

The human sciences, models and metrological mythology

Joshua A. McGrane - Oxford University Centre for Educational Assessment (OUCEA),
Department of Education, University of Oxford

Andrew Maul - Gevirtz Graduate School of Education, University of California, Santa Barbara

Abstract

Measurement concepts and vocabulary have become commonplace in the human sciences. To this end, the field of educational measurement has largely concerned itself with the development and use of a family of mathematical, “psychometric” models collectively known as latent variable models. These models contain parameters that are commonly interpreted as referring to properties of individuals and test items, and estimates of these parameters are accordingly interpreted as measured values of these properties. Such interpretations are based on the erroneous assumption that the correspondence between these parameters and properties has been substantiated, whereas the educational measurement literature and discourse frequently conflates or confuses the two; different instantiations of a pervasive representational fallacy. This fallacy drives an unscientific approach to modelling where the nature of educational phenomena and their causal role in measurement is ignored. To advance beyond the mythology of “latent variables”, the human sciences need to eradicate this fallacy and establish scientific models for measurement.

The human sciences, models and metrological mythology

“The tendency has always been strong to believe that whatever receives a name must be an entity or being, having an independent existence of its own...” (Mill, 1878, p.5)

“It is no longer a question of imitation, nor duplication, nor even parody. It is a question of substituting the signs of the real for the real...” (Baudrillard, 1994, p.2)

1. Introduction

In a time when the detection of gravitational waves and the discovery of new fundamental particles make headline news, and when the instrumentation involved in such discoveries seems closer to science fiction than everyday life, it is easy to lose sight of the fact that measurement is foundationally a human, epistemic and representational¹ process (Mari, 2005). In the physical sciences, this representational epistemology has developed from antiquity to the modern day, including the ongoing refinement of physical theory, concepts, laws, models, units and instrumentation, which serve to substantiate the scientific and commercial esteem granted to measurement processes and the objective and standardized information they provide (Mari, 2003; Maul, Torres Irribarra, Mari, & Wilson, 2018). More recently, measurement concepts and vocabulary have also become commonplace in the human sciences, including in educational systems and research, for purposes including assessment, evaluation, and accountability. Part of this adoption has been motivated by a desire among psychologists, educational researchers, and other human scientists to have their activities perceived as being legitimately scientific (Michell, 2000). In education, there may be the additional desire to imbue decisions regarding educational attainment and placement and evaluation of the quality of teachers, schools, and

¹ The term “representational” is intended here in the general semiotic sense, as distinct from the more specific sense of the representational theory of measurement.

educational programs with a sense of the same objectivity and standardization that is attained through physical measurement.

To this end, the field of educational measurement has largely been concerned with the development and refinement of methods to summarise students' performances on tests and responses to survey questions using a family of mathematical, "psychometric" models collectively known as *latent variable models*². This is an encompassing category that includes item response models (including the Rasch model), which will be the focus of this article, as well as factor, latent class and latent profile models as special cases (see, e.g., Borsboom, 2008). These models contain parameters that are commonly interpreted as referring to properties of individuals (e.g., their "abilities") and test items (e.g., their "difficulties"), and therefore parameter estimates are interpreted as measured values³ of these properties. However, despite the impression that may be given by the many volumes of educational measurement books and journals, the appropriateness of these interpretations is far from a settled matter.

Specifically, repeated concerns have been raised that psychometricians may have in fact assumed some of the mathematical and statistical trappings of physical measurement⁴, as formalised in their latent variable models, without sufficiently considering the substantive nature of the phenomena under investigation and their relation to the parameters of these models (e.g., Borsboom, 2006; Briggs, 2013; Guttman, 1991; Humphry, 2013; Kyngdon, 2008, 2015; Maraun, 1996; Maul, 2017; Michell, 2000, 2017)⁵. In the case of educational measurement,

² Latent variable models are bound by the common assumption of a latent (indirectly observed) structure which is probabilistically related to observed variables. In essence, they are statistical models that relate this hypothesised latent structure to expectations of observable outcomes by some regression function (Borsboom, 2008).

³ The phrase "measured values" is intended here in the sense given by the International Vocabulary of Metrology (JCGM, 2012) as a "quantity value representing a measurement result".

⁴ In particular, that measurement in the human sciences by way of numerical assignments is analogous to quantitative measurement in the physical sciences, including the numerical representation of quantitative properties, thereby justifying the further quantitative operations carried out on the assigned numbers.

⁵ We will not repeat the prior arguments presented in these various exemplar references, but rather, stress that despite the differences in the details of these writers' arguments, there is a common thread of criticism regarding the lack of substantiation of psychometric models and modelling.

which typically involves the assessment of students' knowledge, skills and/or abilities via standardized tests, the phenomena under investigation would include the various cognitive, metacognitive, affective and motivational properties of students that are researched throughout the educational and psychological literature, as well as the related content knowledge and skills described in educational systems' curricula. If measurement-oriented work in the human sciences were following the lead of physical metrology, one might expect to see primacy given in such work to developing and refining theory, concepts, laws, models, units and instrumentation with primary consideration of these phenomena. However, standard practices in the human sciences, and educational measurement in particular, often instead gives primacy to the formalisms of latent variable (and, in particular, item response) models, which are applied to item and test scores, as *prescriptions* for the phenomena under investigation (Maraun, 2007; Sijtsma & Emons, 2013). For example, item response models prescribe that test behaviour is necessarily a stochastic process involving a quantitative property of students (e.g., reading comprehension ability).

Through this prescription of a parallelism between the properties of the models and the properties of educational phenomena, this standard practice begs the question of whether the models *accurately* represent the relevant educational phenomena. Moreover, they invert the conventional scientific modelling approach where models are developed to *approximate* real phenomena for different purposes, including measurement (Morrison & Morgan, 1999). This inversion is alarming, as when psychometricians claim to measure properties of students such as educational achievement and growth, educational researchers and practitioners (as well as the general public) may quite reasonably interpret the putative measurement results as objective representations of these phenomena within and across students, analogous to, for example, measured values of the students' heights. However, as per the criticisms previously referenced, such interpretations go well beyond the scientific status of latent variable models. In fact, Michell (e.g., 2009, 2017) and Maraun (e.g., 1996, 2007) argue that such interpretations

are based on mythology underwritten by the concealed, fallacious belief that latent variable models are tools for discovering unobserved causal entities, and more specifically quantitative properties of individuals.

In this article, we extend their critiques to argue that this mythology is underwritten by an epistemic fallacy where properties of latent variable models are routinely confused or conflated with properties of the phenomena they purport to represent, and models are given epistemic primacy instead of the properties purported to be measured. In the first of the following sections, we overview the logic of representation and exemplify how this logic has been distorted in epistemic activities, both in general terms and specifically in the field of educational measurement. In the next section we overview the concept of modelling, including different types of models, to elucidate how this representational fallacy perpetuates an unscientific approach to the use of latent variable models by obfuscating basic ontological questions. Finally, we suggest a path forward for users of these models, starting with closer adherence to the logic of representation and closer attention to the theoretical and empirical substantiation of the relations between the different components of a scientific model of measurement.

2. The logic of representation

For millennia, philosophers have been fascinated by the study of the relationships between aspects of reality and our representations thereof (now generally referred to as semiotics), including words, numbers, and models. Much of their work has focused upon the psychological and/or linguistic aspects of semiotics, but the seminal work of Charles Sanders Peirce (1913; see also Ogden, Richards, Malinowski, & Crookshank, 1923) explicated the logical⁶ underpinnings of representation, and argued that representations are necessarily composed of

⁶ Logic is intended here in the classical philosophical sense built upon the three 'laws of thought', rather than alternative paraconsistent or intuitionist logical systems.

three distinct but interrelated parts: a sign, an object, and an interpretant⁷. Under this three-term (i.e., ternary) relationship, the sign plays the symbolising role in the representation, the object is the state of affairs being represented, and the interpretant is an individual's (or any non-human cognizer's) concept or understanding regarding the sign-object relation (see Figure 1). This relationship is logically irreducible because if any one of these three terms is removed, there is no representation.

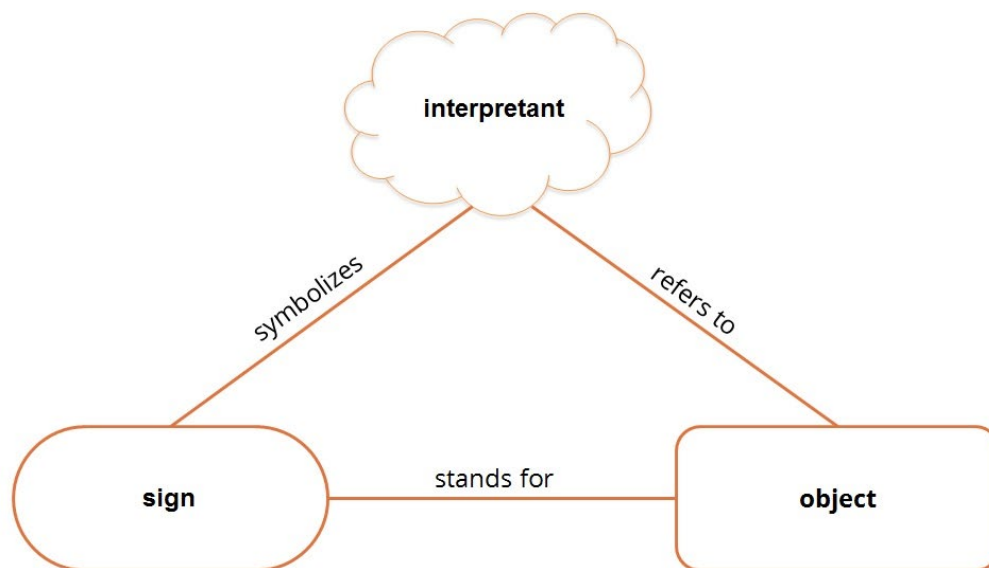


Figure 1. Representation as a logically irreducible, three-term relation.

This irreducibility is a feature of all kinds of signification differentiated by Peirce, including symbols (where the object-sign connection is purely conventional), indexes (where there is a causal connection between the sign and the object), and icons (where the connection between the sign and object is based on a mutual resemblance, imitation or isomorphism) (Mari, 2005). For example, the word “rabbit” (i.e., a symbol) bears no meaningful relationship to a type of actual four-legged, long-eared, hopping mammal in the absence of an understanding (i.e.,

⁷ This ‘semiotic triangle’ is a simplification of Peirce’s (1913) more complete semiotic theory, in which many further distinctions are made between different classes of objects, signs and interpretants.

interpretant) of this conventional word-animal (i.e., sign-object) relation that is specified in the English language. Similarly, the sight of a rabbit limping cannot signify that the animal is in some way injured (i.e., an index) unless one has an understanding of rabbits' typical movements and how these are causally related to their biological soundness. Moreover, a two-dimensional diagram of a rabbit (i.e., an icon) may only signify an actual rabbit (or its features) if there is an understanding of their structural resemblance.

2.1. Representational fallacy

A representational fallacy occurs when any two terms of the representational relation (or properties of these terms) are confused or conflated with one another (Dyke, 2008). In the case of representations of real phenomena (e.g., a rabbit), a representational fallacy occurs when, for example, the signs standing for reality (e.g., the word-symbol "rabbit") are conflated with reality itself (e.g., an actual rabbit). While this error may strike the adult reader as an unlikely one to make, Piaget (1979) observed that fallacious reasoning of this kind occurs as a normal part of human cognitive development, as children in the preoperational stage of development (2 to 7 years of age) tend to conflate signs meant to stand for objects with the objects themselves—in this case, the names of objects are taken to be intrinsically located in the objects themselves; a phenomenon he termed *nominal realism*. While this phenomenon typically resolves itself with further cognitive and linguistic development, it does lay the basis for similar fallacies to play out in adults in contexts where more abstract and/or complex representational systems are employed (e.g., Rozin, Markwith & Ross, 1990).

On this point, a number of philosophers have noted this kind of fallacious reasoning in people's use of representational devices. For example, Mill's (1878, p.5) observation in the opening quote of this paper refers to the epistemic intuition that if a name exists, it *must* stand for some concrete entity or object (i.e., the fallacious inference that the presence of a sign and interpretant necessitates the presence of a real object or entity that is represented). Moreover,

he noted that people tend to treat general or abstract terms, e.g., “species” or “genus”, as names referring to objects rather than, as in the case of these examples, names referring to relations between objects (or classes of objects). Given this confusion and the absence of concrete referents for such terms, Mill argues that they are often mystically understood as standing for entities of some “hyperphysical realit[y]” (p.5). Similarly, Baudrillard (1994) argued that in many aspects of modern society, signs have come to precede and even be *substituted* for the aspects of reality they were originally intended to signify, e.g., money was originally intended as a concrete *symbol* of value, and now is itself used in many contexts to *define* value. Baudrillard used the term *simulacra* to refer to signs that masquerade as representations of some aspect of reality, which take on a “hyperreal” status.

Finally, of particular relevance to the current article, Dyke (2008) argued that contemporary work in metaphysics often commits a form of the representational fallacy when philosophers draw conclusions about the nature of reality solely from the examination of linguistic representations of reality. For example, with respect to the nature of time, the conclusion that time is tensed is drawn from the grammatical tenses used to describe time. Dyke terms this the “language-to-reality approach” and argues that it has permitted linguistic questions to masquerade as ontological questions in contemporary metaphysics. Moreover, Dyke points out that the truth or adequacy of a representation cannot be conflated with its meaningful use in discourse; for example, we can meaningfully discuss unicorns irrespective of their actual (lack of) existence. The primacy given to language in this approach is analogous to the primacy given to latent variable models in measurement in the human sciences—what may be termed a “model-to-reality” approach. Given the overlap, it is unsurprising that the representational fallacy is also present in the educational measurement literature, as well as the human sciences more generally.

3. The representational fallacy and educational measurement

Throughout the educational measurement literature, item response models and their mathematical properties (e.g., parameters) are routinely discussed as if they were either identical to or unproblematically mappable onto educational phenomena and their substantive properties. Moreover, the results of data modelling activities are interpreted, in and of themselves, as a basis for claiming successful measurement of students' knowledge, skills and abilities, irrespective of whether they have been validated or not. This model-to-reality approach to educational measurement (and the human sciences more generally) is highlighted by Maraun (1996, 2007) who argues that latent variable models should not actually be considered models, as users of these statistical tools do not specify the correspondence between the mathematical symbols of their formulae and real phenomena in advance, and so these symbols cannot be interpreted as representational in a scientific sense, a point we will return to in the following section. Maraun goes on to echo Mill's observations addressed in the previous section in arguing that it is often simply assumed that the mathematical latent variable *must* represent some phenomena.

This assumption is common in discourse throughout the educational measurement literature, particularly with respect to item response model parameters. In formal terms, these parameters, or 'latent variables', are real numbers (i.e., they may continuously vary between positive and negative infinity) and the label 'latent' simply means that the values for the parameters are stochastically estimated (i.e., with error) based on manifest scores (real or simulated) that take only integer values. However, as these models were created to summarise patterns of test scores, the parameters are also *assumed* to be 'latent' in the sense of representing ontological properties of test-takers and items (c.f. Borsboom, 2008) belonging to some unobservable, quantitative property-filled hyperreality (i.e., the 'latent realm'). Given this assumption, the parameters are often named, accordingly (and deceptively), as "person" (or more specifically, "ability", "proficiency", etc.) and "item" (or more specifically, "difficulty", "severity", etc.) parameters, and as they are specified as real numbers, their estimated values

are *assumed* to be measured values (and more specifically, quantitative representations) of these properties.

This ‘model-out’ assumption of a representational relation and the fallacious reasoning regarding ontological issues that often follows it is well exemplified in Rasch’s (1977) original formulation of his model of reading ability. Rasch explicitly states that, “in a concrete formulation of this problem I *imagined* - in good statistical tradition - the possibility that the reading ability of a student at each stage, and in each of the two...dimensions, could be *characterized in a quantitative way...by a positive real number* defined as regularly as the measurement of a length” (p.59, emphasis added). In this quote, he is candid that the model (including the quantitative specification of the parameters) was constructed “in an imaginary way” based on statistical theory of countable outcomes, and not as a model *of* the relevant properties of students, i.e., the psychological properties that contribute to reading performance, based on theorising about these properties. From this basis, he goes on to state regarding the test and person parameters in his model that it is “reasonable to speak of [the former] parameter *as the* ‘degree of difficulty of the test’ [and of the latter] *as the* ‘degree of disability of the student’” (p.62, emphasis added).

While Rasch may have intended such statements as convenient linguistic shorthands rather than literally (similar to how it is customary in physics to use the same symbol for a physical quantity and the parameter used to model that quantity), by introducing this conflation in a disciplinary context where the distinction between quantities and their parameters is not well acknowledged and their representational relationship is not well substantiated, such an explanation of the conception of his model still leaves the door open to the fallacious approach to modelling critiqued in this article. Moreover, while all models do involve imagination in some sense (for example, in the form of abstraction, idealization, and analogical reasoning about some domain in terms of another domain), it is ultimately the scientist’s responsibility to establish that such acts of imagination do not provide a distorted or simply inaccurate

representation of a phenomenon -- that is, to ensure that their models are “responsible to reality”, in the language of Hilary Putnam (e.g., 2000). Finally, taken literally, Rasch’s conflation of the parameters of a mathematical model with the properties of tests and individuals they purport to represent is not reasonable at all, as it is a conflation of the logically independent sign and object terms of the representational relation.

Because of this foundational blur between model parameters and actual properties of tests and students in the formulation of item response models, it is now commonplace in educational measurement conferences and literature to hear and read examples of psychometricians referring to ‘theta(s)’ (the Greek letter typically used to symbolise the person parameter in item response models), in place of (hypothesised) quantitative properties of individuals. This is reminiscent of the nominal realism phenomenon identified by Piaget in children at the pre-operational stage of development, as psychometricians speak as if model parameters are indistinguishable from properties of individuals, and by extension, the properties of the former (e.g., that values of theta are strictly ordered) are intrinsic to the latter, thus committing a logical error that might be termed *parameter realism*. This error then propagates out to the discourse around ‘parameter estimation’, whereby psychometricians often discuss ‘theta estimates’ as if they were unproblematically identical to estimates of the magnitudes (i.e., measured values) of psychological properties of individuals.

Most recently, the representational fallacy has been prominent in the educational measurement literature and conferences in debates around the appropriate representation of the psychological properties of students, particularly to justify quantitative interpretations of educational growth (Briggs, 2013, 2015; Briggs & Domingue, 2013; Castellano & Ho, 2015; Thissen, 2016; van der Linden, 2015). These debates have taken place in response to Michell’s (e.g., 2000, 2009) criticisms that a form of methodological thought disorder is pervasive

throughout psychometrics and the scientific disciplines that make use of it⁸. This thought disorder involves a two-level breakdown of critical inquiry whereby, firstly, psychometricians routinely *assume* the existence of psychological quantities (by way of the quantitative parameters in their models) without seriously testing that hypothesis, and secondly, the failure to test such a fundamental hypothesis is ignored by the mainstream of the discipline.

While these exchanges have valuably contributed to breaking down the second tier of Michell's criticism, they seem to a large extent to have missed the crux of the argument. Specifically, given the pervasive influence of model-to-reality thinking in psychometrics, these debates have often been framed in terms of whether the scales produced by item response models (i.e., linear representations of the person parameter estimates from a particular dataset) have 'interval properties'. In the vein of Dyke's (2008) criticism discussed previously, this is an example of a model question masquerading as an ontological question: when interpreted as a model question, it is trivially true that person parameter estimates from a Rasch model are on an interval scale, as this is a property of the model itself (Thissen, 2016). However, the actual ontological issues at stake are, firstly, whether the postulated psychological property exists (in the sense that the property is instantiated by the individuals purported to be modelled), and secondly, whether it is quantitative in nature (Maul, Torres Irribarra & Wilson, 2016; Michell, 2000, 2005). To truly grapple with these long outstanding ontological questions in educational measurement, psychometricians need to adopt a more scientific approach to modelling. This, in turn, requires some amount of clarity regarding what a model is, what it aims to accomplish, and how modelling activities fit into larger schemes of human activities.

⁸ Michell's claim that psychometrics suffers from a scientific pathology additionally suggests that the representational fallacies discussed here may be more deliberately (but not necessarily consciously) motivated, both at the level of individual scientists (cf. Gigerenzer & Garcia-Retamero, 2017) and communities of practice (cf. Boag, 2015). However, given the speculative nature of this thesis, we remain agnostic regarding potential motivations of these fallacies, as the required analysis to support such conjecture goes beyond the intended scope of this article.

4. On modelling

The term “model” is ubiquitous throughout the literatures on psychometrics and measurement in the human sciences more generally, but it is not always clear in what sense it is being used. The term has a wide range of uses in the general literature—the Stanford Encyclopaedia of Philosophy entry on the topic refers to no less than 22 senses of the term (Frigg & Hartmann, 2018). In the present paper, we limit ourselves to discussing three broad types of models that relate, in different ways, to the measurement process, which we term (a) *mathematical* models, which exist independently of data, (b) *data* models, here meaning both models of immediately observed data and models that involve generalization to unobserved data, and (c) *substantive* models, which are models of real phenomena, such as models of cognition or learning in a specific domain. A model of the measurement process, like many other kinds of scientific models, will necessarily have all of these components (Mari, Carbone, Giordani, & Petri, 2017).

4.1. *Mathematical models in the absence of data*

Equations (or systems of equations) are sometimes referred to as models. In the absence of data, calling equations “models” may strain the usual definition of the latter term, as it is unclear what if anything is being modelled -- in general, models are by definition models *of* something. Nevertheless, one finds reference to many such mathematical models throughout the literature in the human sciences, particularly in journals dedicated to psychometrics or applied statistics, though sometimes with the accompanying (possibly tacit) claim that such models could be used for more traditional modelling purposes when combined with appropriate forms of data. Examples include generalized linear or log-linear models, which are often cast as being appropriate for application across an array of contexts, independently of the specific substance (Freedman, 1985). Within psychometrics, the mathematical equations that comprise latent variable models (including the Rasch model) are sometimes (confusingly) also referred to as

measurement models, independently of how or even if they are applied to specific areas of cognitive theory or application.

4.2. Data models

The mathematical models described previously are often combined with data, either for the sole purpose of representing immediately observed (actual) data⁹ or additionally for making inferences to unobserved (possible) data. In the human sciences, statistical models are used for both descriptive purposes (e.g., a representation of the distribution of some set of observed data) and for inferential purposes (e.g., a representation of the distribution of some population of data estimated on the basis of a set of observed data, plus some set of rules governing how the inferential process should take place, including estimation procedures and assumptions regarding the population distribution). In such instances, the generalization to other possible datasets is made on the basis of statistical theory, rather than substantive theory. Figure 2 diagrammatically illustrates such a situation, highlighting three elements: 1) data model, which itself subsumes a mathematical model (e.g., the Rasch model) and rules for how variation in observed data is transmitted into model parameter estimates (e.g., maximum-likelihood estimation); 2) the observed data (e.g., item responses from the administration of a test to a group of individuals); and 3) the resulting parameter estimates. Additionally, statistical sampling theory may be invoked to make inferences regarding possible (i.e., unobserved) data, reflected in (for example) error bands for the parameter estimates.

⁹ The concept of ‘observed data’ is ambiguous, as is the concept of ‘observation’ more generally (e.g., Bogen, 2017). In psychometric applications the term “observed data” often refers to behaviors (e.g., item responses) that have been coded (e.g., in terms of correctness of the response, or degree of endorsement of a statement), such “observations” are thus already (and inevitably) theory-laden, and thus arguably already involve some sort of substantive model. We use the term “observed data” here simply to refer to data that are *treated* as unproblematically observed.

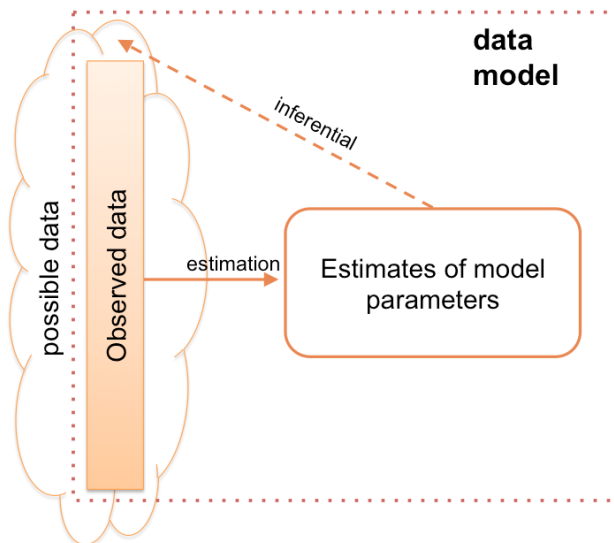


Figure 2. Schematic of a data model used for inferential purposes, showing the connections between the observed data, parameter estimates, and possible data.

Another prominent example of models used to describe immediately observed data comes from the work of scholars such as Krantz, Suppes, Luce, and Tversky in their seminal work on representational measurement theory (e.g., Krantz et al., 1971), in which empirical (i.e., observed) relationships between objects (e.g., “Barnaby is older than Beattie, who is older than Teddy”) are mapped onto numerical relationships (e.g., Barnaby is assigned a higher numeral than Beattie, who is assigned a higher numeral than Teddy), in such a way as to preserve a specified amount of information about the empirical relationship (in this case, ordinal but not quantitative information).

The representational theory and other mathematical theories of measurement are often collectively referred to simply as “measurement theory” (Tal, 2015). However, this is arguably a misnomer, as such theories address only purely formal aspects of measurement (Mari, 2003), starting with the presupposition that data have been unproblematically ‘observed’ (Borsboom, 2005, ch.4), and from there exploring the conditions under which relations among numbers can be used to express relations among objects. Such theories are silent about the actual act of

measuring, and in particular say nothing about the structural features of the measurement process that serve to secure its epistemic authority (Maul, Torres Iribarra, Mari, & Wilson, 2018)—or, as put by Kyburg (1984), “the theory of measurement is difficult enough without bringing in the theory of making measurements”.

4.3. *Substantive models*

Substantive models—for example, models of cognition or learning in a specific domain—comprise a third class of entities also referred to as “models” in the human sciences. For example, learning progressions, possibly represented in the form of construct maps (e.g., Wilson, 2005), are often used to represent a theory about how learning (typically and/or optimally) takes shape in a given area of study. Such models are specified at a range of levels of formality and of generality but share in common that they seek to represent some aspects of natural (including psychological, social, etc.) phenomena while ignoring others. In the human sciences, substantive models may or may not have mathematical content in and of themselves, but generally do not, outside of a few small sub-disciplines such as mathematical psychology. Substantive models also may or may not have explicit empirical content, and thus may or may not be testable.¹⁰

Figure 3 diagrammatically illustrates a particular kind of substantive model. Here, there is substantive theory specifying how (variation in) a real property of persons (e.g., reading comprehension ability) is connected to (variation in) possible data, which in the context of the human sciences, usually take the form of observable behaviours of some sort. In the more specific context of educational assessment, the substantive model may be used to specify the hypothesized way in which variation in the property causes variation in the specific subset of

¹⁰ The actual development of substantive models is, of course, often a long, messy, and iterative process, and outside the scope of our present commentary. In general, multiple independent forms of evidence and means of confirmation are involved in the development of substantive models,

possible behaviours from which the actual test items could plausibly be considered to have been sampled (sometimes termed a “universe of generalization” by, e.g., Cronbach, Gleser, Nanda & Rajaratna, 1972). In principle, this substantive model could include both (a) specification of the range of ways in which a general property (e.g., reading comprehension ability) could instantiate itself in a specific individual (e.g., reading comprehension ability of this student), which in turn helps establish the degree of *definitional uncertainty*, and (b) quantitative theory regarding the property and its causal and/or functional relations (e.g., to indications of a measuring instrument), and identification of influence properties that need to be experimentally or statistically controlled, which in turn helps establish the degree of *instrumental uncertainty* (see, e.g., Mari et al., 2017; Mari & Petri, 2017; Maul et al., 2018). In practice, while such aspects of a substantive theory are commonly encountered in physical metrology, they are generally absent in the human sciences (c.f., McGrane, 2015; Maul et al., 2018; Maul, Mari, & Wilson, 2019).

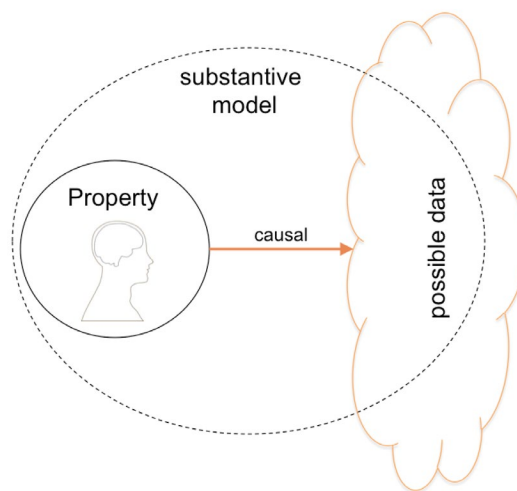


Figure 3. Schematic of a substantive model including the theory of the human property and its causal connection to hypothetical observations.

4.4. Combining the three types of models

Models used for scientific purposes frequently have theoretical, mathematical, and empirical components, and as such combine the three previously-described types of models. Indeed, most entities referred to as models throughout science aim to approximately represent natural phenomena or theories about natural phenomena, and from this perspective, as per Maraun's (2007) objection, it is especially jarring for exclusively mathematical entities such as equations to be referred to as "models" in and of themselves.

In the context of a proposed measuring instrument, the combination of the data model and the substantive model yields what might be termed a scientific model of the measurement process (see Figure 4). Compared to the data model in isolation, the most crucial addition to the scientific model is the explicit and antecedent specification of a real property (of a human being, in this case) thought to be causally responsible for the observed data. This broader epistemic framework logically determines the appropriateness of a specific data model.

Not all scientific models have explicit mathematical content. Those that do generally seek to represent natural phenomena in the form of equations, such as $F = MA$, which describes the relationship between the physical properties of force, mass, and acceleration in the form of a linear equation, and is testable given appropriate sources of data. In such cases, the mathematical content of the model is derived from substantive theory, and there is no obviously transferable mathematical content of the model, over and above what is supplied by the discipline of mathematics itself (e.g., in this case, the principles that govern how linear equations work).¹¹

¹¹ It does sometimes occur that the mathematical content of a model is used analogically in a domain other than the one for which the model was developed, as when some types of empirical relations are generalized from one phenomenon to another, such as the generalization of the inverse-square law from the model of gravity to electromagnetism. But this is ultimately an analogy between *phenomena* (e.g., using electromagnetism as a model for gravity) rather than between the models themselves.

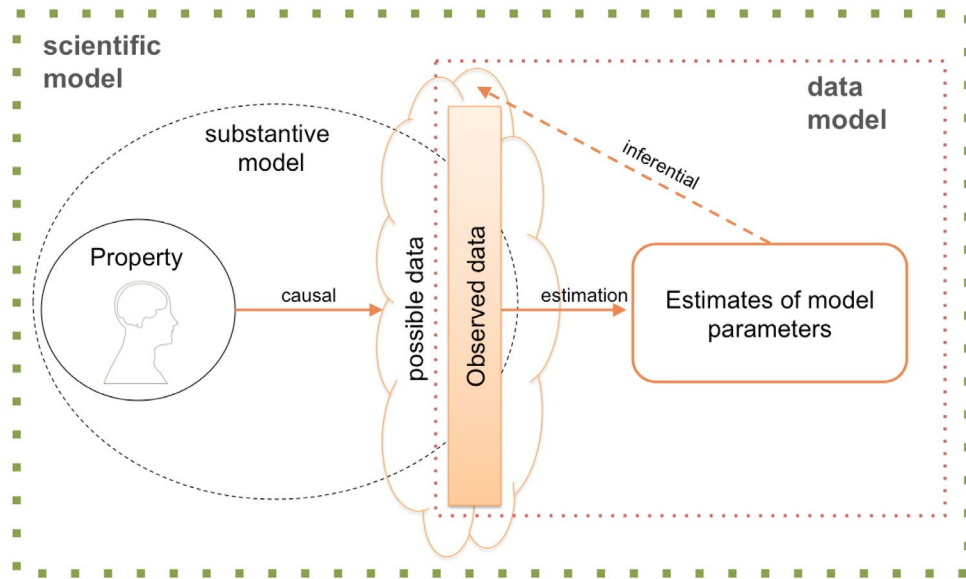


Figure 4. Schematic of a scientific model of the measurement process that combines both the substantive and data models.

By contrast, in educational measurement, and the human sciences more generally, the mathematical content of models of (for example) learning and cognition is usually supplied by “off-the-shelf” sources – mathematical models developed in isolation from that (or indeed possibly any) substantive context. Tellingly, even after tests have been developed and data have been collected, it is frequently the case that researchers and practitioners do not know which mathematical model to use, and often do not even have a particularly strong understanding of under what conditions it would be necessary or useful to employ a mathematical model at all (c.f., Borsboom, 2006). As a corollary, when mathematical models are employed to ‘validate’ instruments, they are often applied in *pro forma* fashion without consideration for whether and how the statistical criteria that often serve as indicators of success (e.g., unidimensionality, psychometric reliability, high correlations among response patterns on different test items, etc.) connect to hypotheses one might have about the actual phenomena under investigation (Maul, 2017). As a result, answers to questions about statistical

models and representational devices such as scales are often confused with answers to questions about real phenomena, and the anomalous direction of the relationship between mathematical models and theories in educational measurement (i.e., the model-to-reality approach) only makes it more difficult to notice such distinctions.

4.5. An example: quantitative properties and interval scales

As introduced in the previous section, claims regarding measurements being “on an interval scale” or scales “having interval properties” are made throughout the human sciences, and often such a claim is treated as interchangeable with the claim that a measured property is quantitative. However, this glosses over or buries the distinction between representational devices and the properties being represented: the terms “nominal”, “ordinal”, “interval”, and “ratio” apply (as originally developed by, e.g., Stevens 1946, and throughout the literature on representational measurement theory) to the former, whereas the terms “qualitative” and “quantitative” apply to the latter.¹² Examples of the uncritical mixing of these vocabularies are ubiquitous in the human sciences; for example, the term “variable” is often used to refer both to real properties and to the mathematical entities that aim to represent such properties (Markus, 2008), and, similarly, the term “construct” is used to refer both to “real but unobservable objects of study in psychological research” and “conceptual heuristics that function both to summarize potentially large classes of observables and foster ease of communication across members of a research community” (Slaney & Racine, 2013, p.4)

¹² Further confusing matters is the fact that the term “scale” is also used throughout the human sciences to refer to measuring *instruments* (e.g., questionnaires), as in the expression “a scale was developed to measure self-esteem...” This terminology is jarring from the perspective of the wider literature on measurement, which reserves the term “scale” for the (usually numerical) representational system on which measurement results are reported (e.g., the Celsius temperature scale). Calling instruments “scales” blurs the distinction between a method for gaining information about a property and the way this information is represented, an epistemic fallacy related to but distinct from the representational fallacy discussed in this paper.

Further confusing matters is the fact that in principle, one could make claims regarding scale types in terms of mathematical models or data models without reference to substantive theory or real properties of persons; from there, it may be a short leap to assume that what holds true for the model also holds true for the modelled property. Within the context of the mathematical model *qua* mathematical model, one could say, for example, that values of the person parameter (i.e., “thetas”) in the Rasch model are on an interval scale with respect to values of the item parameter (i.e., “deltas”) and vice versa (which, again, is true by construction). Within a data model, one could say that a scale constructed for a given dataset has interval properties in that it preserves the empirical relations between the data up to affine transformations. Alternatively, one could say that, based on demonstration of adequate fit of a given dataset to the Rasch model, the estimates of person and item parameters for that dataset are on interval scales with respect to one another. All of these claims could be true independently of whether the measured property really exists.

4.6. On the issue of model fit

It is common in psychometric practice to evaluate the extent to which the mathematical model fits the (observed) dataset. This is sometimes described as evaluating the plausibility of the hypotheses that the data were generated by a process described by the mathematical model. In practice, this fit information can be interpreted and utilized in several distinct ways. According to what might be described as an “item response theory paradigm” (see, e.g., Andrich, 2004), if misfit is detected, model constraints are removed to better accommodate the data, e.g., the constraint on the so-called discrimination parameter in a 2-parameter item response model (see, e.g., van der Linden, 2016) may be removed to achieve better model fit to the data. Conversely, in what might be described as the “Rasch paradigm” (Andrich, 2004), if model misfit is detected, the observed data are changed in some way to better accommodate the model. This normally occurs by deleting misfitting items, removing aberrant responses, or, rarely,

through the collection of new data with enhanced control over the source(s) of misfit. In general, according to this latter “paradigm”, the mathematical model is treated as immutable because it possesses desirable mathematical properties that are taken to be *prescriptions* for measurement.

However, neither of these strategies explicitly or necessarily involve the interpretation of misfit as a form of feedback regarding the substantive model (c.f., Michell, 2000; Heene, 2013). Rather, both strategies concentrate on the relationship between the data and the mathematical model, i.e., they concentrate on the right-hand side of Figure 4. Thus, it is possible to conduct such model-fitting exercises, even consistently with much of the literature on “best-practices”, without consideration of the connection to the substantive theory of the property—and indeed, even in the complete absence of any substantive theory or property (c.f., Wood, 1978; Maul, 2017). This, in turn, may reinforce the perception that formal features of models *are* features of reality—the representational fallacy.

If model-fitting activities are instead explicitly interpreted in the context of a broader scientific model, as suggested in Figure 4, the detection of model misfit may lead to several outcomes. First, as in the Rasch paradigm, the observed data may be altered if the scientist suspects that the causal transduction process between the human property and the data was perturbed in some way. This would necessitate the collection of new data to test the revised view of transduction with better control over such perturbations. Second, the detection of misfit may lead to an alteration of the theory of the human property within the epistemic framework of the model, e.g., one may change the assumption that the relevant property is quantitative. Such a change to the epistemic framework may, in turn, lead to changing the data model, e.g., if the person property is no longer theorized to be quantitative, then the Rasch model, given the quantitative formulation of its parameters, would no longer be an appropriate data model. In other words, the data model would be altered to better approximate the substantive phenomena its formal properties are purported to represent; a reality-to-model approach.

5. Summary and conclusions

The pervasive presence and influence of the representational fallacy in applications of measurement in the human sciences highlights the need to undertake the necessary substantive theoretical and empirical work required to establish the representational relations between the properties of (mathematical) models and the properties of humans.¹³ In particular, in many applications the ontological referents of latent variable model parameters remain largely unsubstantiated, and instead of addressing this scientifically dire situation, standard psychometric practice continues to fallaciously assume that these “latent variables” must represent some human property, or worse, encourages a fallacious discourse where the (often deceptive) names for these parameters are substituted in as simulacra for such properties. This then obscures the basic ontological assumption that these properties exist, and more specifically exist in the quantitative form assumed by most latent variable models.

It seems difficult to escape the conclusion that this standard practice is grounded in an unscientific approach to modelling where, rather than developing and refining models to more accurately represent human phenomena, the mathematical formalisms of latent variable models are given epistemic primacy and presented by psychometricians as unqualified prescriptions for measurement. This model-to-reality approach encourages confusion between model questions and ontological questions, including confusions between their distinct vocabularies, and directs measurement efforts toward “off-the-shelf” model-fitting activities. These activities may be

¹³ It is sometimes noted that there is a distinction between discovery and application, or between measurement in the service of basic research (as is often found in various fields of psychology) and measurement in the service of societal goals (as is often found in educational and commercial applications), each with distinct goals; it may be further argued that cases that fall into the latter category goals may not have goals related to scientific discovery or questions pertaining to truth. However, even in such cases the “adequacy and appropriateness” (Messick, 1989) of the instrument for its intended uses rests on the critical hypothesis that the measured property does exist (see, e.g., Maul & McGrane, 2017), and thus that the kind of work called for in this section remains necessary.

successfully carried out independently of any substantive model of the property purported to be measured, and so fail to motivate the substantiation of the model parameters.

To place measurement in the human sciences on a scientific foundation, modelling practices need to be inoculated against the representational fallacy. Firstly, users of latent variable models need to be aware of the logic of representation and be vigilant in their practice and discourse to not confuse or conflate the logically independent terms, and by extension, not fallaciously assume that the model parameters must represent some phenomena. Secondly, psychometricians need to understand that a scientific model of the measurement process necessitates more than a data model and should also provide a substantive model, including an adequate definition of the measured property and specification of how information is causally transduced from the property to the data. Such a model gives epistemic primacy to the actual phenomena of the human sciences, rather than mathematical formalisms, and establishes (instead of assumes) a representational correspondence between them. By eradicating the representational fallacy and giving primacy to its own substantive phenomena through scientific modelling practices, the human sciences may begin to advance beyond the metrological mythology of latent variable models to imbue their measurement practices with the same objectivity and standardization as physical metrology.

References

- Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42, 17-116.
- Baudrillard, J. (1994). *Simulacra and simulation*. Ann Arbor, MI: University of Michigan Press.
- Boag, S. (2015). Repression, defense, and the psychology of science. In S. Boag, S., L. Brakel, V. Talvitie (Eds.). *Philosophy, Science, and Psychoanalysis : A Critical Meeting*. London: Karnac Books.

- Bogen, J., "Theory and Observation in Science", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = [<https://plato.stanford.edu/archives/sum2017/entries/science-theory-observation/>](https://plato.stanford.edu/archives/sum2017/entries/science-theory-observation/).
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425-440.
- Borsboom, D. (2008). Latent Variable Theory. *Measurement: Interdisciplinary Research and Perspectives*, 6, 25-53.
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204-226.
- Briggs, D. C. (2015, 4). *Debate: Equal interval scales in educational testing: Attainable goal or myth?* Presentation at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Briggs D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, 38, 551–576.
- Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics*, 40, 35–68.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. New York: Wiley.
- Dyke, H. (2008). *Metaphysics and the representational fallacy*. New York, NY: Routledge.
- Freedman, D. A. (1985). Statistics and the scientific method. In *Cohort analysis in social research* (pp. 343-366). New York, NY: Springer.
- Frigg, Roman and Hartmann, Stephan, "Models in Science", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = [<https://plato.stanford.edu/archives/sum2018/entries/models-science/>](https://plato.stanford.edu/archives/sum2018/entries/models-science/).

- Gigerenzer, G., & Garcia-Retamero, R. (2017). Cassandra's regret: The psychology of not wanting to know. *Psychological Review*, 124(2), 179-196.
- Guttman, L. (1991). *'Louis Guttman: In Memoriam'*, Chapters from an unfinished textbook on Facet Theory. Jerusalem: The Israel Academy of Sciences and Humanities and the Hebrew University.
- Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, 4:246. doi: 10.3389/fpsyg.2013.00246.
- Humphry, S. M. (2013). A middle path between abandoning measurement and measurement theory. *Theory & Psychology*, 23(6), 770-785.
- JCGM (2012). International Vocabulary of Metrology (VIM) – Basic and General Concepts and Associated Terms (2008 edition with minor corrections), Joint Committee for Guides in Metrology, <http://www.bipm.org/en/publications/guides/vim.html>.
- Kyburg, H. E. (1984). *Theory and measurement*. Cambridge University Press.
- Kyngdon, A. (2008). Treating the Pathology of Psychometrics: An Example from the Comprehension of Continuous Prose Text. *Measurement: Interdisciplinary Research and Perspectives*, 6, 108-113.
- Kyngdon, A. (2013). Descriptive theories of behaviour may allow for the scientific measurement of psychological attributes. *Theory & Psychology*, 23(2), 227-250.
- McGrane, J. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, 6:431. doi: 10.3389/fpsyg.2015.00431.
- Maul, A., & McGrane, J. (2017). As pragmatic as theft over honest toil: Disentangling pragmatism from operationalism. *Measurement: Interdisciplinary Research and Perspectives*, 15(1), 2-4.
- Maraun, M. D. (1996). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research*, 31(4), 603-616.

Maraun, M. (2007). *Myths and confusions: Psychometrics and the latent variable model*.

<http://www.sfu.ca/~maraun/Mikes%20page-%20Myths%20and%20Confusions.html>

Mari, L. (2003). Epistemology of measurement. *Measurement*, 34, 17-30.

Mari, L. (2005). Principles of semiotics as related to measurement. In P. Sydenham & R. Thorn (Eds.), *Handbook of measuring system design* (pp. 134–139). New York, NY: John Wiley and Sons.

Mari, L., Carbone, P., Giordani, A., & Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Studies in History and Philosophy of Science*, 65, 46-56.

Markus, K. A. (2008). Constructs, concepts and the worlds of possibility: Connecting the measurement, manipulation, and meaning of variables. *Measurement: Interdisciplinary Research and Perspectives*, 6(1-2), 54-77.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2), 51-69.

Maul, A., Mari, L., & Wilson, M. (in press). Intersubjectivity of measurement across the sciences. *Measurement*. doi: 10.1016/j.measurement.2018.08.068

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, 79, 311-320.

Maul, A., Torres Irribarra, D., Mari, L., & Wilson, M. (2018). The quality of measurement results from a structural perspective. *Measurement*, 116, 611-620.

Messick, S. (1989). Validity. In R. L. Linn. (Ed.). *Educational measurement, 3rd edition*. (pp. 13-103). New York: American Council on Education/Macmillan.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, 10(5), 639-667.

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, 38(4), 285–294.

- Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp.111-133). Charlotte, NC: Information Age Publishing.
- Michell, J. (2017). On substandard substantive theory and axing axioms of measurement: A response to Humphry. *Theory & Psychology*, 27(3), 419-425.
- Mill, J. (1878). *Analysis of the phenomena of the human mind* (2nd Ed.). Volume 2. London, UK: Longmans, Green, Reader, & Dyer.
- Morrison, M., & Morgan, M. (1999). Models as mediating instruments. In M. Morgan & M. Morrison (Eds.), *Models as Mediators: Perspectives on Natural and Social Science* (Ideas in Context, pp. 10-37). Cambridge: Cambridge University Press.
- Ogden, C. K., Richards, I. A., Malinowski, B., & Crookshank, F. G. (1923). *The meaning of meaning*. London: Kegan Paul.
- Peirce, C. S. (1913/1998). *The essential Peirce: Selected philosophical writings* (Vol. 2). Bloomington, IN: Indiana University Press.
- Piaget, J. (1979). *La représentation du monde chez l'enfant* [The child's conception of the world] (J. Tomlinson & A. Tomlinson, Trans.), New Jersey, NJ: Littlefield, Adams & Co.
- Putnam, H. (2000). *The threefold cord: Mind, body and world*. New York City: Columbia University Press.
- Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14, 58-94.
- Rozin, P., Markwith, M., & Ross, B. (1990). The sympathetic magical law of similarity, nominal realism and neglect of negatives in response to negative labels. *Psychological Science*, 1(6), 383-384.
- Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, 20(7), 2-16.

- Sijtsma, K., & Emons, W. H. (2013). Separating models, ideas, and data to avoid a paradox: Rejoinder to Humphry. *Theory & Psychology*, 23(6), 786-796.
- Slaney, K.L. & Racine, T.P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology* 13, 4-12.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15-21.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677-680.
- Tal, E. (2015). "Measurement in Science", *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/measurement-science/>.
- Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, 41(1), 81-89.
- Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Wood, R. (1978). Fitting the Rasch model—A heady tale. *British Journal of Mathematical and Statistical Psychology*, 31, 27-32.
- van der Linden, W. J. (2015). *Debate: Equal interval scales in educational testing: Attainable goal or myth?* Presentation at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- van der Linden, W. J. (2016). Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume One* (pp. 41-58). New York, NY: Chapman and Hall/CRC.