PSYCHOMETRIKA—VOL. 71, NO. 3, 425–440
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1447-6

THE ATTACK OF THE PSYCHOMETRICIANS

DENNY BORSBOOM

UNIVERSITY OF AMSTERDAM

This paper analyzes the theoretical, pragmatic, and substantive factors that have hampered the integration between psychology and psychometrics. Theoretical factors include the operationalist mode of thinking which is common throughout psychology, the dominance of classical test theory, and the use of "construct validity" as a catch-all category for a range of challenging psychometric problems. Pragmatic factors include the lack of interest in mathematically precise thinking in psychology, inadequate representation of psychometric modeling in major statistics programs, and insufficient mathematical training in the psychological curriculum. Substantive factors relate to the absence of psychological theories that are sufficiently strong to motivate the structure of psychometric models. Following the identification of these problems, a number of promising recent developments are discussed, and suggestions are made to further the integration of psychology and psychometrics.

Key words: Psychometrics, modern test theory, classical test theory, construct validity, psychological measurement

In a recent overview of the psychometric developments of the past century, Embretson (2004, p. 8), noted that "[. . .] at the end of the 20th century, the impact of IRT on ability testing was still limited" and that "[t]he majority of psychological tests still were based on classical test theory." This conclusion applies with equal force to many other areas of mainstream experimental and quasi-experimental research, such as research on personality, attitudes, cognitive development, and intelligence. In fact, throughout psychology, one rarely encounters serious psychometric modeling endeavors.

Thus, even though psychometric modeling has seen rapid and substantial developments in the past century, psychometrics, as a discipline, has not succeeded in penetrating mainstream psychological testing to an appreciable degree. This is striking. Measurement problems abound in psychology, as is evident from the literature on validity (Cronbach & Meehl, 1955; Messick, 1989; Borsboom, Mellenbergh, & Van Heerden, 2004), and it would seem that the formalization of psychological theory in psychometric models offers great potential in elucidating, if not actually solving, such problems. Yet, in this regard, the potential of psychometrics has hardly been realized. In fact, the psychometric routines commonly followed by psychologists working in 2006 do not differ all that much from those of the previous generations. These consist mainly of computing internal consistency coefficients, executing principal components analyses, and eyeballing correlation matrices. As such, contemporary test analysis bears an uncanny resemblance to the psychometric state of the art as it existed in the 1950s.

The question that arises is why psychologists have been so reluctant to incorporate psychometric modeling techniques in their methodological inventory. The goal of the present paper is to answer this question by identifying and analyzing the factors that have hindered the incorporation of psychometric modeling into the standard toolkit of psychologists. A second goal is to offer some suggestions for improving the integration of psychological theory and psychometric techniques. However, to communicate the urgency of this situation, I will first consider some examples where things have gone seriously wrong.

This research was sponsored by NWO Innovational Research grant no. 451-03-068. I would like to thank Don Mellenbergh and Conor Dolan for their comments on an earlier version of this manuscript.

Requests for reprints should be sent to Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: d.borsboom@uva.nl

Crisis? What Crisis?

There might be people who are wondering whether the incorporation of psychometric theory is really all that important for psychology or, perhaps, there are even some who are of the opinion that things are actually going rather well in psychological measurement. This is not the case. The daily practice of psychological measurement is plagued by highly questionable interpretations of psychological test scores, which are directly related to the lack of integration between psychometrics and psychology. The following examples serve to substantiate this.

Principal Components and Latent Variables

Many investigations into the structure of individual differences theorize in terms of latent variables, but rely on Principal Components Analyses (PCA) when it comes to the analysis of empirical data. However, the extraction of a principal components structure, by itself, will not ordinarily shed much light on the correspondence with a putative latent variable structure. The reason is that PCA is not a latent variable model but a data reduction technique (e.g., Bartholomew, 2004). This is no problem as long as one does not go beyond the obvious interpretation of a principal component, which is that it is a conveniently weighted sumscore. Unfortunately, however, this is not the preferred interpretation among the enthusiastic users of principal components analysis.

Consider, for instance, the personality literature, where people have discovered that executing a PCA of large numbers of personality subtest scores, and selecting components by the usual selection criteria, often returns five principal components. What is the interpretation of these components? They are "biologically based psychological tendencies," and as such are endowed with causal forces (McCrae et al., 2000, p. 173). This interpretation cannot be justified solely on the basis of a PCA, if only because PCA is a formative model and not a reflective one (Bollen & Lennox, 1991; Borsboom, Mellenbergh, & Van Heerden, 2003). As such, it conceptualizes constructs as causally determined by the observations, rather than the other way around (Edwards & Bagozzi, 2000). In the case of PCA, the causal relation is moreover rather uninteresting; principal component scores are "caused" by their indicators in much the same way that sumscores are "caused" by item scores. Clearly, there is no conceivable way in which the Big Five could cause subtest scores on personality tests (or anything else, for that matter), unless they were in fact not principal components, but belonged to a more interesting species of theoretical entities; for instance, latent variables. Testing the hypothesis that the personality traits in question are causal determinants of personality test scores thus, at a minimum, requires the specification of a reflective latent variable model (Edwards & Bagozzi, 2000). A good example would be a Confirmatory Factor Analysis (CFA) model.

Now it turns out that, with respect to the Big Five, CFA gives Big Problems. For instance, McCrae, Zonderman, Costa, Bond, & Paunonen (1996) found that a five factor model is not supported by the data, even though the tests involved in the analysis were specifically designed on the basis of the PCA solution. What does one conclude from this? Well, obviously, because the Big Five exist, but CFA cannot find them, CFA is wrong. "In actual analyses of personality data [. . .] structures that are known to be reliable [from principal components analyses] showed poor fits when evaluated by CFA techniques. We believe this points to serious problems with CFA itself when used to examine personality structure" (McCrae et al., 1996, p. 563).

I believe this rather points to serious problems in psychologists' interpretation of principal components; for it appears that, in the minds of leading scholars in personality research, extracting a set of principal components equals fitting a reflective measurement model (or something even better). The problem persists even though the difference between these courses of action has been

clearly explicated in accessible papers published in general journals like *Psychological Bulletin* (Bollen & Lennox, 1991) and *Psychological Methods* (Edwards & Bagozzi, 2000). Apparently, psychometric insights do not catch on easily.

Group Comparisons

The interpretation of group differences on observed scores, in terms of psychological attributes, depends on the invariance of measurement models across the groups that figure in the comparison. In psychometrics, a significant array of theoretical models and associated techniques has been developed to get some grip on this problem (Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993). In practice, however, group differences are often simply evaluated through the examination of observed scores—without testing the invariance of measurement models that relate these scores to psychological attributes.

Tests of measurement invariance are conspicuously lacking, for instance, in some of the most influential studies on group differences in intelligence. Consider the controversial work of Herrnstein and Murray (1994) and Lynn and Vanhanen (2002). These researchers infer latent intelligence differences between groups from observed differences in IQ (across race and nationality, respectively) without having done a single test for measurement invariance. (It is also illustrative, in this context, that their many critics rarely note this omission.) What these researchers do instead is check whether correlations between test scores and criterion variables are comparable (e.g., Lynn & Vanhanen, 1994, pp. 66–71), or whether regressing some criterion on the observed test scores gives comparable regression parameters in the different groups (e.g., Herrnstein & Murray, 2002, p. 627). This is called prediction invariance. Prediction invariance is then interpreted as evidence for the hypothesis that the tests in question are unbiased.

In 1997 Millsap published an important paper in *Psychological Methods* on the relation between prediction invariance and measurement invariance. The paper showed that, under realistic conditions, prediction invariance does not support measurement invariance. In fact, prediction invariance is generally indicative of *violations* of measurement invariance: if two groups differ in their latent means, and a test has prediction invariance across the levels of the grouping variable, it must have measurement bias with regard to group membership. Conversely, when a test is measurement invariant, it will generally show differences in predictive regression parameters. One would expect a clearly written paper that reports a result, which is so central to group comparisons, to make a splash in psychology. If the relations between psychometrics and psychology were in good shape, to put forward invariant regression parameters as evidence for measurement invariance would be out of the question in every professional and scientific work that appeared after 1997.

So what happens in psychology? In 1999, two years after Millsap's paper appeared, the American Psychological Association is involved in the publication of the 1999 *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). With regard to the problem of test bias, we read: "[u]nder one broadly accepted definition, no bias exists if the regression equations relating the test and the criterion are indistinguishable for the groups in question" (AERA, APA, & NCME, 1999, p. 79). Another influential source, the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial Organizational Psychology, 2003), quite explicitly favors predictive invariance over measurement invariance as a method for investigating test bias (pp. 31–34). Perhaps, then, it should not come as a surprise that the fact, that prediction invariance is hardly ever violated, leads Hunter and Schmidt (2000, p. 151) to conclude that "the issue of test bias is scientifically dead." Unfortunately, on the basis of Millsap's work, one would rather say that, in the absence of violations of prediction invariance, the issue of test bias is in acute need of scientific scrutiny.

The Psychometric Connection

Of course, the impression resulting from these admittedly extreme examples cannot simply be generalized to the field of psychology as a whole. On the other hand, the involvement of influential theorists in psychology, as well as some of our most important professional organizations, indicates that we are not merely dealing with a few exceptions that prove the rule. In fact, it suggests that these examples are symptomatic of a serious problem in psychological measurement. One side of the problem is that psychologists have a tendency to endow obsolete techniques with obscure interpretations. The other side is that psychometricians insufficiently communicate their advances to psychologists, and when they do they meet with limited success. The result is a disconnect between psychometric theory on the one hand, and psychological research on the other. As a consequence, scientific progress in psychology is slowed. The questions that now arise are: (1) Why does this situation exist; and (2) What can we do about it? I address these questions in turn.

Obstacles to a Psychometric Revolution

The obstacles to a successful integration of psychometrics and psychology fall into three broad classes. The first class is theoretical in nature, and has to do with the dominant philosophical views on psychological measurement. The second concerns a loosely connected set of practical, social, and pragmatic factors that limit the amount of methodological innovation that the researcher in psychology can afford. The third relates to a shortage of substantive theory that is sufficiently detailed to drive informed psychometric modeling.

Theoretical Factors

Operationalism Rules

One of the main breakthroughs of the past century in psychometric thinking about measurement consists in the realization that measurement does not consist of finding the right observed score to *substitute* for a theoretical attribute, but of devising a model structure to *relate* an observable to a theoretical attribute. An essential precondition for this realization to occur is that, either intuitively or explicitly, one already holds the philosophical idea that theoretical attributes are, in fact, distinct from a set of observations, i.e., that one rejects the operationalist thesis that theoretical attributes are synonymous with the way they are measured (Bridgman, 1927).

Although one would expect most psychologists to subscribe to the thesis that theoretical attributes and measures thereof are distinct—after all, the rejection of operationalism was one of the driving forces behind the cognitive revolution—the standard research procedure in psychology is, ironically, to pretend that it is not true. Both in textbooks on psychological methods and in actual research, the dominant idea is that one has to find an "operationalization" (read: observed score) for a construct, after which one carries out all statistical analyses under the false pretense that this observed score is actually identical to the attribute itself. In this manner, it becomes defensible to construct a test for, say, self-efficacy, sum up the item scores on this test, subsequently submit these scores to analysis of variance and related techniques, and finally interpret the results as if they automatically applied to the *attribute* of self-efficacy because they apply to the *sumscore* that was constructed from the item responses.

This would be a relatively minor problem if psychologists widely realized that such an interpretation is crucially dependent on the assumption that the summed item scores can serve as adequate proxies for the actual attribute, and, perhaps more importantly, that violations of this

strong assumption present a threat to the validity of the conclusions reached. But this realization seems to be largely absent. The procedure mentioned is so common that it must be considered paradigmatic for psychological research. It brings with it the idea that properties that pertain to the sumscore somehow must also pertain to the attribute under study, so that, for instance, attributes are presumed to induce a linear ordering of people because sumscores do. But of course the assumption that an attribute induces a linear ordering cannot be derived from the fact that sumscores have this property; for the latter are linearly ordered by definition, while the former are not. Moreover, for many psychological attributes, alternative structures (like latent class structures or multidimensional structures) are no less plausible. However, the default strategy in psychological research precludes the consideration of such alternatives. And nobody knows how often these alternatives may actually be more accurate and truthful.

Classical Test Theory

It is an unfortunate fact of psychometric life that every introductory textbook on psychological research methods starts and ends its section on measurement with ideas and concepts that are based on classical test theory. Classical test theory may be considered the statistical handmaiden of the philosophy of operationalism that, at least as far as actual research practice is concerned, dominates psychology.

The reason for this is that the central concept of classical test theory—the true score—is exhaustively defined in terms of a series of observations; namely, as the expectation of a test score over a long run of replicated administrations of the test with intermediate brainwashing (Lord & Novick, 1968; Borsboom & Mellenbergh, 2002; Borsboom, 2005). Thus, the connection between the theoretical attribute (the true score) and the observation (the test score) is, in classical test theory, fixed axiomatically (Lord & Novick, 1968, Chap. 2). Therefore, within classical test theory, this relation is not open to theoretical or empirical research. This is in stark contrast with modern test theory models, in which the relation between test scores and attributes (conceptualized as latent variables) can take many forms (Mellenbergh, 1994).

Instead, classical test theory draws the researcher's attention to concepts such as reliability and criterion validity. The latter concept is especially important because it suggests that what is important about psychological tests is not how they work, but how strongly they are correlated with something else. This shift of attention is subtle but consequential. In an alternative world, where classical test theory never was invented, the first thing a psychologist, who has proposed a measure for a theoretical attribute, would do is to spell out the nature and form of the relationship between the attribute and its putative measures. The researcher would, for instance, posit a hypothesis on the structure (e.g., continuous or categorical) of the attribute, on its dimensionality, on the link between that structure and scores on the proposed measurement instruments (e.g., parametric or nonparametric), and offer an explanation of the actual workings of the instrument. In such a world, the immediately relevant question would be: How do we formalize such a chain of hypotheses? This would lead the researcher to *start* the whole process of research by constructing a psychometric model. After this, the question would arise which parts of the model structure can be tested empirically, and how this can best be done.

Currently, however, this rarely happens. In fact, the procedure often runs in reverse. To illustrate this point, one may consider the popular Implicit Association Test (IAT), which was developed by Greenwald, McGhee, and Schwartz (1998). This test is thought to measure so-called implicit preferences. A typical IAT application involves the measurement of implicit racial preferences. Subjects are presented with images of black and white faces and with positive and negative words on a computer screen. They are instructed to categorize these stimuli as quickly as possible according the following categories: A: "either a white face or a positive word," B: "either a black face or a negative word," C: "either a white face or a negative word," and D: "either a

black face or a positive word." The idea is that people who have an implicit preference for Whites over Blacks will be faster on tasks A and B but slower on C and D; the reverse is the case for those who have an implicit preference for Blacks over Whites. The IAT-score is computed by subtracting the log-transformed average response latency over compatible trials (A and B) from that over incompatible trials (C and D). Higher values on the resulting difference score are then considered to indicate implicit preference for Whites over Blacks.

Note the following facts about the psychometric work under consideration. First, the original paper puts forward no psychometric model for the dynamics underlying the test whatsoever. Second, even though the test is described as a measure of individual differences, the main evidence for its validity is a set of mean differences over experimental conditions, and no formal model explicating the link between these two domains is offered. Third, a Web of Science search reveals that the paper has been cited in over 420 papers. Some of these publications are critical, but most involve extensions of the test in various directions as well as substantive applications to different topics, which indicates that the IAT is a popular measurement procedure despite these points. Fourth, it took no less than eight years for a detailed psychometric modeling analysis of the proposed measure to see the light (Blanton, Jaccard, Gonzales, & Christie, 2006); and that analysis suggests that the scoring procedures used are actually quite problematic, because the various possible psychometric models on which they could be predicated are not supported by the data.

This illustrates a typical feature of the construction of measurement instruments in psychology. Let us say that in the ideal psychometric world, nobody could publish a test without at least a rudimentary idea of how scores are related to attributes, i.e., the outline of a psychometric model, and an attempt to substantiate that idea empirically. From the IAT example it is obvious that our world differs in two respects. The first concerns theory formation: the construction of a formal model that relates the attribute to its indicators is not necessary for a measurement procedure to be published and gain substantial following. The second concerns data analysis: Psychologists do not see psychometric modeling as a necessary tool to handle data gathered with a newly proposed testing procedure; running observed scores through ANOVA machinery, and computing correlations with external variables is perceived as adequate.

It is important to consider how these aspects are conceptually related, because quite often psychometricians try to sell the data analytic machinery to psychologists who have never asked themselves what the relation between the attribute and the test scores might be in the first place. It is obvious that such psychologists have no use for these modeling techniques; they may even perceive them as a silly mathematical circus. Psychometric modeling is only relevant and interesting to those who ask the questions that it may help answer. And because classical test theory axiomatically equates theoretical attributes with expected test scores, it has no room for the important and challenging psychometric question of how theoretical attributes are related to observations. Therefore, researchers who think along the lines of classical test theory simply do not see the need to ask such questions.

The Catch-All of Construct Validity

Operationalist thinking and classical test theory are mutually supportive systems of thought, which sustain a situation in which researchers habitually equate theoretical attributes with observational ones. However, although such practices may conceal the measurement problem, they do not make it go away; and many researchers are, at some level, aware of the fact that, with respect to psychological measurement, there is something rotten in the state of Denmark.

Now, psychologists can do a fascinating sort of Orwellian double-think with respect to the measurement problem: They can ask good psychometric questions, but then relegate them to a special theoretical compartment, namely that of "construct validity," instead of trying to answer

them. Relevant questions that are routinely dropped in the catch-all of construct validity are: What is it that the test measures? What are the psychological processes that the test items evoke? How do these processes culminate in behaviors, like marking the correct box on an IQ-item? How do such behaviors relate to individual differences? What is the structure of individual differences themselves? What is the relation between such structures and the test scores? In fact, looking at this list, it would seem that a question is considered to concern construct validity at the very instance that it becomes psychometrically challenging.

Construct validity functions as a black hole from which nothing can escape: Once a question gets labeled as a problem of construct validity, its difficulty is considered superhuman and its solution beyond a mortal's ken. Validity theorists have themselves contributed to this situation by stating that validation research is a "never-ending process" (e.g., Messick, 1988), which, at most, returns a "degree of validity" (Cronbach & Meehl, 1955; Messick, 1989), but can by its very nature never yield a definitive answer to the question whether a test measures a certain attribute or not. This effectively amounts to a mystification of the problem, and discourages researchers to address it. In addition, this stance must be fundamentally ill-conceived for the simple reason that no physicists are currently involved in the "never-ending process" of figuring out whether meter sticks really measure length, or are trying to estimate their "degree of validity"; nevertheless, meter sticks are doing fine. So why should "construct validity" be such an enormous problem in psychology?

The general idea seems to be based on the conviction (taken from the philosophy of science, and especially the work of Popper, 1959) that all scientific theories are by their nature "conjectures that have not yet been refuted"; i.e., tentative and provisionally accepted working hypotheses. Whether one subscribes to this idea or not, it is evident that it cannot be specifically relevant for the problem of validity, because this view concerns not just validity, but every scientific hypothesis, and, by implication, applies to every psychometric hypothesis. Thus, if validity is problematic for this particular reason, then so are reliability, unidimensionality, internal consistency, continuity, measurement invariance, and all other properties of tests, test scores, and theoretical attributes, as well as all the relations between these properties that one could possibly imagine. But this is thoroughly uninformative; it merely teaches us that scientific research is difficult, and that we hardly ever know anything for sure. While this may be an important fact of life, it has no special bearing on the problem of test validity and most certainly cannot be used to justify the aura of intractability that surrounds the problem of "construct validity."

It can be argued that, if the construction and analysis of measurement instruments were done thoroughly, this process would by its very nature force the researcher to address the central questions of construct validity before or during test construction (Borsboom et al., 2004). Not being able to do so, in turn, would preclude the construction of a measurement instrument. Thus, the fact that basic questions such as "What am I measuring?" and "How does this test work?" remain unanswered with respect to an instrument, which is considered fully developed, implies that we cannot actually take such an instrument seriously. In fact, a discipline that respects its scientific basis should hesitate to send tests, for which such basic problems have not been solved, out for use in the real world. A reference to the effect that such problems concern construct validity, and therefore their solution is impossible, cannot be taken as an adequate justification of such practice. So used, construct validity is merely a poor excuse for not taking the measurement problem seriously.

Pragmatic Factors

Although the ideological trinity of operationalism, classical test theory, and construct validity forms an important obstacle to the incorporation of psychometric modeling into the standard

methodology of psychology, there are also more mundane factors at work. These concern a hodgepodge of factors relating to the sociology of science, the research culture in psychology, practical problems in psychometric modeling, and the poor representation of psychometric techniques in widely used computer software. We will consider these factors in turn.

Psychometrics Is Risky

Suppose that an unconventional thinker in psychology were to stumble across a psychometric model, and recognize its potential. Suppose also that the psychologist were to use the model to analyze data that were gathered using the average psychological test. The researcher would quickly encounter a problem. Namely, psychometric models have a tendency to disprove commonly held assumptions, like unidimensionality and measurement invariance. The researcher then gets involved in fundamental problems concerning the structure of the attributes under investigation and the relation that they bear to the observations. Such questions are among the most fascinating and important ones in any science, but they are not popular in psychology. So, even if the psychologist has some success in answering these questions and coming up with a reasonable model for the observations, it will turn out difficult to get these results published, because many editors and reviewers of scientific journals are not overly familiar with psychometric models, and will often suggest that these results be published in a psychometric journal rather than a psychological one. This, of course, is not what the researcher necessarily wants; moreover, psychometric journals may refuse the work for the reason that it is not sufficiently psychometrically oriented, so that the researcher gets stuck between a rock and a hard place. Thus, career-wise, turning to psychometric modeling techniques is risky.

It Shouldn't Be Too Difficult

This problem is compounded by the research standards that are currently accepted in psychology. Even though the research topic of psychology—human behavior and the mental processes that underlie it—is perhaps the most complicated ever faced by a science, the contents of scientific papers that deal with it are required to be below a certain standard of difficulty. I have seen at least one case where a manuscript that used psychometric modeling was rejected by a major journal because, according to the editor, it was too difficult for the journal's audience since it contained some basic matrix algebra (i.e., addition and multiplication). That a scientific journal should reject a paper for being difficult is almost surrealistic; yet, the use of equations, in general, is discouraged in many psychological journals. This is detrimental to the development of psychology. If physics journals had existed in the seventeenth century and had adhered to this policy, it would have been impossible to achieve the break with Aristotelian theory that is now known as the Scientific Revolution. The current research culture in psychology, however, actively works against the formalization of theories and the use of mathematical modeling techniques, which include psychometric models.

Educational Issues

Psychologists typically do not receive a substantial amount of mathematical training. In this respect it is illustrative to compare the average educational program of psychologists with that of, say, economists. Every trained economist understands basic calculus, for instance, while trained psychologists often do not know what calculus is in the first place. The reason for this difference is clear. It is simply impossible to read advanced economics texts without substantial mathematical knowledge, and hence mathematical training is a bare necessity. Evidently, no such baggage is required in psychology. As Lykken (1991, p. 7.) stated, "there are many courses in the psychology curriculum, but few have real prerequisites. One can read most psychology texts

without first even taking an introductory course." It is not strictly necessary for a student to have even a rudimentary understanding of mathematics in order to complete a degree in psychology; and neither is such understanding required on the part of the researcher who studies psychology's advanced texts. The consequence of this situation in the present context is that psychologists often lack the necessary skills to understand what psychometric models do or what they can be used for, which hampers the dissemination of advances in psychometrics.

But It's Not in SPSS!

The reason that, say, Cronbach's alpha and principal components analysis are so popular in psychology is not that these techniques are appropriate to answer psychological research questions, or that they represent an optimal way to conduct analyses of measurement instruments. The reason for their popularity is that they are default options in certain mouse-click sequences of certain popular statistics programs. Since psychologists are monogamous in their use of such software (most in my department are wedded to SPSS) there is little chance of convincing them to use a model—any model—that is not "clickable" in the menus of major statistical programs. For reasons that defy my understanding, psychometric models are not well represented in such software. In fact, for a long time the psychometric modeler was typing in obscure commands next to a command prompt on a black and white screen. Considerable improvement on this point has been made (e.g., Muthén & Muthén, 2001), but this improvement is of a relatively recent date.

Thou Shalt Not. . .

Psychometric models are often considered to have a normative component. People who subscribe to this point of view, see psychometrics as a set of golden rules that the researcher should live by. I am thinking of the "no Rasch, no good" philosophy and associated doctrines. The presentation of psychometrics in terms of strictures (instead of opportunities, for instance) is damaging to its public image; for it is a law of human psychology that people whose behavioral repertoire is limited to "you are not allowed to do that" do not get invited to parties. Thus, it is important to get psychologists to see that psychometric modeling gives them new possibilities, instead of presenting them with strictures and limitations; good examples of such a positive strategy can be found in De Boeck and Wilson (2004).

Sample Size Issues

The estimation and testing of psychometric models is not always feasible due to the fact that one often needs large data sets for this purpose. In experimentally oriented research, for instance, sample sizes typically involve dozens rather than hundreds of subjects, and in such cases the use of psychometric models with latent variables is often hard to justify. In contrast, treating variables as "observed," i.e., as recorded without measurement error, returns the possibility of doing science with 15 to 30 subjects per cell in a standard factorial design. So why would one then even consider the use of psychometric techniques in psychological research? Simply adding up some scores and calling the sumscore "self-efficacy" does the job just as well, but with much fewer subjects. The question is, of course, whether this is true.

Coombs (1964) has said that we buy information by assumption. In many cases, however, one instead needs information to buy assumptions. For instance, if one knows one is using a good measurement instrument, one can use this information to "buy" the very useful assumption—common to all observed score regression techniques, including the various species of ANOVA—that one models "observed variables." This effectively means that one can drop the measurement problem and directly estimate population differences in the theoretical attribute on the basis of observed scores. However, there is no uncertainty concerning the question whether psychologists

have the knowledge needed to buy such assumptions: they do not. Hence, any argument against psychometric modeling on the basis of the larger sample sizes needed, when compared to analysis of variance and related observed score techniques, is in fact based on a wager. This wager involves the question whether observed score techniques are still trustworthy, if one does not buy the assumptions needed, but steals them. That is, does the run-of-the-mill research design plus analysis still work, if one pretends to have solved the measurement problem, while one has in fact ignored it? I do not know the answer to this question, but given the small and notoriously unstable effects that psychologists usually find, I would not like to bet on psychology's chances in this gamble.

Substantive Factors

It will be obvious by now that the integration of psychology and psychometrics faces significant obstacles of various natures. When ideological, theoretical, practical, and sociological factors conspire against the incorporation of a method, it is no surprise that such a method has trouble getting off the ground. However, we have not yet devoted attention to what may be the single most important problem that faces psychometric modeling. This is the almost complete absence of strong psychological theory.

The problem is best illustrated with an example. Suppose we are interested in personality and want to construct a measurement instrument for, say, conscientiousness. Like most people, we have a common-sense idea about the characteristics of conscientious people. For instance, they tend to be in time for appointments, do their best to succeed on a job, feel guilty when they fail to meet obligations, etc. Suppose that we assess these characteristics through a self-report questionnaire; for ease of exposition, assume we construct a set of items in the spirit of "I feel guilty when I fail to meet obligations," and score them dichotomously in a yes/no format. How do we then relate the item responses to the attribute of conscientiousness?

Consider the following list of options. We could view the items as a sample from a domain of behaviors, and define the attribute as the proportion of the behaviors in that domain that any given person exhibits, which would lead us toward generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). We could also view conscientiousness as a causal function of the behaviors assessed in the items (people are conscientious because they meet appointments). This would lead us toward a formative model, like a PCA model (Bollen & Lennox, 1991). Alternatively, we could reverse the causal relation (people meet appointments because they are conscientious), which would lead us toward a reflective latent variable modeling scheme (Edwards & Bagozzi, 2000). We could further envision the causal relation to govern person-specific changes or as a relation between sets of individual differences (Borsboom et al., 2003; Hamaker, Dolan, & Molenaar, 2005; Molenaar, 2004).

Within any of these schemes of thinking, we still have many choices to make (see also Mellenbergh, 2001). Suppose, for instance, that we opt for the reflective latent variable approach. Should we then conceive of conscientiousness as a continuous variable (pushing us toward IRT) or as a categorical variable (pushing us toward a latent class approach)? If we suppose 0conscientiousness is continuous, what is the form of the relation between the item responses and the trait? Is it a monotonic function or not (Stark, Chernyshenko, Drasgow, & Williams, 2006)? If it is monotonic, is the function smooth, like a logistic function (which would suggest a one- or two-parameter logistic model; Rasch, 1960; Birnbaum, 1968), or erratic (which would suggest a nonparametric alternative; Mokken, 1970)? If it is smooth, can we then assume that zero and one are the right asymptotes for the Item Response Function or are different asymptotes more realistic (Hessen, 2004)? Is a logistic function at all appropriate?

This is a psychometric embarrassment of riches. Even within this concise set of questions to be answered we encounter no less than four radically different conceptualizations of the relation between conscientiousness and conscientious behaviors: a universe-sampling relation (generalizability theory), a formative causal relation (formative model), a reflective causal relation with the latent variable categorical (latent class model), and a reflective causal relation with the latent variable continuous (IRT). Moreover, as the IRT example shows, within each of these conceptualizations there are many more fine-grained choices to be made before we truly have a candidate model. Literally none of these choices are dictated by substantive theory.

Of course, *researchers* make such choices all the time—otherwise they could do nothing with their data. For instance, personality traits are usually taken to be continuously structured and conceived of as reflective latent variables (even though the techniques used do not sit well with this interpretation). The point, however, is that there is nothing in personality *theory* that motivates such a choice, and the same holds for the majority of the subdisciplines in psychology. Thus, the crucial decisions in psychological measurement are made on the basis of pragmatic or conventional grounds, instead of on substantive considerations.

This may be the central problem of psychometrics: psychological theory does not motivate specific psychometric models. It does not say how theoretical attributes are structured, how observables are related to them, or what the functional form of that relation is. It is often silent even on whether that relation is directional and, if so, what its direction is. It only says that certain attributes and certain observables have something to do with each other. But that is simply not enough to build a measurement model.

The Light at the End of the Tunnel

This paper has sketched a rather grim picture of the role of psychometrics in psychology. Fortunately, however, several positive developments have also taken place in the last decade or so. Three developments are especially noteworthy.

The first concerns the increasing number of conceptually and practically oriented introductions to psychometric modeling that have appeared since the late 1980s. Important examples, among others, are the books by Bollen (1989), Frederiksen, Mislevy, and Bejar (1993), Embretson and Reise (2000), Embretson and Hershberger (1999), Hagenaars (1993), Heinen (1996), Kaplan (2000), and Sijtsma and Molenaar (2002). These works present the subject matter in a relatively accessible way, thereby facilitating a transition to psychometric modeling in psychology.

A second promising development is the increase of user-friendly software for psychometric modeling. Of course, the one program based on psychometric ideas that has, in the past decades, made something of a breakthrough is the LISREL program by Jöreskog and Sörbom (1996). This prepared the road for various more recent latent variable modeling computer programs including the versatile Mplus program by Muthén and Muthén (2001) and Vermunt's and Magidson's Latent Gold (2000). The increasing popularity of freeware statistical computing programs like R (Venables, Smith, & The R Development Core Team, 2005) and Mx (Neale, Boker, Xie, & Maes, 2003) is also promising. Finally, the group of Paul de Boeck (e.g., De Boeck & Wilson, 2004) has worked out effective ways to do IRT modeling through the program SAS by specifying IRT models as mixed regression models. One hopes that such developments will necessitate the most widely used program in psychology, SPSS, to incorporate latent variable modeling options in its basic test analysis section. In all, these developments are certainly going in the right direction, and hopefully will result in a situation where psychometric modeling is a realistic option for the average researcher in psychology within a decade or so.

A third positive development is the increasing number of psychometrically informed research papers that have been appearing in the past decade. The recently introduced section

on applied psychometrics in *Psychometrika* presented some good examples of such papers (e.g., Bouwmeester & Sijtsma, 2004; Van Breukelen, 2005). Substantial psychometric literatures are further building up on process-based psychometric modeling of intelligence (sub)tests (Embretson, 1998; Mislevy & Verhelst, 1990; Süss, Oberauer, Wittmann, Wilhelm, & Schulze, 2002) and cognitive development (Jansen & Van der Maas, 1997, 2002; Dolan, Jansen, & Van der Maas, 2004). Formal modeling approaches are also gaining momentum in personality theory (Fraley & Roberts, 2005), emotion research (Ferrer & Nesselroade, 2003), and social psychology (Blanton et al., 2006). In a more general sense, the framework of explanatory IRT highlights the potential of psychometrics in substantive contexts (De Boeck & Wilson, 2004).

These developments are creating a set of background conditions against which major changes in psychometric practice become possible. The development of accessible introductory works on psychometrics and the development of user friendly computer programs remove some significant practical obstacles; and the fact that substantively driven psychometric modeling endeavors are published in both technically oriented and substantive journals may create the momentum that is needed to establish the breakthrough of psychometric modeling. Hopefully, this will lead to a change in the psychological research culture and ultimately further progress in psychology. It is important that psychometricians support and, where possible, accelerate this process. There is room for improvement on several fronts.

Write Good Introductory Textbooks

Psychologists' first, and at some places only, contact with the theory of psychological measurement occurs through introductions to psychological research methods. Such books usually contain a section on psychological measurement. This section is always outdated and often flawed. The student is mesmerized through the formula $X = T + E$ but not given the tools to understand what it means. The caricature of classical test theory so induced is invariably accompanied by the misinterpretation of "true scores" as "construct scores" (Borsboom & Mellenbergh, 2002; Borsboom, 2005), which sets the stage for the operationalist mode of thought described earlier in this paper. Also, there is usually an explanation of reliability (as test–retest reliability or internal consistency) and a section that emphasizes, but does not elucidate, the difficulties of validity. If one is lucky, there is a treatment of the so-called convergent and divergent validity coefficients. This teaches one how to eyeball correlation matrices, which cannot be considered an optimal research procedure but is better than nothing. That is where it stops. No attention is given to the crucial questions how psychological attributes are structured, how they are related to observed scores, how one can utilize substantive theory in test construction, or how one can test one's ideas through the construction and evaluation of measurement models. It would be a significant advance if these books were to update their sections on measurement, so that a psychologist no longer has to unlearn earlier ideas when she/he decides to take the measurement problem seriously. It seems psychometricians are the right candidates for this job.

Read and Publish Widely

The founding fathers of the Psychometric Society—scholars such as Thurstone, Thorndike, Guilford, and Kelley—were substantive psychologists as much as they were psychometricians. Contemporary psychometricians do not always display a comparable interest with respect to the substantive field that lends them their credibility. It is perhaps worthwhile to emphasize that, even though psychometrics has benefited greatly from the input of mathematicians, psychometrics is not a pure mathematical discipline but an applied one. If one strips the application from an applied science one is not left with very much that is interesting; and psychometrics without the "psycho" is not, in my view, an overly exciting discipline. It is therefore essential that a psychometrician keeps up to date with the developments in one or more subdisciplines of psychology. This

is not to say that the purely conceptual and mathematical study of psychometric models is unimportant. On the contrary. However, as a discipline, psychometrics should consciously and actively avoid a state of splendid isolation. This requires regular reading of psychological journals and visits to substantively oriented conferences. Psychometricians should, in my view, also actively promote such behavior in their (PhD) students, who often cannot see the substantive forest for the mathematical trees. Substantive involvement ideally leads to a greater number of psychologically oriented publications by psychometricians, and hence to a more prominent presence of psychometrics in psychological research; this, in turn, may facilitate the acceptance of the ideas of modern psychometric theory in psychological circles.

Psychometrics and Theory Formation

Traditionally, psychometricians develop mathematical models, but leave the development of the substantive theory that is supposed to motivate these models to psychologists. As such, psychometrics takes the attitude of an ancillary discipline, which helps psychology with the formalization of theory into statistical models, and with the analysis of psychological data. I think that, with respect to the measurement problem, a century of experience teaches us that this procedure does not work very well. As has been discussed above, psychological theories are often simply too vague to motivate psychometric models. Most psychologists appear neither capable of, nor interested in, constructing more precise theories. I suggest that the more adventurous psychometricians would do well to take matters into their own hands, and start developing psychometric theories (as opposed to psychometric models) with a substantive component.

As it happens, there is another, quite similar discipline that is also waiting for a revolution that never happened (Cliff, 1992), namely, mathematical psychology, where a good deal of experience and know-how on the development of formal psychological theories is available. The question of how attributes may be structured, for instance, has received ample attention in the work on fundamental measurement theory (Krantz, Luce, Suppes, & Tverky, 1971). However, for reasons that elude me, and that are probably historical in nature, there is very little communication and collaboration between the fields of psychometrics and mathematical psychology, even though they manifestly have so much in common. Much could be gained by a further integration of these disciplines. Some psychometricians and mathematical psychologists (e.g., Scheiblechner, 1995; Tuerlinckx & De Boeck, 2005; Falmagne, 1989; Doignon & Falmagne, 1999) have already explored some of the common ground with promising results. That common ground may harbor significant further opportunities to promote the development of formal psychological theorizing.

Discussion

This paper has traced the lack of successful integration between psychometrics and psychology to a number of theoretical, pragmatic, and substantive factors that obstruct necessary changes in the research practices of psychologists. These factors are wide-ranging and distinct in nature, and thus render the task of breaking the resistance of psychology to psychometric modeling a formidable one. However, the incorporation of psychometrically sensible thinking in psychological research is important, not just for the progress of psychology as a science, but also for society as a whole. For, insofar as psychological research remains purely scientific, the lack of psychometrically defensible analyses may obstruct progress; but apart from that it is mostly harmless. However, psychological testing also has a significant and direct impact on people's lives—for instance, through the use of tests in psychiatric diagnoses or for the selection of employees—and at present such applications do not always stand on firm grounds, to say the

least. Thus, we face a pressing obligation to improve the practice of psychological research, and this obligation is not merely of a scientific nature.

There is no question that there is ample room for such improvement. The current practice of psychological measurement is largely based on outdated psychometric techniques. We should not sit around while the psychometric procedures of our fellow psychologists slip into obsolescence. Psychometricians may prevent this through active participation in the education of psychologists, the dissemination of psychometric insights among researchers, but also through the development of formalized psychological theory. At the very least, the psychometricians of the twenty-first century should strive to play a more pronounced role in substantive psychological research than is currently the case.

Max Planck stated that it hardly ever happens that scientists radically change their established ideas: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." Perhaps this holds to an even stronger degree for methodological innovations, because these necessitate not just the revision of theoretical ideas, but also require researchers to learn new skills. Several promising processes, like the development of versatile computer programs and the increasing number of successful psychometric investigations in substantive psychology, suggest that the tide may be turning. I suggest we work as hard as possible to facilitate the emergence of a new generation of researchers who are not afraid to confront the measurement problem in psychology.

References

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

Bartholomew, D.J. (2004). *Measuring intelligence: Facts and fallacies*. Cambridge: Cambridge University Press.

Blanton, H., Jaccard, J., Gonzales, P.M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*, 192–212.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

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

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

Borsboom, D., & Mellenbergh, G.J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*, 505–514.

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

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Bouwmeester, S., & Sijtsma, K. (2004). Measuring the ability of transitive reasoning, using product and strategy information. *Psychometrika, 69*, 123–146.

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

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

Coombs, C. (1964). *A theory of data*. New York: Wiley.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

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

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Doignon, J.P., & Falmagne, J.C. (1999). *Knowledge spaces*. New York: Springer-Verlag.

Dolan, C.V., Jansen, B.R.J., & Van der Maas, H.L.J. (2004). Constrained and unconstrained normal finite mixture modeling of multivariate conservation data. *Multivariate Behavioral Research, 39*, 69–98.

Dolan, C.V., Roorda, W., & Wicherts, J.M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence, 32*, 155–173.

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

Embretson, S.E. (1998). A cognitive design system approach for generating valid tests: Approaches to abstract reasoning. *Psychological Methods, 3*, 300–396.
Embretson, S.E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement, 2*, 1–32.
Embretson, S.E., & Hershberger, S.L., Eds. (1999). *The new rules of measurement: What every psychologist and educator should know.* Mahwah, NJ: Erlbaum.
Embretson, S.E., & Reise, S., Eds. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Falmagne, J.C. (1989). A latent trait theory via stochastic learning theory for a knowledge space. *Psychometrika, 54*, 283–303.
Ferrer, E., & Nesselroade, J.R. (2003). Modeling affective processes in dyadic relations via dynamic factor analyses. *Emotion, 3*, 344–360
Fraley, R.C., & Roberts, B.W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*, 60–74.
Frederiksen, N., Mislevy, R.J., & Bejar, I.I. (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.
Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.
Hagenaars, J.A. (1993). *Loglinear models with latent variables*. Newbury Park: Sage.
Hamaker, E.L., Dolan, C.V., & Molenaar, P.C.M. (2005). Statistical modeling of the individual: Rationale and application of multivariate time series analysis. *Multivariate Behavior Research, 40*, 207–233.
Heinen, T. (1996) *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks: Sage.
Herrnstein, R.J., & Murray, C. (1994). *The Bell curve*. New York: The Free Press.
Hessen, D.J. (2004). A new class of parametric IRT models for dichotomous item scores. *Journal of Applied Measurement, 5*, 385–397.
Hunter, J.E., & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests. *Psychology, Public Policy & Law, 6*, 151–158.
Jansen, B.R.J., & Van der Maas, H.L.J. (1997). Statistical tests of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.
Jansen, B.R.J., & Van der Maas, H.L.J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383–416.
Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8 User's reference guide* (2nd ed). Chicago: Scientific Software International.
Kaplan, D. (2000). *Structural equation modeling. Foundations and extensions*. Thousand Oaks, CA: Sage.
Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*, Vol. I. New York: Academic Press.
Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Lykken, D.T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W.M. Grove (Eds.), *Thinking clearly about psychology, Vol. 1*. Minneapolis, MN: University of Minnesota Press, pp. 3–39.
Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.
McCrae, R.R., Costa, P.T., Jr., Ostendorf, F., Angleitner, A., Hrebickova, M., & Avia, M.D., et al. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology, 78*, 173–186.
McCrae, R.R., Zonderman, A.B., Costa, P.T., Jr., Bond, M.H., & Paunonen (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552–566.
Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.
Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300–307.
Mellenbergh, G.J. (2001). Outline of a faceted theory of item response data. In: A. Boomsma, M.A.J. Van Duijn, & T.A.B. Snijders (Eds.), *Essays in item response theory*. New York: Springer-Verlag.
Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525–543.
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequence of measurement. In H. Wainer, & H.I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.
Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.
Millsap, R.E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248–260.
Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement, 17*, 297–334.
Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.
Mokken, R.J. (1970). *A theory and procedure of scale analysis*. The Hague: Mouton.
Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201–218.
Muthén, L.K., & Muthén, B.O. (2001). *Mplus user's guide* (2nd ed.). Los Angeles, CA: Muthén & Muthén.
Neale, M.C., Boker, S.M., Xie, G., & Maes, H.H. (2003). *Mx: Statistical modeling* (6th ed.). Box 980126 MCV, Richmond, VA 23298, USA.
Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchinson Education.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60*, 281–304.
Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
Society for Industrial Organizational Psychology (2003). Principles for the application and use of personnel selection procedures. Bowling Green, OH: Society for Industrial Organizational Psychology.
Stark, S., Chernyshenko, O.S., Drasgow, F., & Williams, B.A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25–39.
Süss, H., Oberauer, K., Wittmann, W.W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—And a little bit more. *Intelligence, 30*, 261–288.
Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*, 629–650.
Van Breukelen, G.J.P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359–376.
Venables, W.N., Smith, D.M., and The R Development Core Team (2005). *An introduction to R, Version 2.2.0*. R-Project, 2005. URL: http://CRAN.R-project.org
Vermunt, J.K., & Magidson, J. (2000). *Latent GOLD user's manual*. Boston, MA: Statistical Innovations Inc.

Manuscript received 6 JAN 2006
Final version received 20 APR 2006
Published Online Date: 23 SEP 2006

Editor's Note

Denny Borsboom was the 2004 winner of the Psychometric Society Dissertation Prize. This essay and the subsequent commentary grew out of conversations following Borsboom's presentation of his work at the International Meeting of the Psychometric Society 2005, Tilburg, The Netherlands.

PSYCHOMETRIKA—VOL. 71, NO. 3, 441–445
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1491-2

IN PRAISE OF PLURALISM. A COMMENT ON BORSBOOM

MICHAEL KANE

NATIONAL CONFERENCE OF BAR EXAMINERS, MADISON, WI

I tend to agree with Professor Borsboom that psychology, and more generally the social sciences, could benefit from better psychometric modeling. However, if psychometric developments are to have more effect on everyday practice in psychology, psychometricians probably need to pay more attention to the substantive and methodological problems in various areas of psychology. For example, Professor Borsboom is critical of the *Standards for educational and psychological testing* (AERA, APA, NCME, 1999) for suggesting that group differences in test-criterion relationships are relevant to test bias. He bases his criticism on the finding that predictive invariance is not the same as measurement invariance. However, he fails to acknowledge the social, political, and ethical problems associated with failures of predictive invariance in high-stakes contexts (e.g., employment and admissions testing). The *Standards* are designed to provide test publishers and test users with guidance on a range of practical measurement problems. In this context, predictive invariance is a major issue in itself. If we want psychologists to pay more attention to psychometric analyses, these analyses need to recognize the psychologists' problems and goals.

Before getting into a discussion of Professor Borsboom's analyses, it is probably useful to consider his desired state of affairs. He mentions several options for the interpretation of attributes, but seems to prefer what he calls a "reflective latent variable modeling scheme," in which a latent attribute is assumed to cause the test behavior, and the psychometric model reflects this causal relationship. In an earlier paper, Professor Borsboom and his colleagues argued that a test is valid as a measure of 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. (Borsboom, Mellenbergh, & Van Heerden, 2004, p. 1061.)

Under the reflective model, the latent attribute is the variable of interest, and the test is developed to reflect the attribute. I am comfortable with this kind of interpretation. It is certainly a common model for the interpretation of measurements in the social sciences. However, as illustrated below, it is not the only viable kind of score interpretation, and it is not the best interpretation in many cases.

Operational Definitions, Classical Test Theory, and Construct Validity

Professor Borsboom considers three theoretical obstacles to the integration of psychometrics and psychology: operational definitions, classical test theory, and construct validity. My remarks will focus on these three basic concerns.

Requests for reprints should be sent to mkane@ncbex.org.

Operational definitions were introduced into physics in reaction to the overthrow of traditional, common-sense assumptions about space and time by Einstein's theory of relativity. Bridgeman (1927) and the logical positivists sought to eliminate this kind of upheaval by eliminating implicit assumptions in science. They did not so much equate theoretical constructs with observable attributes, as strive to eliminate theoretical assumptions from their descriptions of observations. They tried to be absolutely clear about what they were doing (generally a good habit), but they also tended to downplay or eliminate theory altogether (not generally a good strategy).

Following this lead, some psychologists decided to define some theoretical attributes (e.g., intelligence) in terms of specific measures. Ironically, this simple replacement of theoretical attributes by observable attributes, called "operationism," had effects diametrically opposed to Bridgeman's goal (Ennis, 1973). Instead of eliminating unwarranted theoretical assumptions, "operationism" assigned all of the assumptions associated with a theoretical construct to the scores on a particular test, thus importing unwarranted assumptions by the carload. To define a theoretical term like intelligence narrowly in terms of a specific measure, while interpreting it broadly in terms of the traditional notion of intelligence, is clearly unwarranted.

However the operational specification of measurement procedures is certainly legitimate, if not essential. The operations used to collect data and to generate scores should be clearly described. Measurement procedures should be operationally defined, but theoretical attributes cannot be operationally defined.

Professor Borsboom sees the "true scores" of classical test theory as reinforcing operationist tendencies in psychology. The *true* score, which is defined as the expected score over replications of the measurement procedure, is clearly dependent on the operational definition of this procedure. However, true scores are used mainly as a basis for analyzing the precision, or reliability, of measurements, and in classical test theory, reliability is paired with validity, which examines the relationship between the true scores and the variable of ultimate interest. By focusing on the distinction between the true score and the variable of interest, validity theory tends to run counter to operationism.

The theory of validity has a long and checkered history, but by the 1980s, a general conception of construct validity provided a unified framework for validity (Messick, 1989). In the original formulation of construct validity (Cronbach & Meehl, 1955), substantive theory was assumed to provide a "nomological" network of relationships among theoretical constructs and observable attributes, and the meanings of the constructs were determined by their roles in this network. The validity of a measure of a theoretical construct would be evaluated in terms of how well its scores satisfied the relationships in the network.

Initially, the nomological networks were conceived of as formal theories (e.g., Newton's laws), but because such theories are rare to nonexistent in psychology, the requirement was relaxed to include open-ended collections of relationships involving the construct of interest. There was a shift from what Cronbach (1989) called the "strong form" of construct validity to what he called the "weak form" of construct validity.

Under the weak form of construct validity, the tight networks envisioned by Cronbach and Meehl (1955) were replaced by collections of relationships involving the construct. For constructs of any generality, such collections could be both vast and ill-defined, making it very difficult to evaluate the measure's fit to the network. Professor Borsboom's conclusion that construct validity functions as "a black hole from which nothing can escape" overstates the case, but by rolling all of the issues inherent in justifying a proposed interpretation into one big ball, many discussions of construct validity have tended to discourage would-be validators.

Nevertheless, the basic question addressed by validity theory, how to justify claims based on test scores, is of fundamental importance. I have suggested that validation can be simplified without being trivialized by requiring that the inferences and decisions to be derived from test

scores be spelled out and evaluated (Kane, in press). This approach allows for a variety of possible interpretations and uses for test scores, with the caveat that any proposed interpretation or use be justified by appropriate evidence. So, operationally defined variables are fine as long as we recognize them for what they are, and do not slide any theoretical claims in under the radar. A claim that the score resulting from a measure can be interpreted as an estimate of a latent attribute that causes the observed performances is also acceptable as long as the claim can be justified. A theory-based interpretation is admissible as long as the theory is specified and the measure's fit to the theory is established.

Professor Borsboom argues that construct validity "must be fundamentally ill-conceived for the simple reason that no physicists are currently involved in the 'neverending process' of figuring out whether meter sticks really measure length" (Borsboom, 2006, p. 431). Of course, it is also hard to find physiometric models (corresponding to our psychometric models) that specify a causal relationship between the latent attribute of length and the observed extension of objects in space. Length once provided a classic example of an operationally defined attribute (remember the platinum–iridium bar in a temperature-controlled chamber in Paris—the standard meter). Now, it can be considered a theoretical attribute within the special theory of relativity. The operational definition was adequate at one time and is still adequate in many contexts. The newer, theory-based definition is used when it is needed. Having methodologists tell scientists what they can and cannot do would limit progress, if the scientists paid any attention to this advice; luckily, they generally don't pay much attention.

The Role of Theory in the Development and Validation of Measures

Professor Borsboom suggests that a psychologist who proposes a measure for a theoretical attribute should spell out the relationship between the attribute and the proposed measure and that this would lead the researcher, "to start the whole process of research by constructing a psychometric model" (Borsboom, 2006, p. 429). This is all well and good, but how does the researcher go about defining the attribute, the measure, and the relationship between the two? Presumably, the process is not arbitrary. We would generally not propose a measure of intelligence that was based on speed in running the mile, just as we would not propose a measure of physical fitness based on completing verbal analogies. However, on a finer level, what tasks should be included in a measure of intelligence (or fitness) and what should be left out? And, how do we evaluate how well the measure is working?

One important determiner of how this process is likely to work is the state of theory development in the area under investigation, with little or no substantive theory at one end of the continuum and a formal theory that can guide test development at the other end.

Development and Validation of Measures without Theory

Assuming that no formal theory exists (the usual case), the test-development process is necessarily ad hoc, guided by general conceptions of the attribute of interest. For example, intelligence is assumed to promote success on a wide range of cognitive tasks, and measures of intelligence generally consist of samples of such tasks. The model is causal but not detailed or formal.

In the absence of a formal theory that specifies a particular form for the psychometric model, the researcher who follows Professor Borsboom's advice is likely to adopt some standard unidimensional IRT model. We have a test consisting of a set of tasks that are thought to reflect the latent attribute, and the IRT model is applied to responses to these tasks. The estimation

of the model parameters requires large sample sizes, which are not available in many areas of psychology, but if the model can be applied, it yields estimates of a latent ability for each person and of one or more parameters for each task.

The resulting latent ability scale derives most, if not all, of its substantive meaning from the sample of tasks on which it is defined. A formal IRT model can be applied to different kinds of tasks to define different scales; like all formal systems, it does not, in itself, contribute substantive content. The population on which the scaling is conducted can also influence the proposed interpretation, but in most applications of IRT, it is assumed that the scale is invariant across persons and groups. So, the scale derived from a standard IRT analysis is largely determined by the observations used to estimate the model parameters. The latent scale is operationally defined in terms of the domain of tasks on which it is based (not that there is anything wrong with that). The fact that the responses have been scaled using a psychometric model does not turn the scale into a theoretical construct. Bridgeman's (1927) conception of operational definitions allowed for the use of sophisticated mathematical models (Benjamin, 1955).

The choice of which of the currently available IRT models to use is often dictated by the preferences of the modeler; some like fewer parameters and some like more parameters. The choice may also be influenced by fit statistics, with preference going to the model that provides the best fit to the data. In the absence of strong substantive theory, this is a reasonable basis for evaluating models, but it also reflects the dependence of the scale on the observations generated by the measurement procedure and not on an a priori conceptualization of the theoretical attribute.

In the absence of theory, we do not know how the latent attribute has its effect (although we may have some hypotheses), and we do not know how this attribute is related to other variables (although we may have some hypotheses). In order to develop our understanding of these issues, we will need to do some empirical research and some theorizing. To draw conclusions about the attribute (e.g., that test results can be generalized to new contexts) simply because we scaled the responses and assigned a trait label to the scale (e.g., "intelligence") would be unwarranted.

Assuming that we want to use our scale scores to make some predictions about future performance on nontest tasks in nontest contexts, it would be prudent to examine how well these predictions turn out. Assuming that we want to make causal claims about how the attribute affects performance on the tasks included in the measure or on other tasks, we would need to develop support for these inferences, and the support for such causal inferences generally involves both empirical research and theory development. The procedures used to develop support for such inferences have been discussed under the heading of construct validity (Messick, 1989).

It is tempting to interpret the latent ability estimates generated by the IRT model as a real, causal attribute, and if no claim is to be made beyond this immediate causal claim (i.e., that some otherwise unspecified latent attribute causes the observed performances), the causal attribution does not make much difference. However, if the hypothesized causal claim is used to justify other inferences (e.g., predictions about future performance on other tasks or in other contexts), then these additional claims need to be examined.

Development and Validation of Measures Based on Theory

If solid theory exists, it can be used to guide test development in a way that builds support for a reflective interpretation in terms of a causal, latent attribute. In particular, if the theory provides a causal explanation of the relationship between the latent attribute and performance on some set of tasks, performance on the tasks can be used to draw conclusions about the causal attribute. This approach works well in areas with highly developed quantitative theories, but it cannot be implemented otherwise. A specification of how a theoretical attribute produces certain effects

is possible only after the theory is in place; it is not the first step, but one of the last steps in a research program.

An alternative approach that has been applied to several kinds of test performance is to take a well-established measure of some attribute (e.g., intelligence) and develop causal models for the performances included in the measure (Embretson, 1998). This research tradition takes performance on the measure as an observable variable of interest and seeks to explain the performance in terms of latent abilities and task characteristics. The dependent variables in these analyses, performances on specific tasks, are operationally defined. The independent variables used to specify task characteristics are also operationally defined. Latent ability, which is determined by overall performance on the test as a whole, is also initially operationally defined, but can be interpreted as a latent, causal attribute after the causal model of performance has been developed.

The Role of Psychometricians

At the end of his paper, Professor Borsboom suggests that psychometricians read widely and that they get involved in the development of substantive theory. This is a great suggestion. It can foster the development of both substantive areas and psychometric theory. I think that it would be especially useful for psychometricians to join research groups as full participants who are actively engaged in all aspects of research projects. I am not talking about a consulting gig or an elegant reanalysis of some existing data set (not that there is anything wrong with either of these activities), but rather about participating in the preliminary analysis, hypothesis development, study design, and data collection, as well as the analysis and interpretation of the results.

Psychologists are likely to make use of psychometric models if they perceive these models to be helpful to them in achieving their goals. Getting them to make greater use of psychometric models is partly a function of making the models accessible (through education, better textbooks, computer programs), as Professor Borsboom suggests, but it is also a function of getting their attention, and a great way to get their attention is to show them what you can do for them.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Benjamin, A. (1955). *Operationism*. Springfield, IL: Charles C. Thomas.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

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

Cronbach, L.J. (1989). Construct validation after thirty years. In R.E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.

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

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

Ennis, R. (1973). Operational definitions. In H. Brody, R. Ennis, & L. Krimerman (Eds.), *Philosophy of educational research* (pp. 650–669). New York: Wiley.

Kane, M. (in press). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Manuscript received 15 MAR 2006
Final version received 20 MAY 2006
Published Online Date: 23 SEP 2006

PSYCHOMETRIKA—VOL. 71, NO. 3, 447–450
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1500-5

WHEN A PSYCHOMETRIC ADVANCE FALLS IN THE FOREST

LEE ANNA CLARK

UNIVERSITY OF IOWA

Borsboom (2006) attacks psychologists for failing to incorporate psychometric advances in their work, discusses factors that contribute to this regrettable situation, and offers suggestions for ameliorating it. This commentary applauds Borsboom for calling the field to task on this issue and notes additional problems in the field regarding measurement that he could add to his critique. It also chastises Borsboom for occasionally being unnecessarily perjorative in his critique, noting that negative rhetoric is unlikely to make converts of offenders. Finally, it exhorts psychometricians to make their work more accessible and points to Borsboom, Mellenbergh, and Van Heerden (2003) as an excellent example of how this can be done.

Key words: psychometrics, psychological measurement, construct validity, critique

Borsboom (2006, p. 435), in his own words, "has sketched a rather grim picture of the role of psychometrics in psychology." He deplores the fact that advances in psychometric modeling have had little impact on psychological testing, and illustrates his point with two examples, taking the field to task for treating factors that have been identified via principal components analysis (PCA) as latent variables and for ignoring measurement invariance when interpreting group differences on psychological tests. He then discusses three domains in which major obstacles exist: theory/philosophy of science, pragmatic factors, and substantive theory.

As one who considers measurement of fundamental importance in psychology and who has developed several psychological instruments, I found myself alternatively cheering, indignant, and concerned as I read Borsboom's paper: Cheering, when his critique reflected my own dismay at the lack of respect and attention that many psychologists afford measurement; indignant, when I felt that he belittled—whether fairly or unfairly—substantive measurement psychologists, a group in which I include myself; and concerned, when he criticized practices that I feared that, in my own ignorance, I myself might have engaged in. The last I felt strongly enough that I read half-a-dozen of the papers Borsboom cites before beginning this commentary, in which I elaborate on the basis for each of these reactions.

I begin by cheering Borsboom's critique. There is no question that most psychologists receive too little training in mathematics and thus lack the skills needed to understand fully various complex psychometric issues and even to use advanced psychometric techniques in their work. Some recognize and compensate for this deficiency by collaborating with those who have the needed skills, but many too often choose the easier path of analyzing their data using only what they know already—and that is if they even conduct research; to wit, most psychologists are not researchers at all. There also is no question that many studies have too few participants and that the field would be well-served by reviewers and editors insisting that, to use Borsboom's language, researchers "buy" rather than "steal" the assumption that one can "directly estimate population differences in the theoretical attribute on the basis of observed scores" (Borsboom, 2006, p. 433). Additionally, the lack of substantive theory to guide psychological measurement is also beyond doubt. Psychology is still a young science and I am optimistic that appropriately explicit, testable

I wish to thank Frank Schmidt for his help in preparing this paper.
Requests for reprints should be sent to la-clark@uiowa.edu.

theories will be developed eventually, but it is disappointing that the field generally does not seem eager to embrace that future.

In some ways, Borsboom does not go far enough in his critique. Except as implied by his critiques of the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998) and the common practice of interpreting scores on an ad hoc paper-and-pencil test of X "as if they automatically applied to the *attribute*" of X (emphasis in original, p. 428), he does not criticize what likely are thousands of published studies in which the outcome of an experimental manipulation or the difference between two naturally occurring groups is assessed with an instrument or procedure developed for that particular study, with the resulting scores treated as a psychological construct (i.e., attribute), with no apparent thought given to the measurement issues involved. When others later review the literature, such studies are categorized by whatever label the researchers selected for their purported construct, be that comparative optimism, social attraction, or self control. At least those who use PCA know they should—and make an explicit attempt to—address measurement issues.

A related pet peeve of mine that he also could have mentioned is the far-too-common justification for the selection of study measures amounting to little more than a mantra: "According to the literature, measure X's reliability is good and it has been shown to be valid (citation)," with no numbers for reliability, nor even a "valid for Y" nod to criterion validity (about which I share Borsboom's concerns). As a reviewer, I often request that the authors either take reliability and validity at least somewhat seriously and provide some relevant data, or simply eliminate the mantra, which has no added value. Perhaps this sad practice reflects helplessness in the face of the "aura of intractability that surrounds the problem of 'construct validity' " (2006, p. 430), but I concur that this is "merely a poor excuse for not taking the measurement problem seriously" (2006, p. 431).

Nor does Borsboom take on projective testing with its theoretically rich but empirically questionable relation between responses and attributes. Given the recent official statement by the Board of Trustees of the Society for Personality Assessment (BTSPA) that "the Rorschach possesses reliability and validity similar to that of other generally accepted personality assessment instruments, and its responsible use in personality assessment is appropriate and justified" (BTSPA, 2005) including it in his critique would have been timely, but perhaps he judged projective testing not to be worth the time and journal space.

On the other hand, portions of Borsboom's critique aroused my indignation, and illustrated another reason, which he did not discuss, that also contributes to the general failure of psychologists "to incorporate psychometric modeling techniques in their methodological inventory" (2006, p. 425), specifically, a tendency of the mathematically inclined to denigrate the efforts of those who lack their measurement skills. At times this emerges as sarcasm: "Once a question gets labeled as a problem of construct validity, its difficulty is considered superhuman and its solution beyond a mortal's ken" (2006, p. 431). He goes on to note that physicists aren't "involved in the 'never-ending process' of figuring out whether meter sticks really measure length" and wonders, "So why should construct validity be such an enormous problem in psychology"? (2006, p. 431).

I have acknowledged that this point has some validity. Nevertheless, those who take construct validity seriously and also recognize the difficulty of fully understanding psychological constructs, don't deserve to be mocked, any more than we might mock theoretical physicists who acknowledge that the Big Bang Theory or General Relativity are not finished products and that their validation is similarly a "never-ending process" (Messick, 1988, cited in Borboom, 2006, p. 431), even duly noting the difference between validating a theory and a measure. Perhaps I should fight fire with fire and deplore that there apparently are psychologists who are not able to distinguish the qualitative difference between validating a measure of intelligence and a measure of length?

At other times this "superior" attitude takes the form of simplifying complex issues and then using this simplification to make those who do not attend to the complex issue look foolish. For example, Borsboom uses the argument developed by Millsap (1997) to illustrate how psychologists routinely ignore important measurement issues. He summarizes Millsap's argument, "if two groups differ in their latent means, and a test has prediction invariance across the levels of the grouping variable, it must have measurement bias with regard to group membership" (2006, p. 427). He goes on to say that if psychologists were good psychometricians, then "to put forward invariant regression parameters as evidence for measurement invariance would be out of the question in every professional and scientific work that appeared after 1997" (2006, p. 427). Finally, he documents that this has not happened, citing both official publications and Hunter & Schmidt (2000), concluding that "test bias is in acute need of scientific scrutiny (Borsboom, 2006, p. 427)."

When someone takes issue with Frank Schmidt, my antennae go up, so I took a close look (admittedly for the first time) at Millsap's argument, which made the narrower point that in measuring two groups on a test and a criterion that share a single common factor, it is not possible for there to be simultaneously "prediction invariance" (i.e., the same regression slope and intercept), "measurement invariance" (i.e., the same factor structure, which in the case of a single factor means equivalent factor loadings), and different observed-score distributions (specifically, differences in the variance of the common factor). So why would Hunter and Schmidt (2000, p. 151) declare that "the issue of test bias is scientifically dead"? From the arguments that Hunter and Schimdt make in their paper, I suspect that the answer lies in the fact that most psychological constructs used in applied settings are broadly specified rather than narrowly precise, so that in cases relevant to Millsap's argument—and with factor invariance evaluated via significance tests—trivial differences in factor variation emerge as significant. In other words, whereas Millsap's point, strictly speaking, may be true, it makes no practical difference when applied in real world settings and thus has been largely ignored.

To his credit, Borsboom (2006, p. 428) acknowledges "that psychometricians insufficiently communicate their advances to psychologists." Ironically, Millsap (1997) is an excellent case in point. Borsboom implies that the paper is "clearly written" (p. 427) and he may find it so, but I did not, which directly relates to his point that psychology journals may reject papers for being too difficult (i.e., containing mathematics). This problem is only partly with journal standards—it is not unreasonable for journals to want their contents accessible to their readers—but also that mathematical psychologists typically do not write in a way that is accessible to "mainstream" journal readership.

If Millsap (1997) were written so I could assign it in a first-year graduate class (and I believe it could be), it then would have a better chance to advance the field. Perhaps the currently popular push to "translate" research findings into practically usable forms will spill over into measurement, and mathematically inclined psychologists will begin to make their advances readily available to those working in more applied areas such as personality assessment and clinical practice. This might not hasten psychology becoming a mathematically grounded science, it might even delay it by at least partially obviating the need for mainstream psychologists to learn the relevant mathematics, but it also might fill the gap in the meanwhile.

Finally, Borsboom's paper engendered my concern that I was missing something important in my own work, so I read a number of the paper he cites in his criticism of the current status of psychometrics in psychology. I was relieved to find that—whereas the papers were quite useful for sharpening my thinking and learning some psychometric terminology—I did not need to totally revamp my research program. As I hope I made clear in my cheering section, I do not dispute that many of Borsboom's critiques are justified, but at the same time, some of them are exaggerated (to make a point?) in ways that ultimately distract from his message. First, his critique appears to be addressed to all psychologists who are not psychometricians, whereas actually it is aimed at only

a subset (though admittedly a large subset). Although, to his credit, he admits that his "extreme examples cannot simply be generalized to the field of psychology as a whole" (2006, p. 428), this point runs counter to the general tenor of his paper. Regrettably, his critique is most directly aimed not at those who will read his paper, but at those who should, but won't. My point here is that important psychometric advances must be communicated to mainstream psychologists, so the question is how to preach to the congregation and not just the choir? To use another analogy, if an important psychometric advance falls in a forest where there are no psychologists to hear it, does it make a sound? Apparently not.

Second, I believe Borsboom exaggerates the gulf between schools of thought. He has few kind words for classical test theory and yet cites Mellenbergh (1994) who states that "a Rasch-type model can be formulated as a special case of classical test theory" (p. 300, citing Moosbrugger & Müller, 1982). When you are trying to make a persuasive argument, it is more helpful to bridge from what one's audience understands or believes already to the point that one wants them to embrace. Alienating one's audience by simply rejecting key elements of their current understanding as fallacious is more likely to meet with resistance than open arms.

Let me end with more cheering. As mentioned, I read several papers Borsboom cites, and the most valuable paper from my perspective was one of his own (Borsboom et al., 2003, p. 205) which laid out quite clearly (indeed, although there are a few path diagrams with Greek letters, the most complex equation is "$2 + 2 = \ldots$"), the problem with what they call a "uniformity-of-nature assumption," that "the relation between mechanisms that operate at the level of the individual and models that explain variation between individuals is often taken for granted, rather than investigated" (p. 215). The paper served as an important reminder—and provided a well-reasoned argument to support it—that a complete science of psychology necessarily will require understanding of both processes and structures as well as their interrelations. For this, we will need psychologists across our broad discipline to bring their special expertise to bear on common problems, so that together we can find uncommon solutions.

References

Board of Trustees of the Society for Personality Assessment (BTSPA) (2005). The status of the Rorschach in clinical and forensic practice: An official statement. *Journal of Personality Assessment*, *85*, 219–237.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440.

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

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Hunter, J.E., & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, *6*, 151–158.

Millsap, R.E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

Manuscript received 20 MAY 2006
Final version received 20 MAY 2006
Published Online Date: 23 SEP 2006

PSYCHOMETRIKA—VOL. 71, NO. 3, 451–455
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1497-9

PSYCHOMETRICS IN PSYCHOLOGICAL RESEARCH: ROLE MODEL OR PARTNER IN SCIENCE?

KLAAS SIJTSMA

TILBURG UNIVERSITY

This is a reaction to Borsboom's (2006) discussion paper on the issue that psychology takes so little notice of the modern developments in psychometrics, in particular, latent variable methods. Contrary to Borsboom, it is argued that latent variables are summaries of interesting data properties, that construct validation should involve studying nomological networks, that psychological research slowly but definitely will incorporate latent variable methods, and that the role of psychometrics in psychology is that of partner, not role model.

Key words: construct validity, latent variable models, role of psychometrics in psychology, status of latent variables.

Introduction

Borsboom (2006) addresses a topic that probably has occupied psychometricians' minds for quite some time: Why is it that psychology takes so little notice of the modern developments in psychometrics? According to Borsboom, these developments are, in particular, latent variable models such as item response models, confirmatory factor models, and latent class models. Borsboom suggests three causes for this lack of interest.

First, their operationalistic orientation leads psychologists to ignore multidimensional attribute structures or nominal attributes as possible causes of test scores. Instead, psychological attributes are equated with test scores such as the number-correct score. This results in assigning, more or less habitually, the same properties to attributes than to test scores. An example is the linear ordering of individuals on a single dimension.

Second, their reliance on classical test theory prevents psychologists from seeing the distinction between observable variables and psychological attributes. This is due to the definition of the true score as the expectation of an observable variable, the test score. Thus, research is led in the direction of investigating reliability and validity of test scores and away from studying relationships between psychological attributes and behavior elicited by the test.

Third, psychologists have embraced the Popperian idea that theory is continuously under construction, thus accepting beforehand that it cannot ever be fully determined what a test measures. Thus, the process of construct validation (i.e., finding the meaning of a test score) becomes an enterprise that is always underway, and the discouraging thought of endlessly investigating the meaning of measurement is often taken as an excuse for not trying at all.

Borsboom's way out of this dead-end street is to be found in latent variable modeling. Although I think that the propagation of latent variable methods will stimulate the construction of better instruments, I do have four critical comments. They concern the status of statistical latent variables relative to psychological attributes, construct validation, the state of psychological research, and the role of psychometrics in psychology.

Requests for reprints should be sent to Klaas Sijtsma, Department of Methodology and Statistics, FSW, Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands. E-mail: k.sijtsma@uvt.nl.

Latent Variables and Psychological Attributes

Like Borsboom, I believe that the widespread use of latent variable models will stimulate researchers to think about:

(1) the possibility that their attributes may be multidimensional rather than unidimensional, and that an attribute may be represented by nominal categories rather than a continuum;
(2) the distinction between observable item scores and latent traits, factors, and latent classes so that more thought is given to the causes of item responses; and
(3) using item response models with restrictions on the item parameters, cognitive diagnosis models that combine features of multidimensional item response models with such restrictions, and latent class regression models and other latent class structures to better understand what tests measure.

Unlike Borsboom, however, I think that latent variables—latent traits, factors, and latent classes—are summaries of the data and nothing more, and that, compared to Borsboom's ambitions, this seriously limits the possibilities of latent variable models. Let me explain my point of view. Suppose a psychologist uses 20 items for measuring the attribute of inductive reasoning, an attribute that is at the heart of the intelligence construct. He hypothesizes that the items elicit responses based on the same cognitive mechanism; that a higher level of inductive reasoning (or, similarly, a better degree of functioning of the cognitive mechanism that is labeled inductive reasoning) increases the likelihood that a respondent solves the items correctly; and that the attempts to solve an item do not in any way affect the probability that subsequent items are solved correctly. In Borsboom's and my perfect worlds, the researcher puts his hypotheses to the test by means of a unidimensional, monotone, locally independent item response model.

This is where the strength of latent variable models resides, as far as I am concerned: in the possibility to test the assumptions of the models through their observable consequences. Classical test theory does not provide the researcher with the means to do this and leaves him with "measurement by fiat." However, the possibility of testing of a model for the data—indeed a mighty weapon—is also where the strength of these models ends. The latent trait in an item response model is a mathematical summary of variation in the data between cases. Its direct meaning derives from a mathematical model, not a psychological theory, and it is estimated from the data, not the cognitive processes behind the data. The best that can happen, indeed, is that the item response model fits the data, which may then be taken as support for the hypothesis that the test is driven by inductive reasoning.

I use the word "support," not "proof," because there always remains a "gap" between fitting a model to data and the final piece of evidence that the test indeed is driven by the hypothesized attribute. This gap can only be "crossed" by inference or human judgment, and the hypothesis or the theory becomes more likely the more evidence is collected in their support. However, there is no way in which it can be decided that at a certain point the evidence is complete, other than that different researchers of good reputation agree that it does (Hofstee, 1980). How might additional evidence look in our inductive reasoning example?

It might have the form of other tests for inductive reasoning that use a different kind of item, yet be hypothesized to elicit the same or nearly the same cognitive processes as the previous test. Different cognitive skills or item properties might be distinguished using the new test(s), and a componential item response model (e.g., De Boeck & Wilson, 2004)—actually, akin to a nonlinear regression model—might be fitted to new data. Both the fit and the misfit of such models can contribute valuable knowledge to theory formation for inductive reasoning. A new set of items may give rise to another—that is, not exactly the same as the first—latent trait or even a set of latent (sub)traits, which is not at all unlikely in an item set designed to elicit different

skills or to let different item properties exercise their influence on cognitive processes. This need not worry anyone, as long as one sees latent variables as tools for summarizing data, not entities independent of the data on which they are fitted. My conclusion is that statistical latent variables help describe variation in data that is consistent with a putative psychological attribute; but, in isolation, goodness of fit of a latent variable model to data does not illuminate the existence or functioning of the attribute.

Construct Validation

Based on the assumptions that psychological attributes exist and exercise a causal influence on item responses and that latent variables represent psychological attributes, Borsboom proposes to limit the process of construct validation to latent variable modeling of item response data alone and discard studying relationships with other variables in a nomological network. I just explained that I disagree with the second assumption; below I will comment on the first.

The first assumption boils down to a conception of construct validation that entails the use of a substantive theory about the attribute of interest for predicting the pattern of responses to a set of items and, reversely, using latent variable modeling of these responses for establishing construct validity as a property of a test (Borsboom, 2006; Borsboom, Mellenbergh, & Van Heerden, 2004). This conception excludes tests for the vast majority of psychological attributes that are not supported by the kind of detailed and established theory that Borsboom seems to have in mind. For these attributes the "theory" will make inaccurate predictions and, as a result, the latent variable model will not fit. But what will one do next?

Taking substantive theory as a starting point for test construction is an excellent idea that has existed for a long time but is not widely practiced. The reason probably is that much theory is still in its puberty, infancy, or even at the fetal stage. Given this state of affairs one often has no other choice than to cling onto about every piece of evidence available in learning about test validity, including relationships with other interesting variables. There is no reason to exclude well-developed theories about attributes and their tests. For example, transitive reasoning is an example of a theoretically well-developed attribute (Bouwmeester, 2005), but its relationship with several verbal abilities may be interesting in its own right. Such studies are justified when different researchers of good reputation disagree about this relationship, and may shed more light on transitive reasoning but also, perhaps unexpectedly, on verbal intelligence.

Borsboom's assumption about the ontology and causality of psychological attributes seems to lead to a very restrictive conception of the process of construct validation: Elegant in its rigor but impractical for psychology (and many others areas). It seems to me that we still know so little about the functioning of the human brain in general and cognitive processes including those underlying personality traits and attitudes in particular, that it is difficult even to say what an "attribute" is. In the absence of such knowledge, I prefer to consider psychological attributes as organizational principles with respect to behavior. Thus, my point of view is that psychological attributes define which behaviors hang together well and are useful to the degree in which tests sampling these behaviors play a role in predicting interesting psychological phenomena.

The State of Psychological Research: A Case Study

In his own words, Borsboom sketches a grim picture of psychological academic research. I agree that occasionally psychologists are capable of wild adventures but not unlike any breed of academicians—including those involved in psychometrics, I would like to add. However, I believe psychology is in a better state than Borsboom suggests. I also think that psychometrics

has much help to offer, but perhaps less spectacularly than Borsboom would hope for. Here is what I see in present-day test and questionnaire construction.

At the time of writing this reaction, I was involved in several projects together with researchers from education, psychology, marketing, and medicine. Each of them uses questionnaires to measure an attribute: attitude toward homework and study (education), self-concealment (i.e., keeping things secret from others; psychology), service-quality (of computer helpdesks and restaurants; marketing), and perceived educational climate in Dutch hospitals (medicine). Each of the researchers is trying hard to work with a good definition of the attribute of interest that is well founded in the relevant literature; to find a useful operationalization of the attribute into a set of items; to be aware of item wording, use of both positive and negative item phrasing, and the threat of response sets; and to think about the composition of the sample and the way in which the data should be collected. They all use item response models or other modern statistical methods, such as latent class models and multilevel models. Of course, they do many of these things acting on my advice but what counts is that they are motivated and will carry on their knowledge to others. Thus there is progress which, however, proceeds slowly.

I have to admit that it is difficult to explain to my colleagues why latent variable models are better methods than Cronbach's alpha, the item-rest correlation, and principal components analysis. After all, what item response theory does is model the dimensionality of one's data and represent persons and items on a scale, but isn't this exactly what principal components analysis and classical test theory also do? A psychometrician shakes his head in disbelief about so much naivety but a psychological researcher who has not been trained to see the difference thinks this is "much ado about nothing." An effective recipe to make people see the—admittedly often subtle—differences is to do the classical and modern analyses next to one another and report both. For example, what convinced my fellow researchers to use Mokken scale analysis was that it allows the investigation of dimensionality by means of user-friendly software and without the artifacts of principal components analysis and factor analysis caused by discrepancies in the frequency distributions of the item scores. Notice these are practical arguments. Given these experiences, I think that researchers from substantive disciplines will accept modern psychometric methods if they are convinced of their practical advantages over classical methods and if results can be obtained without much trouble (which could mean including a psychometrician in the project).

The Role of Psychometrics in Psychological Research

Borsboom spurs his fellow psychometricians to take the lead in psychological research and use their latent variable models as blueprints for psychological measurement devices. This is motivated by the lack of fine-grained psychological theories that define exactly what a particular attribute stands for in terms of cognitive processes and functions, and how it should be operationalized in terms of items. The question then is whether in the absence of substantive theory an "empty" statistical model can fill the void and determine how attributes are measured. For example, unidimensionality, monotonicity, and local independence are necessary to have at least an ordinal scale, but they do not imply the kinds of tasks and data psychologists may find ideal for the assessment of a particular attribute.

As I see it psychometrics cannot replace substantive theorizing about intelligence and personality for designing good measurement instruments; it can only provide support. To learn about intelligence and personality, more and more research has to be done in the best traditions of these fields. Good substantive theories are the basis for good operationalizations and measurement procedures. The role of psychometricians is to make researchers more aware of the importance of sound theory and its operationalization, an appropriate research design, a correct definition

of the population and corresponding stratification of the sample, and the pitfalls in designing a test or questionnaire. Important additional questions, several mentioned by Borsboom, are: Do I expect a unidimensional or a multidimensional latent structure underlying the data? Are dimensions continuous are categorical? Should the items be questions, statements, tasks, games, or assignments? Should responses be oral, in writing, or sensomotoric? Should the data be correct/incorrect scores, ordered rating scale scores, category membership scores, or response times? Such choices determine which method should be used for data analysis.

Borsboom's approach is different in that he would take a psychometric model as point of departure and say: For a measurement instrument to have these particular properties, this is how your test should look like and these are the kinds of data you have to collect. It looks as though this view is somewhat at odds with Borsboom et al. (2004) who posit a substantive theory as point of departure. However, given that they assume an ontological status for the attribute and assume that a latent variable represents an attribute (also, Borsboom, 2006), it follows that latent variable models indeed provide blueprints for theory about attributes.

Instead of blueprints for theory, I see latent variable models as tools for analyzing data. They perform best, like any statistical method, when the data result from a well-established substantive theory. Nothing beats a good theory: if one knows which strings to pull, the expected data structure will stand out clearly and statistical analysis will be simple. Test construction should always be based on substantive theory, no matter how primitive, because only then does one know what one is looking for by means of statistical analysis, and only then can expectations be refuted or supported. Absence of theory leads to data beset with many weak signals and an overdose of noise, and the outcome of data analysis depends to a high degree on the statistical model used instead of substantive theory. Alternatively, running many models will usually not contribute greatly to theory formation other than, for example: "It looks like your data are primarily unidimensional but this may depend largely on the items you used." Thus, the role of psychometricians in psychological research is to be found in propagating the formation of theory, the operationalization of the attribute, the construction of the test, and the choice of the appropriate psychometric methods for analyzing the data, in that order.

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Bouwmeester, S. (2005). *Latent variable modeling of cognitive processes in transitive reasoning*. Tilburg, Tilburg University (PhD dissertation).

De Boeck, P., & Wilson, M. (2004) (Eds.). *Explanatory item response models. A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Hofstee, W.K.B. (1980). *De empirische discussie. Theorie van het sociaal-wetenschappelijk onderzoek. (The empirical discussion. Theory of social science research*). Meppel, The Netherlands: Boom.

Manuscript received 7 MAY 2006
Final version received 19 MAY 2006
Published Online Date: 23 SEP 2006

PSYCHOMETRIKA—VOL. 71, NO. 3, 457–461
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1501-4

MEASUREMENT WITHOUT COPPER INSTRUMENTS AND EXPERIMENT WITHOUT COMPLETE CONTROL

WILLEM J. HEISER
LEIDEN UNIVERSITY

1. Introduction

One basic reason that measurement in psychology requires statistics is that psychologists do not use copper instruments anymore, as they used to do in the nineteenth century. Instead, they determine test or total scores on the basis of miniature experiments with discrete outcomes, and use a variety of standard statististical techniques for reaching conclusions on the basis of observed data.[1] Borsboom (2006) wants us to believe that psychologists are seriously misled in their hope that they can make progress this way, and recommends an invasion of psychometricians carrying IRT missiles and SEM guns into psychology. My prediction is that such an invasion would simply be ignored. That is not to say that whenever psychometric modeling really makes a difference, no attempts should be made to reach the mainstream of psychology. Indeed, many psychometric contributions that are obsolete according to Borsboom, like Cronbach's alpha and exploratory factor analysis, in fact entered into the mainstream of psychology only because they tend to provide sensible answers to real problems, which cannot be easily surpassed. We should be more proud of them (even when a bit vulgarized), and carefully foster our accumulated knowledge base. Apart from strictly psychometric contributions, it has always been a task of psychometricians to introduce relevant new developments in the broad domain of mathematics and statistics into psychology. I am convinced that we should continue to do so, even when it concerns "observed score techniques" that are so detested in the focus article.

Borsboom is right in pointing out that the impact of IRT modeling in academic psychology is limited, and that problems of measurement invariance and test bias are ill-understood and neglected (but of course the IRT movement always had a primary focus on its major successes in another discipline, educational testing). Of the factors that he mentions as hindering the fruitful interplay between psychologists and psychometricians, I have no quarrel with the substantive and with most of the pragmatic ones, but I fail to see the relevance of the theoretical factors. In the following, I will try to explain why, and offer an important factor overlooked by Borsboom, which has to do with the changing relation between Cronbach's (1957) "two disciplines of scientific psychology."

2. Measurement without Copper Instruments

It is undoubtedly true that the single most important and typical contribution of psychometrics to both psychology and statistics is the latent variable. We have true scores, Thurstone's

Requests for reprints should be sent to Willem J. Heiser, Department of Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands. E-mail: Heiser@fsw.leideuniv.nl.

[1]The fact that psychologists so often have to deal with discrete outcomes (categorical data with only a few categories, ordinal data, counts), often with repeated measures, and sometimes with typically structured designs, defines the niche for psychometrics in the larger domain of statistics.

discriminal processes, factor analysis, multidimensional scaling models, structural equation models, item response models, and so on. Latent variables are hypothetical entities that may be fixed (parameters) or random (stochastic variables), and may also be classified in other ways. They are the basic elements of measurement models. This framework permits us to transform discrete psychological responses into numerical outcomes with known quality characteristics. Thus, the measurement model is a hypothetical copper instrument.

2.1. Role of the Latent Variable

It is often said that measurement models specify theoretical relations between concepts (latent variables) and observables. That sounds innocent enough, but as Borsboom convincingly shows in his example of measuring conscientiousness, it leads to an embarrassment of riches. I concur with Borsboom in that psychological theory is not strong enough to allow a motivated choice between specific psychometric models. But I also hold that this weakness is exactly why research psychologists use the sensible strategy to avoid reliance on hypothetical copper instruments whenever they can. As I have tried to argue elsewhere (Heiser, 2003), they are not interested in measurement per se, but in the establishment of (cause and effect) relations.

2.2. Attack Against Classical Test Theory

What can a psychologist do if she wants to report or use a concrete person score? It seems to me that calculation of scores unavoidably depends on the data. Under the condition of an embarrassment of riches, suppose she arbitrarily chooses to count those responses that are positive manifestations of the concept she wants to score. There are old results in psychometric theory that ensure that she cannot be too far off, compared to other scoring rules. Differences between weighted and unweighted counting tend to be small if the items being combined are correlated (Gullikson, 1950, p. 355). Also, under the same type of regularity conditions, there is a slightly nonlinear, but monotonic functional relation between a latent trait testimate and a true score estimate (Lord and Novick, 1968, p. 386), implying that the order of the person scores is identical. Recently, Warrens, De Gruijter, and Heiser (2006) showed that a similar relation exists between the latent trait person score and the optimal scaling person score.

It turns out that by far the most important consideration is that the items form a homogeneous set. There is again an embarrassment of riches in the choice of methods for finding a homogeneous set, but once found (approximately), different methods to calculate person scores are close to equivalent for research purposes (more care may be required for individual selection decisions). Classical test theory is not obsolete. Psychometric models at the item level are a refinement. Throwing classical test theory out of the window would only impair the credibility of psychometrics, and increase the gap with psychology.

2.3. Questionable Interpretations of Test Scores

The interpretation of principal components as "biologically based psychological tendencies," endowed with causal forces, as cited in Borsboom, is indeed a long stretch of the imagination. But one cannot blame principal components for this type of wishful thinking. Would it be any better with latent traits or factor scores? I do not think so. Perhaps we should put part of the blame on ourselves, since the idea that we can have causal models for correlated data without controlled interventions arose in our own field. Would it be possible that the language of variables with arrows pointing to each other in supposedly meaningful directions is giving the psychologist the false hope that he could discover causal relations instead of just relations?

2.4. Operationalism Rules

Borsboom is right that psychologists operationalize a lot, but that does not imply that they believe in philosophical operationalism. Neither do they need to believe in the psychometric dream of the hypothetical copper instrument, in which observables are related to theoretical attributes. Rather, like other scientists, psychologists tend to believe in the more general idea of approximation. As long as some protocol of data collection and/or method of data analysis can be justified, as providing an approximation of the psychological variable under study, it will do. Psychometricians might make fine distinctions between true scores and latent traits, or between formative and reflective models, for psychologists these are just two brands of approximations. They have to take a leap of faith anyhow, and it requires clear evidence of superiority in a variety of aspects for one particular method to become the preferred brand.

3. Experiment Without Complete Control

Apart from the measurement problem, there is a second reason why psychological research needs statistics: it is usually impossible to keep irrelevant variables fully under control. This difficulty is sometimes called *the third variable problem*. Picking up and ruling out third variables is the driving force of research design. Important methods of control are randomization (adding chance to the process!), factorial crossing, blocking, introducing covariates, and so forth. Although third variables can also be controlled ex post facto by regression methods, the Fisherian style of experimental thinking has caused a revolution in psychological research (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989, Chap. 6). What are the consequences?

3.1. Samples Size Issues

Under this heading, Borsboom launches what at first appears to be a side attack against "experimentally oriented research" and the "various species of ANOVA," which would involve betting on observed score techniques and "stealing" assumptions. It is one thing that Borsboom cannot see the blessing of small sample statistics, but in any case, with this part of his argument he is not going to win the hearts of psychologists, who are discovering that the great thing of explanatory or independent variables is that you do not need to measure them.

3.2. The Rise of the Independent Variables

Borsboom underestimates the enormous impact of the movement toward more experimental research in psychology, with its emphasis on bringing situational variables under tight control, its attention for causal mechanisms by formulating "cause–effect hypotheses," and its tendency to regard individual differences as a nuisance since they increase within-treatment variance.[2] In contrast to what Borsboom believes, psychologists are much less frequently eye-balling correlations than they used to do. Fifty years ago, Cronbach (1957) could still write,

> In contrast to the Tight Little Island of the experimental discipline, correlational psychology is a sort of Holy Roman Empire whose citizens identify mainly with their own principalities. The discipline, the common service in which the principalities are united, is the study of correlations by Nature. While the experimenter is interested only in the variation he himself creates, the correlator finds his interest in the already

[2]Ironically, psychometricians specialized in IRT call the latent variable measuring individual differences a nuisance variable or, collectively, the nuisance parameters. Although the reason is different, the low esteem for the variations of Nature is equally disconcerting.

> existing variation between individuals, social groups, and species. (Cronbach, 1957, p. 672)

Cronbach also thought that "the tide of separation in psychology has already turned," and mentions as a prime example Meehl's introduction of construct validity in test theory, "capitalizing on the methodological and philosophical progress of the experimentalists" [*sic*!]. Nevertheless, I believe that fifty years later all signs are telling us that the experimentalists are winning big-time over the correlationists. The Holy Roman Empire is falling apart, and the Tight Little Island is growing into an archipelago where the sun never goes down.

The experimental method is triumphing in many areas outside traditional experimental psychology, like social, developmental, clinical, and even organizational psychology. Time and space do not permit going deeply into the reasons and effects of this revolution. But a major methodological aspect (and advantage for the psychologist) is that the independent variables need not be measured, but instead are manipulated, in which process they are reduced to attributes. If you study the effect of fear, there are many possible manipulations to create fear that can be compared with a control condition, but in any case the fear treatment is "on" or "off." Reduction to attributes implies that the theoretical model or reasoning to predict the outcome of the experiment can be qualitative instead of quantitative. Under the experimental method, psychology can reason in attributes, which explains what Borsboom calls "the almost complete absence of strong psychological theory." There is no need for quantification, except for the dependent variable.

3.3. Traits and States

Traits (either manifest or latent) can be dependent variables only in quasi–experiments, for instance when we compare monozygotic twins raised together and raised apart, since they are stable person characteristics. When the aim is to change the dependent variable by experimental manipulation, it must be a *state*. Quantitative state variables almost always involve counts, rates, ratings, or time, and are rather intricately related to the substantive research paradigm. They are never standardized with respect to some population, because effects of experimental manipulation are measured with respect to each other or with respect to a control group. They can also be more "quick and dirty" than standardized tests, because measurement error will only increase within-treatment variance, but not change between-treatment variance. Although effect size is negatively affected by measurement error, that involves a calculated risk and not a blind gamble; after all, a research paradigm comes into wider use only if its originator demonstrates that under typical circumstances of the setup reasonable to large effect sizes can be achieved.

4. Conclusion

Fisherian methodology rules in psychology, while the homeland of psychometrics is correlational methodology. Borsboom's timely discussion paper forces us to think hard about strategies that could save us from isolation and irrelevancy. Some of Borsboom's suggestions (write textbooks, publish widely) are fine. But from my analysis it should be clear that his suggestion to become an active psychological researcher is underestimating what it takes to be part of the psychological research community. For almost all of us, it would take an irreversible career change. When correlational psychology was still strong, one could imagine that psychometricians were a special kind of psychologist, since—after all—test theory was its infrastructure. But this is no longer the case.

There are signs that cognitive psychologists finally face individual differences as a serious factor and start modeling them. We should of course support this development whenever we can. It also appears that the mathematical psychology community finds itself in similar dangers as we do. I would strongly favor an attempt for rapprochement. It would add mass and focus if we had a united platform for the whole of quantitative psychology, following the motto on the cover of this journal. In the recent past, some of our colleagues have made a career change to statistics, but such a move only rarely increased their impact. Psychometrics is a discipline by itself, with a body of results that stood the test of time, but now it has to find a new balance between psychology and statistics.

Latent variables are important, but we should not try to push them at all costs, and they can no longer be the dominant instrument in our repertoire. Psychology needs a new generation of statistical techniques adapted to its current challenges. Physiological outcome measures are hot, so there is a need for functional data analysis. More hierarchical data are collected, so there is a need for multilevel analysis. There is increased interest in moderator variables, so we should work on regression trees in (quasi-) experimental setups. These are just a few examples to broaden the scope of psychometrics.

Finally, psychometrics should care more about its image in the outside world. What we do not do enough of—and I blame myself, too—is propagating and defending our heritage in the larger scientific and public community. We should follow the example of people like John Carroll, who took a brave stand against Stephen Jay Gould's biased views on mental testing and factor analysis (Carroll, 1995). Join forces with friends, and attack that enemy!

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440.
Carroll, J.B. (1995). Reflections on Stephen Jay Gould's *The mismeasure of man* (1981). *Intelligence, 21*, 121–134.
Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance*. Cambridge: Cambridge University Press.
Gullikson, H. (1950). *Theory of mental tests*. New York: Wiley.
Heiser, W.J. (2003). Trust in relations. *Measurement: Interdisciplinary Research and Perspectives, 1*, 264–269.
Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Warrens, M.J., De Gruijter, D.N.M., & Heiser, W.J. (2006). A systematic comparison between classical optimal scaling and the two-parameter IRT model. *Applied Psychological Measurement*, *30*, 1–15.

Manuscript received 30 JUN 2006
Final version received 30 JUN 2006
Published Online Date: 23 SEP 2006

PSYCHOMETRIKA—VOL. 71, NO. 3, 463–467
SEPTEMBER 2006
DOI: 10.1007/s11336-006-1502-3

CAN WE BRING ABOUT A VELVET REVOLUTION IN PSYCHOLOGICAL MEASUREMENT? A REJOINDER TO COMMENTARIES

DENNY BORSBOOM

UNIVERSITY OF AMSTERDAM

First, I thank Sijtsma, Clark, Kane, and Heiser for taking the time and effort to provide a commentary on my paper, and the Editor for allowing me to respond to them. In general, the commentators agree with the thesis that psychometrics and psychology are, to an extent that must be deemed problematic, disconnected. They further agree with the upshot of this diagnosis: Psychometricians need to work harder to make a difference in psychology, and psychologists need to develop a greater awareness of important psychometric developments. However, the commentators also raise several points of disagreement and criticism. I focus on four important topics that concern:

(a) the theoretical factors influencing the current state of affairs;
(b) the question of how to proceed in the absence of theory;
(c) the communication between psychometricians and psychologists; and
(d) the relation between experimental psychology and psychometrics.

Theoretical Factors Revisited

In an attempt to analyze the theoretical factors contributing to the present state of affairs, I identified as likely perpetrators: Operationalism, classical test theory, and construct validity. Kane, Sijtsma, and Heiser disagree with various aspects of this diagnosis.

Both Kane and Sijtsma argue that, in the absence of detailed formal theory, measurement models are necessarily driven by pragmatic factors and global considerations about the attribute of interest. Latent variables that occur in such models are therefore, at least initially, operationally defined. As such, operational definitions play a natural and important role in scientific progress. For this reason, to demand full-blown formal theoretical definitions *before* the first attempt at measurement is to put the cart before the horse. In an important sense "[a] specification of how a theoretical attribute produces certain effects is possible only after the theory is in place; it is not the first step, but one of the last steps in a research program" Kane (2006, p. 444). Sijtsma (2006, p. 453) makes a similar point in arguing that the specification of the causal relation between an attribute and the observations, as advocated in Borsboom, Mellenbergh, and Van Heerden (2004), while perhaps desirable, may be premature because "we still know so little about the functioning of the human brain in general and cognitive processes including those underlying personality traits and attitudes in particular."

Thus, Kane and Sijtsma argue that an orienting ideal, that is, knowledge of how a test actually works, should not be categorically forced upon psychology as a necessary requirement for test use; researchers need breathing space to work with operational definitions, general conceptualizations

Requests for reprints should be sent to Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam. E-mail: d.borsboom@uva.nl

of attributes, and the liberal forms of construct validity theory. I agree with much of this—as long as everybody knows what they are doing and it is clear that we are proceeding in the right direction. Now, perhaps Kane and Sijtsma are in the good part of psychology, the one that makes all the quick and indisputable progress. In the part of psychology where I live, however, all too often researchers seem to have little sense of psychometric direction. They follow dubious guidelines, like "if alpha is higher than .80 the test is reliable," "if the scree-plot shows a good drop the test measures only one thing," and "if the test scores show some reasonable pattern of correlations the test must have some degree of validity"; Clark mentions and criticizes some similar practices. It seems to me that, in such cases, operational definitions, construct validity, and classical test theory are not functioning in the healthy way that the methodology books describe, but instead have become instances of dogmatic thought that hinder progress in psychology. Hence, while I recognize and acknowledge that many of the ideas in classical psychometrics have a legitimate role to play in scientific practice, as Kane, Heiser, and Sijtsma stress, at the same time I would urge psychometricians to take a closer look at how these concepts are functioning in actual scientific practice. I think they will find that much of psychology is not following the idealized template of scientific progress that underlies Kane's and Sijtsma's nuanced points on measurement, but is rather, quite simply, lost when it comes to dealing with psychometric questions. I also think that there is serious room for improvement, because, in contrast to fifty years ago, when the foundations of classical test theory and construct validity were laid down, we now have the computing power and mathematical tools to make modern psychometric models work.

That operationalism, construct validity, and classical test theory are not currently playing a productive role in steering psychology forward does not mean that the ideas and concepts in these theories are *generically* useless. Heiser rightly criticizes such a point of view with respect to classical test theory, but misattributes it to me. I actually think classical test theory is quite useful, for instance,

(a) as a statistical theory of composite test scores;
(b) as a first-order approximation to a restricted class of latent variable models (Holland & Hoskens, 2003); and
(c) as a source of established theorems that can shed light on the relation between theoretical attributes, item response models, and observed test scores (Ellis & Junker, 1997).

However, classical test theory *is* useless as a *measurement model*. The reason is that measurement requires one to relate attributes to test scores, and classical test theory does not have the expressive power to do so (see Borsboom, 2005, pp. 32–33, for a detailed argument). If one is only interested in test scores, classical test theory is fine; but to deal with the measurement problem, one needs a latent variable model.

What To Do in the Absence of Theory?

Latent variable models are most useful if there exists a reasonably detailed view of the structure of attributes and the way that response processes connect this structure to the observations. Sijtsma and Kane both address the problem that this sort of theory is often lacking. Sijstma (2006, p. 453), for instance, states that "[t]aking substantive theory as a starting point for test construction is an excellent idea that has existed for a long time but is not widely practiced. The reason probably is that much theory is still in its puberty, infancy, or even at the fetal stage." Kane (2006, p. 453) states that "[a]ssuming that no formal theory exists (the usual case), the test-development process is necessarily ad hoc, guided by general conceptions of the attribute of interest." For precisely this reason, I argued that it is important to develop substantive psychological theory as it relates to measurement, and suggested that psychometricians should play a more pronounced

role in this process. Without such theory, validation research will indeed be never-ending, as has been argued elsewhere (Borsboom, Mellenbergh, & Van Heerden, 2004), because it means that one is stuck with a black box model of the relation between attributes and test scores. Sijtsma (2006, p. 453), however, thinks this point of view is "impractical," Kane appears to view the ad-hoc development of measurement instruments as something that we simply have to live with, and Heiser (2006, p. 459), states that psychologists "have to take a leap of faith anyhow" so that it does not really matter to which measurement model their leap leads.

This is a remarkable reaction to the current state of affairs. If we all agree that stronger substantive theory is needed to get a grip on the measurement problem, and such theory does not presently exist, then is it really the proper course of action to settle for second best, and just model whatever data substantive psychologists come up with, so that we basically assist them in fitting ad-hoc statistical models to ad-hoc constructed tests? Or should we attempt to develop the required theory? I take it to be rather preliminary to admit defeat at this early stage of theory development, and hence would argue that we take the second option seriously. And if we do this, then why should psychometricians not pay attention to the formalization and development of psychological theory insofar as it relates to measurement? Do psychometricians take the "psycho" in psychometrics seriously, or has psychometrics really become a subdivision of statistics, so that the subtitle of this journal should be changed from "a journal for quantitative psychology" to "a journal for data analysis"? I hope that, in this case, to ask the question is not to answer it.

Communication Breakdowns

One of the causes of the lack of integration between psychometrics and psychology undoubtedly lies in communication problems. Now, with respect to the communication breakdown that is currently in place, Clark reacts strongly to my suggestion that psychometric papers are often ignored in psychology, and lays the blame partly on psychometricians. In reference to Millsap's work on measurement invariance, for instance, she states that "[i]f Millsap (1997) were written so I could assign it in a first-year graduate class, [it] would have a better chance to advance the field" (Clark, 2006, p. 449). I think Clark is correct here, but at the same time feel that this illustrates one of the major problems that hinder psychometric advances in psychology. Think of it again: *If one wants to communicate a psychometric advance, the paper has to be written at a level that allows it to be assigned to a first-year graduate class*. The reason that this is true is exactly what makes it food for thought. Are there no important ideas in psychometrics that cannot be understood by the average first-year graduate student?

Now, all this should not be taken to mean that psychometricians are somehow "superior" to psychologists, or that substantive psychologists "deserve to be mocked," as Clark states (2006, p. 448). As far as I can see, we are all dwarfed by the problem of explaining human behavior; hence nobody is justified in taking a "superior" attitude—and it was certainly never my intention to do so. Further, I agree that it is important that psychometricians write accessible papers that are as clear as possible; and there certainly is much room for improvement on this score. However, I also think that, when the content of a psychometric paper is as manifestly important as Millsap (1997), some effort can reasonably be expected on the part of the reader.

But how important, exactly, *are* such psychometric insights as communicated in Millsap's paper? Clark (2006, p. 449) expresses doubts: "whereas Millsap's point, strictly speaking, may be true, it makes no practical difference when applied in real world settings and thus has been largely ignored." However, it is important to note that, in the present context, the point is not that the requirement of measurement invariance should be raised to the level of a *sine qua non* for research in general—that would be wrong, because its importance clearly depends on pragmatic

dimensions of the research context, i.e., what the data are used for, as well as on the severity and nature of violations of measurement invariance, as Clark correctly indicates. The point is that Millsap's work shows *that prediction invariance cannot be adduced as evidence for measurement invariance*. This implies that the duty to investigate to what extent the inconsistency between measurement invariance and prediction invariance is important to real world settings lies *not* with those who think measurement invariance should be more central; *it lies with those who want to adduce prediction invariance as evidence that there is no test bias*. And it is in *this* perspective that I find it astounding that the official APA and SIOP testing standards and guidelines, which should light the researcher's path on such issues, uncritically embrace an invalid line of argument without issuing a serious qualification or warning on this point.

Psychometrics and Experimental Psychology

Psychometric theory has largely been developed in the context of individual differences research and correlational psychology. Heiser argues that the importance of psychometric theory has become limited, because contemporary psychology is predominantly experimentally oriented; and, according to Heiser, in experiments psychometric theory is not needed. Hence, current psychology has no need for psychometric modeling.

In my opinion, this diagnosis is incorrect. Experimental psychology can benefit enormously from psychometric input, as is demonstrated by some of the best recent work in the field. In fact, one of the examples I used comes directly from experimental social psychology, where Blanton, Jaccard, Gonzales, and Christie (2006) demonstrated the importance of psychometrics for the interpretation of the implicit association test. Other examples include Raijmakers, Dolan, and Molenaar (2001), who used finite mixture modeling to investigate discrimination learning; Visser, Raijmakers, and Molenaar (in press), who applied hidden Markov models to study sequence learning; and Wicherts, Dolan, and Hessen (2005), who used multigroup confirmatory factor analysis to investigate the origin of stereotype threat effects. That these studies were published in top journals suggests that, whatever Heiser may think of the Archipelago of experimental psychology, its inhabitants actually find the psychometric approach to their problems quite refreshing.

Heiser (2006, p. 459) also thinks that "the great thing of explanatory of independent variables is that you do not need to measure them." This may be a truism if one equates explanatory variables with experimental conditions, but that is not what researchers are typically interested in. Just like questionnaire users are not interested in items, but in the attributes that these items measure, so experimentalists are not interested in experimental conditions, but in the processes that these conditions manipulate. In fact, questions like "Do manipulations x and y affect the same process?" are quite common in experimental psychology, and such questions are amenable to psychometric treatment. It is, of course, true that the psychometric theory of experimental manipulation is not well developed, but this is precisely the sort of venue that psychometricians are excellent candidates for exploring. In conclusion, I would say that the experimental Archipelago that, according to Heiser, renders psychometric theory obsolete, is actually a significant growth market for creative and substantively interested psychometricians.

Conclusion

Many interesting and important questions in psychology are either of a psychometric nature, or have a strong psychometric component. With respect to such questions, psychometrics has much to offer to psychology, but has so far not realized its potential. The time is right for

this to change. Many versatile modeling approaches have been developed in the past decades; with the current computing power, using such approaches is quickly becoming a realistic option for researchers. This is illustrated by the fact that, in the past few years, several excellent psychometric modeling approaches to empirical problems have surfaced in various areas of psychology. Psychometric modeling is gaining momentum; and we should use this momentum to create as much educational and research opportunities as possible for researchers who are able and willing to take a psychometric approach to their problems. When psychological researchers see that psychometric modeling allows them to investigate problems that are otherwise inaccessible, that it is not a dead-end street but a source of interesting problems and ideas, that it need not be all that much more difficult to execute than the current default methods of analysis but yields more insight into measurement problems, I think that psychometrics may become firmly entrenched in large parts of psychology fairly quickly. With good tactical planning, no shots need be fired, and we may witness a velvet revolution in the next few years.

References

Blanton, H., Jaccard, J., Gonzales, P.M., & Christic, C. (2006). Decoding the implicit association test: Implication for criterion prediction. *Journal of Experimental Social Psychology*, *42*, 192–212.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychology Review*, *111*, 1061–1071.

Clark, L.E. (2006). When a psychometric advance falls in the forest. *Psychometrika*, *71*, 447–450.

Ellis, J., & Junker, B.W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495–523.

Heiser, W. J. (2006). Measurement without copper instruments and experiment without complete control. *Psychometrika*, *71*, 457–461.

Holland, P.W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, *68*, 123–149.

Kane, M. (2006). In praise of pluralism. A comment on Brosboom. *Psychometrika*, *71*, 441–445.

Millsap, R.E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

Raijmakers, M.E.J., Dolan, C.V., & Molenaar, P.C.M. (2001). Finite mixture models for simple discrimination learning. *Memory & Cognition*, *29*, 659–677.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika*, *71*, 451–455.

Visser, I., Raijmakers, M.E.J., & Molenaar, P.C.M. (in press). Characterizing sequence knowledge using online measures and hidden Markov models. *Memory & Cognition*.

Wicherts, J.M., Dolan, C.V., & Hessen, D.J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*, 696–716.

Manuscript received 19 JULY 2006
Final version received 19 JULY 2006
Published Online Date: 23 SEP 2006