Queen's University
Belfast

How secure is a Newtonian paradigm for psychological and educational measurement?

Cantley, I. (2015). How secure is a Newtonian paradigm for psychological and educational measurement?
Theory and Psychology, 25(1), 117-138. DOI: 10.1177/0959354314561141

Published in:
Theory and Psychology

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© The Author(s) 2014.

Download date:15. Feb. 2017

How secure is a Newtonian Paradigm for Psychological and Educational Measurement?
Ian Cantley
Queen's University Belfast

Biography

IAN CANTLEY is a lecturer in education at Queen's University Belfast. His research interests include the philosophy of education and the mathematical/philosophical foundations of educational measurement models. ADDRESS: School of Education, Queen's University Belfast, 69-71 University Street, Belfast, UK, BT7 1HL. Email: [i.cantley@qub.ac.uk]

Abstract

From the early 1900s, some psychologists have attempted to establish their discipline as a quantitative science. In using quantitative methods to investigate their theories, they adopted their own special definition of measurement of attributes such as cognitive abilities, as though they were quantities of the type encountered in Newtonian science. Joel Michell has presented a carefully reasoned argument that psychological attributes lack additivity, and therefore cannot be quantities in the same way as the attributes of classical Newtonian physics. In the early decades of the 20th century, quantum theory superseded Newtonian mechanics as the best model of physical reality. This paper gives a brief, critical overview of the evolution of current measurement practices in psychology, and suggests the need for a transition from a Newtonian to a quantum theoretical paradigm for psychological measurement. Finally, a case study is presented that considers the implications of a quantum theoretical model for educational measurement. In particular, it is argued that, since the OECD's Programme for International Student Assessment (PISA) is predicated on a Newtonian conception of measurement, this may constrain the extent to which it can make accurate comparisons of the achievements of different education systems.

Keywords

Introduction

From the advent of modern psychology in the 19th century, some psychologists have sought to establish their discipline as a quantitative science. The 19th century German philosopher and physicist Gustav Theodor Fechner invested considerable effort in attempting to establish the relationship between nature and spirit (Geist). In studying the links between the mental and physical realms, Fechner (1860) devised methods for describing the relation between external physical stimuli and mental states, such as sensations. Through his study of the so-called psychophysics, Fechner was able to gain insights into the philosophical mind-body problem. However, as Heidelberger (2004, p. 2) notes, "the most significant and renowned outcome of his [Fechner's] endeavour was psychophysics, which became the foundation for quantitative empirical psychology." Fechner's pivotal role in the establishment of quantitative psychology was acknowledged by Boring (1929, p. 286):

> Of course, it is true that, without Fechner or a substitute which the times would almost inevitably have raised up, there might still have been an experimental psychology. ... There would, however, have been little of the breath of science in the experimental body, for we hardly recognize a subject as scientific if measurement is not one of its tools. Fechner, because of what he did and the time at which he did it, set experimental quantitative psychology off upon the course which it has followed.

Therefore, the establishment of psychology as a scientific discipline was dependent upon the ability of its pioneers to devise valid and reliable methods for measuring psychological phenomena. This viewpoint was endorsed by, amongst others, Spearman (1937, p.89):

> But great as may be the potency of this [experimental method] ... there is yet another one so vital that, if lacking it, any study is thought by many authorities not to be scientific in the full sense of the word. This further and crucial method is that of measurement, or rather of mathematics; for this latter is what science really needs.

The current paper describes the key milestones in the development of measurement theory in psychology and critically evaluates the theoretical basis of current approaches to psychological measurement. The relevance of the early views of Euclid on measurement are considered in addition to the contributions of more recent prominent thinkers, such as Joel Michell, who has undertaken extensive work on the foundations of psychological measurement. It is argued that Michell's (1990, 1997, 1999, 2000, 2003a, 2003b, 2008, 2011, 2012) work demonstrates there is no overlap between the principles underpinning psychological measurement and those underpinning the measurement of dynamic attributes of macroscopic objects in Newtonian physics. Rather, it is suggested that a quantum theoretical framework may provide a more secure basis for psychological measurement. Furthermore, a case study is presented that considers the implications of a quantum theoretical model for educational measurement. In particular, it is argued that, since the OECD's Programme for International Student Assessment (PISA) is predicated on a Newtonian conception of measurement, this may constrain the extent to which it can make accurate comparisons of the achievements of different education systems.

Newtonian versus quantum theoretical measurement

In Newtonian physics, macroscopic objects possess their dynamic attributes intrinsically, and measurement is a mechanism for checking up on the values of those attributes. For example, when an apple falls to the ground, it possesses a definite velocity at each point on its path irrespective of whether the velocity is actually being measured. A measurement of the velocity simply yields a description of a pre-existing reality. One of the striking features of quantum theory is the pivotal role played by participation: the physicist no longer stands back and offers an objective description of what unfolds when an apple falls to the ground. Rather, measurement influences and does not merely check up on something that already exists.

Heisenberg (1958/2000, p. 22) asserts: "In classical physics science started from the belief – or should one say from the illusion? – that we could describe the world or at least parts of the world without any reference to ourselves". However, in contrast to Kantian philosophy, the role played by human beings can never be purged from the fundamental conceptual problems of quantum theory. Heisenberg makes frequent reference to the role of the observer when discussing the measurement problem in quantum theory:

> This again emphasizes a subjective element in the description of atomic events, since the measuring device has been constructed by the observer, and we have to remember that what we observe is not nature in itself but nature exposed to our method of questioning. ... In this way quantum theory reminds us, as Bohr has put it, of the old wisdom that when searching for harmony in life one must never forget that in the drama of existence we are ourselves both players and spectators. (Heisenberg, 1958/2000, pp. 24-25)

Bohr's appeal to the notion that the physicist is both player and spectator is a reference to a fundamental quandary in quantum theory whereby the observer, in effect, influences, at least in part, what he or she observes. This does not, however, render quantum theory subjective. Rather, the strong objectivity of classical Newtonian mechanics gives way to quantum theoretical weak objectivity or inter-subjectivity. Wheeler (1996), who worked with Bohr on the explanation of nuclear fission, claims: "We are inescapably involved in bringing about that which appears to be happening" (p. 120) and "Useful as it is under everyday circumstances to say that the world exists 'out there' independent of us, that view can no longer be upheld. There is a strange sense in which this is a 'participatory universe'" (p. 126).

Newell (1973), in his paper "You can't play 20 questions with nature and win", argues that much research in cognitive psychology consists of asking binary questions and doing experimental work to study various psychological phenomena. In other words,

researchers consider a particular behavioural phenomenon and posit theories about the organisation of the mind required to produce the behaviour, generally leading to binary questions concerning whether theory A or theory B is appropriate. Newell stresses that, although such research has important contributions to make to our understanding of human behaviour, the focus needs to shift from answering questions to developing good models of how the whole system of the mind works.

Wheeler (1996) makes use of the game of "20 questions" to contrast classical and quantum measurement and to highlight the role of participation. The orthodox approach to the game illustrates Newtonian measurement. A group of people send one of their number outside a room while they select a word. The object of the game is for the person to return and try to identify the word using at most 20 questions (each soliciting one bit of information) such as: "Is it an animal?", "Is it a mineral?", "Is it pink?" The unorthodox version of the game illustrates the participatory nature of quantum measurement. In this case a word is not selected in advance by those in the room.

Each person in the room can answer as he or she pleases provided the word that the person thinks of is compatible with the responses to all prior questioning. For example, suppose the first person questioned is asked, "Is it expensive?" If that person responds in the negative, the second person must think of something which, whatever its other properties are, is not expensive. Assume that the second person thinks of the word "sock" and the questioner asks, "Is it something you wear?" The third person must then think of something which isn't expensive but can be worn. As the questions continue the game becomes increasingly demanding for the people in the room. In the "Newtonian" version of the game, there is a word in the room to be discovered. In the "quantum" version of the game, there is not a word in the room. In this case the word is determined by the sequence of questions the person asks

when he or she re-enters the room. The person playing the Newtonian game is a mere observer while the person in the quantum version is an actor as well as an observer.

In the Newtonian case, the word is already in the room waiting to be discovered and the questioner who succeeds in finding the word is merely unearthing something which pre-exists his or her questions. In the quantum version there is no word in the room and the questioner participates in the selection of the word; changing his or her question sequence will change the word. This sheds light on Heisenberg's (1958/2000, p. 25) aphorism: "what we observe is not nature in itself but nature exposed to our method of questioning". It also explains why psychology will search in vain for what intelligence *is*, for example. The psychologist is both actor and spectator. When the psychologist selects some items for IQ tests and rejects others, he or she is, in part, already defining intelligence. This is the psychologist as participator. Subsequently the psychologist studies response patterns in search of consistencies which may throw light on the nature of intelligence. This is the psychologist construed as mere observer. Unfortunately, this dual role condemns the psychologist to a fruitless search for that pre-existing "word in the room". A Newtonian model for the measurement of psychological predicates therefore appears to be under strain in the context of this example.

The classical concept of measurement

The classical concept of measurement is predicated on the fact that all measurable attributes are quantitative. One of the earliest treatments of measurement was given in Book V of Euclid's *Elements* (Heath, 1956).

Aristotle divided quantities into multitudes (discrete quantities), e.g., number of people in a room, and magnitudes (continuous quantities), e.g., length of a field. Aristotle defined a quantity thus: "We call a quantity that which is divisible into constituent parts of

which each is by nature a one and a "this". A quantity is a *multitude* if it is numerable, a *magnitude* if it is measurable." (cited in Stein, 1990, p. 164)

One quantity was said to be a measure of another if the latter was a whole-numbered multiple of the former. Clearly this notion of measure was appropriate for multitudes (which can, for example, be regarded as whole-numbered multiples of one) but, in general, it is not applicable to magnitudes since pairs of magnitudes may be incommensurable, i.e., no whole-numbered multiple of one of them may be equal to some whole-numbered multiple of the other (e.g., the lengths of the side and the diagonal of a square are incommensurable). Therefore, Euclid's challenge was to generalise the concept of measure to render it applicable to an arbitrary magnitude relative to any unit. Euclid achieved this generalisation by introducing the concept of ratio: "A *ratio* is a sort of relation in respect of size between two magnitudes of the same kind." (Book V, Definition 3) (Heath, 1956, p. 114) He augmented this definition by adding: "Magnitudes are said to *have a ratio* to one another which are capable, when multiplied, of exceeding one another." (Book V, Definition 4) (Heath, 1956, p. 114)

Euclid stated what it means for magnitudes to be in the same ratio in Book V, Definition 5:

> Magnitudes are said to *be in the same ratio*, the first to the second and the third to the fourth, when, if any equimultiples whatever be taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively taken in corresponding order. (Book V, Definition 5) (Heath, 1956, p. 114)

This rather antiquated explanation of what it means for magnitudes to be in the same ratio can be more fully understood by realising it is equivalent to stating that two ratios of magnitudes, $w:x$ and $y:z$, are the same if and only if, for all pairs of natural numbers m and n,

1. $n\,w > m\,x$ if and only if $n\,y > m\,z$

2. $n\,w = m\,x$ if and only if $n\,y = m\,z$

3. $n\,w < m\,x$ if and only if $n\,y < m\,z$

Therefore, any specific ratio of magnitudes is completely characterised by three classes of rational numbers, i.e., for any pair of magnitudes, w and x, of the same quantity, $w : x$ is characterised by the three classes:

1. $C_1 = \left\{ \dfrac{m}{n} : n\,w > m\,x \right\}$

2. $C_2 = \left\{ \dfrac{m}{n} : n\,w = m\,x \right\}$

3. $C_3 = \left\{ \dfrac{m}{n} : n\,w < m\,x \right\}$

If w and x are commensurable, the class C_2 is non-empty and any member of this class is equal to the ratio $w : x$. If, on the other hand, w and x are incommensurable, the class C_2 is empty and, in this case, $w : x$ is not characterised by a particular numerical ratio but, rather, by the classes C_1 and C_3 in the sense that it is the ratio that falls between these two classes.

Therefore, if a particular magnitude, x, is taken as a unit, any other magnitude of the same kind, w, can be characterised relative to x by one of the following two approaches:

1. If $\dfrac{w}{x} = \dfrac{m}{n}$ where m and n are natural numbers, then $\dfrac{m}{n}$ is the measure of w in units of x, or

2. If $\dfrac{w}{x} \neq \dfrac{m}{n}$ where m and n are natural numbers, then the measure of w in units of x is only approximated by a rational number, although the approximation can be to an arbitrary degree of precision.

Euclid's work in Book V of the *Elements* therefore provided a basic framework for understanding the concept of measurement, albeit in terms of rational approximations for incommensurable magnitudes.

Quantitative structure and measurement in Newtonian physics

In Newtonian physics, attributes such as length possess a quantitative structure and they are known as quantities. A specific value of a quantity is known as a magnitude of that quantity. Magnitudes of a quantity are taken to be measurable since, due to their inherent quantitative structure, they can sustain ratios to one another that are expressible as real numbers.

Although quantitative science has been in existence since at least the time of Euclid, the explicit characteristics of quantitative structure were only formulated at the end of the 19th century and the beginning of the 20th century. Hölder (1901) devised a set of seven axioms that define the concept of a continuous measurable quantity. Hölder stipulates that an attribute Q is a measurable quantity if and only if it satisfies the following seven conditions (Michell, 1999, pp. 52-53):

1. Any two magnitudes of Q are either identical or different and, if they are different, one is always greater than the other. This means that, given any two magnitudes, w and x, of Q, exactly one of the following is true:

   i. w is identical to x ($w = x$, $x = w$), or

   ii. w is greater than x and x is less than w ($w > x$, $x < w$), or

   iii. x is greater than w and w is less than x ($x > w$, $w < x$).

2. For every magnitude of Q there exists one that is less. In other words, for every magnitude, w, of Q, there exists a magnitude, x, of Q such that $x < w$.

3. For every pair of magnitudes of *Q* there exists another magnitude, their sum, which is well-defined. Thus, for every ordered pair of (not necessarily distinct) magnitudes, *w* and *x*, of *Q*, there exists a magnitude, *y*, of *Q* such that $w + x = y$.

4. Every sum of two magnitudes of *Q* is greater than each individual magnitude involved in the sum, so that for all magnitudes, *w* and *x*, of *Q* $w + x > w$ and $w + x > x$.

5. If one magnitude of *Q* is less than another, then there exists a third magnitude of *Q* which makes up the difference between them. Hence, for any magnitudes, *w* and *x* of *Q*, if $w < x$, then there exist magnitudes *a* and *b* of *Q* such that $w + a = x$ and $b + w = x$.

6. The sum of three magnitudes of *Q* is identical irrespective of whether it is the addition of the third to the sum of the first two, or the addition of the first to the sum of the last two. In other words, for all magnitudes, *w*, *x* and *y*, of *Q*, $(w + x) + y = w + (x + y)$.

7. Given any two non-empty classes of magnitudes of *Q*, an 'upper' and a 'lower' class, such that each magnitude of *Q* belongs to either class but not to both, and each magnitude of the upper class is greater than any of the lower, there must exist a magnitude of *Q* that is no greater than any in the upper class and no less than any in the lower class, i.e., there must be a least upper bound of the lower class. Hence, there must exist a magnitude *z* of *Q* such that every $z' < z$ belongs to the lower class and every $z' > z$ belongs to the upper class.

If the above seven conditions are satisfied by *Q*, then *Q* is a quantity and, as such, it can be measured. Quantities can be subdivided into extensive and intensive quantities. An extensive quantity is one whose additive nature (i.e., conditions 3 to 6 above) is self-evident from the behaviour of some objects that demonstrate magnitudes of the quantity. Length is an obvious example of an extensive quantity since its additive structure can be illustrated directly by using a set of rigid, straight rods. If two of the rods are combined end to end, the

concatenated length of the rods is equal to the sum of their individual lengths. An intensive quantity, such as density, is one whose additivity is not obviously demonstrable by concrete means.

Hölder's seven axioms render it possible to prove that every magnitude of a quantity is in fact measurable relative to an arbitrary magnitude as the unit of measurement. Consider any magnitude, w, of a quantity, Q, and define $2w = w + w$, $3w = (w + w) + w$, $4w = ((w + w) + w) + w$ and so on, so that, for a general natural number, n, $nw = (n-1)w + w$. Then, for a given pair of magnitudes, w and x, and a given pair of natural numbers, m and n,

1. $\dfrac{m}{n}$ is what Hölder terms a "lower fraction" in relation to the ratio $w : x$ if and only if

   $nw > mx$, and

2. $\dfrac{m}{n}$ is what Hölder terms an "upper fraction" in relation to the ratio $w : x$ if and only

   if $nw \leq mx$.

In other words, the measure (ratio) of magnitude w relative to magnitude x as the unit is located via an ordered sequence of positive rational numbers. At first sight this appears to be the same method as Euclid used for specifying when two ratios of magnitudes are equal. However, Hölder's approach goes much further since his method of locating ratios of magnitudes actually matches the ratios uniquely with positive real numbers. Dedekind (1872) observed that every positive real number is a least upper bound, or cut, of a non-empty class of rational numbers. Ratios of magnitudes correspond to classes of lower fractions and, according to Dedekind (1872), each class of lower fractions corresponds to its own least upper bound (cut). Accordingly, therefore, Hölder proved that every ratio of magnitudes is associated with a positive real number and, in this way, he showed that the measure of one magnitude relative to another magnitude as unit corresponds to a unique positive real number.

Therefore, for any magnitudes, *w* and *x*, of the same quantity, the magnitude of *w* relative to *x* may always be expressed by a positive real number, *r*, where $w = rx$. In other words, the ratio of *w* to *x* is a positive real number, *r*, which gives the measure of *w* in units of *x*.

Only continuous quantities, i.e., those that satisfy Hölder's axioms, are measurable in the traditional sense of the term. The process of identifying if a particular attribute is measurable is straightforward for extensive physical quantities, such as length and weight. However, for intensive physical quantities such as density and for psychological attributes, the process for identifying measurability is much more complex.

The measurability of psychological attributes

There is a widespread view within psychology that some psychological attributes can be subjected to the process of measurement (Michell, 1990, 1997, 1999). The 14th century French scholar Nicole Oresme considered the possibility of measuring psychological variables (Michell, 1990, p. 7), and there was a renewed focus on the measurability of psychological phenomena after the scientific revolution. For example, in 1725, Francis Hutcheson published his mathematical theory of the psychological basis of moral behaviour and judgment, which was discussed by Brooks and Aalto (1981). Hutcheson posited that the "moment of good" of an individual (a measure of the positive impact on the public of the individual's actions) is related to his benevolence and abilities thus:

$$M = B \times A$$

where M = moment of good of individual

B = measure of individual's benevolence

A = measure of individual's abilities. (Brooks and Aalto, 1981, pp. 347-348)

In a similar manner, Hutcheson related the "moment of evil" of an individual (a measure of the amount of evil produced by the individual) to the strength of his malice and abilities:

$$\mu = H \times A$$

where μ = moment of evil of individual

H = measure of individual's malice

A = measure of individual's abilities. (Brooks and Aalto, 1981, p. 348)

Fechner, however, was the first to propose actual methods for measuring psychological attributes, which is why he is regarded as the father of modern quantitative psychology.

Michell (1999) argues that there are five main reasons for psychologists adopting the stance that certain psychological attributes are measurable:

1. The inclination of psychologists to model their discipline on quantitative natural science, and physics in particular.

   After the scientific revolution, quantitative physics was considered to be the gold standard of scientific success because of its capacity to accurately model and predict the behaviour of physical systems. Emerging sciences such as psychology were therefore modelled upon physics because of its obvious success. The German psychologist, Ebbinghaus, commented on this tendency: "The brilliant results produced in natural science by measurement and calculation readily suggested the idea that something similar might be done for psychology." (Ebbinghaus, 1908, p.13)

2. An inherent belief that the attainment of precision and exactness could only be obtained through measurement.

   Some prominent psychologists, such as Cattell, saw the exactness of measurement as a reason for incorporating it into their discipline: "Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement" (Cattell, 1890, p. 373).

3. Pythagoreanism, the metaphysical view attributed to Pythagoras, that all things are composed of numbers.

This doctrine, which persisted for many centuries as a central tenet of European thought, can be taken to mean that all attributes are essentially quantitative but their quantitative nature is sometimes veiled by human perception. If Pythagoreanism is accepted, it follows that psychological attributes are quantitative and, theoretically, measurable.

4. The quantitative imperative: the view that measurement is vital in a discipline if it is to be considered scientific.

   The quantitative imperative originated from the belief that all science is quantitative and, as an aspiring science, psychology must therefore be quantitative. This stance was supported by the English psychometrician Francis Galton, who claimed that "until the phenomena of any branch of knowledge have been submitted to measurement and number, it cannot assume the status and dignity of a science" (Galton, 1879, p. 149).

5. The need to market psychology as a quantitative discipline due to the widespread acceptance of Pythagoreanism and the quantitative imperative in the 19th century scientific community.

   Michell (1999) opines that, to effectively promote psychology as a science in the 19th century, it was necessary to market it as a quantitative discipline.

The Ferguson committee

In 1932, a committee of nineteen scientists was established by the British Association for the Advancement of Science to investigate the validity of psychological measurement practices. The Ferguson Committee, as it became known, was chaired by the physicist A. Ferguson, and consisted of a number of psychologists and others from outside the ranks of the psychology profession including the physicist N. R. Campbell.

The contribution of J. Guild was a cornerstone of the Ferguson Committee's deliberations. Guild (1938) set the scene by giving an account of Campbell's theory of fundamental and derived measurement. Guild indicated that fundamental measurement involves the numerical representation of an empirically determined analogue of numerical addition, while derived measurement entails the discovery of constants that are functionally related to fundamental measures through numerical laws. Guild proceeded to argue that, in experimental psychophysics, psychologists did not establish an analogue of numerical addition for sensory intensities and, consequently, measurement of sensation intensities did not constitute fundamental measurement. Guild also stipulated that, since there were no fundamental measurements implicated in psychophysical measurement, neither could there be any derived measurements. Guild (1938, p. 328) therefore concluded that psychophysical measurement did not actually exist, and he applied his arguments to both Fechner's sensation intensities and to a later psychophysical concept known as sense-distances: "We must conclude therefore that sensation intensity is not measurable ... It is not measurable in any sense of the term". Guild's arguments in relation to Fechner's work were accepted by the Ferguson Committee, but there was greater resistance to his criticisms in relation to sense-distances.

Both the interim and final reports of the Ferguson Committee, published in 1938 and 1940 respectively, were somewhat equivocal in their conclusions. Campbell and his supporters clearly won the debate, but there remained uncertainty about the validity of the measurement of sense-distances. However, despite the uncertainties pertaining to some aspects of psychophysical measurement, psychologists were left in no doubt overall that their measurement practices were somewhat dubious. For example, in the final report of the Ferguson Committee, Guild commented:

To insist on calling these other processes measurement adds nothing to their actual significance but merely debases the coinage of verbal intercourse. Measurement is not a term with some mysterious inherent meaning, part of which may have been overlooked by physicists and may be in course of discovery by psychologists. It is merely a word conventionally employed to denote certain ideas. To use it to denote other ideas does not broaden its meaning but destroys it: we cease to know what is to be understood by the term when we encounter it; our pockets have been picked of a useful coin ...

(Ferguson et al., 1940, p. 345)

The Ferguson Committee had considered the Sone scale for the measurement of perceived loudness, which was devised by S. S. Stevens and, therefore, Stevens took a special interest in the work of the Committee. Stevens responded to the dilemma that psychology was faced with by proposing, in 1946, a new definition of measurement in psychology.

Stevens' definition of measurement in psychology

Stevens (1946, p. 677) redefined measurement as "the assignment of numerals to objects or events according to rules". This definition is now widely accepted within psychology. For example, Michell (1997) indicates that, in a survey of psychology books published between the early 1950s and the early 1990s, he found that, of 44 books which included a definition of measurement, 39 of those gave a definition either identical to, or similar to, Stevens' 1946 definition. Michell also confirmed that none of the 44 definitions he considered were remotely like the traditional scientific concept of measurement. Michell (1997, p. 360) opines: "These observations confirm that psychology, as a discipline, has its own definition of measurement, a definition quite unlike the traditional concept used in the physical sciences".

According to the traditional view of measurement, when an attribute is measured there is an attempt to determine ratios between magnitudes of the attribute. Therefore, measurement in the traditional sense is only possible for attributes which possess a quantitative structure, i.e., those which satisfy Hölder's axioms. However, according to Stevens' definition, measurement of non-quantitative attributes would be distinctly possible because numerical assignments could be made to the attributes using an arbitrary rule. Ellis (1966, p. 39) commented upon the unsatisfactory nature of Stevens' definition of measurement:

> [I]t is doubtful whether this definition of measurement [Stevens' definition] is really satisfactory. There is no doubt that measurement always involves the assignment of numerals to things according to rule, but if no restrictions are placed on the nature of the rule, it seems to admit far too much.

Stevens (1946) stressed that numerals can be assigned to objects or events using different types of rules, and he proceeded to develop his notorious theory of the four possible types of measurement scales: nominal, ordinal, interval and ratio. Stevens (1946, p. 677) stipulated that: "The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale."

In Stevens' opinion, measurement is possible since there is an isomorphism between empirical relations pertaining to the attributes of objects and properties of numerical structures: "The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world." (Stevens, 1946, p. 677)

The Ferguson Committee was critical of psychophysical measurement since it was not predicated upon the demonstration of an additive relation between sensory intensities. Stevens' introduction of his four types of measurement scales made the Committee's demands look unnecessarily restrictive. The result was that Stevens' definition came to be accepted as the authoritative definition of measurement by the psychological community. Stevens had therefore successfully fended off the criticisms levelled at measurement practices in psychology and, simultaneously, legitimised his own psychophysical research methods through his new definition of measurement.

The case for a quantum measurement paradigm in psychology

The certain knowledge of classical Newtonian physics emerged as the favoured paradigm for psychology in the late 19th century. In particular, the pioneers of experimental psychology believed that by modelling the discipline on Newtonian mechanics it would be accepted, in Clark Hull's words, as "a fullblown natural science" (Hull, C.L., 1943, p. 273). The laws of Newtonian mechanics are deterministic and they can be used to predict, with certainty, the subsequent motion of a macroscopic object if the relevant data are known for the object.

Gigerenzer (1987) questions why experimental psychologists failed to adopt quantum theory as their new paradigm after it emerged in the early part of the 20th century. Such a move would have been a natural progression since psychologists were fascinated by classical Newtonian physics in the 19th century. The laws of Newtonian physics fail to provide a model that can accurately predict the behaviour of microentities such as subatomic particles and photons. However, quantum theory provides a framework that is capable of accurately accounting for the dynamic attributes of such microentities. Quantum theory accounts for a realm that is inherently uncertain; there is no objective reality in the quantum world in the sense that microentities do not have dynamic attributes until they are measured. In stark

contrast to classical physics, the measurement process actually influences the measured values of the attributes. Furthermore, the microentity and the measuring device form a unified and inseparable system such that the measured attribute is a joint property of both.

The quantum pioneer Niels Bohr referred to structural parallels between the study of psychological attributes and the study of quantum entities, although he never developed his ideas. For example, Bohr (1998, p. 90) expressed the "hope that the epistemological attitude which had led to the clarification of the much simpler physical problems [of atomic physics] could prove itself helpful also in the discussion of psychological questions". Bohr believed that quantum physics and psychology share a common goal: to use ordinary language to communicate unambiguously about what transcends direct experience. The constructs of interest to quantum physicists, such as electrons, only manifest themselves in macroscopic measuring instruments that can be read by the human eye. Analogously, the constructs of interest to psychologists, such as the ability of a child, are not visible to the human eye but, rather, must be inferred from what the child writes in a test, for example.

Gigerenzer indicates that although quantum theory was considered in the 1940s and 1950s within psychology, it was "unequivocally rejected as a new ideal of science" (Gigerenzer, 1987, p. 11). Gigerenzer (1987) argues that quantum theory was rejected as a paradigm for psychology because it appeared to contravene two facets of psychology's quest for certain knowledge: determinism and objectivity. Psychologists realised that if they adopted quantum theory as their dominant paradigm, the determinism of Newtonian physics would be replaced by indeterminism. Gigerenzer (1987) argues that psychologists refused to allow objective probabilities to enter their thinking at a fundamental theoretical level as in quantum mechanics. Uncertainty in the quantum realm is irreducible in the sense that it does not arise because of ignorance on the part of the observer. "Classical ignorance" may occur in Newtonian physics when some of the parameters relating to the motion of a macroscopic

object are unknown despite the fact they do actually exist, thereby culminating in the need to invoke subjective probabilities to describe the outcomes of the motion. Consider, for instance, the situation when a fair die is tossed. If the precise initial position and orientation of the die, its velocity of projection, the coefficient of restitution between the die and the surface upon which it lands, etc. were known, it would be possible to calculate with certainty the outcome of the experiment. However, in practice these variables, whilst they exist in reality, will be unknown to the casual observer, meaning that the experiment does not have a deterministic outcome but, rather, statistical predictions of the possible outcomes of the experiment in terms of subjective probabilities must suffice.

Psychologists were quite prepared to apply probabilities in the following contexts:

1. To explain the "classical ignorance" inherent in traditional psychological measurement models that are predicated on the Newtonian paradigm of classical physics, and

2. To test hypotheses concerning psychological attributes, e.g., using analysis of variance.

Psychologists also found it difficult to accept that objectivity would be undermined by the fact that, in quantum theory, the measurer and the measured interact to influence the outcomes of measurement. As Gigerenzer (1987, p. 12) puts it, "knowledge is not about reality; it is about reality *and* the knower". Bohr (1934/1987, p. 119) similarly opines "we are both onlookers and actors in the great drama of existence", which implies that quantum mechanics is a "participatory" discipline.

Gigerenzer (1987) posits that, since the early days of experimental psychology, measurements of psychological attributes such as intelligence, perceived loudness, etc. have been considered to be independent of the actual instruments used to measure them. This assumption of independence is exemplified by the approach taken to the measurement of

20

perceived loudness by the psychophysicist Stevens. In the Newtonian tradition, Stevens attempted to measure perceived loudness as a thing-in-itself and he "considered the measurement instrument [to be] ... of no theoretical relevance" (Gigerenzer, 1987, p. 17). He used two different methods to measure perceived loudness:

1. Magnitude estimation, which entailed judging the ratio of the loudness of two tones, and

2. Category rating, which entailed judging the interval.

Stevens used both methods for all subjects involved in the study, but he was shocked to discover that the two different measurement methods repeatedly produced inconsistent results. If measurements of psychological attributes were independent of the measuring instrument used, the results produced by the two methods should have been linearly related. Rather than accept the flawed assumption upon which his research was predicated, Stevens favoured the measurements produced by the magnitude estimation method and simply ignored the values produced by the category rating approach. If Stevens had not taken this approach, he would need to have accepted that the measurement methods used were theoretically relevant rather than just data capture strategies. As Gigerenzer (1987, p. 17) observes, "subjective values, subjective strategies, etc., would be considered as processes that were elicited by or dependent upon certain tasks rather than independent of them".

Stevens' work demonstrates that a psychological attribute such as perceived loudness cannot be described as a thing-in-itself but, rather, it depends upon the method used to measure it. As in the quantum realm, it is only meaningful to refer to a measurement with respect to a particular measuring instrument. In Stevens' work on the measurement of perceived loudness, the problem pertaining to the conflicting values generated by the two different methods could have been resolved by including a reference to the method used to measure a particular value in any measurement report. However, such an approach would

have undermined psychologists' quest for objectivity in their discipline. It would thus appear that quantum theory provides a better paradigm for psychological measurement than classical Newtonian mechanics since, in quantum theory, the entity measured and the measuring instrument form an indivisible whole.

A quantum measurement paradigm for psychological predicates

The essential characteristics of a quantum measurement paradigm for psychological attributes will be described with reference to the measurement of cognitive abilities. In his later philosophical writings, Ludwig Wittgenstein provided an extensive analysis of the nature of intentional psychological predicates such as learning, understanding, thinking, remembering, and so on. The current paper uses facets of Wittgenstein's later philosophy to argue that a quantum theoretical model is actually more appropriate than a Newtonian model for the measurement of cognitive abilities.

Bruner (1996, p. 129) defines learning as following rules to "go beyond the information given", and it is therefore appropriate to consider the philosophical foundations of rule-following. Wittgenstein (2009) argues that the source of an individual's ability to follow a rule cannot be a finite object in the individual's mind, such as a formula or an image. According to Wittgenstein, a rule by itself gives rise to the paradox that, under some interpretation of its requirements, any answer can be brought into accord or into conflict with the rule:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be brought into accord with the rule. The answer was: if every course of action can be brought into accord with the rule, then it can also be brought into conflict with it. And so there would be neither accord nor conflict here.
>
> (Wittgenstein, 2009, §201)

For example, if a child is asked to evaluate the expression $2x$ when $x = 3$, they could attach the orthodox interpretation to the instruction and proffer the answer "6". Alternatively, they could, for example, attach an unconventional interpretation, e.g. that evaluating the expression $2x$ entails concatenating 2 and the numerical value of *x*, and thus give the response "23".

Perhaps the inability of a rule by itself to guide a child in its use could be resolved by positing that, in addition to the rule, the child must be able to attach the correct interpretation to the rule. Alas, a rule plus an interpretation stands logically at the same level as the rule by itself, and so this simply leads to an infinite regress:

> If it [the rule] requires interpretation, that could be done in lots of ways. So how do I tell which interpretation is correct? Does that, for instance, call for a further rule – a rule for determining the correct interpretation of the original – and if so, why does it not raise the same difficulty again, thereby generating a regress? (Wright, 2001, p. 163)

A further possibility is that the simplest interpretation of the rule is privileged. However, reasoning based on Gödel's (1931) incompleteness theorem and Chaitin's (2007) Algorithmic Information Theory undermines this potential route out of paradox: "You can *never* be sure that a computer program is what I like to call *elegant*, namely that it's the most concise one that produces the output that it produces. *Never ever*!" (Chaitin, 2007, p. 121) It is tempting to propose that the difficulties associated with interpretations could be avoided if there is a Platonic mechanism in the child's mind, that requires no interpretation, but which gives the child access to all future uses of a rule. Wittgenstein was vehemently opposed to this possibility, as exemplified by his outright rejection of mathematical Platonism: "The mathematician is an inventor, not a discoverer" (Wittgenstein, 1978, I, §168).

Wittgenstein's extensive analysis of rule-following leads to the conclusion that, prior to a child offering an answer to a rule-governed question (such as the elementary algebraic

substitution problem considered above), there are no criteria for determining if the relevant rule has been followed correctly. The child is both correct and incorrect prior to saying or writing their actual answer: they are in a superposition of two states simultaneously. This is analogous to the situation that occurs in quantum theory when, for example, the position of a microentity, such as an electron, is measured. Prior to measurement, the electron is in a superposition of different states, corresponding to the possible outcomes of the measurement process but, when a measurement is made, this superposition collapses to give one actual measurement result.

Wittgenstein argues that it is not possible to follow a rule in one's mind and that a well-established custom or practice, into which one must be trained, is the ultimate arbiter between correct and incorrect applications of the rule:

> '[F]ollowing a rule' is a practice. And to *think* one is following a rule is not to follow a rule. And that's why it's not possible to follow a rule 'privately'; otherwise, thinking one was following a rule would be the same thing as following it. (Wittgenstein, 2009, §202)

In the algebraic substitution example considered previously, it is when the child offers the answer "6" or the answer "23" that the criteria associated with the practice of evaluating algebraic expressions are invoked to adjudge their response to be either correct or incorrect. At the instant the child gives their answer to the question, there is a transition from being in a superposition of two states (correct *and* incorrect) to being in a single state (correct *or* incorrect). This is similar to notion of "wave-function collapse" in quantum theory where, for example, the probability wave-function, which incorporates information on all possible measurement outcomes and their associated probabilities, collapses to yield a single value when the position of an electron is measured.

In quantum theory, microentities do not possess their dynamic attributes (such as position and velocity) intrinsically but, rather, the act of measurement influences the attributes. A microentity and the device used to measure one of its dynamic attributes form a non-separable system, such that the measured attribute is a joint property of both the microentity and the measuring device. An analogous situation occurs when the ability of a child to respond correctly to a mathematical problem, such as the algebraic substitution example considered above, is measured. According to Wittgenstein (2009), the child's ability does not exist as a thing-in-itself but, rather, it is only meaningful to refer to the ability relative to the measuring instrument: the practice of evaluating algebraic expressions. The child's ability to solve the problem is non-separable from the relevant mathematical practice.

In contemporary psychological measurement models, the probabilities that an individual associates with another person's intentional predicates (such as learning) are perceived to be subjective, because the individual does not have direct access to the private mental states of the other person. Those aligned with the Cartesian conception of the mind posit that there would be no uncertainty, and therefore no need to resort to probabilities, if it were possible for the individual to have direct access to the mental states of the other person. To put it differently, if God were to look into the mind of the other person, there would be no uncertainty; the uncertainty would yield to certainty. Wittgenstein (2009) contends that, when a person expresses a thought, for example, they are not describing an inner state with which the expression can be checked for accuracy. According to Wittgenstein, mental states cannot be construed as mental objects that are analogous to objects in the physical world. Therefore, the uncertainty pertaining to the mental predicates of another person cannot be reduced by inspecting their mental states, since those states do not exist as things-in-themselves which bear comparison with what the person subsequently says, writes or does. The uncertainty is not a consequence of ignorance.

The distinctive feature of the inner seems to be that it has to be guessed at from the outer of the other person and is known only from within. But when through accurate consideration this conception vanishes into thin air, the inner indeed has not become the outer, but for us there is no longer direct inner evidence and indirect outer evidence for the inner. (von Wright, 1982, p. 33)

While uncertainty in measuring physical attributes of objects arises from instrument fallibility and human limitations, the uncertainty associated with measuring psychological predicates is constitutive, and not a shortcoming of any sort. Uncertainty in predicting the weather reflects shortcomings in the instrumentation used, but the vagueness of psychological predicates is not a shortcoming or deficiency. Wittgenstein explains the constitutive nature of the uncertainty in psychological measurement using the notion of "thermometer pain":

One could imagine that to determine whether someone is in pain a kind of clinical thermometer is used. If a human being cries or moans, they take his temperature and only if this shows such and such a sign, they start to pity the one who suffers and to treat him the way we treat the one who is 'clearly in pain'. (von Wright, 1982, p. 50)

Clearly Wittgenstein is using this bizarre notion to make a point. The thermometer is designed to reduce uncertainty; if someone is feigning pain, the thermometer reading will find them out. However, the same predicate is not being measured with greater accuracy using the thermometer; rather an entirely different predicate from our everyday conception of pain, pain*, say, is being measured. Therefore, the uncertainty in psychological measurement is irreducible and probabilities associated with measurement values are objective rather than subjective. This resonates with the irreducible nature of uncertainty in quantum theoretical measurements.

The implications of this transition from a Newtonian to a quantum theoretical measurement paradigm for one particular type of psychological measurement, educational measurement, are considered in the following section.

Implications for educational measurement: PISA

Psychological measurement models are used extensively to underpin research and policy formulation in education. For example, the Rasch (1960) model is the basis of the OECD's PISA, which aims to evaluate the efficacy of the education systems of OECD member nations. PISA entails a tri-annual assessment of the skills of 15 year-olds in reading, mathematics, science, problem solving and financial literacy. National governments recognise the importance of monitoring and evaluating their education systems to facilitate public accountability and, ultimately, to improve the quality of education. International comparisons such as PISA have become a key part of this process. Policymakers are significantly influenced by such comparative studies, and a direct relationship between a country's educational achievement and its economic potential is frequently posited: "Students who demonstrate high achievement levels are more likely to be productive workers and members of society when they leave the education system." (OECD, 1996, p. 193)

The use of international league tables such as those generated by PISA has the potential to, for example, influence foreign direct investment in a nation and it is therefore not surprising that PISA results have the capacity to throw a country's schooling regime into turmoil. This is exemplified by the widely publicised German "PISA shock" which occurred in the aftermath of Germany achieving comparatively low scores in PISA 2000. Heated debate ensued regarding necessary changes to the country's education system: "The PISA results and their reception in the German context led to appeals for a reform of secondary education from almost all relevant social groups, including political parties, employers, trade

unions (including teachers' associations), parents' associations and academics." (Ertl, 2006, p. 621) Similar controversy occurred when the Japanese government used the PISA 2003 results to legitimise contentious education policy decisions (Takayama, 2008).

Since educational testing is a powerful activity that has the potential to impact considerably on society, it is evident that it must be predicated upon a secure measurement paradigm. There have been numerous critiques of PISA (see, for example, Bonnet, 2002; Prais, 2003; Goldstein, 2004; Grisay and Monseur, 2007; Dohn, 2007), which raise doubts about underlying theoretical and methodological issues in the assessment paradigm. The current paper has questioned the mathematical and philosophical foundations of psychological measurement models such as that utilised in PISA, and the implications for PISA of a transition to a quantum theoretical approach to psychological measurement are now elucidated.

Item response theory attempts to relate the level of a psychological construct, such as ability, possessed by an individual to their performance on the discrete items of a test designed to measure the construct. According to Raykov and Marcoulides (2011, p. 247), the dominant assumption of item response theory is that "the responses on items of a test under consideration (and consequently overall test performance) can be accounted for by one or more latent abilities or constructs".

Psychologists/educationalists posit that an individual possesses an intrinsic ability, θ, which is the source of the individual's performance on a given test item. Since the intrinsic ability cannot be measured directly, item response models have been developed as a mechanism for estimating the intrinsic ability level. These are mathematical models relating the probability of an individual responding correctly to a test item to the individual's ability level and properties of the item. Raykov and Marcoulides (2011) consider a number of different item response models, such as the one-parameter logistic model:

$$P_g(\theta) = \frac{e^{\theta - b_g}}{1 + e^{\theta - b_g}}.$$

In this model, which is equivalent to the Rasch (1960) model that underpins PISA, the following notation is used:

$P_g(\theta)$ = probability of individual giving a correct response to item g

θ = ability level of individual

e = base of the natural logarithm function

b_g = difficulty level of item g, i.e. value of θ for which $P_g(\theta) = 0.5$.

Mathematical techniques are used in conjunction with item response models to estimate the ability level, θ, of an individual on the basis of their responses to the items on a test. In so doing, it is assumed that the ability level, θ, of the individual is a thing-in-itself which can be abstracted away from the measuring instrument.

If a Newtonian model of psychological/educational measurement is rejected in favour of a quantum-theoretical paradigm, as suggested in the current paper, it is meaningless to refer to the ability of an individual as a thing-in-itself that is the source of the individual's responses to the items on a test. Rather, it would only be meaningful to refer to the individual's ability with respect to each discrete test item at the instant when he/she actually responds to the item. It is evident, therefore, that a quantum theoretical basis for psychological measurement gives rise to serious conceptual problems for item response theory. Since PISA utilises the Rasch (1960) model as its theoretical basis, the proposed transition from a Newtonian to a quantum theoretical measurement model would have critical implications for the validity and reliability of PISA data and thus render the PISA study to be a meaningless exercise. A measurement project such as PISA is simply untenable under a quantum theoretical framework.

Conclusion

Michell (1999) specifies two conditions, which he attributes to Helmholtz, for empirically testing if an attribute is a continuous quantity, i.e., if it is measurable in the Newtonian sense:

1. There must be a method for comparing objects which permits the determination of whether or not two objects are the same with respect to the attribute, and

2. The attribute must possess an additive structure in the sense that there must be a method for combining objects in a way that demonstrates additivity of the attribute concerned.

   According to Michell (1999, p. 71), additivity of an attribute is demonstrated if there is an actual physical process for combining magnitudes of the attribute so that:

   i. $w + x = x + w$ for any two magnitudes, w and x, of the attribute, and

   ii. $(w + x) + y = w + (x + y)$ for any three magnitudes w, x and y of the attribute, and

   iii. The combined magnitude is unchanged if equivalent objects are substituted – objects which have the same magnitudes as the individual objects that are being combined.

Michell (1999, p. 74) stresses that, as an alternative to the above direct tests for quantitative structure, there are indirect methods for testing if an attribute is a continuous quantity, such as conjoint measurement theory (Luce and Tukey, 1964). Michell (1997, p. 358) cautions, however, that "the practice of measurement requires getting some grip, either directly or indirectly, upon the additive structure of the attribute in order that ratios between magnitudes of the attribute may be discovered or estimated".

In Michell's view, the psychological community failed to demonstrate that psychological attributes actually possess an additive structure, and simply imposed their own definition of measurement through their journals and textbooks. Michell's thesis is that

psychologists simply pronounced their discipline to be a quantitative one and subsequently stifled debate on the issue. Michell (1997, p. 374) portrays psychological measurement as a "methodological thought disorder" and he expresses serious concerns about the scientific basis of current measurement practices in psychology. Michell's extensive work on the foundations of psychological measurement demonstrates that the measurement of psychological predicates and, by extension educational predicates, does not conform to a Newtonian paradigm. However, I contend that Michell is making the mistake of holding psychology to an inappropriate standard. Michell is wrong to assume that measurement in psychology is a mechanism for *checking up* on the values of pre-existing mental attributes. This type of measurement is the preserve of Newtonian physics.

Michell failed to consider the possibility that a quantum theoretical rather than a Newtonian paradigm may actually offer a more secure basis for psychological measurement. The current paper has initiated the debate concerning this lacuna in Michell's work. However, further research at a fundamental philosophical level is necessary to more fully investigate the structural parallels between quantum measurement in physics and the measurement of psychological predicates. In particular, given the significance attached to high stakes assessments such as PISA by policymakers, the author believes that a re-appraisal of the philosophical foundations of psychological/educational measurement is necessary forthwith.

References

Bohr, N. (1934/1987). *The philosophical writings of Niels Bohr volume 1: Atomic theory and the description of nature.* Woodridge, CT: Ox Bow Press.

Bohr, N. (1998). *The philosophical writings of Niels Bohr volume 4: Causality and complementarity* (Supplementary papers edited by J. Faye & H.J. Folse). Woodridge, CT: Ox Bow Press.

Bonnet, G. (2002). Reflections in a critical eye: on the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9(3)*, 387-399.

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

Brooks, G.P. and Aalto, S.K. (1981). The rise and fall of moral algebra: Frances Hutcheson and the mathematization of psychology. *Journal of the History of the Behavioural Sciences*, *17*, 343-356.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Cattell, J.McK. (1890). Mental tests and measurements. *Mind*, *15*, 373-380.

Chaitin, G.J. (2007). *Thinking about Gödel and Turing*. Hackensack, NJ: World Scientific.

Dedekind, R. (1872). Continuity and irrational numbers. In *Essays on the theory of numbers*. English translation by W. W. Beman (1901). Chicago, IL: Open Court.

Dohn, N.B. (2007). Knowledge and skills for PISA – assessing the assessment. *Journal of Philosophy of Education*, *41(1)*, 1-16.

Ebbinghaus, H. (1908). *Psychology: An elementary textboo*k. Boston, MA: Heath. (English translation by M. Meyer.)

Ellis, B. (1966). *Basic concepts of measurement.* New York, NY: Cambridge University Press.

Ertl, H. (2006). Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32(5)*, 619-634.

Fechner, G.T. (1860). *Elemente der psychophysik*. Leipzig: Breitkopf & Hartel. (English translation by H.E. Adler, *Elements of psychophysics*, vol. 1, D. H. Howes & E.G. Boring (Eds.). New York, NY: Holt, Rinehart and Winston.)

Ferguson, A., Meyers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, N.R., Craik, K.J.W., Drever, J., Guild, J., Houstoun, R.A., Irwin, J.O., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby, J.H., Smith, T., Thouless, R.H. and Tucker, W.S. (1940). Quantitative estimates of sensory events. *The Advancement of Science: Report of the British Association for the Advancement of Science*, *2*, 331-349.

Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology*, *2(2)*, 149-162.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M.S. Morgan (Eds.), *The probabilistic revolution – Volume 2: Ideas in the sciences* (pp. 11-33). Cambridge, MA: The Massachusetts Institute of Technology Press.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, *38*, 173-198.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11(3)*, 319-330.

Grisay, A. and Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33(1)*, 69-86.

Guild, J. (1938). Are sensation intensities measurable? *British Association for the Advancement of Science*, *108*, 296-328.

Heath, T.L. (1956). *Euclid: The thirteen books of the elements*, *vol. 2 (books III-IX)*. Mineola, NY: Dover Publications.

Heidelberger, M. (2004). *Nature from within: Gustav Theodor Fechner and his psychophysical worldview*. Pittsburgh, PA: University of Pittsburgh Press.

Heisenberg, W. (1958/2000). *Physics and philosophy.* London, England: Penguin Classics.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaftern zu Leipzig, Mathematisch-Physische Klasse, 53*, 1-46. (English translation in Michell and Ernst (1996, 1997).)

Hull, C.L. (1943). The problem of intervening variables in molar behavior theory. *Psychological Review*, *50(3)*, 273-291.

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

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Mahwah, NJ: Lawrence 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, England: Cambridge University Press.

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

Michell, J. (2003a). Epistemology of measurement: the relevance of its history for quantification in the social sciences. *Social Science Information*, *42*, 515-534.

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

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspectives*, *6*, 7-24.

Michell, J. (2011). Qualitative research meets the ghost of Pythagoras. *Theory & Psychology*, *21*, 241-259.

Michell, J. (2012). "The constantly recurring argument": Inferring quantity from order. *Theory & Psychology*, *22*, 255-271.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W.G. Chase (Ed.), *Visual information processing* (pp. 283-308). New York, NY: Academic Press.

OECD (1996). *Education at a Glance*. Paris, France: OECD.

Prais, S.J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education*, *29(2)*, 139-163.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Raykov, T. and Marcoulides, G.A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Spearman, C. (1937). *Psychology down the ages*, *vol. 1*. London, England: Macmillan.

Stein, H. (1990). Eudoxos and Dedekind: on the ancient Greek theory of ratios and its relation to modern mathematics. *Synthese*, *84*, 163-211.

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

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education*, *44(4)*, 387–407.

von Wright, G.H. (1982). *Wittgenstein* (catalogue of unpublished material). Oxford, England: Blackwell.

Wheeler, J.A. (1996). *At home in the universe*. Woodbury, NY: American Institute of Physics Press.

Wittgenstein, L. (1978). *Remarks on the foundations of mathematics* (Edited by G.H. von

    Wright, R. Rhees & G.E.M. Anscombe; translated by G.E.M. Anscombe; 3rd edition).

    Oxford, England: Blackwell.

Wittgenstein, L. (2009). *Philosophical investigations: The German text with an English*

    translation (Edited by P.M.S. Hacker & J. Schulte; translated by G.E.M. Anscombe,

    P.M.S. Hacker & J. Schulte; 4th edition). Chichester, England: Wiley-Blackwell.

Wright, C. (2001). *Rails to infinity*. Cambridge, MA: Harvard University Press.