity theorists, it is important to rule out alternative hypotheses for such data. However, to what hypothesis should such alternative hypotheses be considered rivals? Obviously, they are rivals to the hypothesis that general intelligence exists and causes variation in the measurement outcomes. What, then, is to be seen as the defining feature of validity if not exactly the truth of that hypothesis? And if this is correct, then where does this leave the instrumentalist, positivist, and empiricist? Consider, for example, instrumentalism. This view does not invoke truth but usefulness as the primary criterion for the adequacy of scientific theories and measurements. However, we are surely not seriously considering the idea that we have to rule out rivals to the hypothesis that intelligence tests are useful. The Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) comes in a big heavy box, which is very useful to hit people on the head with, but the hypothesis that the WAIS is valid for inflicting physical injury is certainly not the kind of hypothesis we are interested in. Clearly, from the viewpoint of ruling out alternative hypotheses, the hypothesis that the test is useful is neither intended nor relevant, except for the very special hypothesis that it can be used to measure intelligence *because variation in intelligence produces variation in IQ scores*.

However, the flip side of this coin should not be overlooked: If no attribute answers the referential call, the test is not valid for measuring that attribute, no matter how useful the test may be for prediction or selection or how well it may fulfill other functions. As Kagan (1988) said, this "urges care in the use of descriptive terms" (p. 619) because if such a term is treated as referential but has no referent, then one is reifying terms that have no other function than that of providing a descriptive summary of a set of distinct attributes and processes. For instance, one then comes to treat a name for a group of test items as if it were the common cause of the item responses. That, of course, is a mistake. Kagan (1988) further noted some difficult problems in constructing referential connections for theoretical terms. Most important is the observation that this connection may not work in the same way or even refer to the same attribute or process in different contexts, situations, or persons. In a latent variable context, this problem is also noted by Borsboom et al. (2003), who discussed the possibility that the latent space underlying a set of item responses may not have the same dimensionality or structure across persons. This is a complicated issue, but that does not mean that psychologists may therefore neglect it. To state that one measures an attribute but that that attribute does not exist is not to put forward some sophisticated philosophical position but to make an empty gesture to evade the difficulties involved.

In conclusion, a positivist or instrumentalist reading of construct validity requires a descriptive theory of meaning that must invoke nomological networks. Cronbach and Meehl (1955) tried to construct an account of validity on this basis. However, the nomological network interpretation of construct validity is inadequate, as has been recognized in the literature. Dropping the nomological network from consideration simply means that one has to go back to a realist interpretation of psychological attributes. In a realist interpretation, however, the crucial issue is reference and not meaning. Therefore, a question like *Are IQ tests valid for intelligence?* can only be posed under the prior assumption that there does exist, in reality, an attribute that one designates when using the term *intelligence*; the question of validity concerns the question of whether one has succeeded in constructing a test that is sensitive to variations in that attribute.

Causality Versus Correlation

Although construct validity theory is, in its original form, inadequate, it does represent a serious attempt to forge a validity concept that has an account of meaning and a function for theory and that stresses that there is no essential difference between validation research and research in general. Moreover, if one removes the nomological network from consideration, replaces meaning with reference, and reintroduces the realist perspective, much of what is said in construct validity theory remains consistent and plausible. Also, the idea of construct validity was introduced to get rid of the atheoretical, empiricist idea of criterion validity, which is a respectable undertaking because criterion validity was truly one of the most serious mistakes ever made in the theory of psychological measurement. The idea that validity consists in the correlation between a test and a criterion has obstructed a great deal of understanding and continues to do so.

The concept continues to exert such a pervasive influence on the thinking of psychologists because many are under the impression that construct validity is really criterion validity, with the criterion replaced by the construct (this fallacy cannot be attributed to

construct validity theorists, as is evident from the writings of Cronbach & Meehl, 1955; Kane, 2001; and Messick, 1981, 1989). However, the inadequacy of this view does not depend on whether one views the criterion as a variable to be predicted from test scores, as an infallible measure of the theoretical construct to be measured, or as the theoretical construct itself. The crucial mistake is the view that validity is about correlation. Validity concerns measurement, and measurement has a clear direction. The direction goes from the world to psychologists' instruments. It is very difficult not to construct this relation as causal. Criterion validity uses correlation and similarity, whereas it should use direction and causality.

Of course, causality is a laden term, and many researchers seem afraid to use it. The platitude *correlation is not causation* is deeply inscribed in the conscience of every researcher in psychology, and in the literature, the word *causes* is often replaced by euphemisms like *determines*, *affects*, or *influences*; in measurement, we see traits "manifesting" or "expressing" themselves. What is meant is that traits cause observed scores. It is perfectly all right to say this because hypothesizing a causal account does not mean that one interprets every correlation as a causal relation. This, again, is the epistemological side of the issue, which remains as problematic as ever—although progress has been made in this respect, as is evidenced in the work of writers like Pearl (2000) as well as in the development of latent variable models. The primary power of causality lies in the theoretical opportunity to think directionally rather than in terms of similarity or correlation (see, for some good examples, Glymour, 2001; Pearl, 2000). Now, measurement is a causal concept, not a correlational one, and validity is so too. For clarification, some absurdities to which any theory based on a correlational account of validity leads are pointed out. The criticisms must be explicitly understood as applying not just to the criterion validity view but to any view that does not invoke a causal arrow pointing from the attribute to the measurement outcomes.

First, it has been observed by Guilford (1946) that the idea of criterion validity leads to the conclusion that a test is valid for measuring many things, as epitomized in his famous statement that a test is valid for anything with which it correlates. However, the likelihood of encountering zero correlation in real life is exceedingly small, and especially in the social sciences, everything tends to correlate with everything (Meehl, 1978). Therefore, the upshot of any line of thinking that sees correlation as a defining feature of validity is that everything is, to some degree, valid for everything else. This absurdity does not arise in a causal theory because it is not the case that everything causes everything else.

Second, the idea has the unfortunate consequence of equating degrees of validity with the value of the correlation coefficient: The higher the correlation, the higher the validity. The limiting case is the case in which two variables correlate perfectly, which would imply perfect validity. That is, if one views validity as correlational, one is bound to say that if two constructs have a perfect correlation, then "they are really the same construct under two different labels" (Schmidt & Hunter, 1999, p. 190). This is very problematic. For instance, suppose one is measuring the presence of thunder. The readings will probably show a perfect correlation with the presence of lightning. The reason is that both are the result of an electrical discharge in the clouds. However, the presence of thunder and the presence of lightning are not the same thing under a different label. They are strongly related—one can be used to find out about the other—and there is a good basis for prediction, but they are not the same thing. When one is validly measuring the presence of thunder, one is not validly measuring the presence of lightning for the simple reason that one is not measuring the presence of lightning at all. The limiting case of the correlational view implies that perfect correlation is perfect validity, and this leads to the idea that deterministically related attributes are the same thing. This absurdity does not arise in a causal theory because variations in the presence of lightning do not play a causal role in producing variations in the presence of thunder.

Third, the correlation is a population-dependent statistic; that is, it is sensitive to the amount of variability in the attribute to be measured across populations. A well-known instance is the attenuating effect of restriction of range in the presence of imperfect relationships between variables. Any correlational view must therefore hold that validity itself is by necessity variable over populations. Corrections for unreliability and restriction of range (Lord & Novick, 1968) are going to solve some of the trouble here but not all of it. In particular, there is one important, well-established case of valid measurement in which the population dependence of correlations raises serious problems. This is the case of extensive measurement (Campbell, 1920; Krantz, Luce, Suppes, & Tversky, 1971). This is very troubling because extensive measurement is more or less the paradigm example of measurement in general (Narens & Luce, 1986). To take a familiar example of extensive measurement, suppose that one is measuring the length of rods and that the measurement apparatus used is a meter stick. Further suppose that one is measuring without error. The correlation between the measurement outcome and the real length will be unity in most populations, as it should be, but there is an important class of populations in which it will be zero. This is the class of populations of rods of equal length. Therefore, one must conclude that in such a population, the meter stick is not valid for measuring length. This is a strange result. In extensive measurement, it is quite meaningful to say that all objects in such a subpopulation are, say, 4.2 meters long and that this measurement is valid. In the causal account, this absurdity does not arise. This is because causality is directional and conditional: The causal account says that if there are differences in the attribute, then these will produce differences in the measurement outcome. However, if there are no differences in the attribute, no differences in the measurement outcomes are expected. This in no way precludes the validity of the measurement outcomes themselves, which is exactly as it should be. Do note that the important issue here is not that the correlation, as a statistic, is in some sense inadequate but that a conceptualization of validity in terms of covariation, rather than causality, is flawed.

In conclusion, correlations are epistemologically relevant because they are sometimes indicative of causality, but they are not, and cannot be, constitutive of validity. Perhaps we have refuted this view in somewhat greater detail than is strictly necessary, as criterion validity has been considered inadequate at least since Cronbach and Meehl's (1955) introduction of construct validity (Kane, 2001; Messick, 1989). We considered a thorough refutation important, however, because it is our impression that many people who do not subscribe to the criterion validity perspective still entertain a correlational conception of validity—the only difference is that they have replaced the criterion with the construct

itself. However, it is clear that if attribute differences do not play a causal role in producing differences in measurement outcomes, then the measurement procedure is invalid for the attribute in question. Correlations are not enough, no matter what their size. Height and weight correlate about .80 in the general population, but this does not mean that the process of letting people stand on a scale and reading off their weight gives one valid measurements of their height. To state otherwise is to abuse both the concept of measurement and of validity. The very fact that a correlational view of measurement allows for this kind of language abuse must be considered a fundamental weakness; any theory of validity that sustains such absurdities should immediately be dropped from consideration. Therefore, not just criterion validity but any correlational conception of validity is hopeless. The double-headed arrows of correlation should be replaced by the single-headed arrows of causation, and these arrows must run from the attribute to the measurements.

Where to Look for Validity

The proposed theory of validity now stands. Validity is a property of tests: A valid test can convey the effect of variation in the attribute one intends to measure. This means that the relation between test scores and attributes is not correlational but causal. A test is valid for measuring an attribute if variation in the attribute causes variation in the test scores. In this case, we say that it is true that the test measures the attribute in question. The concept of validity thus expresses nothing less but also nothing more than that an attribute, designated by a theoretical term like *intelligence*, exists and that measurement of this attribute can be performed with a given test because the test scores are causally affected by variation in the attribute. This conception does the job we want validity to do, and it does it in a simple and effective way.

The analysis has direct relevance for the practice of test construction and analysis. In particular, it seems that the emphasis on the role of constructs in theories, and their place in nomological networks, has prompted validation research to adopt what has been called a top-down strategy (Cervone, 1997). This basically means that much validation research is concerned with creating tables of correlation coefficients and then checking whether these go in the right direction. Although such macrolevel relations are important, it would seem that the primary objective of validation research is not to establish that the correlations go in the right directions but to offer a theoretical explanation of the processes that lead up to the measurement outcomes. That is, there should be at least a hypothesis concerning the causal processes that lie between the attribute variations and the differences in test scores. As in Embretson's (1983) terminology, validation should be concerned primarily with construct representation and only secondarily with nomothetic span.

The upshot of this line of reasoning for test construction is clear. Purely empirical methods, based on the optimization of external correlations, are very unlikely to generate tests that can be considered valid measurements. This is because focusing on predictive properties will destroy, rather than enhance, measurement properties such as validity. The reason for this is simply that items that measure the same attribute will be correlated. Correlated items, however, are relatively useless in prediction because they generate multicollinearity (they do not explain unique variance in the criterion; Lord & Novick, 1968, p. 332; Smits, Mellenbergh, & Vorst, 2002). What one selects when optimizing predictive utility are items that are mutually uncorrelated but highly correlated with the criterion. This is not what one expects or desires in measurement. Note that this does not preclude that tests constructed in this manner may be highly useful for prediction. It does imply that optimizing measurement properties and optimizing predictive properties are not convergent lines of test construction.

What is missing in such empirically oriented methods is a theory on what happens between the attribute and the test scores. Maximizing correlations will not remedy this problem but aggravate it. One has to start with an idea of how differences in the attribute will lead to differences in test scores; otherwise, the project of test construction is unlikely to generate tests that are valid for more than prediction. This may be one of the few instances in which psychology may actually benefit from looking at the natural sciences. In the more exact quarters, nobody starts constructing measurement instruments without the faintest idea of the processes that lead to the measurement outcomes. And, it is interesting to note that the problem of validity appears never to have played the major and general role it has played in psychology. These two observations may well be related: The concept of validity may never have been necessary because the instruments were generally set up on the basis of an idea of how they would work. In that case, the question of what it is, precisely, that is measured can simply be resolved by pointing to the processes that lead to the measurement outcomes.

In contrast, the question of what psychological instruments measure is generally not answered by pointing to the way the instruments work but by pointing to the relation they have with other instruments. This way of working makes the question *What is measured?* a question to be answered after the test has been constructed. Thus, the contrast here is between a conception that sees validity as something that one puts into an instrument and a conception that views validity as something to be discovered afterward. Psychologists have tended to construe validity as an empirical matter; that is, the question of what is measured is to be answered by data. However, a century of experience with test construction and analysis clearly shows that it is very hard to find out where the scores are coming from if tests are not constructed on the basis of a theory of item response processes in the first place. Therefore, we would like to push our validity conception one step further and to suggest not only that epistemological issues are irrelevant to validity but that their importance may well be overrated in validation research too. A large part of test validity must be put into the test at the stage of test construction—a stage of the testing process that has received little attention compared with the enormous emphasis that has been placed on test analysis. Thus, it is suggested here that the issue may not be first to measure and then to find out what it is that is being measured but rather that the process must run the other way. It does seem that if one knows exactly what one intends to measure, then one will probably know how to measure it, and little if any validation research will be necessary. If this is correct, then the problem of validation research is not that it is difficult to find out what is measured; the problem is that it is difficult to find out what one intends to measure.

In this view, validation is not, and cannot be, a purely or even mainly methodological enterprise. This does not mean that methodological and psychometric techniques are irrelevant to valida-

tion research but that the primary source for understanding how the test works must be substantive and not methodological. Thus, we consider it impossible to argue for test validity solely on the basis of a multitrait, multimethod matrix. Such a matrix is helpful, but a favorable matrix configuration is not constitutive of validity. What is constitutive of validity is the existence of an attribute and its causal impact on scores. Therefore, if one does not have an idea of how the attribute variations produce variations in measurement outcomes, one cannot have a clue as to whether the test measures what it should measure. No table of correlations, no matter how big, can be a substitute for knowledge of the processes that lead to item responses. The knowledge of such processes must be given by substantive psychological theory and cannot be based on methodological principles.

There are certainly tests for which a considerable body of knowledge has accumulated in this respect. Examples of research in this direction are, for instance, the cognitive modeling approach in spatial reasoning tests (Embretson, 1994) and the latent class approach in the detection of developmental stages (Jansen & Van der Maas, 1997). Such approaches are distinct from mainstream validation research because they look for evidence of validity in different places. The balance scale task (Inhelder & Piaget, 1958; Siegler, 1981) is a good example. In this task, which is intended to measure the level of cognitive development, children are confronted with a balance scale. The balance scale has weights on both sides, and children have to indicate whether the scale will tip and, if so, to which side. The weights vary in number and are placed on varying distances from the center of the balance scale. The point of departure, in this work, is the formulation of a theory that characterizes the attribute. In this particular case, the theory says that children go through four discrete stages in cognitive development (Inhelder & Piaget, 1958). This theory is translated in a formal model, in this case a latent class model, which conceptualizes developmental stages as latent classes and development as an ordered series of discrete transitions between these classes (Jansen & Van der Maas, 1997).

Second, class membership is related to response behavior. Specifically, the theory formulates how children in different stages will approach the items in the task. For instance, children in the first stage will simply count the number of weights on both sides of the center to reach a decision; children in the second stage will take distances between the weights into account, but only if the number of weights on each side of the balance scale are equal; children in the third stage will incorporate both the number of weights and the distance of weights from the center but will start guessing if these cues conflict; and children in the fourth stage will compare products of weights and distances on both sides of the scale.

Third, these response strategies are linked to item response patterns. This is possible because children in different stages make different kinds of mistakes. Therefore, one can construct items for which the response strategies give conflicting responses, so that on some items, children in an earlier stage outperform children in a later stage. Through this chain of theory, the classes can be characterized with sufficient precision to allow for testing the adequacy of the model against observed response patterns (Jansen & Van der Maas, 1997, 2002). Here, the use of latent variable modeling can provide an excellent method for testing the theory against empirical data. The example shows how much can be achieved through a coordinated pattern of theory, test construction, and data analysis. It also shows how little remains of the validity problem, which is virtually reduced to the question of whether this theory of response behavior is true.

Now, the evidence for the validity of the balance scale test is not conclusive, as evidence hardly ever is; both the theory of discrete stage transitions and the number and character of the response strategies are not settled issues. In the present context, however, the point is not primarily that there is a correct theory of response behavior. The important point is that there is a theory of response behavior at all. And when such a theory is present, the problem of validity loses much of its mysterious quality and elusive character because it is clear what must be the case in reality for this test to be valid for the developmental stages in question. The clarity is achieved because the entire chain of events that leads from the attribute to the test scores is characterized. We know what has to happen between item administration and item response for the test to be valid. In other words, we know how the test is supposed to work.

Contrast this with mainstream validation research, for example, with research on personality tests. We do now have the methodology to test the hypothesis that the covariation between item scores is due to a number of common causes, namely confirmatory factor analysis (Jöreskog, 1971). However, as is often the case in psychology, we have beautiful models but too little theory to go with them. We can formulate the hypothesis that Extraversion is the common cause of the scores on a number of different items of subtests, but there is no good theory available to specify how different levels of Extraversion lead to different item responses. Thus, there is, at present, no detailed hypothesis of how the causal effect of Extraversion on the test scores is being conveyed. The causal model may be set up, but the arrows in it are devoid of interpretation. This does not show that personality tests are invalid or that no such thing as Extraversion exists. However, it does preclude any firm treatment of the problem of validity. The reason for this is that researchers expect to get an answer to the question of what the test measures, without having a hypothesis on how the test works. If one attempts to sidestep the most important part of test behavior, which is what happens between item administration and item response, then one will find no clarity in tables of correlation coefficients. No amount of empirical data can fill a theoretical gap.

It is disconcerting to find that a large proportion of test research is characterized by an almost complete absence of theories of response behavior and that so few researchers recognize that the problem of psychological measurement is not a matter of following the "right" methodological rules but of tackling one of the most challenging problems in psychology: How do psychological characteristics relate to empirical observations? Fortunately, there are various recent developments in theoretically inspired modeling (e.g., Embretson, 1994, 1998; Jansen & Van der Maas, 1997; Mislevy & Verhelst, 1990; Süss, Oberauer, Wittmann, Wilhelm, & Schulze, 2002; Wilhelm & Schulze, 2002) that show how much is gained when one starts to consider the processes involved in item response behavior and to utilize advanced test theory models that have been developed in the past century. However, studies that proceed in this manner are still scarce throughout psychology, and it is therefore no surprise that the problem of validity is so

widespread. What we hope to have shown is that it need not be that way.

Discussion

We have proposed a simple conception of validity that concerns the question of whether the attribute to be measured produces variations in the measurement outcomes. This theory of validity is based on ontology, reference, and causation, rather than on epistemology, meaning, and correlation. Although epistemological issues are central to validation and consequential issues are central to test use, both are considered irrelevant to the concept and definition of validity itself. The conjunction of these theses produces a viewpoint that is almost diametrically opposed to the currently endorsed conceptions of validity, which state that the concept applies to evaluative judgments of test score interpretations, that it depends on nomological networks, that it is complex and faceted, and that social, ethical, and political consequences are relevant to validity. We do not see the need for a unified validity concept (Ellis & Blustein, 1991; Messick, 1989; Moss, 1992; Shepard, 1993) because we think there is nothing to unify. The consequences of the proposed conception are far-reaching, but the overall picture that emerges is consistent and fits the intuitive notions most researchers have about validity quite well. We therefore think that a realist, causation-based concept of validity is a viable alternative to the current consensus in validity theory.

The philosophical assumptions involved in the present conception are strong—stronger, perhaps, than in any previous discussion of validity. Therefore, it may be argued that by invoking realism about psychological attributes and causal relations, we are engaging in metaphysical speculation. We concede this point, but it does not bother us. We think that the very idea that metaphysics and science are necessarily opposed is a relic that stems from logical positivism; in fact, we think that science is the best way of doing metaphysics we know. To the hard-boiled empiricist, we reply that it is naive to think that any scientific theory can get off the ground without introducing an ontological picture of how the world works, which will always contain metaphysical ingredients. Given that this is the case, the metaphysics better be good.

Other objections may come from the postmodern or social constructivist camp. An obvious one is the objection that psychological attributes are social constructions and that we are engaging in an unjustified reification of such constructions. However, that a realist ontology is necessary to apply the concept of validity says nothing about the kind of attributes that can and cannot be invoked. Now, if the supposition that psychological attributes are social constructions is meant in an eliminative way and thus is taken to mean that such attributes do not exist, then they cannot have any causal effects whatsoever, and it is impossible to measure them. However, this is not an argument against the semantics of validity as discussed in this article; it rather expresses the opinion that psychological tests are invalid, and it does so in terms of the very same semantics we have proposed. Another position could be taken by researchers who are prepared to defend the thesis that psychological attributes do exist, although they exist as social constructions. Such researchers may hypothesize that these social constructions in fact do have causal effects, for instance on measurement outcomes. In that case, the concept of validity applies as usual, and it invites constructivists to work out the causal chains involved in the process of social construction. In our view, neither eliminative nor "liberal realist" conceptions of social constructivism provide an argument against the proposed semantics.

There are various other questions about validity that stand in need of further theoretical investigation. For instance, although the present validity concept can be applied directly to reflective latent variable models used in psychological measurement, it seems that formative models (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000) do not allow for such application. In such models, the observed indicators are not considered to be causally affected by the latent variable but, rather, to cause such a latent variable. In this case, it is difficult to see how these observed indicators could be conceptualized as measures of the attribute in question because the arrows between the attribute and the observations run in the opposite direction. Consider, as an example of a construct typically addressed with formative models, socioeconomic status (SES). A formative model conceptualizes SES as a latent variable that is regressed on indicators such as annual income, educational level, and so on. Now, it would be odd to ask whether the question *What is your annual income*? is a valid measure of SES because according to our theory, this question does not measure SES; rather, it measures one of the determinants of SES, namely annual income. And at this level, one can consistently ask the question of validity, namely when one asks whether variation in annual income has a causal effect on variation in the responses to the question.

One may also imagine that there could be procedures to measure constructs like SES reflectively—for example, through a series of questions like *How high are you up the social ladder?* Thus, that attributes like SES are typically addressed with formative models does not mean that they could not be assessed reflectively, in which case the concept of validity, as defined here, can be applied. However, validity does not apply to the relation between formative indicators and the corresponding constructs. Because validity explicitly concerns the relation of measurement, one wonders whether it is appropriate to view formative models as measurement models in the first place. They might be better conceptualized as models for indexing or summarizing the indicators or as causal models that do not involve a measurement structure for which one can ask the question of validity as defined in this article. It would be interesting to further inquire how such models relate to the logic of measurement.

A second issue concerns the distinction between intraindividual and interindividual measurement structures (Borsboom et al., 2003). In our view, the proposed validity concept can be applied to both of these levels, although it is important not to confuse them. If the measurement outcomes are obtained in a group of people at a single time point, then the variation in scores ranges over people, and the only thing that can cause variation over people is something that also varies over people. Conversely, if the variation in scores ranges over time within a person, then the cause of this variation must also vary over time within that person. One's developmental trajectory on a Raven (1938) item does not cause differences between people on responses to that item, and that there exists a given source of differences between people does not cause one's response to the Raven item. The reason that such suppositions do not make sense is not that the Raven is invalid either for measuring developmental processes or for measuring differences between people; it could be valid for both. The problem is rather that there is a mismatch between the domain of

variation in the cause and the domain of variation in the effect. It may be countered that, assuming complete homogeneity of processes across persons, it is possible to make inferences about the structure of intraindividual processes on the basis of the measurement of interindividual differences or the other way around. This is true. However, that one can generalize to another domain does not mean that one has measured something in that domain. It is simply impossible to measure intraindividual variation in attributes with interindividual variation in scores or to measure interindividual variation in attributes with intraindividual variation in scores. This is not an empirical hypothesis that could be refuted by research but a matter of logic. Nevertheless, the validity concept applies with equal force to measurement at the interindividual and the intraindividual level—although not necessarily at the same time or with respect to the same attributes.

A third theoretical consequence of the present article is that it raises the question of whether validity should be conceptualized as a matter of degree. This has become more or less a dogma of construct validity. Cronbach and Meehl (1955) stated that "the problem is not to conclude that the test 'is valid' for measuring the construct variable" but that "the task is to state as definitely as possible the degree of validity" (p. 290). Similarly, Messick (1989) stated that "it is important to note that validity is a matter of degree, not all or none" (p. 13). However, the question of whether an attribute exists and has causal impact on the observations can be answered by a simple "yes" or "no." Thus, the present theory is naturally compatible with the possibility of conceptualizing validity as a qualitative, rather than quantitative, concept, as is commonly done with the similar concept of truth. It would be worthwhile to investigate this possibility further.

These ideas point to another issue that deserves further study. Namely, how does validity line up with other test theoretic concepts like reliability, unidimensionality, measurement invariance, and bias? For instance, if validity is conceptualized as a qualitative concept, it would hardly make sense to say that reliability provides an upper bound for validity (Lord & Novick, 1968). One would rather say that validity is a necessary condition for reliability estimates to make sense. This is not implausible because reliability is an index of measurement precision (Mellenbergh, 1996), and it does seem strange to say that "Test X measures intelligence with a certain precision" but that "The test does not measure intelligence." Similarly, the concept of validity as defined here does not imply absence of bias in tests, which is a commonly held idea (but see Borsboom, Mellenbergh, & Van Heerden, 2002). More generally, our definition suggests a decoupling of the terms *valid measurement* and *optimal measurement*, which are often seen as identical. In our definition, a valid test need not be the best available test, and not all valid tests are on equal footing. Two tests may both be valid, but one may be more reliable. One may consider the possibility that psychometric characteristics like reliability, unidimensionality, and measurement invariance do not provide necessary conditions for valid measurement, as is often thought, but rather presuppose validity as defined in this article. This would mean that validity is not in the same league as other test theoretic concepts. As has been suggested in this article, the reason may be that validity is not so much a methodological as a substantive problem.

The validity concept proposed here has been stripped of all excess baggage. The benefit is that this lends substantial clarity and force to the concept, but the price is that it covers less ground. For instance, when one claims validity, one is not thereby claiming reliability, predictive adequacy, or absence of bias. However, these are important properties, and one may still want to introduce a kind of umbrella term to express one's opinion on the overall quality of a testing procedure. We suggest that to express this judgment, *overall quality* is not such a bad choice. The accompanying plain English terms *better* and *worse* further allow one to express one's opinion that some tests are better than others in a given situation, without having the connotation of precision and objectivity that the jargon of *degrees of validity* suggests. Which testing procedure is best for a person depends on that person's situation and goals, on a very diverse set of test properties that are sometimes desirable and sometimes not, and on the amount of time and money at that person's disposal. But, if one wants to measure something, then one's test must be valid for that something—however suboptimal it may be with respect to properties like reliability. This distinction should not be blurred. Also, the usage of *better* and *worse* does not interfere with the concept of validity as defined here, and perhaps most important, such usage emphasizes that in saying "the WAIS is a good test," one is expressing one's enthusiasm about the WAIS rather than proposing a refutable hypothesis.

In conclusion, the present conception of validity is more powerful, simple, and effective than the consensus position in the validity literature. However, it does seem that it puts conventional thinking on test theory on its head, which necessitates further investigation of the way validity relates to psychometrics, philosophy of science, and substantive psychological theory. Also, the integration of psychological theory, test construction, and data analysis, advocated here as the best way to tackle the problem of validity, needs to be worked out in greater detail. It would be worthwhile to develop theories of response behavior in different domains, which could lead to a greater degree of integration between psychometrics and psychology than exists at present. Psychometric techniques and models have great potential for improving measurement practice in psychology, but only if they are driven by a substantive theory of response processes. We think that, with such theory in hand, the problem of validity will turn out to be less difficult than is commonly thought.

References

Bechtold, H. P. (1959). Construct validity: A critique. *American Psychologist, 14,* 619–629.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26,* 433–450.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Campbell, N. R. (1920). *Physics, the elements.* Cambridge, England: Cambridge University Press.

Cattell, R. B. (1946). *Description and measurement of personality.* New York: World Book Company.

Cervone, D. (1997). Social-cognitive mechanisms and personality coher-

ence: Self-knowledge, situational beliefs, and cross-situational coherence in perceived self-efficacy. *Psychological Science, 8,* 43–50.

Cronbach, L. J. (1980). Validity on parole: How can we go straight. In W. B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement over a decade.* (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Ellis, M. V., & Blustein, D. L. (1991). The unificationist view: A context for validity. *Journal of Counseling and Development, 69,* 561–563.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective.* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods, 3,* 300–396.

Frege, G. (1952). On sense and reference. In P. Geach & M. Black (Eds.), *Translations of the philosophical writings of Gottlob Frege* (pp. 42–55). Oxford, England: Blackwell. (Original work published 1892)

Glymour, C. (2001). *The mind's arrows.* Cambridge, MA: MIT Press.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427–439.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical tests of the rule assessment methodology by latent class analysis. *Developmental Review, 17,* 321–357.

Jansen, B. R. J., & Van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81,* 383–416.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Kagan, J. (1988). The meanings of personality predicates. *American Psychologist, 43,* 614–620.

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

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. I). New York: Academic Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89,* 575–588.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* New York: Cambridge University Press.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–215.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229–258.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, 99,* 166–180.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, England: Cambridge University Press.

Raven, J. C. (1938). *Progressive matrices.* London: H.K. Lewis.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183–198.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16,* 5–8.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46,* 1–74.

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). *The measurement versus prediction paradox in the application of planned missingness to psychological and educational tests.* Manuscript submitted for publication.

Suppe, F. (1977). *The structure of scientific theories.* Urbana: University of Illinois Press.

Süss, H., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence, 30,* 261–288.

Trout, J. D. (1999). Measurement. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 265–276). Oxford, England: Blackwell.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale.* New York: Psychological Corporation.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity. *Journal of Personality and Social Psychology, 84,* 608–618.

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. *Intelligence, 30,* 537–554.

Received September 11, 2002
Revision received November 11, 2003
Accepted November 13, 2003 ■