CHAPTER 7

THE END OF CONSTRUCT VALIDITY

Denny Borsboom, Angélique O. J. Cramer, Rogier A. Kievit, Annemarie Zand Scholten, and Sanja Franic

ABSTRACT

Construct validity theory holds that (a) validity is a property of test score interpretations in terms of constructs that (b) reflects the strength of the evidence for these interpretations. In this paper, we argue that this view has absurd consequences. For instance, following construct validity theory, test score interpretations that deny that anything is measured by a test may themselves have a high degree of construct validity. In addition, construct validity theory implies that now defunct test score interpretations, like those attached to phlogiston measures in the 17th century, 'were valid' at the time but 'became invalid' when the theory of phlogiston was refuted. We propose an alternative view that holds that (a) validity is a property of measurement instruments that (b) codes whether these instruments are sensitive to variation in a targeted attribute. This theory avoids the absurdities of construct validity theory, and is broadly consistent with the view, commonly held by working researchers and textbook writers but not construct validity theorists, that a test is valid if it measures what it should measure. Finally, we discuss some pressing problems in psychological measurement that are salient within our conceptualization, and argue that the time has come to face them.

The Concept of Validity, pages 135–170

INTRODUCTION

Construct validity theory, as Cronbach and Meehl (1955) introduced it, holds that a test score interpretation in terms of a nomological network (a set of laws relating theoretical terms in that network to each other and to observational terms) is valid to the degree that the network itself is supported by the evidence. This idea leaned heavily on the philosophical framework provided by logical positivism, which used the same construction as a general account of the relation between theoretical terms and observations in scientific theory. For the positivists, the nomological network served to endow the terms in it with meaning through so-called implicit definitions (Carnap, 1950). Roughly, the idea was that the meaning of a theoretical term (like 'mass' or 'force') was given, implicitly, by the laws in which such terms play a role. A replica of the positivist idea of fixing the meaning of theoretical terms through the use of implicit definitions occurs in Cronbach and Meehl's (1955) theory of construct validity: the meaning of terms such as 'general intelligence' is to be fixed by pointing at the laws in which these terms play a role, and the validity of interpretations of test scores in terms of such networks is subsequently determined by the degree to which the networks are corroborated by evidence.

At the time—which, methodologically, was thoroughly dominated by behaviorism and operationalism—this move created some leeway for tests that were to be interpreted in theoretical terms (like 'general intelligence' or 'neuroticism') but that had no direct characterization in terms of test content, nor a satisfactory criterion that could function as a 'gold standard.' Unfortunately, the idea of construct validity did not actually work, because there were (and are) no nomological networks involving concepts like general intelligence. Although various models and mechanisms have been suggested for such psychological attributes, these do not resemble the kind of strict laws that could function to build up a nomological network in the positivist sense (e.g., see Suppe, 1974).

To see this, it is useful to shortly discuss what kind of networks the positivists had in mind, because the term 'nomological network' has been used in psychology to indicate everything from a theoretical hunch to a regression model, but rarely to anything that the positivists would have recognized as a nomological network. In the positivists' account of scientific theories, one first requires a division of one's vocabulary into observation sentences ("John sits at home with a book on Saturday night") and theoretical sentences ("John has property i"). Second, observation sentences are to be connected to theoretical sentences by correspondence rules ("a person has i if and only if that person sits at home with a book on Saturday nights"). Third, the terms mentioned in the theoretical sentences are to be connected by laws, for instance $I = \mathrm{f}(e)$, where e is another theoretical term

hooked up to the observations through its own correspondence rules (e.g., "a person has *e* if and only if that person is at a party at least once a week"). The properties *i*, *e*, and others (say, *c*, *a*, and *o*), together with the laws that relate them to each other and the correspondence rules that relate them to observation sentences, then make up a nomological network. The theoretical terms in this network are taken to be symbols that are implicitly defined by their role in the network (i.e., their relation to observation sentences through correspondence rules and to other theoretical terms through scientific laws). Good overviews of this program can be found in Carnap (1950) and especially Suppe (1974).

This analytical scheme has never been successfully applied to any real science, and psychology is no exception to that rule. The reasons for this are many. First, it is not clear that the division between observational and theoretical vocabularies is tenable, because scientific observation usually presupposes theory and almost always requires some theory in the description of observations (this is the so-called problem of theory-ladenness). Second, psychology, like most other sciences, has no unambiguous connections between 'theoretical sentences' and 'observation sentences' in the form of correspondence rules. The representational theory of measurement, as proposed in Krantz et al. (1971), could have served to fill this gap if it were successful (Borsboom, 2005; Stegmüller, 1979), but so far it has not been able to play such a role (an issue we will come back to later; see also Batitsky, 1998; Cliff, 1992; Domotor & Batitsky, 2008). Third, psychology has few to no laws to connect the theoretical terms to each other, thereby limiting the prospects for nomological networks that are actually *nomological*. It sometimes appears that, in construct validity theory, the idea is entertained that a set of correlations or regression functions or loose verbal associations is sufficient to serve as a nomological network, but it is unclear how this could work because loose associations do not uniquely fix the meaning of the theoretical terms (e.g., see Borsboom, Mellenbergh, & Van Heerden, 2004).

Cronbach and Meehl (1955) were clearly aware of the fact that there was something problematic going on here, but deflected the problem by stating that psychology's nomological networks are 'vague' (pp. 293–294). This was not true at the time, and it is not true now. Psychology simply had *no* nomological networks of the sort positivism required in 1955, neither vague nor clear ones, just as it has none today. For this reason, the idea of construct validity was born dead. Contrary to what is widely believed, construct validity as proposed by Cronbach and Meehl (1955) never saw any research action. In accordance, there are few traces of the nomological network idea in current validity theory (Kane, 2006; Messick, 1989). However, it is an interesting historical fact that even though the core of their theory was defective from the outset, several more peripheral aspects of their theory actually did survive, and in fact, are largely constitutive of the construct

validity doctrine as it exists today. These residues of Cronbach and Meehl's vision of construct validity theory are (a) the idea that validity involves the interpretation of test scores; (b) the idea that, as a result, the property of validity is a property of propositions ('test score interpretations') rather than of tests; and (c) the idea that validity is a function of the evidence that can be brought to bear upon such propositions.

We will interpret current accounts of validity consistent with these tenets as instances of construct validity theory, because they have a common origin in Cronbach and Meehl's (1955) paper. Naturally, we recognize that there are significant differences between, say, the accounts of Messick (1989) and Kane (2006), and that not all of the relevant scholars hold the same view on, for instance, the necessity to invoke constructs in test interpretations (e.g., see Kane, 2006, who does not require this). However, in the present context these differences are best viewed as variations on a theme, especially when compared to the radically different view of validity that we propose (see also Borsboom et al., 2004).

The purpose of this chapter is to attack the central elements of the construct validity doctrine and, in doing so, the doctrine itself (for it consists of little more than the conjunction of the above ideas). In what follows, we argue that validity, as *normally understood*—that is, as it is understood by almost everybody except construct validity theorists—does patently *not* refer to a property of test score interpretations, but to a property of tests (namely, that these tests measure what they should measure). We will denote this property with the term *test validity* and take it to coincide with validity as defined and elaborated on in Borsboom et al. (2004). Second, we argue that test validity—in contradistinction to the notion of construct validity—is not a function of evidence, but a function of truth. Third, we argue that to assess test validity, one has to adopt a realist approach to measurement, because one needs to fill in the semantics of measurement for this purpose, and we know of no successful alternatives to realism in this respect.

In addition, we argue that the notion of a 'construct', as used in construct validity theory, functions in two ways that are mutually inconsistent; namely, it is used both to refer to the theoretical term used in a theory (i.e., a symbol), and to designate the (possible) referent of the term (i.e., the phenomenon that is targeted by a researcher who uses a measurement instrument). This double usage has created an enormous amount of confusion, but has become so entrenched in both construct validity theory and methodological language that it has by now become too late to change it; therefore, we propose to do away with the term 'construct' altogether. Instead, we propose to use 'theoretical term' to designate a theoretical term, and 'psychological attribute' to designate the psychological attribute, if there is any, which that theoretical term refers to. Also, we argue that the notion of validity, as normally understood, is both theoretically and practi-

cally superior to the notion of construct validity. Finally, we contend that the construct validity doctrine keeps researchers hidden behind smoke and mirrors, safe from some real problems of psychological measurement we should all deeply care about. Therefore, it fulfills a dubious function in current methodology, because it detracts from the research questions that should be pursued if we are to make any real progress in solving the problem of validity.

VALIDITY AND TEST SCORE INTERPRETATIONS

According to construct validity theory, one cannot obtain evidence for tests, only for propositions. If these propositions involve test scores that are to be interpreted as measures of a psychological attribute, and construct validity is considered to be a function of the evidence for a nomological network involving these test scores and attributes, then it follows immediately that construct validity is also a property of test score interpretations rather than of tests (Cronbach & Meehl, 1955; Messick, 1989). Thus, according to construct validity theory, validity does not refer to the question whether, say, IQ-tests really measure intelligence, but only to the question how well certain IQ-score interpretations are backed by the evidence.

Although this is an integral element of construct validity theory as set up by Cronbach and Meehl (1955) and followed up by writers such as Messick (1989) and Kane (2001, 2006), it clashes with both the ordinary meaning of test validity—whether a test measures what it purports to measure—and with common sense. The reason for this is that the notion of a test score interpretation is too general. It applies to every possible inference concerning test scores—even inferences that have nothing to do with measurement, or that in fact deny that anything is being measured at all.

For instance, suppose we administer the following test, which we will designate as Test X: throw a coin ten times and count how often it falls heads. We conceive of each of the individual trials as an item, and of the number of successes (heads) as the total test score on Test X. Imagine we administer Test X to a sample of people and record the resulting scores. Our interpretation of the test scores is as follows: 'The scores on Test X measure nothing at all'.

To evaluate the construct validity of this test score interpretation, we need to look at the evidence for it. The test score interpretation makes quite a strong, empirically informative claim. For instance, from the proposition considered, one easily derives the following hypotheses: 'Scores on Test X will not show substantive correlations with extra version, intelligence, or attitude tests'; 'The distribution of scores on Test X do not vary across sex, age, or educational level'; 'Scores on Test X are not sensitive to experimen-

tal manipulations involving stereotype threat'; 'Scores on Test X do not suffer from social desirability effects', etc. Note that the list can be lengthened indefinitely; hence, the logical content of the test score interpretation, as for instance Popper (1959) would view it (i.e., the totality of empirical results that the interpretation rules out) is truly enormous, and the potential for falsification is equally impressive. One could venture to gather support for the aforementioned hypotheses, but obviously it is in everybody's interest to refrain from such an undertaking; for we may safely assume that the evidence for the proposed test score interpretation will be overwhelming.

Because construct validity is a function of the evidence for a test score interpretation, we are forced to conclude that the proposed test score interpretation 'scores on Test X measure nothing at all' has an extremely high degree of validity. We therefore establish the following result: *Construct validity applies to test score interpretations that deny that the test in question measures anything whatsoever just as easily as it does to test score interpretations that do claim that something is measured.*

Thus, it is entirely consistent to state that a proposition that *denies* validity as normally understood (i.e., as the claim that the test under consideration measures a psychological attribute) *itself* has high construct validity. This straightforward consequence of the way construct validity theory is set up establishes that construct validity does not cover test validity as normally understood; it is rather entirely orthogonal to it. For many involved in psychometric research, this may be an unanticipated consequence. At the very least, it shows that the question of whether a test measures what it should measure is not necessarily or specifically covered by construct validity theory. If we want a validity theory that addresses our ordinary conception of validity, and thus speaks to the question whether the test measures what it should measure, then validity should not be conceptualized as a property of test score interpretations generally; hence, construct validity theory has to be added to, specified in greater detail, or replaced by something.

EVIDENCE AND TRUTH

One very significant problem that is raised for construct validity theory by the previous example is this: If, as would appear plausible, we would want to limit the notion of validity to 'positive' test score interpretations (i.e., interpretations that *do* state that something is being measured) we are obliged to fill in *what would make such interpretations true.* And in answering this question, it is a serious mistake to answer: *the evidence does.* For propositions are not made true by the evidence that exists for them, but by their conformity to the state of affairs in the world, as is evident from the fact that one can have massive amounts of evidence for a false proposition (e.g., 'time and

space are absolute' or 'energy is continuous' before the early 20th century) and lack any evidence for a true one (e.g., 'time and space are relative' or 'energy is discrete' before the early 20th century).

Construct validity theory, however, has tended to specifically define validity in terms of evidence (Cronbach & Meehl, 1955; Messick, 1989; Kane, 2001). One might therefore be inclined to simply add the requirement of truth to the requirement of evidence. However, if one wants to uphold that validity of the test score interpretation 'IQ-test scores are measures of general intelligence' depends on whether that proposition is true *and* that the validity of the interpretation is a function of the evidence, one encounters serious problems.

To elucidate how this happens, consider the example of phlogiston measurement, as it existed in the 17th and 18th centuries. The theory of phlogiston, proposed by Becher in 1667, held that substances contained a certain amount of phlogiston ('fire-stuff'), which they emitted when burned (Bowler & Morus, 2005). The amount of phlogiston a given material possessed was measured indirectly, by subtracting its weight after burning from its original weight. The phlogiston theory, in its day, was used to devise explanations of various phenomena, such as the fact that some materials burn better than others (they contain more phlogiston), and that burning naturally stops when the burning material is placed in a sealed container (the air in the container has only limited capacity to absorb the phlogiston emitted from the material).

Phlogiston measurement took place against the backdrop of a theory that, compared to most psychological theories currently in existence, was very well worked out. Phlogiston figured in explanatory systems that were at least as aptly described as 'nomological networks' as most current psychological theories, made successful predictions, and therefore was quite strongly supported by 'theoretical rationales' and 'empirical evidence'. (We think that the evidence for phlogiston theory was stronger than *any* evidence for *any* current psychological theory we know—but one need not agree on this to see the force of the point we are about to make.) Alas for phlogiston theory, some materials did not lose, or even gained weight when burned, which conflicted with the prediction of the theory. Although some attempted to save the theory by invoking negative amounts of phlogiston, the curtain fell in 1775, when Lavoisier presented evidence to the French Academy of Sciences to show that burning is a reaction with oxygen, and can be explained without making reference to phlogiston.

Now consider the current doctrine of construct validity theory, which says that the construct validity of a test score interpretation depends on the evidence and theory supporting that interpretation. Let us consider the interpretation 'the difference between the weight of a substance before and after burning is a measure of the amount of phlogiston it contains'. Surely,

the evidence for this interpretation was quite strong before the late 18th century. Therefore, we must conclude that it had a high (in comparison to most cases of psychological measurement, overwhelmingly high) degree of construct validity at that time. In 1775, however, the construct validity of the interpretation sharply dropped when Lavoisier presented his results, and now, at the beginning of the 21st century, its degree of construct validity is zero (or, for those who think that validity is never an all-or-none issue, *close* to zero). Figure 7.1 provides a graphical display of the construct validity of the test score interpretation in terms of phlogiston.

Now, it seems to us (and we assume you will agree) that the weight measures referred to in the figure *never* measured phlogiston, neither in 1670 nor in 1830 nor today. We therefore feel a compelling urge, and we hope you do as well, to accept the conclusion that the test *never* had *any* validity for measuring phlogiston. In fact, it seems to us that, in the graph, the validity of the test score interpretation should be considered to be described by a flat line corresponding to the function f(validity|time) = 0. However, it is clear that current construct validity cannot agree with this conclusion without throwing one of its main tenets—construct validity reflects the strength of the evidence—out of the window.

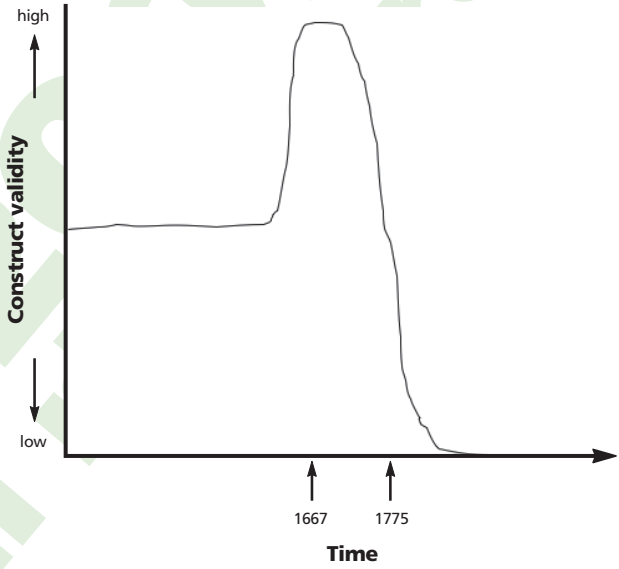

Figure 7.1 Construct validity of phlogiston measures as a function of time. As the figure shows, construct validity was initially undecided, there being no evidence for or against it. A marked increase is visible soon after 1667, when Becher first proposed the theory and evidence started to mount, but validity decreases to nearly zero after Lavoisier published his refutation of the phlogiston theory in 1775.

As a second example, consider the geocentric theory that held that Earth is at the center of the universe, and that the planets and the sun rotated around it, constrained by a system of perfect spheres (Barker & Goldstein, 1992). This theory, at its high time, could accommodate impressively for the observed behavior of the planets. For instance, Gearhart (1985) showed that the predictions of the geocentric system fell within the margins of error of the available data. One of the reasons for this success was that, whenever the empirical discrepancies between observations and predictions became too great, an increasingly elaborate system of epicycles, deferents, equants and eccentrics was proposed to accommodate the anomalous behaviors of certain planets. These ad hoc adjustments were accurate enough that the Ptolemaic system remained widely accepted until well into the Middle Ages, when Copernicus proposed that the Sun might be the center of the universe. The transition was far from complete though; Copernicus still adhered to circles and epicycles, and more important, his view could still not outperform the geocentric view in terms of empirical accuracy (Gearhart, 1985). It was not until Johannes Kepler discarded all received notions and proposed elliptical orbits of the planets around the Sun that we saw the first contemporary model of the Solar System.

Now suppose that a geocentric scholar had used astronomical data to measure the time it takes for Mars to complete its epicycle, while a heliocentric scholar used the same data to measure the time it takes Mars to complete its orbit around the Sun. Each scholar interprets the data in terms of a theory; each scholar's interpretation therefore has a degree of validity. So, whose is higher? As in the case of phlogiston, the answer to this question appears to depend on the time at which it is evaluated—before, during, or after the revolution that Copernicus instigated. According to construct validity theory, one is bound to say that the construct validity of geocentric interpretations of the data was higher than these of heliocentric interpretations at least up to Copernicus, and possibly up to Kepler.

One may object to this that the geocentric theory was grossly overparameterized and therefore scored very badly on the criterion of parsimony at the outset. However, this introduces the delicate question of how to weigh parsimony and empirical adequacy in the evaluation of scientific theories. We do not know what construct validity theory has to say on such matters, but submit that instead of bailing one's way out through the invocation of methodological criteria like parsimony and empirical accuracy, there is a much easier way out of this problem—namely, by acknowledging that interpretations of the data in terms of geocentric theories were simply never valid, so that there was never a valid measure of the time it took Mars to complete its epicycle.

We take these examples to be a *reductio ad absurdum* of the dual thesis that construct validity *both* refers to the truth-value of a proposition, say,

'weight differences measure phlogiston' *and* is a function of the evidence. In our view, one of the two theses has to go. Thus, one must either accept that validity is not a function of evidence but of truth, *or* one must accept the thesis that the state of affairs in the world, insofar as it does not show up in the evidence, is *irrelevant to the determination of construct validity*. There is ample documentation supporting the claim that construct validity theorists explicitly favor an interpretation of validity in terms of evidence (e.g., Cronbach & Meehl, 1955; Kane, 2006 Messick, 1989), thus taking the latter route. This implies that they should take seriously the graph in Figure 7.1 and accept the conclusion that logically follows from it: *The interpretation of weight differences as measures of phlogiston had a high degree of construct validity between 1667 and 1775.*

REFERENCE AND REALITY

The above arguments aim to establish that (a) the idea that validity refers to interpretations of test scores is in need of qualification (namely, we need to consider certain *kinds* of interpretations, not just *any* interpretation is eligible) and (b) the idea that validity is a function of evidence is problematic, if one wants to be able to say such commonplace things as 'test score interpretations in terms of phlogiston never were valid, even though people thought they were'. Construct validity theory is thus underspecified (i.e., too general) and focuses on the wrong thing (i.e., evidence). Consequently it is unable to incorporate test validity as normally understood; the fact that there is evidence for some test score interpretation does not entail that the test in question actually measures the targeted attribute. Thus, the questions 'does the test measure what it should measure?' and 'how much evidence is there for this or that test score interpretation?' are different questions that are conflated in the construct validity literature.

We do not entirely see how this situation came about, but we suspect that it has something to do with the positivist heritage which, in its attempt to exorcise metaphysics from the scientific world view, tried to evade referential connections between theoretical terms ('general intelligence') and the structures that such terms refer to (a single linearly ordered property, if there is one, that causes individual differences on IQ-tests). The positivist program attempted to make sense of measurement without incorporating realist commitments (i.e., metaphysical ones) about the properties targeted by the measurement procedure. So, what the positivists tried to do is to make sense of statements like 'this test measures general intelligence' *without* engaging in the commitment that there actually is *something out there* to measure, i.e., without assuming that such a thing as general intelligence exists independently of the researcher's scaling activities. If successful, this

would allow one to craft a validity concept that does not postulate, at the outset, that the existence and causal efficacy of X are required, if X is to be considered a property that can be measured. Such an approach, however, does not work.

Basically, the problem is that, to evade realism, one has to twist the natural interpretation of the word 'measurement' (which is that one has an instrument that is sensitive to differences between objects with respect to some property, in the sense that it gives different outcomes for different instances of the property) to such a degree that (a) the term no longer means what scientists routinely take it to mean, or (b) the assumptions required for the interpretation of measured properties as constructions are so strong as to imply that measurement is impossible. Two corresponding twists that have been worked out in detail are operationalism and representational measurement theory.

Operationalism famously holds that scientific concepts are synonymous with the procedures used to measure them (Bridgman, 1927), which does not necessarily imply that the properties measured do not exist, but is clearly compatible with that thesis. This doctrine is defective for two reasons. First, because it is incoherent (a linguistic concept such as 'general intelligence' cannot stand in a relation of synonymy to a set of actions, for instance a researcher administering an IQ-test). Second, because it implies that no two measurement procedures could measure the same thing (i.e., a mercury thermometer cannot measure the same property as an electronic thermometer because each defines its own concept), and therefore flies in the face of scientific practice. Thus operationalism twists the meaning of the word 'measurement' to be completely at odds with the way scientists work.

The other alternative, representationalism, holds that measurement consists of the assignments of numerals to objects or events in such a way that the numerical relations between these numerical assignments are isomorphic (i.e., structurally equivalent) to the empirical relations between the objects and events in question (Narens & Luce, 1986). This theory was developed in great detail in the three-volume work *Foundations of Measurement* (Krantz et al., 1971; Luce et al., 1990; Suppes et al., 1989). Representationalism essentially consists of two steps.

First, in the representational step, an empirical relational structure is established by determining the set of objects and the *empirical* relations between objects on some property of interest (e.g., the ordering of minerals according to their hardness). It is important that these relations are qualitative and need no reference to numbers. Also, the relations must be detectable in a straightforward manner by the researcher. If an empirical relational structure conforming to these requirements can be established, then it needs to be proven that a numerical relational structure exists, consisting of a set of numbers and *numerical* relations, that perfectly represents

the empirical relational structure. In the aforementioned *Foundations of Measurement* such proof is provided in the form of axioms that are used to prove representation theorems for many different types of empirical relational structures.

The second step involves showing how the specific numerical scale that was chosen, relates to other possible instantiations (other scales) of the numerical relational structure. The functional relation that relates all possible numerical scales to each other is described in a uniqueness theorem and determines the level of measurement. For example, scales that are isomorphic representations of the hardness of minerals are unique up to monotonically increasing transformations; because minerals can only be ordered according to their hardness, any order-preserving transformation of a 'correct' numerical scale will do equally well in reflecting these empirical order relations.

Representationalism can be used, in principle, to evade realist metaphysics. The way in which this could be done is to establish qualitatively which relations hold between objects (in psychology, these are normally people or items, resulting in relations such as 'John is less intelligent than Jane'), to code these in a matrix, and to find a way to attach a number to each entry in the matrix so that all the empirical relations coded in it are preserved by the numerical relations (say, John's intelligence is represented by the number 95 and Jane's intelligence is represented by the number 120). If this is possible, one can go on to establish the uniqueness of the properties of the representation, and define the measurement level associated with the measurement procedure. One could then use the term 'general intelligence' to refer to the constructed scale, which, evidently, is a human-made construction that one need not necessarily assume exists independently of the researcher or is causally responsible for the observed relations between measured objects (Borsboom, 2005).

Of course representational measurement does not exclude the possibility that the measured property is causally responsible for the variation in measurement outcomes, but it certainly does not require it. Thus, it would seem to provide an excellent way to circumvent realist commitments. But even though the theoretical quality of this work is not in doubt, it is questionable whether realism can be evaded in practice (Batitsky, 1998; Borsboom, 2005; Domotor & Batitsky, 2008). The main problems with representationalism, if it should serve as a way of exorcizing metaphysics, are the following.

First, the empirical relations should be observed without inconsistency and second, the empirical relations between objects as coded in the matrix should be 'observable with the naked eye' (Batitsky, 1998; Domotor & Batitsky, 2008; Van Fraassen, 1980) or something similar. Unfortunately it remains unclear which exact types of observation satisfy this requirement of representational measurement theory and what types of observational

aids would be allowed, if any. We have no choice therefore, but to take the requirement literally. Now the naked eye is an untrustworthy source of information, in the sense that it is unreliable in certain situations; specifically, it is easy to find a level of grain size on which the eye starts behaving erratically, sometimes saying that, for example, A is larger than B and sometimes that B is larger than A (or A is larger than B, B is larger than C, but C is larger than A). If one demands that no reference to the structure of the world is made except to what we can see with the unaided eye, then it directly follows that measurement in the representationalist sense is highly problematic, for what one will get are inconsistent systems of empirical relations that do not allow for the intended isomorphic representations in a numerical system. Hence, we are forced to conclude that measurement is impossible (see also Batitsky, 1998, for an excellent, and more extensive, exposition of such difficulties). This, obviously, flies in the face of scientific practice.

Even if observation is aided (e.g., by microscopes or amplifiers), the empirical relational structure will exhibit some degree of inconsistency. Normally, of course, the will-o-the-wisp behavior of the eye, aided or unaided, is not interpreted to mean that the relation of 'longer than' between A and B shifts randomly from one moment to the next (although, of course, it may in some contexts). Rather, such inconsistencies are routinely interpreted as measurement error. Thus, what one will do, for instance, is introduce the idea that relations between objects are *imperfectly* picked up by the measurement procedure (e.g., the use of the naked eye) and deal with these imperfections by using some statistical theory.

These statistical measurement models can be less or more elaborate in the explicitness, testability and type of measurement assumptions. Some models allow one to test measurement assumptions, be it indirectly; this is, for instance, the case for the Rasch (1960) model. This model is structurally similar to a subtype of representational measurement theory (additive conjoint measurement theory) and therefore it indirectly ensures that measurement assumptions are met, at least, if the model is true. Statistical theory can also be used to form probabilistic versions of representational measurement theory (Karabatsos, 2001). This results in models that capture measurement error but still allow for a relatively direct test of the axioms of representational measurement theory.

However, as soon as we employ any of these statistical measurement models, we immediately fail at our attempt to get rid of metaphysics. For it is exceedingly difficult to make sense of the idea that there is error in the measurements, if one is not allowed to make reference to a true value of these measurements (or true relations that exist between objects). Introducing true values means that one needs something that makes these values true independently of the researcher doing the measurement work. The

prime candidate for being the truth maker here, naturally, is the property targeted by the measurement procedure, and it appears hard (at least, we have not seen successful attempts) to find a plausible alternative candidate for the job at hand. That is the reason why representationalism does not buy one a way out of metaphysics.

Hence, operationalism is inconsistent and representationalism is too strong. There are, as far as we know, no other reasonable candidates to take care of the semantics of measurement—except for realism. Realism, in the context of measurement, simply says that a measurement instrument for an attribute has the property that it is *sensitive* to differences in the attribute; that is, when the attribute differs over objects then the measurement procedure gives a different outcome. This implies that there must be a causal chain that describes the working of the measurement procedure, in which the measured attribute plays a role in determining what the outcomes of the measurement procedure are.

So, a pan balance measures weight because the weight difference between two objects, one placed in each of the pans, determines to which side the balance will tilt. (If gravity is constant, as it is on earth, then the decisive factor would be mass, and we can say that the procedure measures differences in mass). This, it appears to us, is a sensible way to construct the idea of measurement. However, this is not a bargain, as the price paid for these semantics is that one has to make reference to the property measured as a causal force that steers the direction of the measurement outcomes. This is no small matter. It requires a very strong assumption about what the world is like, namely that it contains some property that exist independently of the researcher measuring it. This assumption may be much too strong for many psychological properties. It also obliges the researcher to explicate what the property's structure or underlying process is and how this structure or process influences the measurement instrument to result in variations in the measurement outcomes. This seems to be a very daunting task indeed for many psychological properties that researchers claim to measure.

How does a realist approach to measurement relate to the concept of validity? Well, usually one has the idea that there is some property that determines differences within or between individuals, and one attempts to create an instrument (e.g., an item, test, or observational procedure) that will do one thing if the targeted property has a certain value and will do another if it doesn't. Then one applies the instrument and gets data representing the different outcomes of the measurement procedure. Obviously, these differences have to come from somewhere, i.e., there is a causal antecedent process in which something makes a difference to the outcomes. The hypothesis involved in the question of test validity is that the term one uses to name the property in question (say 'general intelligence') *refers* to the property that causes the differences in measurement outcomes. Or, in an alternative

formulation, the term 'general intelligence' *co-refers* with the description 'the property that causes differences in the measurement outcomes'. Thus, what is at stake in posing the question of validity is the empirical hypothesis that the description 'what is measured by the test' and the term 'general intelligence' designate the same property.

Construct validity theory does not address this type of measurement issue at all. It is, for instance, hard to find a definition of the word 'measurement' in papers on construct validity (for instance, try Messick, 1989, or Kane, 2006). This is remarkable since validity at its base, is a characteristic of measurement, regardless of whether one views validity as a property of test score interpretations or as a property of tests. A clear definition of measurement would seem essential for any hope of a coherent theory of validity. In our view, a realist approach to measurement is the only tenable one. This poses problems for construct validity however, since it must restrict its scope considerably, in order to reconcile itself with this approach. Construct validity can only account for the realist measurement approach by limiting the allowable propositions to one very special test score interpretation, namely that the test scores can be interpreted as measures of the targeted attribute simply because the targeted attribute is causally relevant to the test scores.

Two things follow from this. First, for the proposition 'the test measures what it should measure' to have any truth conditions at all, one needs to fill in the notion of measurement as we have expounded above. Second, what makes the 'test score interpretation' above true is a *property of the test*, namely that it has what it takes to be used *as* a measurement instrument for the targeted attribute. We submit, therefore, that what really matters in validity is *how the test works*, and this is certainly *not* a property of test score interpretations, or even of test scores, but of *the measurement instrument itself* (i.e., of the concrete, physical thing that you can drop on your feet, rather than of a linguistic entity, set-theoretical object, or statistical construction). In fact, we think that all the talk about 'test score interpretations' has not led construct validity theory to a greater level of sophistication, as is commonly assumed, but in point of fact has served to detract the theory from the main issue in test validity and the proper objective of validation research, which is *showing that the test indeed has the relevant capacity to pick up variation in the targeted attribute and that, in point of fact, it actually does so in the typical research settings in which the test is used.*

CONSTRUCTS?

So what about constructs? How should *they* be taken to serve the function of measurement? Are they made of the right stuff? Could they possibly determine the outcome of the measurement process? Or are they post hoc

inventions, figments of the scientist's imagination? Can constructs, in fact, *be measured*?

We do not know the answer to these questions because we do not know what constructs are, that is, we have rarely come across a clear description of what something should be like in order to deserve the label 'construct'. Constructs, as far as we are concerned, are truly shrouded in mystery, and not in the good old scientific sense that we currently don't know what they are, but will know when we're finished doing the relevant research, but in the sense that we don't really know what we are talking about in the first place.

The main problem, as we take it, is this: Construct validity theorists have the habit of using one word for two things, without clearly indicating when they mean what. In particular, the term 'construct' is used to refer to (a) a theoretical term (i.e., the linguistic, conceptual, symbolic entity) that we use as a placeholder in our theories ('general intelligence', '*g*', 'theta', 'the factor at the apex of this hierarchical factor model', etc.), and (b) the property that we think plays a role in psychological reality and of which we would like to obtain measures (i.e., a linearly ordered property that causes the positive correlations between IQ-tests—assuming, of course, that there is such a property).

Now, one cannot measure constructs in sense (a) above. That will not work, no matter how good one's measurement instruments become or how much one learns about the research domain of interest. The reason for this is not that constructs, interpreted in this particular sense, are 'latent' or 'unobservable' or 'vague' or 'complex'. The reason is that trying to measure a construct in sense (a) is very much akin to trying to climb the word 'tree'. That is, one is mistaking a symbol for the thing it refers to (see also Maraun & Peters, 2005). Constructs in sense (a) are purely theoretical terms, symbols concocted by the researcher for the purpose of scientific communication; and these symbols are causally impotent in the measurement process.

One *can* measure constructs in sense (b), that is, the properties our words refer to, but obviously this will only work if these properties are capable of doing causal work. That is, there has to be some property (structure, attribute, entity, trait, process) that listens to its name (the theoretical term) and that actually *does* steer the measurement outcomes in one or the other direction because, whatever else one thinks symbols are good for, they are not up to that particular job.

Loevinger (1957) saw the importance of this issue clearly, when she addressed the semantics of the word 'construct':

> *construct* connotes construction and artifice; yet what is at issue is validity with respect to exactly what the psychologist does not construct: the validity of the test as a measure of traits which exist prior to and independently of the psy-

> chologist's act of measuring. It is true that psychologists never know traits directly but only through the glass of their constructs, but the data to be judged are manifestations of traits, not manifestations of constructs. (Loevinger, 1957, p. 642, italics in original)

Now, if such a property ('trait' in Loevinger's terms) does not exist, then one cannot measure it, however hard one tries. Of course, one still has the theoretical term (all one has to do to bring that into existence is write it down) but one is misdirected if one tries to interpret test scores as measures of the symbolic entity that figures in scientific communication. (Unless, perhaps, if one comes up with a good answer to the question how one measures properties that do not exist, i.e., gives an empiricist interpretation of the measurement process that does not fly in the face of scientific practice or is so strong as to preclude the very possibility of measurement. This project is, to the best of our knowledge, still outstanding.).

Coming back to the issue of phlogiston measurement, we may plausibly conclude that phlogiston did very well as a construct in sense (a)—better, in fact, than most psychological constructs—but not at all in sense (b). The same holds for aether, elán vital, absolute space, and many other obsolete concepts that have been proposed in the history of science. We are currently in a state of ignorance regarding most of the attributes proposed in psychological theorizing, but we suppose that one who hopes to measure such attributes surely does not want them to fall in this particular category of concepts.

Thus, what one needs in measuring a psychological attribute is not just a fitting statistical model, or a theory that offers an explanation of why that model should hold, or a corroborated nomological network; one needs a referential connection between one's construct in sense (a) and one's construct in sense (b). That is, it is important that one's theoretical term designates a property that is sufficiently structured to perform causal work in the measurement process. We think that much of the confusion surrounding the notion of a construct stems from the fact that the term 'construct' is used both to indicate a theoretical term and a property measured.

As long as one clearly recognizes which of the two meanings is intended, there is of course little to worry about. Polysemy is a natural feature of language, and we suppose that physicists have no problem deciding when they take, say, 'charm' to refer to a property of subatomic particles, or to the new secretary. However, in construct validity theory one sees properties that can be attributed to only one of the two types of constructs, being transported to the other one. So, people say that it is important to *find out* what the *meaning* of a construct or test score is. This is clearly confused; one can *find out* things about the property measured—i.e., the sense (b) construct—but it is hard to see why one should do empirical research to find out things about

the symbol that purportedly designates that property—i.e., the sense (a) construct. In contrast, the utilized symbol—the sense (a) construct—clearly has *meaning*, and one could say sensible things about that; but the property referred to—the sense (b) construct—has no more meaning than the tree in your back yard. Similarly, one can *measure* the sense (b) construct, but not the sense (a) construct; one can *rule out alternative explanations* to the hypothesis that the sense (b) construct causes the correlations between observables, but not to the hypothesis that the sense (a) construct does so. One can look at causal effects of some variable on the sense (b) construct, but not on the sense (a) construct. The sense (a) construct may be 'implicitly defined' through a theory, but not the sense (b) construct, because that is a phenomenon in reality and not a theoretical term. And so on.

We hope that the reader is as confused by the use of the terminology of 'sense (a)' and 'sense (b)' constructs as we are. It is clearly a bad idea to utilize such terminology, and we would strongly advise against it. Moreover, the word 'construct' is so thoroughly infected with both meanings that we see no possibility to restrict is usage to indicate either sense (a) or sense (b). The only viable option, we think, is to dispose of talk about constructs altogether and explicitly refer to 'theoretical terms' for sense (a) constructs and 'psychological attributes' for sense (b) constructs. In essence, this means that the theory and vocabulary of construct validity is abandoned entirely.

INTERMEZZO

We could lengthen our critique of construct validity theory almost indefinitely by playing variations of the above themes. However, we prefer to lay the issue to rest here. Those who, at this point, remain unconvinced of the inadequacy of construct validity theory are unlikely to be swayed by further argumentation and are probably beyond salvation.

In the next section, we will therefore turn to some positive remarks regarding the possibilities and challenges provided by a notion of validity that has shaken off the blurry visions of construct validity. As an intermediate conclusion that serves to substantiate these, however, we suggest that the following theses are securely established by the above discussion:

1. Construct validity is about test score interpretations, not about tests. However, regardless of whether one thinks that we need a concept like construct validity to indicate the quality of the evidence for a given test score interpretation, there is a separate issue, important in its own right, that is not specifically covered by construct validity theory, namely whether a *test* is valid for *measuring* an attribute.

2. To say anything sensible about this property, one has to establish the semantics of the word *measurement*. In our view, the only viable candidate for such semantics is realism; that is, the notion of measurement presupposes that there exists some sort of structure that the test is sensitive to, in the sense that the test elicits processes that result in different measurement outcomes dependent on the position of the object or person, subjected to the measurement procedure, on the attribute in question.
3. The question of test validity involves the issue of whether a psychological attribute (e.g., 'general intelligence') exists and coincides with the attribute that the test in fact measures (assuming that there is one). In this case, for instance, the terms 'general intelligence' and 'the attribute measured by the test' co-refer to the same structure. Validating a test is just doing research to figure out whether this is true or not.
4. Whether a test actually has what it takes to serve as a measurement device for the attribute targeted by the researcher is independent of the evidence for test validity. Thus, we may have a valid test without knowing it, and we may have good evidence for validity even though the test in question does not measure the attribute for which it was designed. This is because validity is a function of facts, not of opinions, and our evaluations of the evidence may be mistaken.
5. As a result, construct validity may be high, whereas test validity (as we defined it) is absent, and vice versa. In general, the relation between construct validity and test validity is contingent on the substantive situation examined. Contrary to what is widely thought, one cannot expect construct validity to generally coincide with or imply test validity.

In our opinion, test validity is what researchers are primarily interested in when they talk about validity. Construct validity is relevant only insofar as it concerns the evidential backup of one particular test score interpretation, namely the one that corresponds to test validity: that the test scores can be interpreted as measures of the targeted attribute because the test is sensitive to differences in that attribute.

THE REAL PROBLEMS OF PSYCHOLOGICAL MEASUREMENT

The question of test validity, as conceptualized here, raises many interesting issues about the way tests work that, in our view, receive too little attention in scientific psychology. In this section, we outline some examples of impor-

tant questions that are rarely addressed but, in our view, stand in need of investigation if we are to make serious progress in the realm of psychological measurement.

Reflective measurement models. In the literature on test theory, almost all models that have been considered are variants of what has been called the reflective measurement model (Edwards & Bagozzi, 2000) or the effect indicators model (Bollen & Lennox, 1991). Such models assume that the indicators (item or test scores) depend on a (set of) latent variables through some functional relationship between parameters of the observed score distribution and the position of people and items in the latent space. In the majority of cases (Borsboom, 2008), these models are formally indistinguishable from common cause models; specifically, they require that the latent variable screens off correlations between the observables (Pearl, 2000), a requirement known as local independence in the psychometric literature. Thus, such models assume that there is an underlying variable that affects each of the observables, thereby explaining the correlations between them.

The idea that a set of items depends on an underlying variable appears to be the general motivation for treating the item or test scores as measurements in the first place. This is understandable, as it is hard to see how such scores could be sensibly interpreted as measurements if they did not depend on the same latent variable. Naturally, one can see a latent variable model as a purely heuristic device used to organize the data, or to scale the test scores in some pragmatically useful way, or to reason about one's data; but equally naturally, the fact that a latent variable model can be used in such a manner does not automatically imbue the item scores with the status of measurements; for one can *always* use models pragmatically, regardless of how the data arise. To sensibly interpret the item or test scores as *measurements*, and the instrument that yields them as a *measurement instrument*, the model should not just be useful, but true. That is, it should actually be the case that differences in the item scores depend on the targeted attribute, represented in the model as a latent variable, so that sensibly constructed functions of these item scores (like sum scores or more complicated estimators) can be interpreted as measures of the attribute in question.

In the psychometric developments over the past century or so, the specification and elaboration of such models have come to be viewed as 'technical' issues, to be handled by specialists who are better at statistics than at substantive psychology. This is not always equally productive because, even though the usefulness of latent variable models for scaling and such stands beyond doubt, their importance for the study of test validity is truly monumental and therefore deserves serious attention from substantively interested scholars. The reason for this is that such models code a necessary condition for interpreting the test scores as measures in the first place,

namely, that the item scores measure the same attribute. The fact that latent variable models are underused in the social sciences (Borsboom, 2006) and are hardly ever even mentioned in treatises on construct validity indicates, in our view, that few researchers realize what their importance for validity really is.

In our view, in fact, establishing the *truth* of such a model would *clinch* the question of validity. This may raise some eyebrows between psychometricians and construct validity theorists, and understandably so. Current psychometric dogma has it that the best one can ever do is to ascertain the goodness of fit of a latent variable model against a given dataset, and that such goodness of fit only indicates that *a* latent variable may be responsible for the covariation among observables, but not *which* latent variable this may be. This is, however, only the case if one detaches the measurement model from substantive theory, i.e., if one views a latent variable as an anonymous technicality called 'theta'.

If one does not detach the substantive theory from the measurement problem, then clearly one *can* do better than inspecting goodness of fit indices and eyeballing residual plots, namely by investigating the processes that lead to item responses or test scores and *identifying what element of these processes is causing variation in each of the different items at the same time* (see Borsboom & Mellenbergh, 2007, for some ways in which this may be the case). If successful, such an investigation would clearly end up with a *substantive* specification of the common cause that underlies the item responses, not a purely technical one, and with a *substantive* justification for the structure of the latent space and the form of the item response functions. If successful, such a research program therefore *solves* the problem of test validity, because it by necessity becomes clear *what the items measure* and *how they measure it*. And that is all there is to know regarding validity.

Item response processes versus external correlations. The suppositions that (a) a reflective model is indeed required to sensibly speak of measurement, and (b) that such a model should be true rather than pragmatically useful or consistent with a particular dataset, raise the question of how such a model *could* be true. That is, if test validity concerns the sensitivity of a test to differences in the targeted attribute, then the question that immediately presents itself is how differences in this attribute are causally transmitted into the measurement outcomes. Questions concerning the actual causal relation or process responsible for variation in the measurement outcomes, however, are seldom asked in psychometrics or psychology at large. Perhaps this is because, in many cases, they present extraordinarily difficult problems for the conviction that we are actually engaged in a measurement process when doing our test theoretic work.

For instance, consider the individual differences literature, where factor analysis and its relatives are commonly used to investigate psychological

properties having to do with personality, attitudes and ability. The construct validity doctrine has set the stage for a longstanding and ongoing quest for 'constructs' that are 'measured' by a series of observable 'indicators' in this field. A plethora of psychological constructs have emerged from these endeavors.

In such cases, on typically, one has a theoretical term, say, 'general intelligence' and a set of indicators, say items of the WAIS (Psychological Corporation, 1997), that supposedly measure the property that one thinks the theoretical term refers to. Now, the standard sequence of events is as follows: a principal components or factor analysis is performed, one factor is found to emerge from the analysis, and it is concluded that this factor can be no other than general intelligence (Carroll, 1993; Gustafsson, 1984; Mackintosh, 1998). Subsequently, the test scores are correlated with a number of relevant variables (in the context of intelligence research, these range from reaction times to job performance). If these correlations are in the right direction, this is taken as evidence that the test indeed measures general intelligence. Finally, heritabilities are computed and found to be impressively high. Often, inquiries stop about here, which is not altogether surprising because, within the construct validity framework, there is nowhere further to go (one may want to ponder this for a while). The only resources that current psychometric dogma in general, and construct validation practice in particular, have to offer are (a) inspecting the 'internal structure' of a test by fitting psychometric models or doing some classical test theory, and (b) correlating the test scores with a zillion external variables. Obviously, however, neither factor analysis nor the inspection of external correlations can, by themselves, provide an answer to the question of validity.

To see this, suppose some researcher performs a factor analysis and finds one factor on measurements obtained by balancing people against a set of old, inaccurate mechanical weight scales. Now suppose that for some reason the researcher is convinced that the test scores should be interpreted as a measure of length. If applied to the human population, the test results will show a reasonable correlation with bodily height, as measured with a meter stick—which supports the construct validity of the test score interpretation in terms of height—and will be predictive of a wide variety of phenomena theoretically related to bodily height.

Now imagine that another researcher interprets the test results as measures of weight and presents a comparable correlation with measurements that result from the application of an electronic weight scale. Given the magnitude of the correlation between height and weight in the human population, such a scenario is not unrealistic, especially if we are allowed to play around with reliability a little; furthermore, it is entirely possible that, if the researchers have at their disposal nothing but weak correlational data of the type that, say, intelligence researchers have, the external correlations

of test scores with many other variables would be essentially the same for the electronic scale, the mechanical scale, and the meter stick. In particular, it is not clear that multi-trait-multi-method matrices would be sufficient to decide on the question which two instruments measure the same quantity. In fact, it is likely that the results of most currently fashionable validation strategies would be unequivocal, just as they are in psychology.

How could these scientists resolve their dispute? We can see only one answer: by investigating the processes through which the respective measurement instruments work. If pursued, such a program would likely point to the fact that behavior of the scales depends on objects' mass, while the height measure does not. The researchers thus would choose the interpretation of test scores in terms of weight because they have developed a good idea of how the mechanical and electronic scales work. The fact that we have a broadly correct explanation of how differences in weight result in differences in measurement is thus more convincing than any differences in correlation with external variables (see also Zumbo, 2007, on the importance of explanation in validity theory).

In fact, the entire idea that one can figure out what a test measures merely by looking at correlations (no matter how many) is, we think, mistaken. And what is missing in psychology, when it comes to the question of test validity, is exactly what is present in the above example but absent in validation research concerning tests for intelligence, abilities, and personality; namely, a good theory of how differences in the targeted attribute are responsible for differences in the measurement outcomes. Although such a theory can serve as excellent input for a factor model and for inspecting correlation structures, it is unlikely to be the output of such practices if they are carried out without being informed by substantive theory. Thus, one particularly important issue that we think should receive much more attention in validation research is the development of process theories that connect targeted attributes to test scores.

Interindividual differences and intraindividual processes. Naturally, we are not the first to point to the importance of developing process theories. Calls for such approaches have been repeatedly made by various scholars, including construct validity theorists (e.g., see Embretson, 1994; Leighton & Gierl, 2007; Snow & Lohman, 1989). However, as far as we know there are only a few researchers that have tackled this problem with some success, in the sense that they were able to specify how intraindividual processes may lead to interindividual differences in test scores (for a theoretical example, see Tuerlinckx & De Boeck, 2005; and for an empirical example, Jansen & Van der Maas, 1997, 2002; see also Borsboom & Mellenbergh, 2007, for a discussion of how these examples relate to validity).

For many important tests in psychology, tracking item response processes and relating them to individual differences has proven a tall order. This

has to do with some general and enduring problems in psychology, as discussed in various sources (Borsboom, Mellenbergh, & Van Heerden, 2004; Cronbach, 1957; Lykken, 1991; Meehl, 1978). One of these issues, in our view, is of central importance to the issue of test validity. This concerns the relation, or lack thereof, between the structure of intraindividual processes and of individual differences.

To see this, one may consider the easiest way in which a measurement model could be true, which arises when one assumes that the structure of the time-dependent behavior of a single subject is isomorphic to the structure we find in the analysis of individual differences. Such a situation would obtain if the development of, say, depression in an individual would consist of that individual moving upwards along a latent continuum so as to increase the probability of depression symptoms in accordance with the item response theory model that is found in the analysis of individual differences (i.e., when a model is fitted to data that arise by administering a depression scale to many people at a single time point; see Aggen, Neale, & Kendler, 2005). This would be the case if all individuals functioned according to the same dynamical laws, so that the difference between any two time points for one individual is qualitatively the same as a corresponding difference between two individuals at one time point (see Hamaker, Nesselroade, & Molenaar, 2007; Molenaar, 2004). For many measurement systems in physics, this is clearly the case: for instance, the differences in thermometer readings of a number of substances, which differ in temperature, depend on the same attribute as the differences in thermometer readings of a single substance, when it increases in temperature. Thus, the within-substance model matches the between-substances model. In other words: the attribute that causes thermometer readings to rise or fall for a single substance is qualitatively the same attribute that causes thermometer readings to differ across substances at a single time point—namely, temperature.

There is little reason to assume that such a simple scheme holds for psychological attributes typically subjected to measurement procedures. Certainly, the fit of a model to individual differences data says virtually nothing about the adequacy of that model for individual change: between-subjects differences do not necessarily equate to within-subjects differences. For instance, it could well be that John's performance on IQ measures is determined by both speed of information processing and neural efficiency (i.e., John represents a two-factor model), and Paula's performance is determined by speed of information processing, neural efficiency and working memory capacity (i.e., Paula represents a three-factor model), while the analysis of individual differences between the Johns and Paulas of this world would fit a single factor model nearly perfectly. For as Molenaar, Huizenga, and Nesselroade (2003) demonstrated, different data generating structures in the intra-individual space can easily result in a one-factor model for between-

subjects differences (see also Hamaker, Nesselroade, & Molenaar, 2007, for an explanation of why this is so). Hence, if a one-factor model fits between-subjects data, it is at best premature to conclude that this is evidence for an isomorphic structure 'in the head' of individual people (Borsboom, Kievit, Cervone, & Hood, in press).

Therefore it is a stretch to assume that the fit of a between-subjects model to individual differences serves to substantiate statements like 'extraversion causes party-going behavior in individuals' (McCrae & Costa, 2008, p. 288) or similar claims (see Borsboom et al., in press, for a more extensive discussion of many similar examples). Such a statement implies that Samantha is a passionate party-crasher because she is extraverted while Nicole rather stays at home with her favorite romantic comedy because she is insufficiently extraverted to behave otherwise. However, this likely oversimplifies processes that result in some people being party-lovers and others being party-avoiders. Perhaps Nicole is in fact extraverted (e.g., likes to meet new people, enjoys social interaction, etc.) but does not like to go to parties because she hates loud music and alcohol. Perhaps Samantha forces herself to go to parties in an attempt to overcome her fear of closed spaces packed with people. Certainly the dynamics of their behavior are rather more complicated than a typical measurement model presumes.

With the exception of some basic learning and conditioning theory, we know hardly anything about the time-dependent structure of processes that govern people's behavior. We know even less about how these processes connect to item response behavior, i.e., how people answer items like those administered in typical psychological tests. However, if one pauses to think about these issues for a moment, it appears quite unlikely that the individual differences variables we find in applied psychometrics have isomorphic counterparts in individual people; counterparts that could steer their mental and behavioral processes so as to eventually culminate in, say, ticking a response category on an answer sheet of a personality questionnaire (Borsboom, Mellenbergh, & Van Heerden, 2003). One supposes that something different must be going on.

Now this need not be a problem for psychology; rather it represents a fruitful avenue for theoretical and empirical research. However, it does represent a problem for the interpretation of test scores as measures. Clearly, if such an interpretation is warranted, it must be for different reasons than those that lead us to consider thermometers valid measures of temperature. That is, thermodynamic laws apply generally, to all substances at all time points, and therefore a causal explanation of individual differences in thermometer readings for different substances can be predicated on the hypothesis that, in each of these substances, the same property is responsible for the same behavior in the thermometer. In normal circumstances, the expansion of mercury in a fixed column, which leads us to perceive the

thermometer readings to rise, *always* results from the transfer of kinetic energy from the particles in the measured substance to the particles in the mercury column. There is no other way.

But in psychological testing, there are many different ways of generating item responses. Even in low level skills, like addition or multiplication, one can discern different strategies of coming up with the right answer (left aside those of coming up with the wrong one). In more complex phenomena, such as those that should be expected to underlie the phenomena of interest in the study of attitudes, personality, or psychopathology, as well as complex skills like those involved in playing chess or solving Raven test items, there are myriads of processes that run in parallel, likely to be intertwined in complex ways. It is not clear whether in such cases we can sensibly speak of measurement, especially since many of the more interesting cases of psychological testing might require an altogether different way conceptualizing the relation between test scores and theoretical terms—an issue we turn to next.

Causal networks. In our above ponderings, we assumed that, in some way, a psychometric model adequately pictures the situation in the real world—so that it makes sense to think about the test scores as measures: a one-factor model for general intelligence, for instance, should then refer to a single variable in the real world (e.g., speed of information processing) that causes differential performance on IQ tests, hopefully in the same way between people as within them. But what if we are wrong? What if the psychometric factor in a factor model does in fact not refer to a linearly ordered property that causes variation in IQ-scores? If this is the case, then we have a serious problem on our hands, namely that it may be altogether mistaken to think about the relation between test scores and psychological attributes as one of measurement.

Results presented by Van der Maas et al. (2006) show that, in the case of general intelligence, such a situation may be actual rather than hypothetical: these authors showed that a dynamical model *without any latent variables*, and a fortiori without the factor *g*, easily rivals the theory of general intelligence in explaining empirical phenomena (e.g., the Jensen effect; Jensen, 1998). More specifically, Van der Maas et al. (2006) simulated data based on a dynamical model in which cognitive processes interact beneficially during development. This is called *mutualism*. The mutualism model results in the same positive manifold (i.e., positive correlations between cognitive tasks) that is consistently seen in IQ data and, in a factor analysis, always yields a dominant first factor or principal component. Hence, the mutualism model yields typical IQ data but does not contain an overarching psychometric factor *g* that refers to a construct of general intelligence. This is food for thought, because the measurement of intelligence through IQ-tests is one of the primary and best researched examples of psychologi-

cal measurement—in fact, one could argue that many other measurement systems are copied from the intelligence example.

A similar situation may be true for another broad class of psychological constructs, namely mental disorders (e.g., see Borsboom, 2008; Cramer, 2008). A reflective model is often hypothesized to account for the relationship between a construct, for example depression, and its indicators (e.g., the symptoms of depression). An important consequence of assuming such a reflective model is that the attribute measured explains all systematic covariation between individual symptoms (i.e., local independence; see for instance Lord & Novick, 1968). However, such a model might not paint the most adequate picture of mental disorders (Borsboom, 2008). For example, consider two symptoms of depression, as mentioned in the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1994), sleep disturbance and fatigue. According to a reflective model, a high correlation between those two symptoms is entirely explained by the measured attribute, i.e., depression. However, it would appear rather more plausible to assume that a direct relationship exists between those symptoms: If you don't sleep, you get tired. It is not difficult to see that many such relations may exist between symptoms of mental disorders. If one accepts this possibility, then a *causal network* in which symptoms stand in direct causal relationships toward one another could be the model that best describes the phenomenon of mental disorders. As in the mutualism example, one would have a model without a latent variable and thus, without a unitary psychological attribute that underlies the distinct symptoms. This would be an interesting situation, because it necessitates a complete reconsideration of the way test scores function. In particular, it is not obvious that one should think about the relation between symptoms and syndromes as one of measurement (Borsboom, 2008, proposes instead to view this relation as a mereological one; i.e., the symptoms do not measure the network, but are part of it).

Heritability and phenotypic heterogeneity. Perhaps the most often cited evidence for the reality of traits like general intelligence and extraversion concerns their high heritability. The heritability of IQ-scores, for instance, was found to be 70-80% in adulthood (Bouchard et al., 1990; Posthuma, de Geus, & Boomsma, 2001; Posthuma, 2002), and of liability to depression around 50-70% (Kendler et al., 2001; McGue, & Christensen, 2003). Similar figures hold for many other psychological variables; see Boomsma, Busjahn, and Peltonen (2002) for an overview. It is often thought that such figures provide evidence for the reality of psychological attributes, coded as latent variables in commonly used measurement models, and therefore substantiate the claim that test scores should be interpreted as measures. In our view, however, such evidence is hardly a smoking gun, as it does not

provide us with a sensible answer to the question of how such attributes influence the test scores.

For instance, high heritability of test scores is in no way informative of their homogeneity. One may add up scores on measures of height, IQ, and eye color, and start looking for genetic basis of the newly defined phenotype; all three characters are highly heritable (Bräuer & Chopra, 1978; Bouchard et al., 1990; Silventoinen et al., 2003), so when the scores on their respective measures are added up, the resulting composite is necessarily highly heritable as well—not because its components reflect a single property or are influenced by the same underlying cause, but merely because all three components are highly heritable in themselves. Therefore, the fact that something is heritable does not tell us anything about the homogeneity of its structure—in fact, in most extreme cases, each of the items on a given highly heritable composite could be measuring a different highly heritable characteristic.

The situation is worsened if one assumes that the elements that make up the composite stand in causal relations to each other. In such a case, each element of the composite may send out an effect to the other elements, thereby propagating genetic effects from any one element to all others. In addition, the strength of the causal relations between the different elements may itself be subject to genetic influences; e.g., sleep deprivation may more easily lead to fatigue, loss of concentration, and depressed mood in some people as compared with others, and these individual differences in the strength of the causal links may stand under the influence of genetic structures.

It is interesting to note that such a state of affairs would be in accordance with the fact that researchers have consistently failed to find a noteworthy contribution of any single gene explaining variation in any given psychological trait. For instance, the variation in composite measures such as full scale IQ has been found to be affected by many genes (Gosso, 2007; Plomin et al., 2008). Conversely, no single gene has been found to account for a substantial proportion of the variance in general intelligence (De Geus, Wright, Martin, & Boomsma, 2001; Nokelainen & Flint, 2002; Payton, 2006; Plomin, Kennedy, & Craig, 2006 Posthuma et al., 2005). In light of this, it seems reasonable to consider the possibility that IQ-scores might be complex composites, comprising distinct elements, or measures thereof, that depend on a heterogeneous collection of processes, possibly with mutualistic connections. A similar situation may obtain in the field of psychopathology research, where studies in search of genes associated with depression (association studies of monoaminergic candidate genes, genes related to neurotoxic and neuroprotective processes, studies of gene-environment interactions etc.) have so far failed to come up with a gene that would explain more than a minor part of the variance (Levinson, 2006). Thus, although

heritability estimates are, for most psychological test scores, quite impressive, their evidential strength with respect to the thesis that the tests in question measure a single attribute is limited.

In conclusion, three often cited sources of evidence for the measurement hypothesis—the fit of latent variable models to test scores, the presence of significant external correlations, and the high heritabilities of these test scores—should not be considered definitive on the question of whether we are entitled to interpret our test scores as measures and our tests as measurement instruments. The reason is that none of these lines of evidence addresses the question of how the measurement instrument picks up variation in the targeted attribute and transmits such variation into the measurement outcomes. That, and nothing else, is the smoking gun of test validity. Now, if one ponders the way that tests are structured and the way in which variation in test scores is likely to arise, then it becomes altogether unclear whether we should conceive of the relation between test scores and theoretical terms in terms of measurement. This may seem to be a negative result, but we do not think this is so. It invites us to think about alternative ways of theorizing about the genesis of test score variation.

One supposes that there may be more fruitful alternatives to the theory-observation relation than the almost mandatory methodological outlook in psychology, which has arisen out of the construct validity doctrine coupled with conventional psychometric wisdom. This outlook invariably requires one to interpret one's test scores as measures indicative of a psychological construct, but never makes the parallel requirements of explicating what one means by the notion of measurement and in what way one's psychological attributes may be taken to exist or have causal effects. Therefore, it propagates a situation where researchers entertain shadowy notions of constructs, measurement, and validity, leading them to adopt a monolithic methodological strategy in the analysis of test scores, as coded in conventional test theoretic procedures. This strategy is scarcely motivated by subject matter. Rather, it represents a set of methodological dogmas that may be entirely inappropriate in view of the subject matter of psychology. In our view, it is time to leave these dogmas behind; in this sense, sensible theorizing about test scores is yet to begin.

DISCUSSION

Construct validity theory is at odds with the way in which many, if not most, researchers interpret validity. This raises the question of who has the better validity concept: construct validity theorists, who think that validity is a property of test score interpretations that reflects how strongly these interpretations are supported by the evidence, or the rest of the inhabitants

of the scientific world, who think that validity is a property of tests that signals whether these tests measure what they should measure. In the present chapter, we have argued against the construct validity view and in favor of test validity as it is normally understood.

In our view, and in contradiction to the theoretical mainstream in validity theory (e.g., Kane, 2006; Messick, 1989), it is a mistake to view validity as a property of test score interpretations. Construct validity in no way restricts the type of test score interpretation to be considered, thereby leaving the possibility open to consider the construct validity of *any* test score interpretation that one whishes to make. Such a view easily leads to paradoxical situations where tests score interpretations that deny the validity of tests can nonetheless have high construct validity. Relatedly, it is a mistake to think of validity as a function of evidence. Most important, such a view implies that the construct validity of a test score interpretation is dependent on time: If one adheres to construct validity theory, one would have to agree that a test score interpretation that turns out to be wrong actually was valid up until the moment the falsifying evidence became available. According to the general view that a test is valid if it actually measures what it is supposed to measure, this makes no sense. If a test turns out to measure nothing at all or something completely different from what was once thought, it was never valid in the first place. Therefore a theory of validity that concerns test score interpretations and relies on evidence for these interpretations is inadequate.

In addition, we think it is hardly possible to consider the validity of test score interpretations in terms of measurement without spelling out what one means by the term 'measurement'. And since it appears that the realist, causal interpretation of measurement has, at the moment, no serious rivals, the test score interpretation 'test X measures attribute Y' must be interpreted as requiring the presence of a causal effect of the attribute on the test scores. If this is indeed granted, then it follows that there is one and only one necessary condition for test validity, and that is that the test has the property of picking up variation in the targeted attribute, and transmitting it into variation in the test scores. This, however, is clearly a property of the measurement instrument itself, not of the interpretations of the measurement outcomes. In addition, whether a measurement instrument actually measures what the researcher intends it to measure is ultimately a question of truth, not of evidence. Thus, following this line of reasoning, we arrive at a position almost fully orthogonal to the dominant view in construct validity theory.

To avoid any confusion caused by the ambiguous use of the word construct, we recommend abandoning this term along with the theory it lends its name to. We carve up the world in a way that naturally corresponds to the ordinary semantics of measurement. On the one hand we have a psy-

chological attribute that we hypothesize to exist in the world, and to cause variation in our measurement outcomes. That is the thing we want to measure. On the other hand we have the theoretical term that we use in our theories and that, if we are lucky, in fact picks out the psychological attribute in question. If we can explain how the psychological attribute acts to cause variation in our measurement outcomes, we can truly say something about the validity of our measurement instrument. This requires us to investigate the structures and processes that make up the psychological properties we are interested in and to show that these properties are picked up by the test. In essence, this means that we need to construct a psychometric model that is *psycho*metric rather than psycho*metric*. Rather than substanceless models, preferred for their philosophical or statistical niceties, psychometric models should be formal theories of test behavior. The task of validation then comes down to testing these theories in whatever way necessary.

Thus, in our view, psychometric theories are central to validation. Construct validity theory, however, has it that even though one may fit psychometric models if one so desires, establishing test validity always requires one to do 'something else' as well, and, interestingly, this something else may also be done without ever even raising the question of which measurement model should be considered. This means that one could support construct validity in the *absence* of a measurement model. This, in our view, is predicated on a mystical view of what measurement is—a view that is so vague that it may serve to leave the issue of validity forever undecided. The fact that construct validity theorists rejoice in claiming that validation is 'never-ending', 'open-ended', etc. therefore may not indicate philosophical or methodological sophistication, but rather unwittingly illustrate how deeply misguided construct validity theory really is.

Moreover, the idea that test validity *cannot* be settled, deeply ingrained in the writings of construct validity theorists, blurs the difference between clearly successful and clearly doubtful cases of measurement. Whether one likes it or not, it is a fact of life that one can go to a hardware store and buy a hygrometer for two dollars. If one wants to know how and why that instrument measures humidity, one can look it up in the manual. Clearly such cases of measurement have something that psychology does not have, and we should be interested in what it is. In this respect, hiding behind the complexities of confirmation theory or philosophy of science is in nobody's interest. Surely one can make up all kinds of problems involved in the function of hygrometers, but these are of a completely different order as compared to those in psychology. The difference between the hygrometer manual and the WAIS manual is simply that the former offers an explanation of how differences in humidity are transmitted into differences in the measurement readings, whereas the latter does not offer an explanation

of how differences in general intelligence are transmitted into different IQ-scores.

The legacy of construct validity theory is that people have come to think that theories about how intelligence relates to other properties, or the utility of IQ-scores in predicting college performance, or the correlations between IQ-scores and other IQ-scores, all prevalent in test manuals like that of the WAIS, can be a substitute for the missing corpus of process theories. As long as this remains the case, the validity of psychological tests will be in doubt.

AUTHOR NOTE

Send correspondence to Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands, e-mail: d.borsboom@uva.nl. The work of Denny Borsboom and Angélique Cramer is supported by NWO innovational research grant 452-07-005. The work of Annemarie Zand Scholten is supported by NWO research grant 400-05-055. The work of Sanja Franic is supported by a Nuffic Huygens Scholarship Grant and a Ministry of Science, Education and Sports of the Republic of Croatia grant for postgraduate training abroad.

REFERENCES

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35*, 475–487.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. Washington DC: APA.

Barker, P., & Goldstein, B. R. (1992). Distance and Velocity in Kepler's Astronomy. *Annals of Science, 51*, 59–73.

Batitsky, V. (1998). Empiricism and the myth of fundamental measurement. *Synthese, 116*, 51–73.

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

Boomsma, D. I., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics, 3*, 872–882.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64*, 1089–1108.

Borsboom, D., & Dolan, C. V. (2007). Theoretical equivalence, measurement, invariance, and the idiographic filter. *Measurement, 5*, 236–263.

Bouchard, T. J., Jr., Lykken, D. T., McGue, M., Segal, N. L., & Tellegen, A. (1990). Sources of human psychological differences: The Minnesota Study of Twins Reared Apart. *Science, 250*, 223–228.

Bowler, P. J., & Morus, I. R. (2005). *Making modern science.* Chicago : University of Chicago Press.

Bräuer, G., Chopra, V. P. (1978). Estimating the heritability of hair colour and eye colour. *Journal of Human Evolution London , 9*(8), 627–630.

Bridgman, P. W. (1927). *The logic of modern physics.* New York: Macmillan.

Carnap, R. (1950). *Testability and meaning.* New Haven, CT: Whitlock's.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186–190.

Cramer, A. O. J. (2008). Lack of empathy leads to lack of remorse? Psychopathy as a network. *Newsletter of the Society for the Scientific Study of Psychopathy, 2*, 2–3.

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

Domotor, Z., & Batitsky, V. (2008). The analytic versus representational theory of measurement: A philosophy of science perspective. *Measurement Science Review, 8*, 129–146.

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

Embretson, S. (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.

Gearhart, C. A. (1985). Epicycles, eccentrics, and ellipses: The predictive capabilities of Copernican planetary models. *Archive for History of Exact Sciences,* 207–222.

Geus, E. J. C., de, Wright, M. J., Martin, N. G., & Boomsma, D. I. (2001). Editorial: Genetics of brain function and cognition. *Behavior Genetics, 31*(6), 489–495.

Gosso, M. F. (2007). *Common genetic variants underlying cognitive ability.* Unpublished doctoral dissertation. Vrije Universiteit Amsterdam, The Netherlands.

Gustafsson, J. E. (1984). A unifying model for the structure or intellectual abilities. *Intelligence, 8*, 179–203.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait-state model. *Journal of Research in Personality, 41*, 295–315.

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.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger/Greenwood.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: National Council on Measurement in Education and American Council on Education.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*, 389–423.

Kendler, K. S., Gardner, C. O., Neale, M. C., & Prescott, C. A. (2001). Genetic risk factors for major depression in men and women: similar or distinct heritabilities and same or partly distinct genes. *Psychological Medicine, 31*, 605–616.

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

Leighton, J., & Gierl, M. (Eds.). (2007). Cognitive diagnostic assessment for education: Theory and applications. Cambridge: Cambridge University Press.

Levinson, D. F. (2006). The genetics of depression: A review. *Biological Psychiatry, 60*, 84–92.

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.

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement, Vol. III*. New York: Academic Press.

Lykken , D.T. (1991). What's wrong with psychology anyway? In D.Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology. Volume 1: Matters of public interest*. Minneapolis: University of Minnesota Press.

Maas, H. L. J. van der, Dolan, C., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842–861.

Mackintosh, N. J. (1998). *IQ and human intelligence*. Oxford: Oxford University Press.

Maraun, M. D., & Peters, J. (2005). What does it mean that an issue is conceptual in nature? *Journal of Personality Assessment, 85*, 128–133.

McCrae, R. R., & Costa, P. T., Jr. (2008). Empirical and theoretical status of the Five-Factor Model of personality traits. In G. Boyle, G. Matthews, & D. Saklofske (Eds.), *Sage handbook of personality theory and assessment* (Vol. 1, pp. 273–294). Los Angeles: Sage.

McGue, M., & Christensen, K. (2003). The heritability of depression symptoms in elderly Danish twins: occasion-specific versus general effects. *Behavioral Genetics, 33*(2), 83–93.

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.

Molenaar, P. C. M. (2004). A manifesto on psychology as ideographic science: bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201–218.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical validation of Developmental Systems Theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development:*

Dialogues with lifespan psychology (pp. 339–360). Dordrecht: Kluwer Academic Publishers.

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

Nokelainen, P., & Flint, J. (2002). Genetic effects on human cognition: lessons from the study of mental retardation syndromes. *Journal of Neurology, Neurosurgery and Psychiatry, 72*, 287–296.

Payton, A. (2006). Investigating cognitive genetics and its implication for the treatment of cognitive deficit. *Genes, Brain and Behavior, 5*, 44–53.

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

Cambridge University Press.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York: Worth.

Plomin, R., Hill, L., Craig, I. W., McGuffin, P., Purcell, S., Sham, P., Lubinski, D., Thompson, L. A., Fisher, P. J., Turic, D., & Owen, M. J. (2001). A genome-wide scan of 1842 DNA markers for allelic associations with general cognitive ability: a five-stage design using DNA pooling and extreme selected groups. *Behavioral Genetics, 31*, 497–509.

Plomin, R., Kennedy, J. K. J., & Craig, I. W. (2006). The quest for quantitative trait loci associated with intelligence. *Intelligence, 34*, 513–526.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson Education.

Posthuma, D. (2002). *Genetic variation and cognitive ability*. Unpublished doctoral dissertation. Vrije Universiteit Amsterdam, The Netherlands.

Posthuma, D., Geus, E. J. C., de, & Boomsma, D. I. (2001). Perceptual speed and IQ are associated through common genetic factors. *Behavioral Genetics, 31*, 593–602.

Posthuma, D., Luciano, M., Geus, E. J. C., de, Wright, M. J., Slagboom, P. E., Montgomery, G. W., Boomsma, D. I., & Martin, N. G. (2005). A genome-wide scan for intelligence identifies quantitative trait loci on 2q and 6p. *American Journal of Human Genetics, 77*(2), 318–326.

Psychological Corporation (1997). *WAIS-III WMS-III Technical Manual*. San Antonio, TX: Harcourt Brace & Co.

Rasch, G. (1960). ?. ?, ?, ?-?.

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V., Luciano, M., Martin, N. G., Mortensen, J., Nistico, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G., & Kaprio, J. (2003). Heritability of adult body height: A comparative study of twin cohorts in eight countries. *Twin Research, 6*(5), 399–408.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (pp. 263–311). Washington DC: American Council on Education and National Council on Measurement in Education.

Stegmüller, W. (1979). *The structuralist view of theories: A possible analogue of the Bourbaki programme in physical science*. New York: Springer-Verlag.

Suppe, F. (1974). *The structure of scientific theories*. Urbana: University of Illinois Press.

Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of measurement, Vol. II.* San Diego, CA: Academic Press.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*, 629–650.

Van Fraassen, B. C. (1980). *The scientific image.* Oxford: Clarendon Press.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam: Elsevier.

Queries

Please add the following to References:
Borsboom, Mellenbergh, & Van Heerden (2004), (2003)
Rasch (1960)–Complete ref.
Borsboom (2006)
Borsboom & Mellenbergh (2007)
Meehl (1978)
Cronbach (1957)
Borsboom, Kievit, Cervone, & Hood (in press)