COMMENT

Why Psychometrics is not Pathological

A Comment on Michell

Denny Borsboom and Gideon J. Mellenbergh
UNIVERSITY OF AMSTERDAM

ABSTRACT. This paper comments on an article by Michell (2000), who argues that psychometrics should be qualified as pathological science for two reasons: (a) psychometrics assumes psychological attributes to be quantitative without testing this hypothesis; and (b) the fact that this hypothesis is not tested is disguised. Michell further argues that the hypothesis should be tested using additive conjoint measurement theory. Although relevant to classical test theory, Michell's arguments do not apply to psychometrics in general. In particular, they are largely irrelevant to item response theory models. We show that these models result from introducing probabilistic relations, which are needed to deal with measurement error, and not from a breakdown in critical inquiry, as Michell suggests. Moreover, at least one class of these models can be formulated in terms of additive conjoint measurement theory, which renders Michell's call for the additive conjoint model in need of qualification. Finally, item response theory models are routinely tested against empirical data, and although the assumption that an attribute is quantitative cannot be tested directly, such tests do address the conjunction of this assumption and other model assumptions. We conclude that, although Michell's arguments are important to psychological measurement, they are largely irrelevant to item response theory. In fact, we argue that they can be phrased in terms of this theory in a natural way.

KEY WORDS: empiricism, item response theory, psychometrics, realism, representationalism

In a paper in this journal, Michell (2000) argues that psychometrics, as a discipline, should be classified as an instance of pathological science. Michell defines pathological science as a two-level breakdown in critical inquiry: the first-order breakdown occurs when a hypothesis is 'accepted as true without a serious attempt being made to test it', and the second-order breakdown consists in 'a higher-order attitude, namely that of ignoring the

THEORY & PSYCHOLOGY Copyright © 2004 SAGE Publications. VOL. 14(1): 105–120
DOI: 10.1177/0959354304040200 www.sagepublications.com

first-order breakdown', so that the first order-breakdown 'is not acknowledged or, in extreme cases, is disguised' (p. 641). In psychometrics, this situation occurs because

> . . . (a) a basic, empirical hypothesis (namely the hypothesis that psychological attributes are quantitative) is accepted as true without it ever having been seriously tested for its empirical adequacy, and (b) the fact that this hypothesis has never been satisfactorily tested is disguised. (p. 650)

Michell further suggests that psychological measurement should be dealt with along the lines of additive conjoint measurement (Krantz, Luce, Suppes, & Tversky, 1971; Luce & Tukey, 1964), a measurement system he seems to perceive as the preferred alternative to psychometrics. According to Michell, in contrast to current psychometric models, the additive conjoint measurement model does allow for testing the quantitative structure of attributes in a satisfactory way.

Psychometrics is a very broad discipline, so it is important to examine which parts of it are vulnerable to Michell's (2000) arguments. Especially relevant in this respect is the distinction between classical test theory (CTT; Lord & Novick, 1968) and item response theory (IRT; Hambleton & Swaminathan, 1985; Van der Linden & Hambleton, 1997). In published work, Michell is not entirely clear on whether he intends his arguments to apply to CTT, to IRT, or to both. However, in earlier work (Michell, 1999, p. 12) he explicitly criticizes the Rasch model (which is an IRT model), and in the article which is the subject of this response (Michell, 2000), he observes that, with regard to the quantitative structure of psychological attributes, 'journals such as *Psychometrika* and *Applied Psychological Measurement*, for example, contain little on this issue' (p. 649). As these journals are primarily concerned with developments in IRT modeling, this suggests that Michell does intend his arguments to apply to IRT, and he has confirmed this in personal communications.

While we agree that Michell's (2000) conclusions, although somewhat radical, are relevant to CTT and to a substantial part of the standard measurement practices in psychology, they do not apply to IRT models. There are three reasons for this. First, IRT models result not from a 'breakdown in critical inquiry' (Michell, 2000, p. 641), but from the introduction of a stochastic structure in modeling item responses. Second, at least one subclass of these models, namely the class of additive models such as the Rasch model, comprises a probabilistic variant of the additive conjoint measurement model. Third, IRT models do not take for granted the assumption that psychological attributes are quantitative. In fact, these models are routinely tested against empirical data. And, although these models do not permit directly a test of the hypothesis that attributes are quantitative, they do test this hypothesis in conjunction with other modeling assumptions. From a philosophy of science viewpoint, however, this seems

to be a particular instance of the Quine–Duhem thesis (which holds that hypotheses are never tested in isolation) rather than of pathological science. So, although some of Michell's (2000) arguments are relevant to psychology and psychological measurement, they are largely irrelevant to IRT models.

Deterministic and Probabilistic Models

Michell's (1999, 2000) argument is built around the thesis that the question of whether attributes are quantitative is an *empirical* one. The idea is that one cannot simply declare that one is measuring something on the basis that one is assigning numerals according to rule (Stevens, 1946). On the contrary, whether measurement is possible depends on the fulfillment of certain conditions. These conditions, Michell argues, are empirically testable, but never tested in psychometrics.

What conditions are to be satisfied for an attribute to be quantitative? To answer this question, Michell builds on work in fundamental measurement theory (Krantz et al., 1971), and specifically on the theory of additive conjoint measurement (Luce & Tukey, 1964). This theory shows that, if appropriate conditions are met, measurement scales can be constructed by representing several factors simultaneously on a common dimension. The main condition to be satisfied is additivity: two factors can be scaled simultaneously if they have independent additive effects on (a monotone transformation of) a third variable; hence the term 'additive conjoint measurement'. The axioms of additive conjoint measurement can readily be tested through a condition known as double cancellation, as Michell (2000, p. 658) illustrates. His conclusion that psychometrics uncritically assume that attributes are quantitative, rather than testing this as an empirical hypothesis, seems mainly based on the observation that this approach is not taken in psychometrics.

At this point it is important to observe the distinction between CTT and IRT. It seems relatively obvious that Michell's (1999, 2000) arguments apply to CTT. This theory should be viewed as a tautology rather than a model (Lord & Novick, 1968, p. 48), exactly because it is untestable. Being tautological, classical test theory can be applied to literally every test one could think of. The reason is that the central concept of CTT, the *true score*, is conceptualized as the expected value of the test score for a given subject over an infinite series of independent replications. Since such replications are impossible, CTT has to rely on a thought experiment (Lord & Novick, 1968, p. 29) to establish the true score. This thought experiment can always be performed, and thus poses no empirical restrictions at all. Once the true score is defined as the expected value of the observed score, virtually all theorems of CTT follow smoothly (Lord & Novick, 1968). Because these theorems contain linear relations between the true scores and the observed

scores, the true score must be conceived of as lying on an interval scale, which means it is a quantitative concept. (Note, however, that the relation between true and observed scores is linear by construction, not by assumption; see Lord & Novick, 1968, p. 34.) Also, the classical test theory model is largely untestable unless auxiliary assumptions, such as equal error variances across subjects, are imposed, and it is certainly never tested in actual research. Thus, Michell's (2000) conclusion, that psychometric models assume psychological attributes to be quantitative without testing this assumption, seems relevant to CTT and to research procedures based on it.

However, this is not the case for IRT (which is the central theory in current psychometrics). This theory represents a completely different approach to psychological measurement. It assumes item scores to be a function of an underlying latent variable, and uses this idea to formulate testable measurement models. The IRT approach is very close to additive conjoint measurement; so close, in fact, that we can derive IRT models by loosening the assumptions of fundamental measurement theory. Loosening these assumptions is necessary (and *not* the result of a breakdown in critical inquiry) because the approach taken in fundamental measurement theory, though elegant, is deterministic. As Michell himself has noted (Michell, 1986, 2000), the condition to be met for a series of dichotomous items to form an (ordinal) measurement scale is that they be Guttman scalable (Guttman, 1950). This means that, if a person has a 'correct' answer to a difficult item, then that person must have correct answers to all items that are less difficult. As a result, item scores must vary deterministically with the position on the measured attribute. We can graphically represent this idea by drawing discontinuous item response functions (IRFs) as in Figure 1 (item response functions are functions that relate the probability of a given response to the attribute being measured). The fact that the IRFs in a Guttman scale are deterministic means that, if person A has a position on the attribute which is higher than the item difficulties of item 1 and item 2, but lower than the difficulty of item 3, that person will answer items 1 and 2 correctly (or 'endorse' them, in case of personality or attitude items) with probability 1, and will answer item 3 correctly with probability 0.

The deterministic structure in the Guttman model is very restrictive, and there are few, if any, psychological scales that satisfy it. Does this mean that psychological measurement is impossible? No. It is reasonable to assume that our measurements contain a substantial amount of error. However, if we want to deal with measurement error, we must introduce a stochastic structure by formulating probabilistic, rather than deterministic, IRFs. Now, we can introduce probabilistic IRFs in various ways. One way is to assume that an attribute is quantitative, and to choose a parametric function that describes how the probability of a correct item response changes with the position of the attribute. We could, for example, choose a simple logistic

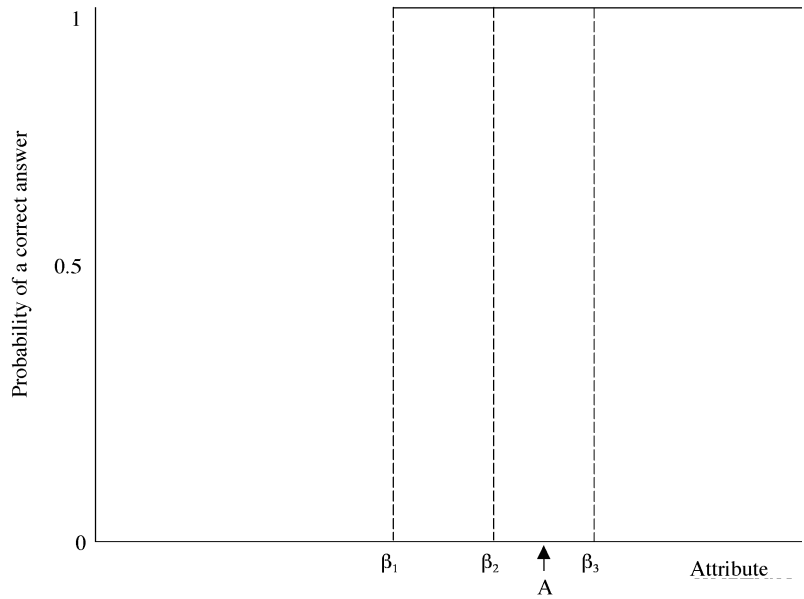

FIGURE 1. Item response functions for three items in a Guttman scale. The functions jump from 0 to 1 at the item locations $ß_1$, $ß_2$ and $ß_3$. Person A's position on the attribute is in between $ß_2$ and $ß_3$.

function with identical slope parameters across items, which would give a graphical representation like the one in Figure 2. In this model, the probability that person A responds correctly to an item decreases parametrically with the item difficulty (conceptualized as the location of the curve): This probability is highest for item 1, lower for item 2, and still lower for item 3.

In this model, the probability of a correct item response is parametrically related to a quantitative attribute. The model is known as the Rasch model (Rasch, 1960). The introduction of a stochastic structure in the model means that the model cannot be represented in the deterministic formulation of additive conjoint measurement theory. Note, however, that this is not because we have assumed the attribute to be quantitative, as Michell (2000) suggests; we could also have introduced a nominal latent variable, which would have yielded a latent class model, in which case quantification is neither achieved nor aspired to (Lazarsfeld & Henry, 1968). The model is not readily formulated in terms of axiomatic measurement theory because it is not deterministic. Hence, latent variable models do not diverge from axiomatic theory because they naïvely incorporate quantitative latent variables, for these models have a considerable degree of flexibility in this respect, but because they incorporate measurement error.

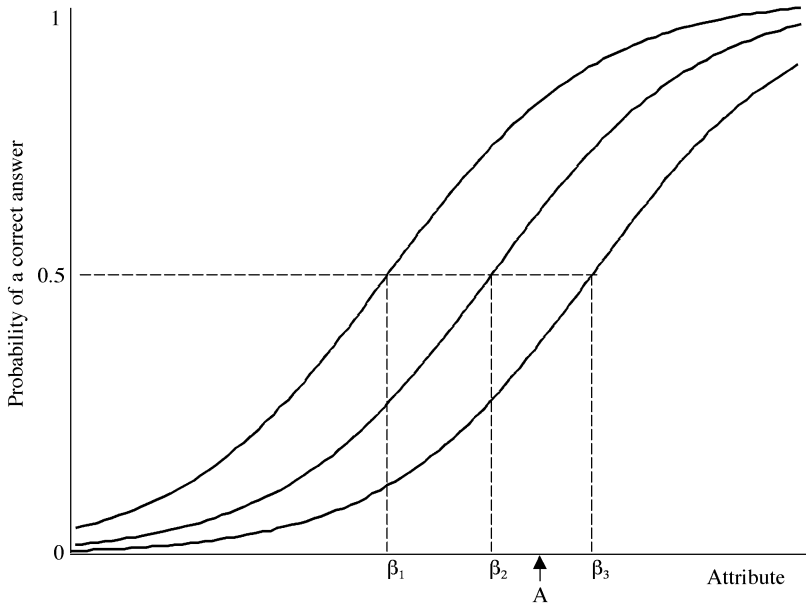

FIGURE 2. Item response functions for three items in a Rasch scale. The item locations $ß_1$, $ß_2$ and $ß_3$ are defined as the position of the attribute for which the probability of a correct answer is .5. Person A's position on the attribute is in between $ß_2$ and $ß_3$.

Interestingly, Michell (1999, pp.12, 166–167) chooses the Rasch model as one of his targets in launching his assault on measurement models in psychology. He apparently makes the assumption that such a model has nothing to do with additive conjoint measurement theory—he repeatedly praises additive conjoint measurement theory as the alternative. However, the fact that the IRFs in Rasch models are parallel means that the data matrix consisting of scores on a number of items, satisfying the Rasch model, can be modeled in terms of main effects of an item factor and an attribute factor, without using terms to represent interactions between these factors (incorporating such interactions would yield Birnbaum's [1968] two-parameter logistic model). Thus, the structure of the model is *additive*. This, of course, suggests a connection between the Rasch model and additive conjoint measurement theory. Indeed, the Rasch model has been argued to be a probabilistic variant of the additive conjoint measurement model (Bond & Fox, 2001; Brogden, 1977; Fischer, 1995; Perline, Wright, & Wainer, 1979; Roskam & Jansen, 1984). In fact, the relation between conjoint measurement and IRT is not limited to the Rasch model; it can be set up for quite a broad class of probabilistic latent variable models. The most general and precise treatment we know of in this context is given by Scheiblechner (1999).

Thus, the very model that Michell (1999) attacks, and wants to see replaced by an additive conjoint measurement model, can be formulated as an additive conjoint measurement model, albeit a probabilistic one. Of course, it may be that Michell disagrees with the development sketched in the papers to which we have referred, is opposed to introducing probabilistic structures, or has found a better way to deal with measurement error. If so, he should clarify his position on these matters. As it stands, however, Michell's argument is insufficient to establish his thesis regarding psychometrics as pathological science.

Hypothesized Attributes versus Constructed Scales

The above arguments show that, formally speaking, the difference between IRT and conjoint measurement models is, at least in some cases, rather small. More importantly, however, the most important source of disagreement between these approaches does not concern whether or not the hypothesis that attributes are quantitative is tested, as Michell (1999, 2000) would have us believe. We think that, at the bottom of this debate, lies a much more fundamental issue. This issue concerns a different outlook on what the proper way of working is, or should be, in psychological measurement. This touches on some deep issues in the philosophy of science.

If one considers carefully the work on axiomatic measurement theory (Krantz et al., 1971), one sees that the general structure of this theory is (a) to assume that qualitative relations between objects hold, (b) to prove that, given these relations, a homomorphic representation in real numbers is possible (this is done in a *representation theorem*), and (c) to prove that this representation is unique up to a well-defined class of transformations of the assigned scale values (this is done in a *uniqueness theorem*). Thus, one observes certain qualitative relations to hold among the objects to be measured, and then one maps these on a scale (quantitative or otherwise), in such a way that all relevant relations between objects are 'mirrored' in the relations between the numerical assignments. For present purposes, we may neglect the technical details of this procedure; what is important is that the scale is explicitly taken to be the construction of the researcher. To ascribe to this scale some kind of independent existence in reality would be to make a category mistake. *We* observe relations in the data; *we* represent these in the number system; and the resulting scale cannot therefore be considered more than our own construction.

In IRT models, this approach cannot be taken in principle, for the presence of a stochastic structure implies that the axioms will not be satisfied by the actually observed relations between objects (i.e. between persons and items). In treatments of the relation between additive conjoint measurement and IRT, the relations to be mapped onto the numerical scale

are always defined as relations between expected values (Scheiblechner, 1999, has provided the most explicit treatment in this respect). However, relations between expected values are not observable. Therefore, it is impossible to construct a representation on the basis of observed relations in the way this is done in axiomatic theory. Item response theorists must therefore take a philosophically important step: they have to assume, a priori, that a latent trait exists, and underlies our observations, for otherwise they cannot construct a model that has testable consequences. This trait, if continuous, is assumed to be measurable and to be quantitative, and in this sense Michell (2000) is correct in stating that some of the most commonly used psychometric models assume that attributes are quantitative.

Thus, the important divide between additive conjoint measurement and IRT is not that the first is an impeccable example of good science, while the second results from a pathological breakdown in critical inquiry; it is that additive conjoint measurement *constructs* a scale on the basis of qualitative observed relations, while the (continuous) IRT model *hypothesizes* a quantitative attribute to exist and to underlie the observed relations. The IRT model, when compared to the additive conjoint model, thus makes a stronger ontological claim. Item response theorists are forced to broaden their ontology in this manner because they cannot maintain that they are representing observed relations; in fact, they do not even want to maintain this, because what they are after is not representation, but explanation. For it is clear that the hypothesis that a latent trait exists, and that variation on this trait is responsible for variation in the observed scores, is the hypothesis that motivates the entire development of IRT; and this is most certainly an attempt at explanation, not representation.

Now, it may very well be this difference that leads Michell (2000) to state that the hypothesis that attributes are quantitative 'is accepted as true without it ever having been seriously tested for its empirical adequacy' (p. 650). However, this formulation is, at best, misleading, and at worst fundamentally incorrect. For there is a world of difference between entertaining the hypothesis that there is a latent variable underlying the item responses *in order to formulate a testable model*, and *accepting this hypothesis as true without testing it*. The former, and not the latter, is what is done in IRT. And if there is something wrong with it, then there is something wrong with science itself. For it seems to us that the deduction of testable predictions from hypotheses concerning the existence of unobservable entities, properties and attributes is not pathological, but common scientific practice.

In conclusion, the important difference between axiomatic measurement theory and IRT models does not lie in the issue of quantification, but in the strategy with which the data are approached. Where axiomatic measurement theory constructs scales, IRT hypothesizes latent variables to exist. Where axiomatic theory assumes the data to behave cleanly according to the axioms laid down in theoretical developments, IRT takes the data to be inherently

noisy. Where axiomatic theory aims to represent observed relations, IRT aims to explain why these relations exist, and where they come from. It does not take an overly sophisticated philosophical outlook here to conclude that, in terms of philosophy of science, axiomatic theory is close to the verificationist scheme of logical positivism, while the approach taken in IRT is more reminiscent of falsificationism. For instance, in axiomatic measurement theory, metaphysical speculations are eschewed; what is assumed to exist is nothing more than observable relations between objects. The theory works its way up from observations to theoretical terms in an axiomatic fashion, and the meaning of these theoretical terms (i.e. scales) derives directly from the observed relations. IRT, on the other hand, hypothesizes a certain state of affairs in the world, translates this hypothesis into testable predictions, and tests these predictions against observed data. The existence of a latent trait cannot be confirmed in the way the existence of a representation is guaranteed if the axioms are satisfied; it can, at most, be tentatively confirmed (i.e. 'corroborated') by data. We think that this difference in philosophical orientation by far outweighs the issue of whether or not attributes are uncritically assumed to be quantitative.

Testability

Michell (2000) repeatedly stresses his conviction that the hypothesis that an attribute is quantitative is an empirical one that should be tested, and one of his main arguments against psychometrics is that this is not done. However, the fact that one hypothesizes a latent variable to underlie one's observations does not imply that the model constructed in this fashion cannot be tested. Once formulated, IRT models can most certainly be tested against empirical data, and, in fact, this is routinely done. Nobody working in IRT, and we dare to make this statement as a universal claim, accepts the hypothesis that attributes are quantitative without testing the model for its empirical adequacy. As a matter of fact, IRT models are regularly rejected because they do not adequately fit the data. Now, it is certainly true that IRT models do not permit a direct test of the hypothesis that an attribute is quantitative. One can see this easily from Figure 2; apart from assuming a continuous latent variable, we also assume that the item responses depend only on this latent variable and on nothing else, an assumption known as unidimensionality. Moreover, we have assumed a function for the item response (in this case, a logistic function), and this function may be inappropriate. It will be clear that, when we assess the model in terms of fit with the data, it is the conjunction of these assumptions that is being tested.

The direct consequence of this practice is that, if the model does not fit the data, it may be difficult to disentangle these assumptions to find out which ones were inappropriate. Certainly, it may be the case that we have

prematurely assumed that the attribute in question is quantitative. On the other hand, maybe it is the form of the IRF that has been mis-specified. Moreover, in commonly used IRT models that allow the slopes of IRFs to vary, one will require an assumption regarding the shape of the population distribution function of the latent variable in question, because in these models no sufficient statistics exist for the person parameters. This distribution function is often assumed to be normal, but of course this may be inappropriate. And now we are merely considering mistaken assumptions 'inside' the formal structure of the model, that is, assumptions that are mathematically explicated; we have not even begun to touch upon the assumptions of a more substantive nature that are required to justify these models in the first place. For instance, are the test items worded appropriately? Could the test be sensitive to cultural differences that we do not intend to measure? Might the items be biased? And does the substantive theory motivate this particular type of model? These questions, which all relate to the problem of validity (Cronbach & Meehl, 1955; Messick, 1989), have not been addressed so far. Therefore, what we are testing against the data is not merely the hypothesis that the latent variable in question is quantitative; it is a much broader system of central assumptions, informed working hypotheses and mere guesses about the structure of the psychological world that is put to the test. And one may truly wonder whether the hypothesis that the attribute in question is quantitative plays such a central role in this scheme as Michell seems to assume.

Michell (2000) is right in stating that common approaches in psychometric modeling do not single out the hypothesis that an attribute is quantitative to put this hypothesis to the test. The reason for this, however, is not that psychometricians are lazy or uninterested in this hypothesis. The primary reason is that it is impossible to test this hypothesis in isolation. The fact that we cannot test a hypothesis in isolation, however, is characteristic not of pathological science, but of science in general. From a philosophy of science viewpoint, it seems a particular instance of the Quine–Duhem thesis (which states that no hypothesis is ever tested in isolation), rather than of pathological science. Moreover, the fact, that we cannot evaluate the hypothesis that an attribute is quantitative in isolation does not imply that we cannot test it at all. The very fact that psychometric models are rejected so often testifies to the opposite.

Where Does a Realist Account of Measurement Lead?

We think that we have proven the allegations of Michell (2000) to be preliminary and largely unfounded. We now have one question to answer, namely why do we disagree with him in the first place? The reason that this question occurs is the following. Michell has gone through a great deal of

trouble to argue for the thesis that the hypothesis that an attribute is quantitative is a hypothesis about a state of affairs in the world (an empirical hypothesis, in his words). This is clear, for instance, from his insisting that measurement is about predication, and not about assignment (Michell, 1999). We agree with this thesis. Michell is, in his own terms (Michell, 1986), closer to the 'classical' conception of measurement than to the representationalist theory as set forward in the classic work by Krantz et al. (1971). The classical position holds that measurement scales are not representations of relations between objects, but that they describe objectively existing magnitudes, where objective magnitudes are *properties* of which objects are the bearers, rather than *representations* of objects. Thus, Michell is a realist about attributes. We think that such realism is, in general, required to give a consistent account of measurement, so we agree again. But this is a strange situation. Given that we are in virtually complete agreement with Michell's proposed ontological claims, why should we reach such different conclusions with respect to the epistemological strategy that follows from them? Why does this position lead Michell to write an article against common psychometric models, while it leads us to write an article in defense of these models? One of us is being inconsistent here, and, unsurprisingly, we think it is Michell.

In our view, a realist position about attributes connects naturally to the general latent variable modeling framework, of which IRT is a particular instance (Mellenbergh, 1994), and only indirectly to representationalist techniques like checking double cancellation. For if one has a realist semantics for attributes, it is plausible to take the hypothesis that an attribute (say, general intelligence) exists in the world as the core hypothesis in one's theoretical system. One then proceeds to derive testable consequences from this system. A plausible way to do this is to set up hypotheses concerning the connection that this attribute may bear to empirical data. And if one does not want to get stuck in a totally unworkable deterministic theoretical system, one will have to formulate probabilistic relations to make this connection. This will lead to some kind of latent variable model, for instance an IRT model. Thus, we have arrived from a realist semantics about attributes to latent variable theory in two or three simple steps. Does it now follow from our ontological framework that the actual data will behave in accordance with the axioms laid down by additive conjoint measurement theory? In other words, would the ontological premise that a quantitative attribute exists lead us to expect that the representationalist's axioms will be satisfied—even when these axioms are weakened so as to hold at the level of expectations of item scores, rather than at the level of observed relations between objects?

It would not. For instance, the hypothesis that general intelligence exists, when conceptualized as a latent dimension which relates to IQ items through a set of probabilistic IRFs, does not entail that the IRFs will be an additive

function of a person parameter (general intelligence) and an item parameter (item difficulty). It may, for instance, be the case that some item *i* is more difficult than item *j* for highly intelligent people, while the reverse is the case for less intelligent people. This could, for instance, occur if one of the items is so constructed that it leads some of the more intelligent people to hypothesize a more complicated answer than the one considered 'correct' by the intelligence tester. One can easily imagine a situation where this would violate additivity, although the item would still be informative about a person's level of intelligence (because its expected score is an increasing function of intelligence, although this function is not strictly parallel to the function of another item). Thus, a realist conception of the attribute does not entail additivity. If there is no additivity, double cancellation does not follow and therefore Michell's epistemological prescriptions do not follow.

Why, then, is additivity so important? There is one, and only one, reason for this. The reason is that, if additivity is not satisfied, one cannot prove a representation theorem in the additive conjoint measurement scheme. For someone who does not have a realist conception of psychological attributes, this is central; for if a representation theorem cannot be proven, no measurement scale can be constructed, and since measurement is scale construction, the very possibility of measurement is thereby precluded. But for the realist, the fact that no representation theorem can be proven is not particularly disastrous. Of course, it would be nice to have one, but it is perfectly all right to speak of measurement if one does not, for measurement consists in finding out people's position on an attribute that exists quite independently of the measurement process. That one cannot construct an additive representation is a pity, but does not frustrate the possibility of measurement at all.

Because additivity is central only for the representationalist, and not for the realist, satisfaction of the double cancellation axiom is a central one for the representationalist, but not for the realist. The realist may use the technique if he or she has good reason to believe that the IRFs are additive, for in this situation it may offer valuable empirical evidence for or against the hypothesis that the attribute is quantitative; but the realist does not give the satisfaction of axioms like double cancellation the status of a definitional *criterion*, that is, a criterion that must be fulfilled in order to speak of 'genuine' measurement. In fact, if one is a realist about attributes, then one is forced to conclude that the probability that one will find a set of items with perfectly parallel IRFs is extremely small in the first place, because this implies the truth of a point hypothesis (the difference between item slopes equals zero) which is defined on a continuous space (because the difference between item slopes, theoretically speaking, varies on the continuum).

So, if one is a realist about attributes, the prediction that double cancellation will hold follows from the hypothesis that a quantitative attribute exists *and* the hypothesis that the IRFs that relate this attribute to the observations

are additive (and from many other background assumptions on which we will not dwell here). Its status is therefore nothing less, but also nothing more, than one of the many epistemological checks that may be carried out to check whether the model is adequate in the special case where one has additive IRFs. For the representationalist, however, the double cancellation tests are not merely epistemologically significant in a few special situations; they are central to the very definition of measurement. For if double cancellation is not satisfied, the representationalist cannot construct the desired representation, and measurement is impossible. We conclude that, if Michell is serious about his realist semantics, he should agree with us, and not with the representationalist; therefore, he should propagate the use of latent variable models, rather than of axiomatic measurement theory; and he should advertise double cancellation for what it is, namely a useful empirical check that works with additive models, while at the same time recognizing that these models comprise but a subset of the models that the realist may entertain as candidates for psychological measurement.

Discussion

The central claim made by Michell (2000), that psychometrics uncritically accepts as true the hypothesis that psychological attributes are quantitative, does not apply to psychometrics in general. It may apply to classical test theory, but it should be noted that the founders of this theory were well aware of the fact that the model is largely untestable (Lord & Novick, 1968, p. 48). IRT models, however, are testable. Although these models do not allow for direct tests of the hypothesis that attributes are quantitative, this does not mean that they uncritically accept it as true. It is the case that the hypothesis at stake is not testable in isolation. One may seriously wonder, however, whether any scientific hypothesis is ever testable in isolation. Therefore, if such a situation would justify the conclusion that psychometrics is pathological, it would probably justify the conclusion that science in general is pathological. This conclusion would seem to go too far even for a highly critical scholar like Michell.

Further, although there are some serious differences between psychometrics and additive conjoint measurement, they do not lie where Michell (2000) puts them. Formally speaking, the models are not all that different. The additive conjoint measurement model is not equipped to deal with error, and if we reformulate it to repair this deficit, we end up formulating an additive IRT model. Nothing in this development justifies using the term ‘pathological science’. In fact, the further exploration of connections between IRT and fundamental measurement theory is an enterprise of considerable importance, and such inquiries can sometimes yield unexpected similarities (see, e.g., Münnich, 1998, and Scheiblechner, 1999). From a

formal perspective, little is gained by portraying these different strategies as antagonistic; much more could be achieved by integration of insights from both fields.

There does exist a large difference between axiomatic measurement theory and psychometric models; but this difference lies in much deeper aspects of philosophical orientation. Where axiomatic theory constructs scales on the basis of observed relations, psychometric models hypothesize the existence of latent variables to explain observed relations. If one has substantial confidence in the adequacy of the formulated psychometric model, one may estimate people's positions on the latent trait on the basis of observed scores. Note that the fact that, in doing this, the term 'estimation' is more appropriate than the term 'measurement' underscores this rather fundamental difference between the approaches. We are of the opinion that axiomatic theory, being deterministic, is unreasonably strict. Loosening it, however, leads to commonly used IRT models. To make this conceptual shift requires a shift in philosophical interpretation from a nearly full-blooded empiricism to the stronger ontology of realism. For if one wants to formulate a probabilistic model, then one needs this stronger ontology. One simply cannot hold that one is representing observed relations, because these relations are relations between expectations, and expectations are not observable. Given that Michell already occupies a realist position, his arguments would gain consistency if he framed them in terms of modern psychometric models. For his ontological position neither implies nor necessitates the importance of axioms like double cancellation. Such procedures can be epistemologically helpful in some cases, and we do not have a quarrel with Michell in this respect. But for the realist they do not, and cannot, play the role of definitional criteria for measurement, as they do for the representationalist. Thus, while Michell is correct in drawing attention to such approaches as possible fruitful strategies to check certain assumptions, he must acknowledge that they do not follow from the hypothesis that the attribute to be measured is quantitative alone. They follow from the conjunction of this hypothesis with many others. The relative importance of these tests must not be neglected, of course, but neither should it be overstressed.

Finally, we would like to highlight some points where we do agree with Michell's claims. Regrettably, in much psychological research, item scores are simply summed and declared to be measurements of an attribute, without any attempt being made to justify this conclusion. Michell (2000) is right in questioning this practice. But he attributes it to the wrong people: virtually all modern psychometric theories analyze response patterns, which is the reason the general approach is called item response theory. In this respect, we submit that every psychometrician working today would endorse Michell's (2000) claim that '[r]esponse patterns are more fundamental than test scores' (p. 659). It is certainly the case that response patterns are

neglected in psychology, and we think Michell is right in calling attention to the fact that summing item scores may sometimes be inappropriate if the required conditions are not met, for example when such mis-specifications lead to spurious results. Although Michell's arguments are largely irrelevant to IRT models, they are important for psychology; in fact, his arguments would gain strength and generality if they were connected to IRT models, rather than aimed at such models. In philosophy, one would classify the arguments in Michell (2000) as misdirected. We would say, however, that they deserve to be redirected.

References

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* (pp. 395–479). Reading, MA: Addison-Wesley.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the social sciences*. Mahwah, NJ: Erlbaum.

Brogden, H.E. (1977). The Rasch model, the law of comparative judgment, and additive conjoint measurement. *Psychometrika*, *42*, 631–634.

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

Fischer, G. (1995). Derivations of the Rasch model. In G. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York: Springer.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stoufer, L. Guttman, E.A. Suchman, P.L. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. IV. Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

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

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

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.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*, 398–407.

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

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*, 639–667.

Münnich, A. (1998). *Judgment and choice.* Unpublished dosctoral dissertation, University of Amsterdam.

Perline, R., Wright, B.D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237–255.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

Roskam, E.E., & Jansen, P.G.W. (1984). A new derivation of the Rasch model. In E. Degreef & J. van Bruggenhaut (Eds.), *Trends in mathematical psychology* (pp. 293–307). Amsterdam: North-Holland.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models. *Psychometrika*, *64*, 295–316.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*, 667–680.

Van der Linden, W.J., & Hambleton, R.K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer.

ACKNOWLEDGEMENTS. The authors would like to thank Maarten Speekenbrink, Guenter Trendler and Ingmar Visser for some useful discussions on conjoint measurement and latent variable theory. We also thank Joel Michell for elucidating his views in various personal communications.

DENNY BORSBOOM is Assistant Professor of Psychological Methods at the University of Amsterdam. His research is at the interface of philosophy of science, psychometrics and psychology. ADDRESS: Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. [email: D.Borsboom@uva.nl]

GIDEON J. MELLENBERGH is Professor of Psychological Methods at the University of Amsterdam. His research interests are in psychometric methods, models and concepts.