Michell, J. (2009) Invalidity in Validity. In Lissitz, R.W. (Eds.). The Concept of Validity: Revisions, New Directions, and Applications (Chapter 6, pp. 111-133). Charlotte: NC. Information Age Publishing. ISBN: 9781-60752-227-0.

CHAPTER 6

INVALIDITY IN VALIDITY

Joel Michell

ABSTRACT

The concept of test validity was proposed in 1921. It helped allay doubts about whether tests really measure anything. To say that the issue of a test's validity is that of *whether it measures what it is supposed to measure* already presumes, first, that *the test measures something* and, second, that *whatever it is supposed to assess can be measured.* An attribute is measurable if and only if it possesses both ordinal and additive structure. Since there is no hard evidence that the attributes that testers aspire to measure are additively structured, the presumptions underlying the concept of validity are invalidly endorsed. As directly experienced, these attributes are ordinal and non-quantitative. The invalidity in validity is that of feigning knowledge where ignorance obtains.

SOME HISTORY: VALIDITY AND THE MYTH OF MENTAL MEASUREMENT

Over the past half-century, I have not only been involved in many facets of the theory and practice of testing but, also, I have been interested in understanding measurement as a scientific method.[1] Measurement has always been an aspiration of testers,[2] however, it has long been an achievement of physical scientists and, so, I reasoned, any attempt to understand it must

The Concept of Validity, pages 111–133

begin with measurement as understood in physics. As a result, I came to see that the way in which tests and testing are discussed (the *discourse of testing*) is not squarely based upon what is known about tests and testing (the *science of testing*). The discourse of testing, in particular presenting testing as if it were a form of measurement, outstrips the science upon which testing is based. There is no evidence either that tests *measure* anything or that the attributes that testers aspire to measure are *measurable*, so presenting tests as if they were instruments of measurement is presenting a myth,[3] the myth of mental measurement, as if it were a known fact. Furthermore, because the concept of validity is integral to this myth, if there is some dissatisfaction with this concept (Lissitz & Samuelsen, 2007), it is not within a mainstream still in thrall of this myth. Only when this myth is jettisoned will the discourse of testing, including talk of validity find a basis in reality. Since this myth is held in place by enduring social factors extrinsic to the science of testing, this is not likely to happen in the short term. We can get some idea of what these factors are by looking at the history of the concept of validity.

Rogers (1995) describes *validity*'s arrival in the mainstream testing movement:

> The year 1921 emerges as pivotal in the emergence of validity as part of the institutional proceedings of the testing field. This is when the idea gained clear acceptance. The Standardization Committee of the National Association of Directors of Educational Research[4] polled its members about the desirability of publishing an official list of terms and procedures. The results of this survey were announced in Courtis[5] (1921). Of particular note is this statement: "Two of the most important types of problems in measurement are those connected with the *determination of what a test measures* [italics added], and of how consistently it measures. The first should be called the problem of *validity* [italics added], the second, the problem of reliability" (p. 80). This is the first institutional definition of validity. (p. 246)

There were powerful pressures at work in education and psychology[6] bringing the burgeoning testing industry under the control of the profession and crafting the concept of *validity* was part of this process. As Courtis' ink was drying, Buckingham[7] (1921) reiterated, "By validity I mean the extent to which they [i.e., tests] measure what they purport to measure" (p.274). *Validity* was soon in the title of journal articles (e.g., Davis, 1922) and the concept was broadcast via McCall's[8] (1922) textbook, *How to Measure in Education*. Kelley's[9] (1927) text, *Interpretation of Educational Measurement*, reaffirmed that "The problem of validity is that of whether a test really measures what it purports to measure" (p.14) and by then this understanding was entrenched in the discourse of testing. It is there in the most recent of texts, such as Furr and Bacharach (2008, p.168), whose "basic definition" of validity, echoes Ebel's[10] earlier "basic notion" that "validity is the degree to which a test measures what it is supposed to measure" (1961, p. 642). Borsboom, Mellenbergh and van Heerden (2004) correctly note that this is *validity*'s core meaning.

By 1921, testers had developed a culture in which tests were marketed as instruments of measurement and test scores called "measures" as a matter of course and the concept of validity was integral to this culture, for to say that validity is the issue of *whether a test measures what it is supposed to measure*, already presumes, first, *that the test measures something* and, second, *that what it is supposed to assess is measurable*. In 1921, neither presupposition was scientifically defensible and testers, in feigning knowledge where none yet existed invoked the myth of mental measurement, the idea that mental traits are quantitative attributes.

However, this myth had its detractors.[11] For instance, in 1920, the *American Journal of Psychology* published a paper by Edwin G. Boring,[12] *The logic of the normal law of error in mental measurement*, arguing that psychological tests deliver only ordinal assessments. It was directed at the "mental test as a newcomer" (p. 1) to quantitative psychology and criticized the fact that testers inferred units of measurement by imposing a distributional form upon observed scores. He noted that distributional forms can only be discovered after units are determined and "The great difficulty is... to find anything that we may properly call a psychological unit" (p. 31). He concluded, "We are left then with rank-orders... and it is with these rank-orders that we must deal. We are not yet ready for much psychological measurement in the strict sense" (p. 32).

This was decades before Stevens[13] introduced his definition of measurement and its associated theory of scales. In distinguishing rank-orders from "measurement in the strict sense," Boring was saying that testers were not able to *measure* in the same sense of *measure* as used in physical science. It was an unpalatable indictment, for testers marketed tests as measurement devices and thought that assessment via tests was "in general the same as measurement in the physical sciences" (McCall, 1922, p. 5). In an address given at the annual meeting of the National Vocational Guidance Association in Atlantic City in 1921, Morris Viteles[14] premised his call for tests to be used in industry as extensively as in education upon the assumption that "Tests are the devices by which mental abilities can be measured" (1921, p. 57). This view was typical and all testers would have understood the tension between it and Boring's reality check.

However, the myth had its champion: Truman Lee Kelley. Described by Boring (1929) as "Thorndike's pupil and Stanford University's copy of Karl Pearson, perhaps now America's leading psychologist-statistician" (p. 528), Kelley (1923) adopted an unashamedly pragmatic view and argued that Boring's agonies over units were unnecessary because "starting with units

however defined, if we can establish important relationships between phenomena measured in these units, we have proceeded scientifically. The choice of unit is purely a question of utility" (p. 418). This was smoke and mirrors. While the fact that test scores relate to important criteria needs explaining, the assumption that they do this *only* because they *measure* something presumes that the relevant attributes are quantitative. Units presume quantity and in science the hypothesis that attributes are quantitative, like any empirical hypothesis, is not made true by wishing.

Despite Kelley's question begging or, actually, because of it, the testing community embraced his stance. While lone voices still asked whether tests deliver measurements (e.g., Adams, 1931; Brown, 1934; Johnson, 1936; Smith, 1938), doubt withered within the mainstream and the issue hibernated. The presuppositions behind the concept of validity were endorsed to advance the reception of tests as instruments of measurement. The process whereby the rhetoric of measurement became entrenched within the testing community was facilitated from the 1930s by the adoption of operationalism.[15] The basic tenet of this philosophy is that within science, the meaning of any concept is synonymous with the corresponding set of operations used to measure it. This was taken to imply that the concepts that testers aspire to measure should be identified with the testing operations involved. Unfortunately, operationalism is based upon the confusion of *what* is measured with *how* it is measured (i.e., the confusion of a concept with how that concept is known (Michell, 1990)) and, so, any attempt to use it to save the myth of mental measurement must be logically defective, no matter how popular it proved to be historically.

The rhetoric of measurement was further reinforced when Stevens' (1946) operational definition of measurement[16] was adopted as part of an emerging, post-Second World War methodological consensus in psychology. This definition appeared to justify the kind of loosening of the term *measurement* from its scientific moorings that had occurred in psychology over the preceding decades: any rule for assigning numerals to things could now be called "measurement" it seemed.[17]

Part of this consensus, as well, was the new concept of *construct validity* (Cronbach & Meehl, 1955). While these authors attempted to correct the confusion involved in operationalism by distinguishing testing operations from the relevant concepts (or *constructs*) that testers aspire to measure, they committed a fallacy of the same kind in confusing the validity of a test with the process of validating it (i.e., they thought of the validity as qualifying an inference from relevant data to the claim that the test measures a nominated construct). As Borsboom et al. (2004) argue, if the concept of validity is to make any sense, it must be understood as a property of tests (i.e., a test, *T*, may have the relational property of measuring attribute, *A*)

and not as a property of the process of showing that test *T* measures *A* (i.e., the means by which *T*'s validity is known).

This new concept did little to resolve the tension between myth and reality. While it raised the issue of whether a test actually measures a nominated construct, and did this in a way that recognized it as an empirical issue, highlighting the inferential gap between what a test *actually* assesses and what it is *intended* to assess, and recognizing that this gap could be bridged by discovery of lawful relationships (Meehl's *nomological networks*), the two presuppositions, *that tests measure something* and *that psychological constructs are measurable* remained securely locked in place. In fact, construct validity's official imprimatur[18] served to further protect these myths and those who attempted to make it the standard concept of validity (e.g., Messick, 1989) were wont to call test scores "measures" as indiscriminately as testers had done half a century before. *Construct validity*, along with other validity concepts (e.g., *predictive*, *concurrent*, and *content*) was understood by the mainstream as just another variation upon the theme of Courtis's core concept.

To summarize to this point: the concept of validity entered the testing profession at a time when the idea that testing was a form of measurement was still questioned within the mainstream; the concept quickly became an important component of testing discourse and its specific function was to reassure testers that even though they might not yet know *what* tests measured, they could sleep easy because their "mental tests measure something" (Kelley, 1929, p. 86); and, so, it was still safe to promote tests as instruments of measurement.

Objectively speaking, the presuppositions upon which the concept of validity is based remain as questionable now as in 1921. Testers are deeply confused about the concept of measurement.[19] They incant Stevens' definition when pressed for a form of words, but when it comes to theorizing, they necessarily, but usually unwittingly presume that the attributes they aspire to assess are quantitative (i.e., they presume that "*To be measurable an attribute must fit the specifications of a quantitative variable*" (Jones,[20] 1971, p.336, italics in original)). However, because they rarely consider these specifications in detail, testers just as rarely come face to face with the extent of their confusion. When these specifications are faced, it is clear that the issue of whether psychological attributes are measurable remains an open question.

SOME PHILOSOPHY: QUANTITATIVE STRUCTURE AND MEASUREMENT

Here is not the place for a detailed description of quantitative structure.[21] Instead, I will briefly indicate the kind of structure necessary for interval

scale[22] measurement, this being the form usually presumed in testing. Our paradigm of quantitative structure is the real number line, which is why we resort to geometric diagrams when theorizing about quantitative attributes. Just as the number line consists of an ordered series of points, so any quantitative attribute consists of an ordered series of *magnitudes.*[23] Order presumes the *mutual homogeneity*[24] of the magnitudes or *degrees*[25] involved. For example, if one degree of an attribute is greater than another, they must be degrees *of the same attribute,* for to be *greater than* is always to be *greater than in some respect.* But order is just one component of quantity.

Quantitative structure also requires that *differences between magnitudes* be mutually homogeneous, for it is necessary that these differences also be ordered. Furthermore, given any two such differences, if the attribute is quantitative, then there always exists a third difference that makes up the deficit between them. For the attribute to be quantitative, this relation of composition must satisfy the *commutative* and *associative* laws of addition.[26] In short, with quantitative attributes, differences between magnitudes are mutually homogeneous and *additively* structured.

Quantitative structure then consists of ordinal structure and additive structure: the class of magnitudes of the attribute and the class of differences between magnitudes are each mutually homogeneous and the differences are composed additively. To complete the picture, it is also usually required that first, there be no least difference, second, no greatest difference, and third, that the ordered sequence of differences contain no gaps (i.e., that the attribute be continuous).

The virtue of this concept of continuous quantity is that it entails that ratios between differences are positive real numbers. The importance of this cannot be stressed too much: it is a door by which real numbers enter science. If any one difference is designated as the unit of measurement, each other difference between magnitudes already possesses a measure, this being the ratio it stands in to the unit. Measurement, in the scientific sense of the term is *the estimation of these ratios.*[27]

Of course, the word "measurement" has a further range of colloquial meanings, a fact Stevens exploited in selling his definition.[28] Only the above meaning matters in science and there is no legitimate role for Stevens' definition. Once that is recognized, it is clear that Boring was right to say, "We are not yet ready for much psychological measurement in the strict sense" (1920, p. 32), for as far as we know, the attributes testers aspire to measure are only ordinal structures.

SOME LOGIC: THEORY OF ORDINAL, NON-QUANTITATIVE STRUCTURES AND AN EXAMPLE

Characterizing quantitative structure indicates the locus of the distinction between quantitative and merely ordinal attributes.[29] This distinction lies in two possible sources. First, a merely ordinal attribute might be such that *the differences between its degrees are not intrinsically greater than, equal to or less than one another.* Even though the degrees of such an attribute are ordered, *the differences between those degrees* might only be the same or different in relation to one another. Then the differences between degrees show *qualitative difference,* but imply no *quantitative distance.*[30] Johannes von Kries (1882)[31] was the first to signal this possibility to psychologists and John Maynard Keynes (1921), through his work on the concept of probability made it more widely known.[32]

Second, even if the differences between its degrees are intrinsically ordered, it does not automatically follow that the attribute involved must be quantitative. For such an attribute to be quantitative, *the compositional relations between differences must be additive* and that possibility does not follow simply from the fact that the differences are ordered. It is an issue that can only be addressed by gaining access, either direct or indirect, to the specific ordinal relations between these differences. The theory of difference structures[33] indicates which relations upon differences entail quantity and which do not. In the philosophy of measurement, Hölder (1901) seems to have been the first to attend to this issue.[34]

Quantitative psychology is notorious for neglecting these two possibilities.[35] From Fechner onwards, the mainstream accepted the "constantly recurring argument" that any ordinal attribute is measurable "because we can speak, intelligently and intelligibly, of 'more' and 'less' of it" (Titchener, 1905, p. lxiii).[36] That is, they believed, erroneously, that mere order entails quantity[37] and ignored the fact that the attributes they aspired to measure might be merely ordinal. This was an egregious oversight because *prima facie* we have more reason to think of mental attributes as merely ordinal than as quantitative.

Differences between merely ordinal and quantitative attributes are significant because they mark a boundary between quantitative and non-quantitative science. A feature that gives quantitative science (e.g., physics) explanatory power is the range of quantitative relationships holding between measured attributes. In his review of the recent book by Cliff and Keats (2003) on ordinal analysis, Luce[38] recommended that our "goal really should be ratio scale measures, and we should not remain content with ordinal scales" because without ratio scales "no strong formal theories re-

not additive structure that is the invariant with monotonically interrelated models. Embretson (2006) nailed it when she concluded, "Model-based measurement, which includes IRT, does not provide a universal metric with zero points and interval widths," noting, "How such metrics could be obtained is difficult to envision for most psychological constructs" (p. 53). Evaluating test data against IRT models, by itself may provide no good reason to think the relevant attributes are quantitative.[42]

A more sensitive approach would be through identifying experimental outcomes specifically diagnostic of additive structure. The theory of conjoint measurement (Krantz et al., 1971) provides an avenue for this.[43] If abilities (for instance) are quantitative attributes, then given any pair of test items assessing the same ability, the difference between them in degree of difficulty will be equal to, greater than or less than the difference in degree of difficulty between any other such pair of items. Furthermore, any such ordinal relationship between differences in degrees of difficulty would not exist without a basis in the items themselves. That is, it would be due to identifiable features of the items involved. If such features exist, it would, in principle, be possible to engineer pairs of items for which it is known in advance that one difference between degrees of difficulty exceeds another. If such item engineering can be achieved, then conjoint measurement theory, with its hierarchy of cancellation conditions (Michell, 1990) could be applied to response data to test whether abilities are quantitative. However, despite the valuable research of Embretson and others[44] into item features, we are still a long way from identifying features systematically linked to the presumed additive structure of abilities and this is because if abilities possess additive structure, we do not yet know enough about this structure or the features of test items related to it. We cannot know whether abilities are quantitative until we have good theories connecting the hypothesized additive structure of abilities to features of test items.

Standard test theories, such as IRT models are less suited to coming to grips with the distinction between quantitative and merely ordinal attributes because they represent an approach in which theories are tailored to fit existing instruments (viz., mental tests) on the assumption that relevant attributes are already known to be quantitative and instruments, already known to be capable of measuring. On these assumptions, theories exist chiefly to justify instruments. Such an approach puts the instrumental cart before the scientific horse. What is really required are instruments tailored to true theories about the psychological processes involved in the attributes we wish to assess. In the absence of such theories we cannot determine whether our attributes are quantitative, no matter how well data fit our models.

The attributes that testers aspire to measure are experienced *only* as ordinal and, furthermore, in so far as we experience differences between degrees of such attributes, we seem to experience them as heterogeneous, not as differences in amounts of some homogeneous stuff.[45] For example, it seems that a person of high mathematical ability, say, does not differ from a person of merely moderate ability by possessing the same kind of knowledge, skills and strategies that distinguish the moderately able from persons of low ability. Instead such a person has a high degree of ability precisely because of the qualitatively different, superior mathematical knowledge, skills and strategies possessed. If such attributes are quantitative, then not only is their quantitative structure yet to be discovered, but also their character as different amounts of some homogeneous quantity is yet to be specified.

SOME IMPLICATIONS: THE SCIENTIFIC AND INSTRUMENTAL TASKS OF ASSESSMENT

In Michell (1997) I distinguished the *scientific* and *instrumental tasks* of measurement. The *scientific task* is to investigate whether the relevant attribute is quantitative; if it is, then the *instrumental task* is to devise measurement instruments by locating standardized operations the outcomes of which are sensitive to the relevant attribute's quantitative structure. While the scientific task is *logically prior* to the instrumental, in practice the two tasks may be undertaken jointly. In relation to the present discussion, the distinction between these two tasks can be extended from the domain of measurement to the more general one of assessment.

So extended, the *scientific task of assessment* would be to investigate the structure of the relevant attribute. Attributes possess their own natural structure and they should not be presumed to only possess structures that we consider desirable, such as quantitative structure. Nor should it be presumed, as is customary in psychology, that all attributes possess one of just two kinds of structure, quantitative and classificatory structures (e.g., Meehl, 1992). Between merely classificatory attributes (like *nationality* or the various diagnostic categories for mental disorders) and quantitative attributes (like *length* and *mass*), there is an array of possible ordinal structures, like the various kinds of partial orders, weak orders and simple orders (Michell, 1990). Since most of the attributes that testers aspire to measure seem at first sight to be in some sense ordinal and since there is no evidence that these attributes are quantitative, it is this array of intermediate, ordinal structures that are most relevant to testing. However, here testers are let down by their education (Michell, 2001), for an introduction to these intermediate structures is not generally included in the testing curriculum. While there seems little point in calling for revision of relevant university course syllabuses,[46] it is obvious that a necessary condition for thinking in

terms of a specific set of concepts is that of first being made aware of their existence.

An attribute is ordinal if for at least some pairs of its degrees, one degree is greater than another (in the relevant sense), where this *greater than* relation is transitive and asymmetric. Identification of the kind of ordinal structure involved in particular cases requires not only defining the relevant attribute explicitly, but also identifying the relevant *greater than* relation. The difficulties involved in doing this will vary from context to context. In the case of achievement testing, where the tester is attempting to assess the level of a person's knowledge in some circumscribed domain, it will be much easier than in, say, the case of ability assessment, where characterizing the relevant ability requires framing hypotheses about cognitive processes and, so, must build upon discoveries in cognitive psychology. However, the scientific task of discovering the attribute's structure cannot be accomplished until testers are able to define the attribute to be assessed and identify the *greater than* relationship for that attribute. Where this has not been done for any attribute, claims to measure it are vacuous because they are made not only in ignorance of what *measurement* means, but also in ignorance of the character of the attribute involved.

The *instrumental task of assessment* is to construct tests, performance upon which reflects aspects of the structure of the relevant attribute. If this task seems like test validation, only by another name, then the distinction needs to be clarified for, while superficially similar, the instrumental task of assessment is different to test validation. In the first place, test validation focuses first on the particular *test* under investigation, while the instrumental task focuses first on the *attribute* to be assessed. Second, given a specific test, validation attempts to discover the *attribute or attributes* underlying test performance, while given a specific attribute, the instrumental task attempts to discover the *features of test items* sensitive to structural properties of that attribute. Third, as already noted in this paper, test validation presumes that something is *measured*, while the instrumental task aspires to *assessment* (a much wider concept than measurement).

If tests are useful for the assessment of certain attributes, then this must be because of the features of the items involved. The instrumental task of assessment is primarily concerned with discovering, for some nominated and explicitly defined attribute (relative to which the *greater than* relation is identifiable), the features of test items that cause individual differences in performance to be sensitive to the ordinal (or possibly, quantitative) properties of the relevant attribute. Only when such item-features are identified and understood will it be possible to engineer tests whose capacity to assess the relevant attribute is known in advance. In this way, the testers' knowledge of the relevant attribute and relevant item-features would make the concept of test validation entirely redundant.

CONCLUSION: INVALIDITY IN VALIDITY

It is the hiatus between what is *known* about psychological attributes (viz., that they are ordinal) and the *myth* (viz., that they are known to be quantitative) that explains certain features of the concept of validity. It explains the presumptions underlying the concept, viz., that psychological attributes are quantitative and tests, instruments of measurement. It explains why the concept emerged when it did. It emerged in time to deflect attention away from the two issues that it presumed answers to. Boring's critique was published in 1920 and the next year the problem of test validity was first announced to the testing profession. And it explains the concept's durability. In the past century, many things have changed in testing, but not the core understanding of validity, as Borsboom et al. (2004) point out. This is because the concept is still useful in disguising the gap between what testers know and the myth of mental measurement. The invalidity in validity is that the concept feigns knowledge where such does not yet exist.

But does not this invalidity hide a genuine problem, if not that of what tests *measure*, then the problem of what they *assess* and, so, might not the concept of validity be rehabilitated in these terms? Can we not ask of any test, *what attributes does that test enable us to assess?* Of course we can, but it is an odd situation to first have constructed an instrument of assessment and then to ask what it assesses. Such a situation would not arise if the character of the target attribute was investigated in advance and when understood sufficiently well, a test aimed at that target then constructed through knowledge of relevant item features. There is no analogue of the problem of test validity in physical measurement.[47] There, instruments are engineered using knowledge of laws relating features of the instrument to the structure of the relevant attribute. In this respect, test theory is an anomalous enterprise:[48] a body of theory constructed for the assessment of we know not what, on the basis of laws yet unproven. Boring (1920) closed his paper with a still relevant admonition:

> But, if in psychology we must deal—and it seems we must—with abilities, capacities, dispositions and tendencies, the nature of which we cannot accurately define, then 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. Just as ignorance will not breed knowledge, so inaccuracy of definition will never yield precision of result. (p. 33)

That is, focus upon the attributes to be assessed; investigate their structure and their lawful connections to features of test items. Few testers realize that knowledge of the structure of attributes is necessary to provide a scientific base for testing practice.[49] Sadly, testing became a profession prior to developing a scientific base and for nearly a century the concept of valid-

ity has obscured this lack. In accepting this concept, testers embraced the "the seeming truth which cunning times put on to entrap the wisest."[50]

APPENDIX 1: THE AXIOMS OF QUANTITY AND THE CONCEPT OF MEASUREMENT

Measurable quantitative attributes have a distinctive structure. It is useful to think in terms of an example, such as length. Quantitative structure is the same in length as in other attributes, only more evident. We experience length via specific lengths, say the length of a pen or a cricket pitch. These specific lengths are *magnitudes* of length. What makes length quantitative is the way in which its magnitudes interrelate. These interrelations may be stated in seven propositions, sometimes called "axioms of quantity" (e.g., Hölder, 190151). The first four state what it is for length to be additive. The remaining three ensure that all lengths are included. Let a, b, c, be any magnitudes of length, then length is additive because:

1. For every pair of magnitudes, a and b, one and only one of the following is true:
   (i) a is the same as b (ie, $a = b$);
   (ii) there exists a magnitude, c, such that $a = b + c$;
   (iii) there exists a magnitude, c, such that $b = a + c$.
2. For any magnitudes, a and b, $a + b > a$.
3. For any magnitudes, a and b, $a + b = b + a$.
4. For any magnitudes, a, b, and c, $a + (b + c) = (a + b) + c$.

Axiom 1 says that any two lengths are either identical or different and if different, there is another length equaling the difference; Axiom 2 says that if a length is entirely composed of two discrete parts, then it exceeds each; Axiom 3 that if a length is entirely composed of two discrete parts, the order of composition is irrelevant; Axiom 4 that where a length is entirely composed of three discrete parts, it is always the composition of *any* one with the remaining two.

The following three axioms ensure the completeness of this characterization:

5. For any length a, there is another b, such that $b < a$.
6. For any two lengths a and b there is another c such that $c = a + b$.
7. For every non-empty class of lengths having an upper bound, there is a least upper bound.

Axiom 5 means that there is no smallest length; Axiom 6 that there is no greatest; and Axiom 7 says that there are no gaps in the ordered sequence of lengths, that is, length is continuous.

These axioms entail that the ratio of any one magnitude of length to any other is a positive real number.[52] For example, one length might be twice another or three and a half times another or the square root of two times another. The *measure* of one length, c, relative to another, d, is the ratio of c to d. In practice, ratios are rarely specified precisely because measurement procedures possess only finite resolution. *Measurement* is the estimation of the ratio of one magnitude of a quantity to another magnitude (the unit) of the same quantity.

When we claim to be able to measure psychological attributes, such as abilities using tests, we are claiming that these attributes have this kind of structure[53] and that tests are sensitive to this kind of structure. Of course, generally, psychologists do not claim to be able to measure on ratio scales. Typically, they want to claim measurement on interval scales. This makes little difference to the issues under discussion. On an interval scale, measures of intervals are on a ratio scale. Hence, what psychologists are claiming is that *differences* between levels of ability possess quantitative structure (i.e., satisfy Axioms 1–7) and that tests are sensitive to this structure upon differences.

APPENDIX 2: THE FUNCTIONAL INDEPENDENCE SCALE

The series of activities comprising the so-called "functional independence measure" as ordered from most to least difficult (Embretson, 2006) is as follows:

1. Climbing stairs;
2. Transferring to bathtub;
3. Bathing;
4. Walking;
5. Dressing upper body;
6. Independent toileting;
7. Transferring to bed;
8. Dressing lower body;
9. Mobility without a wheelchair;
10. Bladder control;
11. Performing personal grooming; and
12. Bowel control.

ACKNOWLEDGMENT

I am grateful to members of the Models & Measurement Seminar, School of Psychology, University of Sydney for comments upon an earlier version.

NOTES

1. My thoughts on this are given in Michell (2005).
2. I use the term "testers" to refer to psychologists, educationalists, and social scientists involved in the construction and use of tests and in the development of associated theories and methods.
3. In this context, by a *myth* I mean a theory or hypothesis believed because it eases our minds.
4. The National Association of Directors of Educational Research (NADER) was the American Educational Research Association (AERA) in embryo.
5. Stuart A. Courtis, a founding member of NADER who "printed tests and sold them to school districts across the country" (Mershon & Schlossman, 2008, pp. 315–316), later became disillusioned with tests as measurement instruments (Courtis, 1928).
6. See Mershon and Schlossman (2008) and von Mayrhauser (1992).
7. Burdette R. Buckingham, also a founder of NADER was then editor of the *Journal of Educational Research* and Director of the University of Illinois' Bureau of Educational Research (Mershon & Schlossman, 2008).
8. William Anderson McCall (1891–1982) had been Thorndike's student and became professor of educational measurement at Columbia University.
9. Truman Lee Kelley (1884–1961) also Thorndike's student, taught at Stanford University from 1920 and was professor of education at Harvard Graduate School of Education from 1931 to 1950 and president of the Psychometric Society in 1938–39 (Stout, 1999).
10. Robert L. Ebel's (Vice President at ETS (1957–1963) and Professor of Education and Psychology at Michigan State University (1963–1982) espoused controversial views on validity (Cizek et al., 2006).
11. While early on, Binet (1905) recognized that his scale "does not permit the measure of intelligence, because intellectual qualities are not superposable" (p.151), doubts about mental measurement had long been voiced in the adjacent field of psychophysics. However, by 1905 most psychologists accepted psychophysical measurement (Titchener, 1905), but doubts persisted (e.g., Brown, 1913).
12. Edwin G. Boring (1886–1968) was then professor of experimental psychology at Clark University. During World War I he had served in the US Army's mental testing program. From 1922, he taught at Harvard, becoming its first professor of psychology in 1934. He is remembered best for writings on the history of psychology (Reed, 1999).

13. Stevens (1946). Stevens was Boring's student (and, incidentally, at one time, Kelley's). His definition and theory of scales shaped the post-war understanding of measurement in psychology (Michell, 1997, 1999).
14. Morris Viteles (1898–1996) was a pioneer of industrial psychology in the United States and authored influential books in the field (Thompson, 1998).
15. Operationalism was a philosophy of science proposed by the Nobel prize-winning physicist, Percy W. Bridgman (1927) and energetically promoted in psychology by Stevens (e.g., Stevens, 1935).
16. He defined measurement as the assignment of numerals to objects or events according to rule.
17. See Michell (1999).
18. The term "construct validity" was introduced in 1954 in the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* published by the American Psychological Association. Paul Meehl was a member of the subcommittee recommending it. The concept reflects the logical empiricist philosophy of science then dominant in the United States. However, the concept was not universally accepted and some influential testers dissented (see, e.g., Ebel, 1961; Horst, 1966).
19. See, for example, Michell (1997, 1999).
20. Lyle V. Jones is one of the few members of the testing community to candidly note this.
21. However, see Appendix 1.
22. Stevens' (1946) distinguished *nominal, ordinal, interval* and *ratio* scales. Only the latter two depend upon the relevant attribute being quantitative in structure.
23. Following time-honored usage, the term *magnitude* refers to any specific level of a quantitative attribute.
24. Euclid noted this as long ago as the fourth century BC in Book V of his *Elements* (Heath, 1908).
25. The term *degree* refers to any specific level of an ordinal or quantitative attribute without any implied commitment to quantity. Hence, all magnitudes are degrees but not all degrees are magnitudes.
26. These correspond to Axioms 3 and 4 of Appendix 1.
27. For example: "Any measured quantity may thus be expressed by a number (the magnitude ratio) and the name of the unit" (Wildhack, 2005, p. 487).
28. As noted, most testers lean towards Stevens' definition. However, in so far as testers believe that the attributes that they aspire to measure are quantitative attributes (and this is what they are committed to in their theories), Stevens' definition amounts to a form of false consciousness and in accepting that definition testers misunderstand what they are about.
29. The concept of an ordinal attribute is wide, ranging from simple orders to partial orders (Michell, 1990). Here, for convenience, I consider mainly strict simple orders (i.e., attributes, the degrees of which are ordered by a transitive, asymmetric and connected relation).
30. As David Hume (1888) noted, "any great *difference* in the degrees of any quality is called a *distance* by a common metaphor... The ideas of distance and difference are, therefore, connected together. Connected ideas are readily

taken for each other" (p. 393). This tendency is a cause of the cognitive illusion I call the psychometricians' fallacy (Michell, 2006, 2009).

31. Johannes von Kries (1853–1928) was a sensory physiologist and critic of psychophysical measurement, proposing the so-called *quantity objection* (Titchener, 1905). Niall (1995) gives an English translation.
32. John Maynard Keynes (1883–1946) wrote his dissertation on the concept of probability. He followed von Kries in thinking that probabilities are only quantitative in very special cases. In most cases, he thought, they are at best ordinal and non-quantitative. Later, this observation influenced his economic thought (e.g., Keynes, 1936).
33. See Krantz et al. (1971).
34. Psychologists ignored Hölder until Krantz et al. (1971). See Michell and Ernst (1996, 1997).
35. However, some quantitative psychologists considered related possibilities. For example, Stevens (1957) distinguished *prothetic* and *metathetic* continua, the former being continuous attributes in which degrees differ according to *how much* and the latter, continuous attributes in which different values of the same attribute differ *in kind.*
36. Titchener was reviewing psychophysics, but similar arguments were used in testing. For example, McCall (1922), echoing Thorndike (1918), argued that "whatever exists at all exists in some amount" and "anything that exists in amount can be measured."
37. Psychologists were not alone in doing this: philosophers, mathematicians, and economists also committed this fallacy (e.g., see Michell, 2007). It was so ubiquitous in psychology that I have called it *the psychometricians' fallacy* (Michell, 2006, 2009). Typically, in psychology, ordinal scales are seen as obstacles to be overcome (e.g., Harwell & Gatti, 2001) rather than as intrinsically worthwhile structures.
38. Along with Patrick Suppes, R. D. Luce is the leading measurement theorist of recent time (e.g., see Luce, 2005).
39. The distinction between quantitative and qualitative methods resides in the character of the attributes investigated (i.e., whether they are quantitative or not) and the recent association of qualitative methods with non-realist philosophies of science is actually a red herring and not an intrinsic feature of such methods (Michell, 2003).
40. The full set is listed in Appendix 2.
41. I use this scale as an example because the attribute assessed is not mental, so it avoids a number of problems intrinsic to psychology but extrinsic to the issues considered here. Functional independence is a social attribute, which the scale indexes via a range of physical abilities, absence of any one of which contributes to a person's dependence upon helpers. Thus, the attribute assessed is actually a rather complex socio-physical one.
42. If the relevant attribute is merely ordinal, then there will always exist subsets of test items that could fit IRT models, such as the Rasch model. In a test construction context where items are successively culled to give a best-fitting set, therefore, the fit of data to such a model is a biased test of whether the relevant attribute is quantitative.

43. Most testers have resisted applying this theory. Cliff (1992) and Borsboom (2005) suggest reasons, but not the primary one, viz., testers presume that the attributes they aspire to measure are already known to be quantitative and, so, have no use for a theory enabling empirical tests of that presumption (Michell, 1999).
44. See, for example, Embretson and Gorin (2001).
45. In this respect, psychological attributes are similar in structure to the kinds of attributes that R. G. Collingwood (1933) called *scales of forms.* Putting Collingwood's metaphysical concerns to one side, it is notable that he explicitly recognised that attributes having this kind of structure cannot be measured. Collingwood's concept had little impact in psychology, although it has received attention in other disciplines (e.g., Allen, 2008).
46. Such calls have been made, vainly, for decades (e.g., Sutcliffe, 1976).
47. Borsboom (2005) notes this. There are, of course, issues of quality control with instruments of physical measurement, such as those of calibration. These are discussed in standard works on metrology (see, e.g., Laaneots & Mathiesen, (2006).
48. I would maintain that this situation exists because the concept of validity is an integral part of the discourse of a pathological science (see Michell, 2000, 2008).
49. One of the few to come close to recognising this was Loevinger (1957) in her emphasis upon trait structure. However, in the time-honoured tradition of psychometrics, she distinguished only two sorts of structure, *quantitative* and *classificatory.* This, of course, anticipated the more recent concern with the distinction between *categories* and *continua* (e.g., Meehl, 1992; De Boeck, Wilson & Acton, 2005). Loevinger showed little interest in characterising explicitly the sorts of structure involved and testers have followed her in this respect.
50. William Shakespeare, *Merchant of Venice III, ii, 100.*
51. The axioms given here are based on Hölder's (Michell, 1999), but are not identical to his. For an English translation of the relevant part of Hölder's classic paper see Michell and Ernst (1996).
52. The *ratio* of one length to another is the magnitude of the first relative to the second (Heath, 1908).
53. Which is not to say that at the same time we claim to know how to test such a proposition. Most of the attributes known to be quantitative are physical but even in physics evidence for quantitative structure is typically indirect, with the exception, of course, of *extensive quantities,* like length, time and weight. For more on this see Michell (1990, 1999, 2005).

REFERENCES

Adams, H. F. (1931). Measurement in psychology. *Journal of Applied Psychology, 15,* 545–554.

Allen, R. T. (2008). Art as scales of forms. *British Journal of Aesthetics, 48,* 395–409.

Binet, A. (1905). *L'Étude éxperimentale de l'intelligence.* Paris: Schleicher.

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

Boring, E. G. (1929). *A history of experimental psychology.* New York: Appleton-Century-Crofts.

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

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

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

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

Brown, J. F. (1934). A methodological consideration of the problem of psychometrics. *Erkenntnis, 4*, 46–61.

Brown, W. (1913). Are the intensity differences of sensation quantitative? IV. *British Journal of Psychology, 6*, 184–189.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium. XIV. *Journal of Educational Psychology, 12*, 271–275.

Cizak, G. J., Crocker, L., Frisbie, D. A., Mehrens, W. A., & Stiggins, R. J. (2006). A tribute to Robert L. Ebel: Scholar, teacher, mentor, and statesman. *Educational Measurement: Issues and Practice, 25*, 23–32.

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

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioural sciences.* Mahwah, NJ: Lawrence Erlbaum.

Collingwood, R. G. (1933). *An essay on philosophical method.* Oxford: Clarendon Press.

Courtis, S. A. (1921). Report of the standardization committee. *Journal of Educational Research, 4*, 78–80.

Courtis, S. A. (1928). Education: a pseudo-science. *Journal of Educational Research, 17*, 130–132.

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

Davis, R. (1922). The validity of the Whipple Group Test in the fourth and fifth grades. *Journal of Educational Research, 5*, 239–244.

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

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640–647.

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

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343–368.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction.* Los Angeles: Sage.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*, 105–131.

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.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont CA: Wadsworth.

Hume, D. (1888). *A treatise of human nature.* Oxford: Clarendon Press.

Johnson, H. M. (1936). Pseudo-mathematics in the social sciences. *American Journal of Psychology, 48*, 342–351.

Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.335–355). Washington, DC: American Council on Education.

Jones, L. V. & Appelbaum, M. I. (1989). Psychometric methods. *Annual Review of Psychology, 40*, 23–43.

Kelley, T. L. (1923). The principles and techniques of mental measurement. *American Journal of Psychology, 34*, 408–432.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: World Book Company.

Kelley, T. L. (1929). *Scientific method: its function in research and in education.* Columbus: Ohio State University Press.

Keynes, J. M. (1921). *A treatise on probability.* London: Macmillan.

Keynes, J. M. (1936). *The general theory of employment, interest and money.* London: Macmillan.

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

Laaneots, R., & Mathiesen, O. (2006). *An introduction to metrology.* Estonia: TUT Press.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Luce, R. D. (2004). Ordinal attributes, ordinal analyses only. Review of Cliff and Keats (2003). *Contemporary Psychology, 49*, 783–785.

Luce, R. D. (2005). Measurement analogies: comparisons of behavioral and physical measures. *Psychometrika, 70*, 227–251.

McCall, W. A. (1922). *How to measure in education.* New York: Macmillan.

Marchel, C., & Owens, S. (2007). Qualitative research in psychology: Could William James get a job? *History of Psychology, 10*, 301–324.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences of kind. *Journal of Personality, 60*, 117–174.

Mershon, S., & Schlossman, S. (2008). Education, science, and the politics of knowledge: The American Educational Research Association, 1915–1940. *American Journal of Education, 114*, 307–340.

Messick, S. (1989). Validity. In R. L. Linn (eds.), *Educational measurement* (3rd ed., pp.13–103). London: Collier Macmillan.

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

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology, 88,* 355–383.
Michell, J. (1999). *Measurement in psychology: a 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. (2003). The quantitative imperative: positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology, 13,* 5–31.
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. (2007). *Bergson's and Bradley's versions of the psychometricians' fallacy argument.* Paper presented at the First Joint Meeting of ESHHS and CHEIRON, University College Dublin, Ireland.
Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives, 6,* 7–24.
Michell, J. (2009). The psychometricians' fallacy: too clever by half? *British Journal of Mathematical and Statistical Psychology, 62,* 41–55.
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.
Michell, J. & Ernst, C. (1997). The axioms of quantity and the theory of measurement, Part II. [An English translation of Hölder (1901), Part II.] *Journal of Mathematical Psychology, 41,* 345–356.
Niall, K. K. (1995). Conventions of measurement in psychophysics: von Kries on the so-called psychophysical law. *Spatial Vision, 9,* 275–305.
Reed, J. W. (1999). Boring, Edwin Garrigues. *American national biography* (Vol. 3, pp. 217–218). New York: Oxford University Press.
Rogers, T. B. (1995). *The psychological testing enterprise: An introduction.* Pacific Grove CA: Brooks/Cole.
Smith, B. O. (1938). *Logical aspects of educational measurement.* New York: Columbia University Press.
Stevens, S. S. (1935). The operational definition of psychological terms. *Psychological Review, 42,* 517–527.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.
Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64,* 153–181.
Stout, D. (1999). Kelley, Truman Lee. *American national biography* (Vol. 12, pp. 491–492). New York: Oxford University Press.
Sutcliffe, J. P. (1976). *Mathematics needed for particular social sciences.* Sydney: Academy of Social Sciences in Australia.
Thompson, A. S. (1998). Morris S. Viteles (1898–1996). *American Psychologist, 53,* 1153–1154.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *Seventeenth yearbook of the national society for the study of education* (Vol. 2, pp. 16–24). Bloomington IL: Public School Publishing.
Titchener, E. B. (1905b). *Experimental psychology: a manual of laboratory practice. Vol. II: Quantitative experiments. Part II: Instructor's manual.* London: Macmillan.
Viteles, M. (1921). Tests in industry. *Journal of Applied Psychology, 5,* 57–63.
von Kries, J. (1882). Über die Messung intensiver Grössen und über das sogenannte psychophysische Gesetz. *Vierteljahrsschrift für wissenschaftliche Philosophie, 6,* 257–294.
von Mayrhauser, R. T. (1992). The mental testing community and validity. *American Psychologist, 47,* 244–253.
Wildhack, W. A. (2005). Physical measurement. In M. D. Licker (Ed.), *McGraw-Hill concise encyclopedia of physics* (pp. 483–487). New York: McGraw-Hill.