Measurement, 6: 7–24, 2008
Copyright © Taylor & Francis Group, LLC
ISSN 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366360802035489

Psychology Press
Taylor & Francis Group

Is Psychometrics Pathological Science?

Joel Michell
University of Sydney

Pathology of science occurs when the normal processes of scientific investigation break down and a hypothesis is accepted as true within the mainstream of a discipline without a serious attempt being made to test it and without any recognition that this is happening. It is argued that this has happened in psychometrics: The hypothesis upon which it is premised, that psychological attributes are quantitative, is accepted within the mainstream, and not only do psychometricians fail to acknowledge this, but they hardly recognize the existence of this hypothesis at all. It is suggested that certain social interests, identifiable within the history of modern psychology, have produced this situation because of the ideological and economic secondary gains derived from presenting psychology as a quantitative science. The question of whether modern item response models are exempt from this critique is considered, and it is concluded that they have not yet faced up to the challenges of seriously testing the relevant hypothesis or even bothered to recognize its existence.

Key words: measurement, order, pathological science, psychometrics, quantity

Science is a cognitive enterprise. That is, scientists want to find out how natural systems work. They employ their characteristic methods, believing that these help uncover the structure and ways of working of the systems of interest to them. A person or social movement (such as science) engaged in cognitive enterprises is itself a natural system. It is a cognitive system. A cognitive system is pathological when it prevents rather than promotes acquisition of relevant knowledge. A familiar example is prejudice: A person believes something, not because of relevant evidence, but for some other reason, say, because it confers secondary gain. A person believing out of prejudice is blocked from knowing what is actually the case, the extent of the blockage depending on the strength

Correspondence should be addressed to Honorary Associate Professor Joel Michell, School of Psychology, University of Sydney, Sydney NSW 2006. E-mail: joelm@psych.usyd.edu.au

of the prejudice. In some measure we all do this and any cognitive system can become pathological. In the case of individual people, the causes are complex, but it may come about because, as Freud (1957) suggested, we are motivated by diverse and conflicting interests, and sometimes our need to believe something in the absence of relevant evidence exceeds our commitment to finding the truth.

As a social movement, science is complex and its character has changed considerably, as it has become increasingly important to other social endeavors, particularly industrial and government organizations, and has become increasingly dependent on them for support. As a result, conflicting social interests motivate science and there is potential for pathologies to arise. There are many famous cases, such as the domination of Soviet genetics by the politically motivated theories of Lysenko (Soyfer, 1994). Psychometrics is a much less dramatic but, nonetheless, clear-cut case (Michell, 2000).

In psychometrics, the prejudice involved is the conviction that psychological attributes—such as cognitive abilities, personality traits, and social attitudes—are quantitative. Survey the psychometric literature: It reveals a body of theories, methods, and applications premised upon the proposition that psychological attributes are quantitative but is devoid of serious attempts to consider relevant evidence for that premise. The theories proposed (such as the factor analytic theories of cognitive abilities and personality) are typically quantitative; mainstream psychometricians typically believe that they are able to measure abilities, personality traits, and social attitudes using psychological tests; and within applied psychometrics, tests are typically promoted using the rhetoric of measurement. Yet, there is little acknowledgment that this premise might not be true: No research programs that I know of are dedicated to testing it; no body of evidence is marshaled in its support (indeed, as far as I know, none exists); and no attempt has been made to devise methods for diagnosing the difference between quantitative and merely ordinal attributes. Psychometrics is premised upon psychological attributes being quantitative, but this premise is rarely treated as raising questions, usually only as answering them.

This alone is not sufficient for pathology, however. Pathology also requires the presence of a positive factor, one deflecting attention from relevant questions (Michell, 2000). Within psychometrics, certain established ideological structures have the effect of discouraging psychometricians from raising the issue of the quantitative structure of psychological attributes. If one dates the origin of psychometrics from the publication of Spearman's (1904) paper, "General Intelligence, Objectively Determined and Measured," it came 3 years after the publication of Hölder's (1901) paper on the axioms of quantity and the theory of measurement. This latter paper provided a clear characterization of quantitative structure and its relation to the real number system, which is the conceptual foundation of measurement. This paper and the discipline it spawned, *measurement theory*, are excluded from consideration in mainstream psychometrics and are missing from the curriculum of psychometrics as typically taught (Michell, 2001). This denies psychometricians the conceptual resources necessary for raising the issue of whether psychological attributes are quantitative.

This exclusion was compounded when psychometricians accepted a definition of measurement implying that their procedures achieve *measurement* without needing to investigate the issue of whether psychological attributes are quantitative. This was Stevens's famous definition of measurement as "the assignment of numerals to objects or events according to rules" (1946, p. 667). This definition is ubiquitous throughout psychology (Michell, 1997). Anyone accepting it will thereby call a procedure *measurement* if that procedure involves assigning numerals to things regardless of whether a quantitative attribute is involved.

Stevens's field was psychophysics, a field aspiring to measure another psychological attribute, viz., sensations, using methods different from those of psychometrics. The issue of whether sensation intensities differ quantitatively had long been raised (e.g., Von Kries, 1882) and Stevens's own scale, the so-called sone scale, putatively for the measurement of loudness (Stevens & Davis, 1938), had recently been criticized on just these grounds in a report on psychophysical measurement for the British Association for the Advancement of Science (Ferguson et al., 1940). This meant that Stevens, in constructing his special definition of measurement, had an interest in deflecting attention from the issue of whether psychological attributes are quantitative. Widespread endorsement of his definition also had just this effect within psychometrics.

In accepting Stevens's definition, psychometricians were being inconsistent because this definition is incompatible with the traditional concept of measurement. Because of their way of theorizing about psychological attributes, psychometricians were committed to this traditional concept. Claiming that a psychological attribute, such as general ability, is related to test scores in the manner supposed in, say, factor analytic theories of ability presumes that general ability is quantitative in structure and quantitative structure entails the traditional view of measurement, viz., that it is the assessment of quantity. The concept of a quantitative attribute and the traditional view of measurement are part of the same conceptual package (Michell, 2005). Psychometricians were engaged in a kind of doublethink, and this could be maintained only because they consistently excluded measurement theory from the knowledge base of psychometrics.

As a substitute, psychometricians endorsed Stevens's (1946, 1951) theory of scales of measurement. At first sight, this seems a positive step because this theory distinguishes four types of measurement scale, *nominal*, *ordinal*, *interval*, and *ratio*, the first two of which are clearly nonquantitative and the second two of which are obviously quantitative. Efforts to determine scale type would seem necessarily to require addressing the issue of whether the relevant attribute is quantitative. Stevens's theory makes most sense when interpreted like this (Suppes & Zinnes, 1963). However, neither Stevens nor psychometricians interpreted it in this way (Michell, 2002). They opted for an *operationist* interpretation, one that makes the issue of scale type turn not on a consideration of the structure of the attribute involved but on the class of admissible transformations stipulated as appropriate for the numerical assignments made. Because linear transformations are routinely applied to test scores, it was argued that test scores must sustain measurement on an interval scale (e.g., Nunnally, 1967; Lord & Novick, 1968). A more coherent interpretation of Stevens's theory says that the class of admissible scale transformations is determined by the structure of the attribute to which numerical assignments are made (Narens, 2002). Turning this on its head, in the way that psychometricians did, meant that the hypothesis that psychological attributes are quantitative need never be raised.

Thus, psychometrics is a science in which the central hypothesis (that psychological attributes are quantitative) is accepted as true in the absence of supporting evidence and this fact is ignored because psychometricians remain ignorant about the concept of quantity; they accept a definition of measurement that deflects attention away from the issue of quantity; and an operationist interpretation is put upon scale type distinctions. That is, psychometricians claim to know something that they do not know and have erected barriers preserving their ignorance. This is pathological science.

Why should pathology emerge in a discipline that proudly proclaims itself a science? It is easy to identify the interests that this pathology serves. There are two sets: First, there are ideological interests and, second, there are economic ones. The ideological interests relate to *scientism*. The term *scientism* denotes an ideologically driven, false image of science. Ironically, it was because psychology proclaimed itself a science that this pathology arose. It is still widely thought that measurement is a necessary feature of all sciences: Knowing something *scientifically* means *measuring* it. This view was so widespread in the 19th century that it had become an idol of the age (Michell, 2003). Modern psychology was a 19th-century invention. So there was pressure on it to find a place for measurement.

Over and above this, however, psychology is a special case. Psychology was not a new science. Psychological phenomena had been investigated for millennia before psychology was excluded from the scientific community during the 17th century when physical science became strongly identified with quantification. It was excluded because its subject matter was deemed to be nonquantitative. The Cartesian doctrine that mental phenomena are completely different from physical phenomena and that physical phenomena are essentially quantitative was the main reason for the 200-year boycott on recognizing psychology as a science. In promoting it as a quantitative science, the psychologists of the 19th century were making a deliberate point: viz., that psychology was a science precisely because it had devised ways of measuring mental phenomena. For this reason, psychologists showcased quantitative psychophysics, despite the fact that psychophysics was of little intrinsic theoretical or practical significance. It was the jewel in psychology's crown because it was taken to mean that psychology was quantitative and that the Cartesian boycott was lifted. Psychophysical measurement was psychology's ticket back into the scientific arena. This was how important measurement was in the eyes of the founding fathers of the discipline, and, as a result, the hypothesis that psychological attributes are quantitative was nonnegotiable.

However, the fact that psychophysics lacked practical application meant that there was pressure to find ways to reconstruct psychology as an *applied* quantitative science. In Spearman's (1904) paper, the founder of psychometrics lamented that as an applied science, psychology had failed and expressed the hope that it might eventually deliver what was needed. From its beginnings in Britain and the United States, the use of psychological tests in the military, education, and industry was wrapped in the rhetoric of measurement because this packaging was thought to secure the place of psychology among the established sciences (Michell, 1999). A century later, psychometricians quite unselfconsciously and without reflection continue to characterize the use of psychological tests as *measurement*, as if no other form of discourse was available.

The second interest sustaining this pathology is economic. I refer not so much to the economic context within which psychological tests are marketed (although economic advantages accrue when tests are advertised as measurement instruments) as to the economic conditions of scientific research. Modern science depends for its existence on research, and research requires financial resources. After World War II, the era of *Big Science* emerged. The social causes were the defense requirements of the Cold War and the increased expectations of Americans and Europeans regarding progress in education, technology, and medicine. The immediate postwar decades witnessed an unprecedented expansion in government investment in scientific research. Research grants became the main vehicle by which not only individual careers but also the aspirations of entire disciplines progressed. This affected disciplines on the margins of the established sciences, such as psychology, as much as the established sciences themselves, because marginal disciplines were forced to compete for smaller shares of the total amount available. This quickly led to a phenomenon called "the new rigorism" (Schorske, 1997, p. 309). Psychology did not actually become more rigorous, but it aped the methodological rigor of the established quantitative sciences as a way of signaling its scientific credentials to granting agencies (Solovey, 2004).

Within the first two postwar decades, a methodological consensus was set in concrete and it has hardly altered since. It included the use of Fisher's theory of experimental design and a hybrid of associated theories of significance testing, Stevens's theory of scales of measurement, classical test theory, and the theory of factor analysis. From a logical point of view, it had little to recommend it (as some now realize), but its real role in the training and practice of psychological researchers was its value as window dressing. It was a device for attracting research funds by attempting to make psychology appear more scientifically rigorous than it is. With so much at stake, there was no room for doubts about whether psychological attributes are quantitative: Economic imperatives dictated that they must be.

The claim that I am making then is (a) that psychometrics is pathological because the hypothesis that psychological attributes are quantitative is accepted as true by mainstream psychometricians, not on the basis of adequate evidence but for extraneous reasons, and (b) that at the same time, (a) is ignored or even disguised. On top of that, I have identified the social interests sustaining this pathology, these being ideological and economic secondary gains derived from presenting psychology as a quantitative science. In different terms, I have presented this critique for more than a decade (Michell, 1990). There I suggested that "the practices called 'psychological measurement' … are, perhaps, nothing more than a pretense" (p. vii), but the term *pretense* might suggest an intention to deceive, which I do not think is present. Later, I used the expression *methodological thought disorder* (Michell, 1997) to characterize psychologists' attitude to measurement and, though that expression is apt, the problem is not exclusively methodological. Subsequently, I prefer the term *pathological science* (Michell, 2000) because the kind of disorder I am thinking of is one that subverts the scientific aim of finding out how things are structured and work.

The response of the psychometric community to this critique has been, largely, silence. A few have objected that it does not apply to a subclass, viz., those using probabilistic, item response models (e.g., van der Linden, 1994; Bond & Fox, 2001; Fisher, 2003; and Borsboom & Mellenbergh, 2004). These models are advocated by a majority publishing in the leading journals (e.g., *Psychometrika*, *Applied Psychological Measurement*). How far the use of such models has penetrated the wider psychometric community—in schools, the military, and industry—is another matter. It would be true to say, however, that a majority of the opinion leaders in psychometrics now favor them.

What do I make of the reply that those using item response models are exempt from my critique? Any group of psychometricians would be exempt were they to admit that they *assume* the empirical hypothesis that psychological attributes are quantitative. If there was some sign that they accept this hypothesis only provisionally, saying something like, "At present we do not know whether this hypothesis is true, but we will assume it recognizing that at some point in the future someone needs to investigate it," then they would no longer be doing science pathologically.

However, explicit mention of the hypothesis that the relevant psychological attribute is quantitative is missing from expositions of item response models (e.g., Bond & Fox, 2001; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Suen, 1990). In the simplest case (dichotomous item responses, say, *correct* or *incorrect*, in ability tests), these models take the probability of a person making one or other kind of response to an item to be a function of a hypothesized quantitative psychological attribute (the so-called *latent trait*). Each person and each item is taken to have a measure on the latent trait, and the probability of a response of a given kind is some specific function of a relation (say, the difference) between these measures. Item response modelers typically state just three assumptions: *unidimensionality* (i.e., there is just *one* latent trait involved); *local independence* (this trait is the only source of systematic individual differences between the responses given to different items); and the *item response function* specified (the specific relationship assumed between the latent trait and the probability of, say, a correct response, on each item, which is often taken to be either a normal or a logistic ogive, differing between items with respect to location on the latent trait). That is, generally, it is not explicitly stated that the latent trait is taken to be a quantitative attribute, although, of course, this is implicit in the character of the response function.

We get a further indication of the thinking of these modelers by considering their reasoning in cases where models do not fit data. Cases of misfit invite the modeler to speculate on what has gone wrong. The typical response is to question one or more of the above three assumptions, and one must look hard for instances where the implicit assumption of quantitative structure is scrutinized. That is, typical of psychometricians generally, the assumption that psychological attributes are quantitative remains unstated, unquestioned, and uninvestigated.

However, there is a minority school among item response modelers advocating so-called *nonparametric models*. These extend the work of Mokken (1971), and Sijtsma and Molenaar (2002) present an introduction. These models weaken the assumption regarding the item response function so that no specific relationship is assumed and it is only required that, for each item, the function be increasing monotonic and that, between items, the individual functions do not intersect. The result is a model in which people or items can be ordered with respect to the latent trait but not measured. As Sijtsma and Molenaar note, for many psychometric applications, ordinal information is sufficient. However, in the presentation of nonparametric models, latent traits are still implicitly taken to be quantitative and the distinction between parametric and nonparametric models is thought to reside only in what we are able to infer about the traits, that is, ordinal versus quantitative information. It would seem that the assumption that psychological attributes are quantitative is so deeply ingrained that it is not questioned even in contexts where the models being considered invite precisely that query.

So, item response modelers are no less pathological in their approach to psychometrics than more traditional psychometricians: The implicit assumption that psychological attributes are quantitative is as pervasive as it is hidden. However, there is one important difference between item response modelers and more traditional psychometricians. The former, now, typically test the fit of their models to data, whereas the latter are less inclined to do this. This prompts a different kind of objection: If the assumption that latent traits are quantitative is a necessary part of item response models (as I maintain it sometimes is), then must it not be the case that this assumption (acknowledged or not) is being tested empirically when the fit of the model to data is assessed? If the answer to this question is yes, then does it not follow that item response modelers should escape the attribution of pathology? Borsboom and Mellenbergh (2004) think so.

This raises the issue of how sensitive goodness of fit indices are to violations of each of the assumptions involved in a model. As Roberts and Pashler (2000) ask, how persuasive is a good fit? Certainly, in the typical psychometric context, it must be said, not very persuasive. Typically, a psychometrician begins with a pool of items and proceeds by discarding those that contribute to poor fit, as if the aim of the exercise was to construct a test having certain psychometric properties, rather than to test hypotheses about the structure of the latent trait. This modus operandi does not provide evidence that the latent trait is quantitative. The reason is that if we try hard enough, test items fitting any model, even Guttman's (1944), can be constructed.

Guttman's is the simplest of all item response models. In the context of ability tests, it is that a person attempting an item will get it correct if and only if the person's measure on the latent trait is not less than the item's. It is an ordinal model in two senses: First, it only requires that the latent trait have ordinal structure; and, second, it only provides an ordering on people and items on that trait. As Borsboom and Mellenbergh (2004, p. 108) note, however, this model is "very restrictive" in the sense that it fits responses to few psychological tests. Most modelers interpret this as evidence for the existence of *error* in test performance. They attempt to accommodate this error in their models, and it is the assumptions made about it, and these alone, that entail a model's specifically quantitative results (Michell, 2004).

The place of error in parametric models leads to an apparent paradox. The term *error* denotes the effects of factors extraneous to the trait under investigation, which are thought to affect individual differences in responses. Let us suppose that parametric modelers are correct about the relationship between a person's and an item's measures being discerned only through a haze of error. Further, suppose that we were able to improve controls in the testing situation and eliminate the effects of extraneous factors, thereby eliminating error. This dramatic improvement in the precision of our procedures would lead to no improvement in the precision of our measurements as such improvements typically do in other sciences. In fact, quite the reverse! Such improvements would entail a Guttman scale, which would only allow people and items to be ordered, whereas, before, observations contaminated by error allowed quantitative measurement. Is it not paradoxical that *improving the precision of our observational conditions decreases the precision of our observations?*

To get this into perspective, think by analogy with procedures in another science, say, astronomy. Suppose we were inspecting some newly discovered star, one that we could only see dimly because of some kind of hazy interference in outer space. Suppose further that through this haze we thought we could detect a system of planets orbiting the star. Then suppose that by some lucky circumstance the haze disappeared and our view of the star improved and that what we had previously thought was a planetary system could no longer be seen. Would we not feel justified in concluding that what we had thought was a planetary system was really only an artifact of the interference? We would feel this because we are suspicious of effects that depend on error. If things that we think are there cannot be detected when the precision of our procedures improves, then we need additional evidence of their existence. Likewise, because the quantitative relationships that we think we can detect via parametric item response models would disappear were error eliminated, we require additional evidence of their existence. We need tests that are specifically attuned to the hypothesis that the relevant trait is quantitative. Item response modeling, as normally carried out, does not involve a serious attempt to test the hypothesis that the relevant attributes are quantitative or any recognition of the fact that such a serious attempt is lacking.

What would such a serious attempt look like? I will not list all of the requirements necessary for a serious test of this hypothesis. One requirement, however, is this: If you are going to seriously test the hypothesis that some latent trait, *X*, is quantitative, then *X* must be specified in sufficient detail for its hypothesized quantitative structure to have a theoretical interpretation in terms of item structures and the psychological processes. As Stevens's mentor, Boring (1920, p. 33), long ago said, "It is senseless to seek in the logical process of mathematical elaboration a psychologically significant precision that was not present in the psychological setting of the problem." At present, we have some idea of what ordinal structure in psychological attributes is like from cognitive theories, like those of Piaget, which give theoretical reasons why one test item should be more difficult than another. However, we have no idea what quantitative structure, over and above mere order, would look like because our psychological theories are uninformative about this. We are not able to say what it is about the structure of the items or the structure of the psychological processes involved that would make the level of ability required to solve one test item exactly double, triple, or, in general, r times that required to solve another. Until our theories do this, the mathematical elaboration of psychometrics will outstrip its theoretical elaboration at precisely the point where ordinal structure progresses to quantitative. The key problem for psychometrics, and the one that it must address to regain scientific credibility, is to find ways of testing the distinction between merely ordinal structure and quantitative structure, over and above that of mere order.

What precisely is the difference between order and quantity? A variety of ordinal structures have been defined (Michell, 1990), but that of a *strict simple order* is the paradigm. If the levels of an attribute are ordered by a transitive, asymmetric, and connected binary relation, then the levels constitute a strict simple order. A binary, *greater than* relation upon the levels of an attribute (symbolized by $>$) is

1. *transitive* if and only if for any levels a, b, and c, if $a > b$ and $b > c$, then $a > c$;
2. *asymmetric* if and only if for any levels a and b, if $a > b$, then *not* $(b > a)$; and
3. *connected* if and only if for any levels a and b $(a \neq b)$, either $a > b$ or $b > a$

(where a, b, and c are any levels of the attribute). Also, there are a variety of quantitative structures, but where measurement is concerned the paradigm is that of an *unbounded continuous quantity* (Hölder, 1901; Michell & Ernst, 1996):

4. for every pair of levels a and b, one and only one of the following is true:

   (i) $a = b$;
   (ii) there exists a level c such that $a = b + c$;
   (iii) there exists a level c such that $b = a + c$;

5. For any levels a and b, $a + b > a$;
6. For any levels a and b, $a + b = b + a$;
7. For any levels a, b, and c, $a + (b + c) = (a + b) + c$;
8. For any a and b, there is a c such that $c = a + b$.
9. for any a, there is a b such that $b < a$.
10. For every nonempty class of levels having an upper bound, there is a least upper bound

(where for any levels a, b, and c, $a + b = c$ if and only if c is entirely composed of discrete parts, a and b).

Clearly, conditions 4–10 entail conditions 1–3, and not vice versa. This is a logical fact, but one worth emphasizing because it is not uncommon to meet psychologists who think that all ordered attributes must be quantitative (Michell, 2006, in press). Arguments to this effect have even been proposed (e.g., Bergson, 1913; Bradley, 1895). It would be nice to put the concepts of mere order and quantity, over and above mere order, onto a common metric, as it were, and assess the magnitude of the difference between them. We can, in fact, do that. Because an unbounded continuous quantity entails a strict simple order, it must be a strict simple order plus something extra. Hence, seeing the difference between the two structures is merely a matter of displaying the structure of an unbounded continuous quantity in such a way that the difference is manifest. We can do this if we project quantitative structure onto order relations between ratios: Then the implications of merely ordinal structure are separated out, and the residue shows what there is to quantity over and above mere order.

Hölder's (1901) concept of quantity was not a new concept. He constructed his axioms so that ratios of pairs of magnitudes defined within his system possessed the structure that Euclid had specified for ratios in Book V of his *Elements* (Heath, 1908). Ratios between magnitudes of a continuous quantitative attribute are now, following Hölder, understood as relations structurally identical to positive real numbers.

Let a, b, c, d, etc., be any magnitudes of the same unbounded continuous quantitative attribute, and let $a{:}b$, $c{:}d$, etc., denote the ratios of a to b (i.e., the size of a relative to b), c to d (i.e., the size of c relative to d), and so on. Consider the order relation between any pair of ratios, $a{:}b$ and $c{:}d$, supposing without any loss of generality that $a \geq c$. The pair of ratios, $a{:}b$ and $c{:}d$, must fall into one and only one of two discrete classes: (a) the class where $b \leq d$; and (b) the class where $b > d$. If it falls into the first class, then $a{:}b \geq c{:}d$. In this case, the order relation between the two ratios follows simply because of the order upon the magnitudes involved. On the other hand, if it falls into the second class, then we cannot tell whether $a{:}b > c{:}d$, given only the ordinal information that we have about the magnitudes. In this case, the order relation between the ratios depends on relations between the magnitudes over and above mere order.

As Euclid indicated in his Definition 7, $a{:}b > c{:}d$ if and only if there exist natural numbers, n and m, such that both $na > mb$ and $nc \leq md$. Then, $a{:}b > m/n \leq c{:}d$. Thus, the extra relations over and above order that must be considered to determine the order relations between pairs of ratios in the second class are the relations of addition that sustain multiples of magnitudes. So the set of all ordered pairs of ratios fall neatly into two classes: those in which the order relation between the pair of ratios is determined by the order of the magnitudes involved (viz., class 1); and those in which the order relation between the pair of ratios is determined by the structure of the magnitudes over and above mere order, what we might call the *additive structure of the attribute* (viz., class 2).

To complete the argument, note two things. First, the complete set of order relations on the pairs of ratios exhausts the content of what it is to be an unbounded continuous quantity. The ordinal structure on the pairs of ratios will be whatever it is if and only if there is an isomorphic mapping, *f*, from magnitudes to positive real numbers such that for any magnitudes, *a*, *b*, *c*, and *d*,

$$a:b \geq c:d \equiv \frac{f(a)}{f(b)} \geq \frac{f(c)}{f(d)}.$$

Hence, *f* must be unique up to multiplication by a positive constant (i.e., is a ratio scale in Stevens's terms). This isomorphism is the same as that achieved by Hölder's axiomatization and, so, the structure defined by Hölder's axioms and the structure given by the set of all order relations on pairs of ratios must be the same structure.

Second, as I have shown elsewhere (Michell, in press), the number of pairs of ratios in class 1 equals that in class 2, and because the order relationships between pairs of ratios exhaust the content of what it is to be an unbounded continuous quantity, it follows that half the structure of such a quantity is due to the merely ordinal relations between magnitudes and the other half is due to additive relations between magnitudes. Putting the point succinctly: Order is half of quantity, and additive structure the other half. Those who would infer quantity from mere order are literally trying to be too clever by half!

It follows, then, that if we want evidence relating specifically to additive structure, as opposed to merely ordinal structure, then we need to look at the order relations between the pairs of ratios in class 2—that is, for any magnitudes, *a*, *b*, *c*, and *d*, order relations between *a:b* and *c:d* where $a \geq c$ and $b > d$. Order relations between the pairs of ratios in class 1 are irrelevant. Noting this enables us to focus attention on just those relations that depend on quantity over and above mere order and not be distracted by other relations.

Interestingly, the above analysis has immediate implications for item response models and to the issue of testing whether latent traits are quantitative. Consider, for example, the model proposed by Rasch (1960) for dichotomous items. This model describes situations in which the probability of a person, *a*, getting an item, *j*, correct ($P(x_{aj} = 1)$) is a function of a single psychological attribute, θ (the relevant latent trait or, in this context, *ability*) as follows:

$$P(x_{aj} = 1) = \frac{e^{(\theta_a - \theta_j)}}{1 + e^{(\theta_a - \theta_j)}}$$

(where *e* is the base of natural logarithms, θ_a is person *a*'s level of ability, and θ_j is the level of ability required to have an even chance of getting item *j* correct and called the item's difficulty). If we transform θ to a new scale, δ, where

$\log_e \delta = \theta$, the relationship to the above discussion becomes obvious. Because $\delta = e^{\theta}$, then Rasch's model becomes

$$P\left(x_{aj} = 1\right) = \frac{\delta_a / \delta_j}{1 + \delta_a / \delta_j} = f\left(\delta_a / \delta_j\right)$$

(Where $f(x) = \frac{x}{1+x}$ and, so, f is an increasing monotonic function mapping positive real numbers into the 0–1 real number interval). That is, according to Rasch's model, $P(x_{aj} = 1)$ is increasingly monotonic with ratios between the relevant levels of the ability attribute, δ. Rasch's model defines the probability of a person getting an item correct as the ratio of the person's and item's abilities transformed by *f(x)* to the 0–1 real number interval.

If these probabilities can be estimated from item response data, information that is diagnostic of whether θ is quantitative is obtainable, not by looking at all order relations between pairs of such estimates but by inspecting order relations between those pairs of estimates corresponding to the pairs of ratios in class 2, that is, by considering the probability of a more able person getting a more difficult item correct versus that of a less able person getting an easier item correct (i.e., for persons *a* and *b* and items *j* and *k*, where $\delta_a > \delta_b$ and $\delta_j > \delta_k$, order relations between estimates of $P[x_{aj} = 1]$ and $P[x_{bk} = 1]$). This result applies not only to Rasch's model but also to any item response model that takes the probability of a person getting an item correct to be an increasing monotonic function of the ratio between the person's ability and the item's difficulty.

However, what exactly would one be looking for in inspecting sample estimates of these probabilities? One would be inspecting them to check that no ensemble of such order relations is incompatible with the hypothesis that θ possesses additive structure. The structure that must obtain on ensembles of order relations between pairs of probabilities in class 2 has, already, been quite clearly specified by Scott (1964) and Krantz, Luce, Suppes, and Tversky (1971), in a quite different context, as part of the general theory of conjoint measurement. The relevant structural condition that these order relations must satisfy is the hierarchy of higher-order cancellation conditions (e.g., double cancellation, triple cancellation). Of course, a number of psychometricians have, in the past, linked parametric item response models with conjoint measurement (e.g., Keats, 1967; Brogden, 1977; Perline, Wright, & Wainer, 1979), and Scheiblechner (1999) has specifically linked the higher-order cancellation conditions with parametric models, apparently without realizing, however, that the hypothesis that the structure of θ is additive is diagnostically linked to these conditions only as they apply to ensembles of order relations between pairs of probabilities in class 2. This is because if θ is merely ordinal, then higher-order cancellation conditions will automatically be satisfied by ensembles of order relations between pairs of probabilities in class 1. In this latter case, these conditions simply follow from satisfaction of the independence condition of conjoint measurement, which is essentially an ordinal condition. So higher-order cancellation conditions are not always diagnostic of quantitative structure, over and above mere order.

The higher-order cancellation conditions of conjoint measurement theory form an infinite hierarchy of conditions. The double cancellation condition is the best known because it figures in the standard axiomatizations of conjoint measurement theory. However, these axiomatizations typically join the double cancellation condition with an Archimedean condition and a solvability condition (e.g., Krantz et al., 1971), and these latter conditions are not directly testable with finite data sets. With finite data sets, their place may be taken by the hierarchy of cancellation conditions identified by Scott (1964). Although infinite with respect to the relevant attribute itself, the testable part of this hierarchy is always finite for any given finite data set. A schema that can be used to generate the finite hierarchy for any finite data set is given by Michell (1990).

Of course, as I have suggested, testing parametric item response models in this way is still only second best. The above tests, although focused on those relations diagnostic of quantity over and above mere order, are still attempting to extract evidence for quantity from the structure of error and not from the theoretically elaborated, psychological significance of the hypothesized latent trait. It is really only when tests of the above form are predicted on the basis of a psychological theory that gives explicit content to quantitative structure and confirmed by data that we will have compelling evidence that the relevant latent trait is quantitative.

The reason I am confident that psychometrics is pathological science is that the theoretical and analytic work necessary to undertake tests of the kind I have just indicated has not yet been done. As far as the analytic work is concerned, at present, we know only what tests of double (Michell, 1988) and triple cancellation (Kyngdon & Richards, 2007) are required, but beyond that, no one has yet specified cancellation tests necessary to diagnose additivity. This, by itself, does not make psychometrics pathological, but it does when conjoined with the presumption that psychological attributes are quantitative.

It should not be assumed that the growing body of psychometric literature discussing the distinction between so-called *categories* and *continua* (e.g., Grayson, 1987; Haslam & Kim, 2002; De Boeck, Wilson, & Acton, 2005) has any direct bearing on the distinction between ordinal and quantitative attributes. Discussions of *categories* and *continua* rarely define these concepts explicitly, although the distinction is sometimes held to be equivalent to that between *qualitative* and *quantitative* attributes. Because of the vagueness inherent in these discussions and the failure of discussants to relate their focal concepts to those of order and quantity as explicitly defined in measurement theory, the relationship remains unclear. However, if by a set of categories is meant a classificatory system defined by a reflexive, symmetric, and transitive binary relation

(Suppes & Zinnes, 1963, p. 23) and by a continuous dimension is meant one satisfying the axiom of continuity (axiom 10), then the relationship is transparent. The attempt to distinguish categories from continua, construed at its sharpest, is the attempt to distinguish attributes involving a classification from attributes that are at least continuous, which, because continuity is a purely ordinal condition, does not distinguish mere order from quantity. Although one might quibble about just what empirically testable implications the axiom of continuity entails, if my characterization of the categories-continua distinction is accurate, then this distinction is not relevant to the one of interest in this paper, which as I have been at pains to point out definitely is empirically testable.

Discussions of the categories-continua distinction employ the further distinction between *manifest* and *latent* attributes. In psychometrics, manifest attributes are typically directly observable features of test performance (such as item scores), and latent attributes (such as abilities, personality traits, and social attitudes), while not directly observable, are hypothesized to underlie differences in test performance. In many instances of psychometric application, ordinal structure in the latent attribute is clearly related to identifiable properties of test items via hypothesized psychological processes (e.g., Kyngdon & Richards, 2007; Luo, Andrich, & Styles, 1998; Michell, 1994, 1998), but theories in the relevant content areas, such as those concerned with cognitive abilities, personality, and social attitudes, are not yet able to connect the hypothesized additive structure of latent attributes to identifiable features of test items, as I have already noted. That is, psychometricians seek a level of complexity in the latent variable "not present in the psychological setting of the problem," to use Boring's (1920, p. 33) terms. Recently, one psychometrician has candidly admitted that while psychological theories sometimes predict latent ordinal structure, they do not generally explain additive structure, or as she puts it, with respect to latent attributes, "there is no natural metric" (Embretson, 2006, p. 51). She proceeds to comment that "how such metrics could be obtained is difficult to envision for most psychological constructs" (p. 53). However, rather than see this situation as a defect, Embretson wants to thereby justify parametric item response theories on what are essentially instrumentalist grounds, such as the fact that when, say, a body of ability test data fits the Rasch model, "the intervals between persons [on the latent trait] have uniform meaning for the (log) likelihood that items are solved" (p. 52). Although instrumentalist considerations are sometimes not unimportant in science, on their own, they never amount to a good reason for accepting a proposition as true when that proposition has empirical content. As I have explained above, the hypothesis that latent attributes possess additive structure is an empirically testable proposition. As Galileo once warned, "we must not ask nature to accommodate herself to what might seem to us the best disposition and order, but must adapt our intellect to what she has made, certain that such is best and not something else" (Crombie, 1994, p. 45).

Because, in the first instance, all that psychologists directly observe about the attributes of interest to them are relations of order, and our theories give psychological content to no more than ordinal structures, the cautious, critical, scientific mind will not conclude that psychological attributes are quantitative because, at present, it has no empirical grounds upon which to do so. This applies as much in areas where item response models are employed as in more traditional psychometrics. It is too early to claim that psychological tests *measure* anything. The term *psychometrics* means psychological measurement. However, the fact is, we do not yet know whether psychometrics actually has a subject. The presumption that it must is the root cause of the error in scientific thinking that I have identified, and the attempt to preserve that presumption untested is what makes this case more than just one of error. It makes it a case of pathological science.

ACKNOWLEDGMENTS

This is a revised version of a paper presented at a symposium on *The Limits of Psychological Measurement*, University of Canterbury, Christchurch, New Zealand, Monday, March 1, 2004. I thank Brian Haig for organizing this event and for his invitation to participate. I also gratefully acknowledge the support of the Federal Government of Australia through an ARC Large Grant to investigate the tensions between quantitative and qualitative methods in psychology; the support of my colleagues, Fiona Hibberd, Agnes Petocz, and David Grayson; the support of our research assistants, Sharon Medlow and James Palethorpe; and comments by reviewers for this journal.

REFERENCES

Bergson, H. (1913). *Time and free will* (F. L. Pogson, Trans.). London: George Allen & Co.

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

Boring, E. G. (1920). The logic of the normal law of error in mental measurement. *American Journal of Psychology*, *31*, 1–33.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology*, *14*, 105–120.

Bradley, F. H. (1895). What do we mean by the intensity of psychical states? *Mind*, *13*, 1–27.

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

Crombie, A. C. (1994). *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts* (Vol. 1). London: Duckworth.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*, 129–158.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, *61*, 50–55.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Marwah, NJ: Erlbaum.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., et al. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science*, *1*, 331–349.

Fisher, W. P. (2003). Mathematics, measurement, metaphor and metaphysics II: Accounting for Galileo's "fateful omission." *Theory & Psychology*, *13*, 791–828.

Freud, S. (1957). Instincts and their vicissitudes. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. *14*, pp. 109–140). London: Hogarth. (Original work published 1915)

Grayson, D. A. (1987). Can categorical and dimensional views of psychiatric illness be distinguished? *British Journal of Psychiatry*, *151*, 355–361.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Haslam, N., & Kim, H. C. (2002). Categories and continua: A review of taxometric research. *Genetic, Social, and General Psychology Monographs*, *128*, 271–320.

Heath, T. L. (1908). *The thirteen books of Euclid's elements* (Vol. 2). Cambridge: Cambridge University Press.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, *53*, 1–46.

Keats, J. (1967). Test theory. *Annual Review of Psychology*, *18*, 217–238.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Kyngdon, A., & Richards, B. (2007). Attitudes, order and quantity: Deterministic and direct probabilistic tests of unidimensional unfolding. *Journal of Applied Measurement*, *8*, 1–34.

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

Luo, G., Andrich, D., & Styles, I. (1998). The JML estimation of the generalised unfolding model incorporating the latitude of acceptance parameter. *Australian Journal of Psychology*, *50*, 187–198.

Michell, J. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology*, *32*, 466–473.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *Journal of Mathematical Psychology*, *38*, 244–273.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (1998). Sensitivity of preferences and ratings to ordered metric structure in attitudes. *Australian Journal of Psychology*, *50*, 199–204.

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

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

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist*, *36*, 211–217.

Michell, J. (2002). Stevens' theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology*, *54*, 99–104.

Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology*, *13*, 5–31.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology*, *14*, 121–129.

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, *38*, 285–294.

Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometricians' fallacy. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *17*, 414–432.

Michell, J. (in press). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*.

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement, Part I. An English translation of Hölder (1901), Part I. *Journal of Mathematical Psychology*, *40*, 235–252.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague: Mouton.

Narens, L. (2002). *Theories of meaningfulness*. Mahwah, NJ: Lawrence Erlbaum.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

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*. Chicago: MESA Press.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

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

Schorske, C. E. (1997). The new rigorism in the human sciences, 1940–1960. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation: Fifty years, four disciplines* (pp. 309–329). Princeton: Princeton University Press.

Scott, D. (1964). Measurement models and linear inequalities. *Journal of Mathematical Psychology*, *1*, 233–247.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Solovey, M. (2004). Riding natural scientists' coattails onto the endless frontier: The SSRC and the quest for scientific legitimacy. *Journal of the History of the Behavioral Sciences*, *40*, 393–422.

Soyfer, V. N. (1994). *Lysenko and the tragedy of Soviet science*. Newark, NJ: Rutgers University Press.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stevens, S. S., & Davis, H. (1938). *Hearing: Its psychology and physiology*. New York: Wiley.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Suppes, P., & Zinnes, J. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 1–76). New York: Wiley.

van der Linden, W. J. (1994). Review of Michell, *An introduction to the logic of psychological measurement*. *Psychometrika*, *59*, 139–142.

Von Kries, J. (1882). Über die Messung intensiver Grössen und über das sogenannte psychophysische Gesetz. *Vierteljahrsschrift für wissenschaftliche Philosophie*, *6*, 257–294.