Conceptual issues in psychological measurement

Borsboom, D.

Citation for published version (APA):
6. THE PROBLEM OF VALIDITY

6.1 Introduction

That the conceptual problems inherent in measurement in general, and psychological measurement in particular, are poorly understood is obvious from the lack of agreement on the meaning of the term 'measurement', the multitude of conceptually different models for implementing it (e.g., Lord & Novick, 1968; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Hambleton & Swaminathan, 1985; Krantz, Suppes, Luce, & Tversky, 1971), and the fact that no psychologist can point to a field where psychological measurement has succeeded without eliciting an immediate claim to the contrary from anotherpsychologist. Given that virtually all aspects of the measurement problem are the subject of ongoing debates (Borsboom & Mellenbergh, 2002; Borsboom, Mellenbergh, & Van Heerden, in press; Lamiell, 1987; Lumsden, 1976; Maraun, 1999; Michell, 1986, 1999, 2000; Schmidt & Hunter, 1999), one would expect these debates to culminate in fierce discussions on the most central question one can ask about psychological measurement, which is the question of validity. It is therefore an extraordinary experience to find that, after proceeding up through the turmoil at every fundamental level of the measurement problem, one reaches this conceptually highest and presumably most difficult level only to find a tranquil surface of relatively widespread consensus (Kane, 2001; Shepard, 1993). In fact, this is not only surprising but slightly worrying, because validity is a largely philosophical topic. Consensus on philosophical problems is rare, and for good reasons: Philosophy is the art of critical thinking and critical thinking generally does not lead to consensus but to debate.

A second remarkable aspect of current validity theory is that the concept validity theorists are concerned with seems strangely divorced from the concept that most researchers have in mind when posing the question of validity. That is, most validity theorists have come to see the validity concept as embracing virtually every test-related problem that may be raised (Cronbach, 1988; Messick, 1989; Shepard, 1993), while many researchers are under the impression that the problem of validity simply concerns the question whether a test measures what it should measure. Moreover, if one regards this simple question as legitimate and crucial, as I do, then one has a very hard time understanding most recent papers on validity. To give but one example that I find puzzling: One can find in Messick (1989; p. 30) the idea that some psychological attributes are real, while others are not, and that in both cases the concept of validity applies. I have great difficulty in understanding the supposition that one can ask what may be called the 'simple' question of validity,
which I construe as the question whether, for example, IQ-tests really measure the attribute we call 'intelligence', in a situation where there is nothing in reality that corresponds to intelligence. That is, if the realist position cannot be taken, I do not understand why the question of validity should apply at all.

Now, when encountering philosophical positions that seem elusive and difficult to understand, one always faces a problem of attribution. In general, there are three possible sources of confusion, and it is often hard to decide which is at play. The first and least attractive possibility is that the confusion arises from one's own limited cognitive resources. The second is that there is a flaw in the position itself. And the third is that the authors in question are analyzing the wrong problem. I think that, in the present case, the third explanation applies. It is my intent to convince the reader that most of the validity literature either fails to articulate the validity problem clearly, or misses the point entirely. I will argue that it is an unfortunate historical accident that the validity concept has been divided into different kinds, torn from its rightful place in science, and reunified by constructing it as an umbrella term intended to cover virtually every thinkable aspect of inference – be it scientific, philosophical, political, or ethical. Validity is not complex, faceted, or dependent on nomological networks. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14), when he stated that a test is valid if it measures what it purports to measure.

The argument to be presented is exceedingly simple; so simple, in fact, that it articulates an account of validity that may seem almost trivial. It is this. If something does not exist, then one cannot measure it. If it exists, but does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether. In these two cases, a test does not possess validity. In all other cases, it is valid. Thus, a test is valid for measuring an attribute if and only if a) the attribute exists, and b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. Now, one may find this unsurprising. In fact, there is a good chance that many readers are inclined to respond that they tell their students this all the time. However, in the validity literature of the past two decades, it is difficult to find a explicit formulation resembling the above. In the writings of leading theorists (i.e., Cronbach, 1988; Kane, 2001; Messick, 1981, 1989, 1998; Shepard, 1993), one will not find much that sustains it; rather, one is likely to find this type of idea in a discussion of historical conceptions of validity (Kane, 2001, p. 319-323). The ontological part of Messick’s (1989; 1998) unificationist conception of validity leans towards it, but it is not clearly articulated and it is questionable whether it is consistent with the other, epistemological, part of his synthesis (Markus, 1998). Moreover, Messick (1989, p.13) views validity as a judgment, and thus conceptualizes the term as applying primarily to the evaluation of evidence bearing on that judgment. Similar views are put forward in Kane (1992), Shepard (1993), and Moss (1992). In keeping with this idea, the current literature conceptualizes validity as applicable to test score interpretations only (Cronbach, 1988; Kane, 2001; Messick, 1989), while the conception stated here is consonant with the older conception that it is a property of tests. Indeed, finding similar conceptualizations requires browsing some of the older archives (e.g., Kelley, 1927; Cattell, 1946; Loevinger, 1957).
However, these treatises are not based on a causal, but on a correlational conception of validity, which I will argue is crucially mistaken. In latent variable theory, one may find causality based lines of reasoning (Bollen & Lennox, 1991; Bollen & Ting, 2000; Edwards & Bagozzi, 2000), but they will not be explicitly linked to validity theory. It thus seems that the validity concept, as formulated above, has not been explicitly proposed in the literature, although it certainly has been hinted at (Cattell, 1946; Loevinger, 1957; Campbell, 1960). An observation that underscores the apparent novelty of the stated conception is that it contradicts most of the conventional wisdom in validity theory, which means that this theory either does not sustain it, or is inconsistent, or both. For example, the above conception implies that a) validity is not a matter of degree, b) the square root of the reliability coefficient is not the upper limit of validity, and c) unreliability, item bias, and other supposedly undesirable characteristics of tests bear no direct relation to validity. This is in contradiction with every paper on validity I know, but I think it is correct.

The aim of the present chapter is to elaborate the implications of this view, and to discuss the ways in which it diverges from, or converges with, both historical and current conceptions. The argument will focus on four points where validity theory seems to have taken the wrong turn. First, it has confused ontological and epistemological claims; second, it has mistaken questions about reference for questions about meaning; third, it has been plagued by a correlational account where there should have been a causal account; and fourth, the idea that validity applies to test score interpretations, rather than to tests, is inadequate. Finally, it will be argued that current validity theory deals with too many issues at the same time, so that it collapses under its own weight. It is proposed that the question of validity must be taken to apply only to the question whether one is measuring the right attribute; not to the question how well one is measuring that attribute. This latter question is left to the technically oriented psychometric literature, which deals with it in a more sophisticated way than the validity literature.

### 6.2 Ontology versus epistemology

If the crucial issue in validity concerns the existence of an attribute that causally influences the outcome of the measurement procedure, then the central claim is ontological, and not epistemological. This is to say that one is claiming something about which things inhabit reality, and what they are doing there. Such claims are about ontology, and as such they are conceptually distinct from the ability to find out about reality, which is the central issue in epistemology. Measurement, of course, is the prototypical epistemological activity in science, and it is therefore easy to make the mistake that we are primarily claiming something on this front. This is because if the ontological claim holds, then the measurement procedure can be used to find out about the attributes to which it refers. Put more simply: If differences in intelligence cause differences in IQ-scores, then the IQ-score differences can be used to find out about the intelligence differences. Thus, in this very special case, the truth of the ontological claim guarantees the epistemological access.
It would seem, then, that to talk about the ontology is to talk about the epistemology, and there surely is a sense in which this is correct. Now it is a small step to conclude that, instead of laying down our ontological claims, which make so abundantly clear what kind of radical assumptions we are making (Borsboom, Mellenbergh, & Van Heerden, in press; Michell, 1999), we could just as well limit our discussion to the epistemological side of the endeavor, which is respectable and familiar. It is another small step to conclude that the question of validity is about particular aspects of this epistemological process we call measurement. The final step leading to some very dark philosophical dungeons from which escape is impossible, is to start talking about some presumed universal characteristics of this epistemological process (usually derived from a few paradigm cases like length or temperature measurement) that, if present, would allow one to somehow be rationally justified in concluding that the ontological claims are true.

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

This line of reasoning quickly gets us nowhere. The reason is that there are no universal characteristics of measurement, except the ontological claim involved. The only thing that all measurement procedures have in common is the either implicit or explicit assumption that there is an attribute out there that, somewhere in the long and complicated chain of events leading up to the measurement outcome, is playing a causal role in determining what values the measurements will take. This is not some complicated and obscure conception but a very, very simple idea. If we, however, fail to take it into account, we will end up with an exceedingly complex construction of superficial epistemological characteristics that are completely irrelevant to the validity issue. And because the measurement processes and models are diverse and complicated, we are likely to buy into the mistaken idea that the concept of validity must also be complicated. So now we get a multiplication of terms. For the human condition is such that someone will inevitably distinguish between ‘kinds of validity’ and ‘degrees of validity’ and so we are bound to come up with a hundred or so ‘validities’, which all come in ‘degrees’, until someone stands up because this is clearly ridiculous, and claims that ‘all validation is one’ (Cronbach, 1980, p.99) so that all kinds of validity can be integrated and subsumed under
one giant umbrella (Messick, 1989). And since we are now thoroughly convinced that we are concerned with characteristics of an epistemological process rather than with an ontological claim, we are going to reach the conclusion that all this time we were really just talking about the one grand epistemological process — scientific research (Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989). However, given that every attempt at drawing a line between 'scientific' and 'unscientific' research either fails or duplicates the distinction between good and bad research, we have now discovered the exciting fact that validation research is research. In other words, we have discovered nothing at all. And the reason for this is that there was nothing to be discovered in the first place.

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

Turning verbs into nouns often leads to understandable classificatory systems but when doing philosophy one does well to stay clear of it. For the union of all possible test related activities of this kind is not validity, but validation. These terms are sometimes used interchangeably in the literature, but they are not the same. This is clear from the fact that validity is a property, while validation is an activity. In particular, validation is the kind of activity we undertake to find out whether a test has the property of validity. Validity is a concept like truth; it represents an ideal or desirable situation. Validation is more like theory testing; the muddling around in the data to find out which way to go. Validity is about ontology; validation is about epistemology. The two should not be confused. Now, I think that most of the validity literature has not dealt with the problem of validity, but with the problem of validation. While there is nothing wrong with describing, classifying, and evaluating validation strategies, such activities are not likely to elucidate the concept of validity itself. In fact, if one concentrates on the epistemological problems long enough, one will move away from the validity concept rather than towards it. Consider, for example, Messick’s (1989) widely cited definition of validity: ‘validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment’ (p. 13; italics in the original). No view could be farther apart from the one being advanced here. Validity, in the present conception, is not a judgment at all. It is the property being
6.3 Reference versus meaning

That the position taken here is so at variance with the existing conception in the literature is largely due to the fact that I have reversed the order of reasoning. Instead of focussing on the epistemological processes and trying to fit in existing test practices, I have started with ontological claims, and I derive the adequacy of epistemological practices only in virtue of their truth. This means that the central point in validity is one of reference: The attribute to which the psychologist refers must exist in reality, otherwise the test cannot possibly be valid. The position here is thus a strongly realist one, in that I construct measurement as involving realism about the measured attribute. This is because I cannot see how the sentences ‘Test X measures the attitude towards nuclear energy’ and ‘Attitudes do not exist’ can both be true. If you agree on this point, then you are in disagreement with some very powerful philosophical movements which have shaped validity theory to a large extent.

One particularly strong variant of these movements once proudly went by the name of logical positivism. Philosophers and scientists endorsing this theory saw it as their mission to exorcise all reference of theoretical terms (like ‘attitude’), because such reference introduces metaphysics, which the logical positivists thought was bad. They therefore constructed theoretical terms as nonreferential. This lead them to focus on the meaning of theoretical terms. Meaning and reference are easily confused, but are very different concepts. To give a classic example (Frege, 1892), ‘the morning star’ and ‘the evening star’ have different meanings (namely ‘the last star still to be seen at morning’ and ‘the first star to be seen at evening’), but refer to the same thing (namely the planet Venus). Because the positivists had a slightly phobic attitude towards metaphysics, they wanted to explain the use of theoretical terms like ‘attitude’ without letting these terms refer to reality.

This was an interesting endeavor but it failed (see Suppe, 1977, for a good overview). However, one of the relics of the approach has plagued validity theory to this day. This is the nomological network. A nomological network is a kind of system of laws relating the theoretical terms to each other and to the observations. For the positivists, this network served to create meaning without reference for the theoretical terms. The idea is that the meaning of a theoretical term is solely determined by the place of that term in the nomological network: the meaning of the term ‘energy’ is fixed by the network and by nothing else – certainly not by a reference to actual energy. Thus, in this view we can have meaning without reference, and can invoke theoretical terms without automatically engaging in ontological claims, which always introduce a lot of metaphysics.

This idea was used by Cronbach & Meehl in 1955 to put forward their idea of ‘construct validity’. Many people think that construct validity is the same as the kind of validity being proposed here, but this is not the case. The construct validity position does not invoke reference (it does not say that the attribute to be measured should exist), and it does not talk about causality (it is not necessary
for the attribute to have a causal role in determining the measurement outcomes). The classic position, as articulated by Cronbach & Meehl (1955), holds that a test can be considered valid for a construct, if the empirical relations, in which the test stands to other tests, match the theoretical relations, in which the construct stands to other constructs. One can imagine this as two path models, one hovering over the other. One model stands for theoretical relations, the other for empirical relations. If the models match, then there is ‘construct validity’ for test score interpretations in terms of the nomological network. For instance, suppose the nomological network says that the construct ‘intelligence’ is positively related to the construct ‘general knowledge’ and negatively to the construct ‘criminal behavior’. Further suppose that one observes a correlation of .5 between an IQ-test and a test for general knowledge, and a correlation of -.4 between the IQ-test and the number of months spent in prison. There is thus a match between empirical and theoretical relations. In construct validity theory, it is this match that constitutes and defines the validity concept.

To define construct validity, no reference to the existence of theoretical entities is necessary, and their causal impact on the measurement outcomes is not even a topic of discussion. Read Cronbach & Meehl (1955) to see how carefully they avoid this issue. As an illustration of the ambiguity of Cronbach & Meehl’s (1955) paper, one may confer Bechtold (1959) and Loevinger (1957), who both discuss construct validity, but are talking about two completely different interpretations of the concept – one positivist, the other realist. In principle, however, within the construct validity perspective there is no friction between ‘Test X measures the attitude towards nuclear energy’ and ‘Attitudes do not exist’. As long as the empirically observed relations, between test X and other tests, match the theoretical relations in the nomological network, all is fine. So, this view has a little bit in it for everyone.

The problem, of course, is that we have few if any nomological networks in psychology that are sufficiently detailed to do the job of fixing the meaning of theoretical terms. To fix this meaning requires a very restrictive nomological network. The reason is that the theory that has to be invoked for construct validity to work is an account similar to the descriptive theory of meaning (Kripke, 1972). This theory does not say ‘intelligence is a real attribute with causal impact on our measurements’, but ‘intelligence is whatever has the relations to other constructs as specified in the nomological network’. Cronbach & Meehl (1955) do not mention the descriptive theory of meaning, but that they rely upon it is evident from statements like ‘a construct is defined implicitly by a network of associations or propositions in which it occurs’ (p. 299-300). It is crucial for the ideas formulated in Cronbach & Meehl (1955) that a descriptive account of meaning is possible, because otherwise one is forced to invoke a reference for intelligence, which brings in the very metaphysics to be avoided through the back door.

In some highly developed theories, like the ones in physics, one could at least begin to consider this account, because they are restrictive enough to single out one particular theoretical term, which is the only one that has all the right relations. In psychology, such an account does not work because we do not have the required theories. That this is not just an academic point, but a decisive argument against
using a descriptive theory of meaning can be immediately seen by considering the intelligence example discussed before. One does not get anywhere by saying that 'intelligence is whatever is positively related to general knowledge and negatively to criminal behavior', because there are too many theoretical terms that will satisfy this description, and many of them will evidently not be the same as intelligence. No theoretical term in psychology can be unambiguously identified in this way. Thus, this theory will not be able to single out theoretical terms by merely describing where they stand in a nomological network. Cronbach & Meehl (1955) do discuss the problem that nomological networks are incomplete and vague in psychology, but they do not mention the most important implication of that problem: It is fatal to any positivist reading of their account, because it shows that reference, and the accompanying realist metaphysics of measurement, cannot be avoided. Instead, they conclude that it leads to vagueness in the construct definitions. This is, of course, true, but not the primary problem. The primary problem is that too many theoretical terms will satisfy construct definitions of the kind Cronbach & Meehl (1955) are discussing (see also Rozeboom, 1960), and that therefore the theory of meaning they use fails to work.

Now, this should not be regarded as a grave problem for psychology in general, because the descriptive theory of meaning is not a very good one anyway (Kripke, 1972). Neither is there any particular problem about not having nomological networks, because one has to start somewhere – and the tight, lawlike relations that make up nomological networks are more likely to be the result of research than a prerequisite for doing it. However, one would expect psychologists to dismiss any account of validity, that requires the existence of nomological networks, as inadequate from the outset because when one thinks the matter through, such a theory would have very undesirable consequences. For example, some psychological tests certainly appear to be measuring something important, and one should be able to say that one thinks, suspects, or hypothesizes that IQ-tests validly measure intelligence even if one momentarily has no nomological network available to fix the meaning of the term 'intelligence'. In a theory of validity that requires the availability of nomological networks this is, strictly taken, impossible. Every psychologist should object to this; not only because the view is unduly restrictive but because it is completely inadequate. Validity does not depend, and has never depended, on the availability of nomological networks because it is not about meaning but about reference.

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

The emphasis that is placed on the importance of ruling out alternative rival hypotheses for corroborating data (Cronbach & Meehl, 1955; Messick, 1989) partly acknowledges this. One can readily see this by introducing the question to what hypothesis the alternative one should be considered a rival. Obviously, to the hypothesis that there is an attribute in reality that produces variation in the measurement outcomes. What, then, is to be seen as the defining feature of validity if not exactly the truth of that hypothesis? And if this is correct, then where does this leave the instrumentalist, positivist, and empiricist? Consider, for example, instrumentalism. This view does not invoke truth, but usefulness as the primary criterion for the adequacy of scientific theories and measurements. However, we are surely not seriously considering the idea that we have to rule out rivals to the hypothesis that intelligence tests are useful. The Wechsler Adult Intelligence Scale comes in a big heavy box, which is very useful to hit people on the head with, but the hypothesis that the WAIS is valid for inflicting physical injury is certainly not the kind of hypothesis we are interested in. Clearly, from the viewpoint of ruling out alternative hypotheses, the hypothesis that the test is useful is neither intended nor relevant, except for the very special hypothesis that it can be used to measure intelligence _because intelligence produces variations in IQ-scores._

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

### 6.4 Causality versus correlation

Although construct validity theory is, in its original form, inadequate, it does represent a serious attempt to forge a validity concept that has an account of meaning, a function for theory, and that stresses the fact that there is no essential difference between validation research and research in general. Moreover, if one removes the
nomological network from consideration, replaces meaning with reference, and rein­
troduces the realist perspective, much of what is said in construct validity theory remains consistent and plausible. Also, the idea of construct validity was intro­duced to get rid of the atheoretical, empiricist idea of criterion validity, which is a respectable undertaking because criterion validity was truly one of the most serious mistakes ever made in the theory of psychological measurement. The idea, that validity consists in the correlation between a test and a criterion, has obstructed a great deal of understanding and continues to do so. The concept continues to exert such a pervasive influence on the thinking of psychologists, because many are under the impression that construct validity is really criterion validity with the criterion replaced by the construct (this fallacy cannot be attributed to construct validity theorists, as is evident from the writings of Cronbach & Meehl, 1955; Kane, 2001; and Messick, 1981, 1989). However, the inadequacy of this view does not depend on whether one views the criterion as a variable to be predicted from test scores, or as an ‘infallible’ measure of the theoretical construct to be measured, or as the theoretical construct itself. The crucial mistake is the view that validity is about correlation. Validity concerns measurement, and measurement has a clear direction. The direction goes from the world to our instruments. It is very difficult not to construct this relation as causal. Criterion validity employs correlation and similarity, where it should employ direction and causality.

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

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

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

Third, the correlation is a population dependent statistic, that is, it is sensitive to the amount of variability in the attribute to be measured across populations. A well known instance is the attenuating effect of restriction of range in the presence of imperfect relationships between variables. Any correlational view must therefore hold that validity itself is by necessity variable over populations. Corrections for unreliability and restriction of range (Lord & Novick, 1968) are going to solve some of the trouble here but not all of it. In particular, there is one important, well-established case of valid measurement where the population dependence of correlations raises serious problems. This is the case of extensive measurement, as discussed in Chapter 4 (Campbell, 1920; Krantz, Luce, Suppes, & Tversky, 1971). This is very troubling because extensive measurement is more or less the paradigm example of measurement in general (Narens & Luce, 1986). In extensive measurement, attributes are not defined solely with respect to individual differences between objects (as is the case in almost all instances of psychological measurement), but with respect to an empirical concatenation operation. In this case, it can be meaningful to say that one is measuring one individual object (which is meaningless with interindividual difference variables). Now suppose we are measuring the length of rods, and that the measurement apparatus used is a meter stick. Further suppose that we are measuring without error. The correlation between the measurement outcome and the real length will be unity in most populations, as it should be, but there is an important class of populations where it will be zero. This is the population of rods of equal length. Therefore, we must conclude that, in this population, the centimeter is not valid for measuring length. This is a strange result. In
extensive measurement, it is quite meaningful to say that all objects in such a sub-population are, say, 4.2 feet long, and that this measurement is valid. In the causal account, this absurdity does not arise. This because causality is directional and conditional: The causal account says that, if there are differences in the attribute, then these will produce differences in the measurement outcome. However, if there are no differences in the attribute, no differences in the measurement outcomes are expected. This in no way precludes the validity of the measurement outcomes themselves, which is exactly as it should be.

Correlations are epistemologically relevant because they are sometimes indicative of causality, but they are not, and cannot be, constitutive of validity. I have dealt with the refutation of this view in somewhat greater detail than is perhaps necessary, as criterion validity has been considered inadequate at least since Cronbach & Meehl’s (1955) introduction of construct validity (Messick, 1989; Kane, 2001). A thorough refutation seemed important, however, because I am under the impression that many people, who do not subscribe to the criterion validity perspective, still have a correlational conception of validity—the only difference is that they have replaced the criterion with the construct itself. I propose that if attribute differences do not play a causal role in producing differences in measurement outcomes, then the measurement procedure is invalid for the attribute in question. Correlations are not enough, no matter what their size. Height and weight correlate about .80 in the general population, but this does not mean that the process of letting people stand on a scale and reading off their weight gives you valid measurements of their height. To state otherwise is to abuse both the concepts of measurement and of validity. In fact, I consider the very fact that a correlational view of measurement allows for this kind of language abuse as a fundamental weakness; and I suggest that any theory of validity that sustains such absurdities should immediately be dropped from consideration. I hope I have convinced the reader that not just criterion validity, but any correlational conception of validity is hopeless.

The causal view of validity is clearly very powerful in comparison to the correlational one. I have not been able to find any implications of it that are remotely near the aberrant behavior of the correlational conception. However, conceptual power always comes at a price, and this price is usually paid in metaphysical currency. That is, I have put causality to work, but this comes at the cost of introducing a heavy assumption into the proposed conception of measurement. Is it plausible that this workhorse will ride in psychology? In the present context, the main danger is that the causal account may seem to be just too strict for psychological measurement. There are measurement experts as well as psychologists who are under the impression that any causal account of psychological measurement is untenable. In particular, I anticipate the following argument.

It may seem that I am proposing that, for instance, John’s intelligence causes his IQ-score. This would require me to introduce an (at best) dispositional attribute as a cause. It is important to make clear that this argument does not apply. For if it did, it would not just be problematic but fatal to my position. However, the argument confuses the two distinct kinds of causal statements that have been discussed at length in Chapter 3. Specifically, the confusion arises from the singular use of the term ‘intelligence’. This usage seems to imply that, for a particular
subject, intelligence plays a causal role in producing test scores. This would indeed be a flawed account because intelligence is not the kind of variable that can be unproblematically introduced as a process variable. Intelligence is an interindividual difference variable, and as such it does not apply at the level of the individual. However, we do not have to assume that intelligence works at this level, because IQ-scores are not intended to measure intelligence in this way. Rather, differences in IQ-scores are intended to measure the effect of differences in intelligence. And in this sense, the causal account surely can be set up.

Thus, we do not have to suppose that intelligence causes IQ-scores in order to claim validity; we merely have to suppose that differences in intelligence cause differences in IQ-scores, which is a much weaker claim. This claim is not refuted by an argument against the use of dispositions as causes. We may remain silent on what happens at the individual level; in fact, we do better to refrain from introducing intelligence as a cause there. Interpreted strictly in terms of differences, I do not think the causal link proposed here is untenable. In fact, it does not seem to be all that controversial. The bold statement ‘intelligence exists’ will give rise to extended discussions among measurement experts, intelligence researchers, and at birthday parties. However, the statement ‘differences in intelligence exist’ is unlikely to elicit more than a faint smile. Similarly, to claim that your intelligence causes your IQ-scores will elicit your denial, and rightly so. But is it really so extraordinary to suppose that differences in IQ-scores are causally determined by differences in intelligence, especially when one considers that such a proposal does not presuppose that intelligence differences are the only cause at work in producing the IQ-score differences?

I think that such claims are not extraordinary at all, and that most measurement practices proceed along just this line of reasoning. In fact, as I have said before, the introduction of a causal line of reasoning is probably one of the few universals in measurement. Of course, latent variable models (Hambleton & Swaminathan, 1985; Bollen, 1989, 2002) explicitly incorporate this idea, because they can be viewed as common cause models (Reichenbach, 1956; Glymour, 2001). It may be less obvious that other approaches also take the causal stance, be it in a more indirect manner. For example, generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is certainly amenable to this analysis. While the idea of tests as samples from a universe of behaviors is not an explicitly causal one, generalizability theory surely assumes that differences in universe scores lead to differences in domain scores; at that level, it is not at all difficult to introduce a causal relation between the two concepts. Similar accounts can be set up for most measurement practices and models.

The only model that truly does not seem amenable to this analysis is the formative model discussed in Chapter 3 (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000), because, in that model, the values of the attribute are determined by the indicators rather than the other way around. However, I do not see this as a problem because the question, whether the weighted sum of the indicators ‘salary’, ‘quality of neighbourhood’, and ‘educational level’ yields a ‘valid’ measurement of SES, seems rather contrived in the first place. It may therefore be a good idea to call the formative model an instance of indexing rather than of measurement. Whether one
The problem of validity

has the right indicators for an index variable would seem a matter of convention and usefulness rather than of validity. This does not preclude that there may be sound arguments for including some indicators and not others; but I think that these considerations do not bear upon the question of validity. Of course, the question of validity may be raised at the level of the indicators (when one asks, for example, whether the question ‘what is your annual income?’ is a valid measure of annual income); but it does not apply at the level of SES because SES is not, properly speaking, measured but constructed.

In conclusion, the causal conception of validity avoids the absurdities of a correlational perspective because it is directional and conditional. So is measurement. And although it is true that the statement, that an attribute like intelligence causes the IQ-scores for an individual subject, is either meaningless or false, this argument does not pose a serious threat to the causal interpretation at hand. Like perception, measurement is about detecting variations, contrasts, and differences; and the only statement we need to make to claim validity is that the measurement instrument will detect the relevant variations, i.e., that variations in the attribute will cause variations in the test scores. Some reflection shows that this line of reasoning is not all that extraordinary, but underlies most or all measurement procedures; and it is certainly the case that most measurement models allow for this interpretation. The only model that does not sit well with this interpretation is the formative model. However, that this model is excluded from consideration in a validity context does not seem to be a problem for, but rather a virtue of, the present conception. This is because the formative model is, in my view, not a model for measurement but for indexing. It seems to me that a causal interpretation of validity is reasonable, and I propose it be considered in the literature.

6.5 Tests versus interpretations

Test theory abounds with unlucky terminology; some of the more infamous examples are ‘true scores’ (Lord & Novick, 1968; Borsboom & Mellenbergh, 2002), ‘admissible transformations’ (Stevens, 1946; Lord, 1953), ‘meaningfulness’ (Suppes & Zinnes, 1963; Michell, 1986), and ‘reliability’ (Lord & Novick, 1968; Lumsden, 1976; Mellenbergh, 1996). All of these terms indicate important concepts, but in every case the label is awkward, because it suggests an unintended meaning. It almost seems as if the one thing that measurement theorists have in common is the curious ability to flawlessly pick the one name for a concept that will guarantee its misinterpretation. And as is the case for most concepts in test theory, the term ‘validity’ is not very well chosen.

In particular, ‘valid’ is an adjective that may naturally be applied to arguments, statements, theories, and judgments, but to apply it to nonlinguistic entities like tests is to stretch the grammatical limits of natural language. Tests are instruments, they serve a purpose and may be useful in this respect, but to say that they are valid seems just as absurd as to say that they are true. Tests do not purport to measure anything, and neither do scores. It is us, the investigators, who desire to measure attributes; it is us who interpret the scores; and it must be us who present the
validity argument. Therefore, it can only be the test score interpretation (Cronbach & Meehl, 1955; Messick, 1989), or else the argument that justifies the interpretation (Kane, 1992; 2001), to which the adjective ‘valid’ may apply – but not the tests themselves. So the argument of modern validity theory goes.

Although most of the steps taken in the above argument are based on plausible ideas, I do not find the argument very convincing on the whole. It is certainly true that to apply the term ‘validity’ to tests is not to use the term in a natural manner, but terminology is just a matter of convention and I cannot help it that this particular terminology has been introduced. It would be a good idea to change it, but history teaches that any attempt to change a term so deeply entrenched as ‘validity’ is guaranteed to fail (Ebel, 1956, is an example in this context). It is, however, my conviction that the old fashioned way of using the adjective ‘valid’, which is to apply it to tests and not to test score interpretations, singles out a very important property of tests. And it is exactly this property that provides such a useful vehicle for saying that we have succeeded in measuring what we set out to measure. To ascribe to the test this property is meaningful and I am willing to defend this practice. The validity literature has taken the other option and has deserted this usage (and, in doing so, left almost the entire research community behind). It is now standard for a validity theorist to say that validity applies only to test score interpretations, with the possible extension to actions based on test scores (Messick, 1989).

There are several reasons why this conception is not optimal. First, one can construct cases where a test score interpretation is valid but the test evidently is not. I have, for example, developed a new test. It is called the number test and contains one question. The question is ‘write down a number between 70 and 130’. I have the following interpretation for the scores: ‘the test scores resulting from administering the number test do not measure any psychological attribute whatsoever’. I have done a number of studies that provide strong evidence for the validity of this interpretation. For example, it turns out that the test scores do not correlate with height, IQ, and extraversion. It seems to me that nobody can reasonably dispute that this interpretation of the scores on the number test is valid, and that the evidence strongly supports this conclusion. It also seems to me that nobody can reasonably dispute that the test is invalid; for it does not measure any interesting attribute whatsoever. It is thus sensible to say that the interpretation ‘the number test is invalid’ is itself valid. This establishes that the validity of tests and the validity of test score interpretations are quite distinct topics. I think that the validity of tests is the topic of interest in psychological measurement.

One may reply that this is an unfair example, because Cronbach & Meehl (1955), Messick (1989), and others meant their definitions to apply to a limited set of test score interpretations, and not to all such interpretations. But how are we to determine which interpretations are eligible for consideration? I think that some reflection will show that there is basically just one class of interpretations in which we are interested. These are interpretations of the form ‘test X measures attribute Y’. Maybe I am missing something here, but it seems to me that this gets us back to square one with a vengeance. For is not the only condition, that unambiguously sustains the validity of the interpretation ‘IQ-tests measure intelligence’, the con-
The problem of validity

dition that the proposition expressing this interpretation is true? But if this is the case, then the term ‘validity’, as applied to test score interpretations, turns out to do nothing more than the concept of truth was already doing (Borsboom, Van Heerden, & Mellenbergh, in press). That is, what prohibits us from saying that IQ-tests are valid for measuring intelligence if and only if the proposition ‘IQ-tests measure intelligence’ is true? And if I am correct on this score, then where does the concept of validity come in at the level of test score interpretations?

The answer is that validity, as applied to interpretations, must be introduced to deal with the epistemological side of the question, and in fact cannot be introduced at any other level. We cannot know whether a test score interpretation is true any more than we can know whether quantum theory is true, for we have no conclusive method of verification. However, what we can do is to evaluate the evidence and theory supporting the interpretation at hand. This is why Messick (1989), Kane (2001), and other theorists see validity as an evaluative judgment, and not as a property of tests. I have no quarrel with the importance of the evidential, theoretical, and consequential issues involved here; but I seriously doubt whether we need to address epistemological issues when defining and delineating the validity concept. In this context, it is important to realize that epistemology is in a quite hopeless state at the present time: The stage is crowded with philosophers subscribing to relativism (Meiland, 1977), scientific realism (Devitt, 1991), constructive empiricism (Van Fraassen, 1980), falsificationism (Popper, 1959; Lakatos, 1978), social constructivism (Latour, 1987), postmodernism (Foucault, 1970), and to smaller movements based on bootstrapping methodology (Glymour, 1980), logical reliability (Kelly, 1996), problem solving (Laudan, 1977), computational approaches (Thagard, 1988), and game theory (Hintikka, 2001). No epistemological criteria for truth or validity are accepted by more than a handful of philosophers of science, and legitimate doubts can be raised as to whether such criteria can be found at all. Therefore, a definition of validity in terms of the quality of the argument put forward, or in terms of the evidence adduced, is unlikely to provide a firm foundation for the concept. Moreover, once we turn to the epistemological side of the problem, we are in no position to claim that we are specifically discussing it with respect to psychological measurement. We are discussing the validity of interpretations in general. So, this is rapidly becoming a very ambitious project. For now we will have to find our way through the epistemological labyrinth that generations of philosophers have so carefully crafted – and given the nature of philosophers, the exit is likely to be missing. Do we really want to go there?

I think that no such expedition is called for. In fact, I submit that the focus on test score interpretations, as opposed to tests, is yet another instance of the emphasis that validity theory has come to place on epistemology, where it should be concerned with ontology. An ontological framework that unambiguously defines what it means for a test to be valid need not be so complicated. I think that I am putting forward an adequate, and yet simple, proposal in this very chapter. The property of tests that we indicate with the admittedly infelicitous term ‘validity’ is just that variations in the attribute to be measured produce variations in the test scores. This is most certainly not a property of test scores or of test score interpretations. It is the test that does the job here, and it is the test to which
the property of validity should be assigned. The test functions as a gateway from the world to us, and if this gateway happens to convey the effect of variations in an intended attribute (like intelligence), and blocks the influence of many other attributes (like height, weight, extraversion, and so on), then that is the property we are interested in. I do not deny that the interpretation of test scores in theoretical terms is important, and there is a nontrivial sense in which modern validity theorists are justified in placing the emphasis here – namely, when one approaches the issue from an epistemological point of view. The argument put forward to convince other researchers that the test has the property of validity (Kane, 1992; 2001) is also important, and worthy of study in its own right. Likewise, the justification for using test scores in selection or placement is important; such justifications should be thoroughly scrutinized, and not just on scientific grounds (Messick, 1989). But not every important aspect of tests and test use is relevant to the validity concept. In fact, I think that most of the issues on which validity theory has focussed in recent years are not directly relevant to the concept of validity. Many of these concerns follow directly from shifting the emphasis from tests to test score interpretations. I think that this move is both unnecessary and inadequate. Validity should therefore be reconceptualized as a property of tests.

6.6 Simplicity versus completeness

The development of validity theory in the course of the 20th century shows a consistent movement towards a greater scope for the concept. The original formulations were more or less technical in nature, stressing primarily the size of validity coefficients. However, at least since the work of Cronbach & Meehl (1955), who made the concept depend on nomological networks, validity theory has aimed at completeness. I think that the concept should be kept as simple as possible. Validity is a central concept in psychological testing, but not in the sense that it embraces and incorporates every important consideration in test use.

The most elaborate attempt to present a complete theory of validity that covers every aspect of tests is the treatise by Messick (1989). This theory is much too big to warrant a fair discussion here, but I would like to highlight some of the major difficulties I see in this kind of unified validity concept. In particular, I want to consider briefly Messick’s progressive reading of his famous faceted conception of validity (Messick, 1989; p. 20-21). This reading suggests that scientific evidence, predictive utility, value implications, and social consequences add up to form a kind of total validity. In Messick’s words, ‘evidence of the relevance and utility of test scores in specific applied settings, and evaluation of the social consequences of test use as well as of the value implications of test interpretation, all contribute in important ways to the construct validity of score meaning’ (Messick, 1989, p. 21). This seems to imply that, say, the use of IQ-scores in personnel selection could be approached through a maximization of the evidential basis, predictive utility, and social benefits of such use. While there may be cases where such maximizations are possible in the additive sense suggested by Messick, in many cases we will be faced with a trade-off instead. To make this point clear, I briefly discuss two cases
where the interests represented by the different cells of Messick's (1989) matrix will juxtapose each other, rather than align.

**Case 1: Measurement invariance, prediction invariance, and fairness**

First, consider the measurement versus prediction invariance paradox as discussed by Millsap (1997). The problem concerns the fact that equal regression lines across groups of, say, job performance on IQ, are generally inconsistent with the hypothesis that the same factor model underlies the IQ-scores in each group. This becomes especially problematic if groups differ in the variances of the latent variable (which will be the rule rather than the exception). In this case, if we have measurement invariance (no bias) across groups, then we will have unequal prediction slopes. If we have equal prediction slopes, however, we will have a test that violates the requirements of measurement invariance. Now suppose we are involved in a selection problem, say, we need to select people for access to an educational program in some medical specialization. Knowing the excellent predictive properties of IQ-tests, we decide to base the selection procedure on IQ-scores. Further suppose that we are selecting people from different ethnic groups. Obviously, we want the best people for the job; and at least from a public health perspective, this is surely in the best interest of society. This requires maximizing predictive utility. We also want to have a good scientific basis, which requires factorial invariance. Finally, we would like a fair and transparent selection procedure, which implies a rule like 'whatever ethnic group you belong to, if your IQ-score exceeds 120, then you are in'. Now, what strategy should we follow to jointly maximize construct validity, predictive utility, and fairness?

It is immediately obvious that no such strategy exists. If we, for example, maximize construct validity, then we ought to use an unbiased measurement instrument, i.e., a test that shows factorial invariance over ethnic groups. If we also want to maximize predictive utility, then we ought to allow the predictive regression lines to differ over ethnic groups. However, if we now set a desired standard on the criterion (job performance), we will have to set a different cut-off score in each of the groups. But this would imply that a person from ethnic group A is selected if her score exceeds, say, 115 on the IQ-test, while a person from ethnic group B is selected if her score exceeds 120. This does not seem fair to us, and it certainly would not seem fair to the general public (in whose interest we are doing this — remember that the goal is to get the best people for the job). It would take a very good psychometrician to explain this practice in court. On the other hand, if we select an instrument with equal regression slopes, we can jointly maximize predictive utility and fairness, but construct validity would have to suffer, because we would have to select a test that violates factorial invariance (i.e., a biased test). Finally, if we jointly optimize construct validity and fairness, by using a factorially invariant test but setting the same cut-off score for all applicants, we will have to give in on the predictive front, and our selection will not be optimal; in this case, we would not get the best doctors. Clearly, then, construct validity, predictive utility, and social consequences do not 'contribute' to the overall validity of the selection procedure, except for the fact that their interplay presents us with a difficult moral dilemma.
Case 2: Homogeneity in measurement is multicollinearity in prediction

The conjunction of scientific concerns and issues of social consequences will give rise to paradoxical situations and moral dilemma’s. This is because the goals of science and society often conflict. However, conflicting interests can also arise when no social consequences are in play. An interesting case concerns the problem of choosing between optimizing measurement or predictive properties of tests. It has often been suggested that these go hand in hand, and that optimizing validity will pay itself back through improved prediction. There are actually strong reasons to suspect that the opposite is the case (Lord & Novick, 1968, p. 332; Smits, Mellenbergh, & Vorst, 2002). Optimizing measurement properties will, in general, lead to suboptimal predictive properties. More seriously, however, optimizing predictive properties will tend to destroy all measurement properties a test might have.

Measurement is typically approached through latent variable models, such as the various item response theory models (Hambleton & Swaminathan, 1985) or common factor models (Jöreskog, 1971; Bollen, 1989). Inspecting the structure of any latent variable model will show that items that measure the same latent variable must be correlated. Now suppose that we want to construct a test to measure extraversion. When optimizing measurement properties, we will construct a homogeneous test, i.e., a test with correlated items. If we now were to use the test for prediction, this property, which is desirable from a measurement point of view, would translate into a problem from a prediction point of view. In essence, any set of items selected on the basis of measurement properties will show multicollinearity in prediction: The items will not add independently to the regression equation. Thus, focussing on the measurement properties by necessity leads to suboptimal predictive properties.

On the other hand, if we focus on prediction instead of measurement, we will turn up with a completely different test. Suppose that we construct a test to maximally predict extravert behavior, for instance, the tendency to engage in group discussions. When optimizing this prediction, we are likely to select items that add to the regression equation independently. That is, we would avoid rather than produce multicollinearity, and therefore we would select uncorrelated items rather than correlated ones (Lord & Novick, 1968, p. 271). Optimizing prediction necessarily produces a set of items that add to the prediction equation independently of one another, and therefore such procedures are likely to select items that measure different things. This problem will occur even if we are selecting items for predicting a highly relevant behavioral domain. Thus, in selecting items on the basis of predictive criteria, we will produce a heterogeneous test and not a homogeneous one. And because all measurement models imply homogeneity, such models will hardly ever fit a test constructed for optimal prediction. The reason for this is that the selected items will not measure the same attribute, and therefore cannot possibly be valid for measuring one attribute.

Measurement and prediction will, in general, not go hand in hand. In fact, they will work against each other. This trade-off occurs because the structure of the prediction problem is simply radically different from the structure of the measurement problem. Optimal predictive properties imply suboptimal measurement properties, and optimal measurement properties imply suboptimal predictive properties.
The problem of validity

Again, it is clear that we cannot have it all at the same time.

Is validity a matter of degree?

The above examples point to a serious problem for any ‘overarching’ conception of validity. Even such simple and commonplace problems as measurement and prediction are not structured in a way that allows for a simultaneous maximization of desirable measurement and prediction properties. It is therefore impossible to improve a test on all the relevant fronts at the same time. There is no inherent problem about this, but a serious problem will occur when we couple these observations with the notion of an overarching validity conception that is supposed to come in degrees. That validity is a matter of degree has become more or less a dogma of construct validity. Cronbach & Meehl (1955, p. 290) state that ‘the problem is not to conclude that the test “is valid” for measuring the construct variable’, but that ‘the task is to state as definitely as possible the degree of validity’. Similarly, Messick (1989, p. 13) writes that ‘it is important to note that validity is a matter of degree, not all or none’. In view of the above examples, I doubt whether this view is adequate.

What exactly does it mean to say that validity comes in degrees? It seems that a theorist who expresses this notion is saying that different tests, or test score interpretations, can be ordered in terms of their validity. But how does this work? Can we say, for example, that the Stanford-Binet is more valid for measuring intelligence than the WAIS? This seems relatively unproblematic, but appearances deceive here. In particular, it is unclear how we should determine this degree of validity. We could imagine that, in a given population, intelligence produces a proportionally larger amount of the variance in WAIS scores than in Stanford-Binet scores. This may be taken to imply that the WAIS is now ‘more valid’. However, the situation here is by no means inconsistent with, for example, the presence of a large biasing effect in WAIS scores. So we could easily have a situation where the WAIS has the larger portion of variance produced by intelligence, but where it is also seriously biased against, say, females. The Stanford-Binet may be more unreliable but unbiased. How are we to weigh these different merits in determining which test has the higher degree of validity? It seems to me that this will be quite difficult. Moreover, the problems are going to multiply very quickly if we now move to different domains of interest. What kind of argument would it take to say that the WAIS is more valid for intelligence than Eysenck’s extraversion scale is for extraversion? To be honest, I do not have a clue. And although one could imagine a kind of ‘validity score’ for each test, which could be considered a weighted sum of desirable characteristics (e.g., reliability, absence of bias, predictive utility, unidimensionality, etc.), the problem becomes insurmountable once we move to a comparison of tests which are intended for different purposes, such as measurement and prediction.

The Minnesota Multiphasic Personality Inventory (MMPI), for instance, has been explicitly constructed with the objective of maximizing its predictive performance with respect to clinical syndromes. The test could reasonably be said to exemplify a predictive instrument, rather than a measurement instrument. But how are we to determine whether the MMPI is more valid for predicting mem-
bership of various categories of mental disorders, than the WAIS is for measuring intelligence? This seems downright impossible. If we have two tests which were developed for the same objective, say, measurement, then we could at least imagine a kind of weighted sum of desirable measurement characteristics (provided that we agree on these). But if we have two tests which were developed for different objectives, like measurement and prediction, we can no longer do this, because there are few if any characteristics which are desirable both in measurement and in prediction. Because the structure of the measurement and prediction problems are different, the desirable test characteristics and their relative importance will be different, and therefore we cannot construct a meaningful comparison. The tests, or test score interpretations, are incommensurable: Attempting to place them on a ‘validity scale’ which comes in degrees is like trying to answer the question whether the U.S. baseball team is better at playing baseball than the Dutch soccer team is at playing soccer.

Saying that validity is a matter of degree implies that one can order tests, or test score interpretations, in terms of their validity. Coupled with the desire to apply the label of ‘validity’ to all possible instances of test use or test interpretation, this proves to be very difficult, if possible at all. It is questionable whether tests that are constructed with different purposes in mind (measurement, prediction, selection, etc.) are scalable on their ‘degree of validity’, and to the best of my knowledge no procedure for doing this has been proposed. I submit that no such procedure can be found at all, simply because of the fact that we are dealing with different problems. Now, the conception of validity proposed here is one concerned with measurement. Therefore, I wish to exclude tests designed for prediction or selection from consideration, except for the case where one is explicitly claiming validity in terms of measurement for such tests (which one is by no means forced to do). This does not mean I want to go back to the ‘tripartite’ scheme, which distinguished different kinds of validity (namely, content, construct, and criterion validity). I think that the question of validity is one of measurement and that, as a result, content and criterion validity are of relatively minor importance. Thus, I strongly endorse Messick’s (1989) suggestion to consider these as questions of content relevance and coverage instead of ‘content validity’, and of predictive utility instead of ‘criterion validity’.

I differ in opinion, however, with respect to the question whether validity should incorporate all these aspects at the same time; I answer this question negatively. In particular, I see no reason to demand that a test, which is intended for prediction, should first be shown valid for measuring an attribute of interest. Thus, I do not endorse Messick’s progressive reading of his validity matrix. Further, I consider the task raised by Cronbach & Meehl (1955), which is to state as precisely as possible the ‘degree of validity’, to be impossible. For a validity concept that comes in degrees requires an ordering of tests, or test score interpretations, with respect to their degree of validity. These tests and test score interpretations, however, will more often than not be incommensurable, so that it is meaningless to speak of their degree of validity. I think that most researchers realize this at some level or another, because nobody ever attempted to develop a construct validity coefficient to capture the implied ordering. Predictive utility does come in degrees, for the
simple reason that it is a direct function of the association between test scores and the variable to be predicted. Validity does not, because it is not a function of any such association.

A simple concept for a simple question

The desire to create a validity concept that comes in degrees seems to result from the fact that there are two questions, both important, that can be asked of any measurement procedure. First, one may ask: ‘does the test measure the intended attribute?’ Second, one may ask: ‘how well does the test measure the intended attribute?’ In my view, the first question can only be answered dichotomously: Either differences in test scores are produced by variations in the attribute of interest, or they are not. The second question addresses the quality of the test, relative to several methodological concepts like reliability, measurement invariance, and unidimensionality. This question is conditional: it is sensible to ask how well a test measures an attribute only if the test does indeed measure it. Thus, asking for an overall evaluation of the quality of the test presupposes its validity for the intended attribute. The question is whether we should intend to cover both questions with one concept. I think it is more sensible to restrict the meaning of validity to apply only to the question whether variation in test scores is produced by variation in the attribute we intend to measure, but not the question how well we measure it. This is a natural consequence of conceptualizing validity in terms of causality: In contrast to a correlation, a causal relation does not come in degrees. I prefer to leave the question how well a test measures an attribute to the various technical approaches that have been proposed in the psychometric literature. I thus divorce validity from various psychometric concepts that are explicitly concerned with the question how well we are measuring the attribute of interest, such as reliability, unidimensionality, and bias.

Reliability

Within the present conception, the problem of reliability is distinct from the problem of validity. The square root of the reliability coefficient is certainly not an ‘upper limit’ for validity, although it does pose an upper limit for predictive utility. But predictive utility is irrelevant to the question of validity as I construe it. The test can be valid in a given setting (i.e., it measures what it should measure), but very unreliable. The traditional question, how much of the total variation in test scores is accounted for by the intended attribute, cannot be an issue of validity because it is dependent on the variation in the population. In a population where there is no variation in the attribute, none of the variation in the test scores is produced by variation in the attribute. As I have argued above, this does not preclude the test from being valid, because the present account of validity is causal, not correlational, and in saying a test is valid, one is saying only that variations in the intended attribute will produce variations in test scores if the attribute variations are present. Validity is a property of the test, and not of the scores, so it should not vary with populations – except for the fact that there may be populations in which the attribute does have causal relevance for the test scores, and populations where it does not. However, while reliability does not place
6.6 Simplicity versus completeness

an upper limit on the validity of a test, validity does place a strong restriction on the applicability of the reliability concept. This is because reliability is an index of measurement precision (Mellenbergh, 1996). Thus, the question that reliability is concerned with is 'how precise are our measurements?'. Obviously, the entire notion of measurement precision presupposes that the test is valid: We cannot say that the IQ-scores measure intelligence with a certain precision, but that they do not measure intelligence. Reliability is not an upper limit for the 'degree of validity', but it is the case that invalidity prohibits any statement concerning the reliability of test scores. If a test is invalid, then the scores cannot reliably measure the attribute of interest, because they do not measure the attribute of interest at all.

Unidimensionality A second question that is highly relevant to the question how well we measure the intended attribute, but not to the question whether we measure that attribute, is the concept of unidimensionality. No psychological test is unidimensional in the sense that the test measures only one attribute in every imaginable situation. All psychological test scores depend on many attributes that, if they were to vary in a population, would cause variation in test scores. This is clear from the fact that we may always create a second source of variation in addition to the one we are studying, thereby creating multidimensionality. For instance, one may measure the genetic quality of seeds by recording the height of the grass they produce. This test is clearly valid because variations in genetic quality, if present, will produce variations in the height of the grass. But if a gardener has just mown half of the lawn, we will in addition measure the effect of his presence, which obviously also produced variation in the height of the grass. The reason is that the test is also valid for measuring the presence of lawn mowers: if there is variation in this presence (i.e., the lawn mower covers some areas, but not others), then this will produce variations in the height of the grass.

Tests themselves are therefore always 'multidimensional' insofar as this term applies to tests at all. IQ-tests may be valid for intelligence; but they will certainly also be valid for dyslexia, motivation, and reading ability. What we must always do when we are trying to single out one attribute, is to secure that, in the population we are working in, there is no variation in other attributes. This means that we must ensure that these other attributes, for which the test is also valid, do not come into play. We may do this in two ways. First, by carefully selecting a population where we can assume zero variation on the other attributes that may cause variation in scores. For the intelligence example, this means that we should exclude people with dyslexia or impaired reading ability from consideration. Second, we can attempt to block the effects from other variables that may cause variation in scores. This is commonly done in personality tests, for example, by telling the subject that there are no correct or incorrect answers. In that case, we are trying to block effects from, say, variations in the tendency to answer in a socially desirable manner. Now, if we succeed in our attempt to exclude other causally relevant variables from operating, then we would expect the resulting scores to fit a unidimensional model. It is clear, however, that whether such a model fits is completely dependent upon our success in blocking effects of other variables.
It is also clear that we can use a test to measure different attributes in different situations. When we try to measure intelligence, it is important to ensure that systematic variation in other causally effective attributes is minimal. However, if variations in IQ-scores also depend on variations in dyslexia, this means that we could, in principle, also use a verbal IQ-test to measure dyslexia. This could be done by administering the IQ-test to populations that are homogeneous in all attributes (including intelligence) except for dyslexia, or by blocking all other effects. Of course, this would be utterly impractical, and I am not recommending any such use of IQ-tests, but the conceptual point is clear: A test can be used, in principle, to measure any attribute that produces variations in the test scores. Unidimensionality can be created, in a sense, by administering tests to populations that are homogeneous with respect to systematic variation in all attributes but the intended one, or by blocking the effects of all unintended sources of variation. However, a test can never be used to measure an attribute that does not causally produce variations in test scores. Thus, the relation between validity and unidimensionality is similar to the relation between validity and reliability: Unidimensional measurement of intelligence presupposes the validity of the test for intelligence, but validity does not—and cannot—presuppose unidimensionality.

Bias The above comments are directly related to the problem of bias, or differential item functioning (Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993). Bias, like unreliability and unidimensionality, refers to the question how well we measure an attribute. In particular, it is concerned with the question whether there exist subpopulations which induce multidimensionality (Shealy & Stout, 1993). An intelligence test is biased against ethnic groups, for example, if the scores depend on an attribute different from intelligence (say, the familiarity with the English vocabulary), on which the ethnic groups score systematically lower. If the presence of bias is established, however, it is not established that the test is invalid. It is shown that the test is sensitive to variations in more than one attribute, but not that the test is insensitive to variation in the intended attribute. If the intelligence test in the above example is biased, this does not imply that the test does not measure intelligence; it may be that the test measures more than intelligence alone, or that it measures intelligence differently in the different groups (Borsboom, Mellenbergh, & Van Heerden, 2002-b). The presence of bias is thus not directly relevant to the question of validity as we have construed it here. This may sound counterintuitive, but it is in accordance with the technical formulation of item bias. This is because the formulation of bias involves an effect of the grouping variable on the expected item response, conditional on the latent variable (Mellenbergh, 1989; Meredith, 1993). However, if no latent variable underlies the test scores at all, the concept of bias cannot even be formulated. Therefore, it seems that the formulation of bias presupposes validity, rather than that validity presupposes the absence of bias.

Validity and psychometrics Validity, as it is conceptualized here, is thus the point of departure rather than of arrival. It is central to test development and use, but
it does not embrace all aspects involved in these practices. It is underlying rather than overarching, simple rather than complicated, and basic rather than unified. Psychometric approaches, which generally deal with the question how well we are measuring, cannot be considered to formulate necessary or sufficient conditions for validity because they presuppose validity. Therefore, it is plausible to separate the question whether we are measuring the right attribute from the question how well we are measuring that attribute. I think that validity theory should refrain from trying to answer both questions with a single concept; not only are they radically different – one is substantive, the other methodological – but, in the past century, the technical issues involved in assessing the latter question have been largely taken over by psychometric approaches. These approaches handle the questions involved in a much more sophisticated manner than conceptual treatments of validity could ever hope for. How precise we are measuring the attribute is a question for theories of measurement precision (Mellenbergh, 1996); whether we measure only one attribute is a question of unidimensionality (Hambleton & Swaminathan, 1985); whether we measure primarily one attribute is a question of essential unidimensionality (Stout, 1991); to what extent our measurements are unbiased is a question of measurement invariance (Millsap & Everson, 1993); and so on. If there is a function here for a unified validity concept employing degrees of validity at all, it would have to involve a method of translating these characteristics into a single number. How this should be achieved is, to my knowledge, unknown. And it is very interesting to see that psychometrics has not developed the need for a validity concept; while concepts like measurement precision, unidimensionality, and invariance flourish, there is almost no psychometric literature which explicitly uses the validity concept itself. The reason for this is that validity theory has no business in psychometrics. Not because validity is irrelevant, but because the entire undertaking of psychometrics presupposes it.

6.7 Discussion

I have proposed a simple conception of validity that concerns the question whether the attribute to be measured produces variations in the measurement outcomes. This concept of validity is based on reference and causation, rather than on meaning and correlation. As a result, it is an all-or-none property. Moreover, it is a property of tests, and not of scores or of test score interpretations. Although epistemological issues are central to validation, and consequential issues are central to test use, both are considered irrelevant to the concept and definition of validity itself. The conjunction of these theses produces a viewpoint that is almost diametrically opposed to the currently endorsed conceptions of validity, which state that the concept applies to test score interpretations, that it depends on nomological networks or is at the very least theory-dependent, that it is complex and faceted, and that social, ethical, and political consequences are relevant to validity. I do not see the need for a ‘unified’ validity concept (Messick, 1989; Moss, 1992; Shepard, 1993; Ellis & Blustein, 1991), because I think there is nothing to unify.

Although the proposed validity concept may be dissonant with the current va-
The problem of validity

In the validity literature, few of its ingredients are truly new. In particular, several related ideas have been put forward by a number of scholars in the previous century (e.g., Cattell, 1946; Campbell, 1960; Loevinger, 1957; Kelley, 1927; Popham, 1997). A realist reading of construct validity also comes very close to the conception proposed here. In addition, I am under the impression that most researchers operate with a validity concept that is highly similar to the one I am proposing. It seems, however, that nobody has yet consistently followed through the consequences of a realist conception of psychological attributes for the concept of validity; and the emphasis on causality as opposed to correlation seems never to have been stressed. As I have argued in the present work, the consequences of such a conception are far-reaching, but the overall picture that emerges is consistent and fits the intuitive notions most researchers have about validity quite well. I therefore think that the proposed validity concept is a viable alternative to the current consensus in validity theory.

The philosophical assumptions involved in the present conception are strong; stronger, perhaps, than in any previous discussion of validity. Therefore, it may be argued that, by invoking real entities and causal relations, I am engaging in metaphysical speculation. I concede this point, but it does not bother me. The very idea, that metaphysics and science are necessarily opposed, is a relic that stems from logical positivism; in fact, I think that science is the best way of doing metaphysics we know. To the hard-boiled empiricist, I reply that it is naive to think that any scientific theory can get off the ground without introducing an ontological picture of how the world works, which will always contain metaphysical ingredients. Given that this is the case, the metaphysics better be good. Other objections may come from the postmodern or social constructivist camp. An obvious one is the objection that psychological attributes are social constructions, and that I am engaging in an unjustified reification of such constructions. To this objection I reply that the position taken here is indispensable for rendering a coherent picture of measurement. It is an ontological attitude one has to take.

To see this, consider the following example. We may measure the degree of aggressive behavior displayed by Donald Duck, Mickey Mouse, and Woody Woodpecker, by rating the number of aggressive acts in a randomly sampled five minutes of film. I am surely not going to deny that Donald Duck, Mickey Mouse, Woody Woodpecker, as well as their aggressive behavior are social constructions; but this is completely besides the point. The point is that, even in this highly contrived situation, the logic we are following is that differences between Donald Duck and Mickey Mouse in their universe scores of aggressive behavior will lead to differences between them in the five minutes of film we sampled. We are thus presupposing the existence of such a universe score, and we are also presupposing a causal relation between this score and the number of aggressive behaviors we have observed: If there are no differences in universe scores, then we expect no differences in the number of aggressive behaviors, but if there are, then we expect these to lead to differences in the number of aggressive behaviors. If we are going to measure something, then we will have to suppose its existence and causal impact. Whether that something is, in itself, more properly conceptualized as a construction or as some kind of natural phenomenon is irrelevant to this issue.
Although I have separated the ontological concerns in psychological measurement, among which is validity, from the epistemological ones, which include validation strategies, the present developments do have some relevance in the area of validation research. In particular, it seems that the emphasis on the role of constructs in theories, and their place in nomological networks, has prompted validation research to adopt what has been called a top-down strategy (Cervone, 1997). This basically comes down to the fact that much validation research is concerned with creating tables of correlation coefficients, and then checking whether these go in the right direction. While I do not deny the relevance of such macro-level relations, it would seem to me that the primary objective of validation research is not to establish that the correlations go in the right directions, but to offer a theoretical explanation of the processes that lead up to the measurement outcomes. That is, there should be at least a hypothesis concerning the causal processes that lie between the attribute variations and the differences in test scores. To use Embretson’s (1983) terminology, validation should be concerned primarily with construct representation and only secondarily with nomothetic span.

In this view, validation is not, and cannot be, a purely or even mainly methodological enterprise. This does not mean that methodological and psychometric techniques are irrelevant to validation research, but that the primary source for understanding how the test works must be substantive and not methodological. Thus, I consider it impossible to argue for test validity solely on the basis of a multi-trait multi-method matrix. Such a matrix is helpful, but I do not view a favorable matrix configuration as constitutive of validity. What is constitutive of validity is the existence of an attribute and its causal impact on our scores. Therefore, if one does not have an idea of how the attribute variations produce variations in measurement outcomes, one cannot have a clue as to whether the test measures what it should measure. No table of correlations, no matter how big, can be a substitute for knowledge of the processes that lead to item responses. The knowledge of such processes must be given by substantive psychological theory and cannot be based on methodological principles. There are certainly tests for which a considerable body of knowledge has accumulated in this respect. Examples of research in this direction are, for instance, the cognitive modeling approach in spatial reasoning tests (Embretson, 1994) and the latent class approach in the detection of developmental stages (Jansen & Van der Maas, 1997). I think we are more likely to find evidence of validity in such explicit attempts to model respondent behavior, than in tables of correlations.

The upshot of this line of reasoning for test construction is also clear. Purely empirical methods, like those used in the construction of the MMPI, are very unlikely to generate tests that can be considered valid measurements. This is because focussing on predictive properties will destroy, rather than enhance, measurement properties such as validity (note that this does not preclude that these tests may be highly useful for prediction). Thus, it seems that one has to start with an idea of how differences in attributes will lead to differences in test scores; otherwise the project of test construction is unlikely to generate tests that are valid for more than prediction. This may be one of the few instances where psychology may actually benefit from looking at the natural sciences. In the more exact quarters, nobody
starts constructing measurement instruments without the faintest idea of the processes that lead to the measurement outcomes. And, interestingly, the problem of validity appears never to have played the major and general role it has played in psychology. These two observations may well be related: The concept of validity may never have been necessary because the instruments were generally set up based on an idea of how they would work. In that case, the question what it is, precisely, that is measured, can simply be resolved by pointing to the processes that lead to the measurement outcomes.

In contrast, the question what psychological instruments measure is generally not answered by pointing to the way the instruments work, but by pointing to the relation they have with other instruments. This way of working makes the question 'what is measured?' a question to be answered after the test has been constructed. Thus, the contrast here is between a conception that sees validity as something that one puts into an instrument, and a conception that views validity as something to be discovered afterwards. Construct validity theorists have tended to construe validity as an empirical matter, that is, the question what is measured is to be answered by data. However, a century of experience with test construction and analysis clearly shows that it is very hard to find out where the scores are coming from, if tests are not constructed on the basis of a theory of item response processes in the first place. Therefore, I would like to push the proposed validity conception one step further, and to suggest not only that epistemological issues are irrelevant to validity, but that their importance may well be overrated in validation research too. A large part of test validity must be put into the test at the stage of test construction (see also Schouwstra, 2000), a stage of the testing process that has received little attention compared with the enormous emphasis that has been placed on test analysis. Thus, it is suggested here that the issue may not be first to measure, and then to find out what it is you are measuring, but rather that the process must run the other way. It does seem that, if one knows exactly what one intends to measure, then one will probably know how to measure it, and little if any validation research will be necessary. If this is correct, then the problem of validation research is not that it is difficult to find out what is measured; the problem is that it is difficult to find out what we intend to measure.