Cartwright, Nancy, Jordi Cat, Lola Fleck, and Thomas E. Uebel. 1996. *Otto Neurath: Philosophy between Science and Politics*. New York: Cambridge University Press.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. New York: Oxford University Press.

Chang, Hasok, and Nancy Cartwright. 2008. "Measurement." In *The Routledge Companion to Philosophy of Science*, edited by Stathis Psillos and Martin Curd, 367–75. London and New York: Routledge.

Efstathiou, Sophia. 2012. "How Ordinary Race Concepts Get to Be Usable in Biomedical Science: An Account of Founded Race Concepts." *Philosophy of Science* 79: 701–13.

Feigl, Herbert. 1970. "The 'Orthodox' View of Theories: Remarks in Defense as Well as Critique." In *Analyses of Theories and Methods of Physics and Psychology*, edited by Michael Radner and Stephen Winokur, 3–16. Minneapolis: University of Minnesota Press.

Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. New York: Faber and Faber.

Hill, Craig, and Krishna Winfrey. 1996. *The 1996 Index of Hospital Quality*. Chicago: NORC at the University of Chicago.

Lemoine, Maël. 2013. "Defining Disease Beyond Conceptual Analysis: An Analysis of Conceptual Analysis in Philosophy of Medicine." *Theoretical Medicine and Bioethics* 34: 309–25.

Northrop, F. S. C. 1947. *The Logic of the Sciences and the Humanities*. New York: Meridian Books, Inc.

Stegenga, Jacob. 2015. "Measuring Effectiveness." *Studies in History and Philosophy of Biological and Biomedical Sciences* 54: 62–71.

Stevens, Stanley S. 1951. "Mathematics, Measurement, and Psychophysics." In *Handbook of Experimental Psychology*, edited by S. S. Stevens, 1–49. New York: Wiley.

Suppes, Patrick. 1998. "Theory of Measurement." In *The Routledge Encyclopedia of Philosophy*, edited by Edward Craig. London: Taylor and Francis. https://www.rep.routledge.com/articles/thematic/measurement-theory-of/v-1.

Weber, Max. 1949. "'Objectivity' in Social Science and Social Policy." In *The Methodology of Social Sciences*, translated and edited by E. A. Shils and H. A. Finch, 50–112. Glencoe, IL: Free Press.

Chapter 6

Psychological Measures, Risk, and Values

Leah M. McClimans

Person-centered measuring instruments, the kind currently in vogue for use in medical and public health research, are a form of psychological measurement. To understand better their possibilities and limitations it is useful to turn to the literature in psychometrics, the study of the theory and methods of psychological measurement. Leading trade journals for psychological measurement, such as *Psychometrika*, *Theory and Psychology*, and *Measurement*, publish philosophical debate, debates conducted mainly by psychometricians. One aspect of this debate centers on the ontology of psychological attributes and what measurement theories and methods befit them. Interestingly, most of the psychometricians participating in these debates agree in general that only a realist ontology is suitable for psychological measurement (Borsboom 2006; Maul 2013; Michell 2005). Their debate, in terms of ontology, is thus less directed toward one another than it is toward psychologists and others who continue to employ measuring instruments built out of theories and methods that cannot sustain a realist ontology (Borsboom 2006; Michell 1999).

In this chapter I discuss two perspectives on what is at stake in the choice between measuring instruments. Psychometricians who argue for realism tend to rely on epistemic values, such as rigor and truth, to make their case. On this view, the integrity of psychological measurement as a science is what is taken to be at stake in the choice of measure. Contrasting this viewpoint, I argue that nonepistemic ethical values, such as trust and harm, are equally important in measurement choice and particularly figure into the person-centered psychological measures often used in health care. I further argue that because these measures require nonepistemic considerations in their development and application, they challenge a strict insistence on realism in measurement. What is at stake is not the integrity of science so much as the usefulness of these measures as tools for medical and public health research.

ONTOLOGY AND PSYCHOLOGICAL MEASUREMENT

There is much debate in philosophical discussions of psychological measures (e.g., what kind of measurement methods should be used to produce psychological measures, the role of validity and whether psychology is even capable of genuine measurement). But standing alongside these debates is an emerging consensus among psychometricians that psychological measures require realism. Different proponents of this view make different arguments for it, but to provide a flavor of this position, consider two leading proponents, Joel Michell and Denny Borsboom.

In the same year that Michell (2005) published his article "The Logic of Measurement: A Realist Overview," Borsboom (2005) published his monograph *Measuring the Mind*. Both publications argue for realism in measurement and situate this position as an alternative to antirealism (e.g., mathematical theories of measurement, such as representational measurement theory [RMT], and operationalist theories of measurement, such as classical test theory [CTT]). Michell advocates for entity realism, which has at its core a correspondence theory of truth whereby a proposition is true if and only if circumstances are as the proposition states (Michell 2005). Thus, if the scale says I weigh 59 kg, then this is true only if I exist at a particular time and space, the attribute weight exists, and my weight is 59 kg. To say that weight exists is to say that it exists at a particular time and space as a feature of something else (e.g., a person, an object). Moreover, attributes such as weight exist as a range, for instance, I might weigh 59 kg now and 61 kg in a year.

Michell's realist ontology strives to align measurement with scientific discovery; indeed, he argues that only realism can do justice to scientific discovery (Michell 1997). For Michell, measurement fundamentally relies on the discovery of ontologically real quantities. Accordingly, metrologists, like other scientists, make hypotheses, gather evidence, and then use that evidence to draw conclusions about the likelihood of their hypotheses. In the case of measurement, the basic hypothesis should always be the same: attribute X is a quantity. This hypothesis refers to the internal structure of an attribute. For an attribute to be quantitative, it must be structured such that the values of the variable stand in certain algebraic relations to one another. Specifically, the relations must be ordered (e.g., transitive and additive), and conforming to the properties of addition (e.g., commutativity). Another way to put this point is that different instances of the same attribute must sustain ratio relations (Michell 2005). On Michell's realist perspective, these ratios are real numbers that represent different levels of the attribute. The job of metrologists is to gather evidence (e.g., concatenation or conjoint additivity) to determine whether this hypothesis is true and thus whether measurement of a specific attribute is possible.

Like Michell, Borsboom's realism concerns both the independent existence of attributes and a truth correspondence between the affairs postulated by a theory and those in reality. But Michell and Borsboom differ in that the importance of realism for Borsboom is not the discovery of a quantity. For Borsboom measureable attributes can be either quantities or qualities; what matters for measurement is whether an attribute is causally relevant. Specifically, he is concerned that there is a causal connection between an attribute represented by a latent variable and between-subject responses to items.[1] To this end, his argument for realism invokes latent variable theory.

Latent variable theory, or modern test theory, as it is sometimes called, is a measurement theory that uses mathematical models to hypothesize the attribute and parameters needed to explain the empirical data from psychological tests. As we saw above with Michell, measurement is here again taken to be an instance of scientific discovery. We use these models to test their adequacy against observed test data (i.e., the extent to which the observed data "fit" the predictions of those responses from the latent variable model, within acceptable uncertainty). If the model fits, then we have some evidence to suggest the observed data are a function of the model. But model fit is underdetermined (i.e., multiple models may fit the observed data within acceptable uncertainty [Borsboom 2005]). Thus, "fit," although necessary, is not sufficient to pick out the correct latent trait for any data set.

Borsboom's argument for realism is that some ontology is needed in addition to "fit" to motivate the choice of model in latent variable theory. In practice psychologists using latent variable theory choose reflective models to explain the empirical data from tests. Borsboom argues that reflective models presuppose realism. Reflective models specify that the pattern of covariation between observed item responses (e.g., answers to reading questions) can be explained by a regression on the latent variable (e.g., reading ability). The idea is that item responses vary as a function of the latent variable (i.e., differences in reading ability affect differences in item responses). Borsboom contrasts reflective models with formative models (Borsboom 2005). With formative models, which are common in sociological and economic modeling, the latent variable is regressed on the observed item responses. Put differently, responses to questions about, for instance, income and education affect the latent variable (e.g., socioeconomic status). In reflective models the latent variable (e.g., reading ability) is understood as determining our measurements, and in the formative model, the latent variable (e.g., socioeconomic status) is a summary of them (Borsboom 2005).

In principle Borsboom suggests there is no reason why psychologists should choose one model over the other; nonetheless they consistently choose reflective models. Why? His answer is that in this choice, psychologists reveal an ontological commitment to entity realism; the choice of a reflective

model presupposes an agent-independent latent variable that causally affects observed responses between subjects (i.e., at the population level). For the purposes of this chapter, let us accept the implications Borsboom draws from the choice of a reflective.

Michell and Borsboom differ in their accounts of measurement; yet they share a commitment to entity and theoretical realism as well as a commitment to scientific inquiry as orientated toward discovery. For Michell, measurement is the estimation of the ratio between two instances of a quantitative attribute, and because quantitative attributes must first be discovered before measurement can take place, this requires that the attributes are ontologically real. For Borsboom, measurement also requires an ontologically real attribute with a determinate structure, but that structure may be qualitative or quantitative. Additionally, Borsboom requires a measuring instrument that is sensitive to variations in the attribute and able to reflect those variations.

No matter what we think about the success of these arguments for realism and the accounts of measurement they underpin, I want to draw attention to the ontological position with which these arguments are meant to contrast: antirealism. Antirealist positions are taken in this debate to be represented by two distinct measurement theories: (1) representational measurement theory (RMT) and (2) classical test theory (CTT). Although RMT is more widely known within the philosophical literature, my focus is on CTT because this is the measurement theory most widely used in psychology. Nonetheless, let me briefly touch on RMT first.

Representational Measurement Theory

RMT seeks to map numerical relations onto qualitative empirical relations in such a way that the information in the empirical set is preserved in the numerical set. The creation of this homomorphism is a measurement scale. Consider for instance the empirical set of rigid rods and the relations between them (e.g., longer than). Representational measurement's goal is to specify numbers and mathematical relations (e.g., addition) that map onto the empirical structure of the rigid rods. Much of the literature of RMT concerns the identification of axioms that hold between objects, the development of representational theorems that specify when a homomorphism is possible, and uniqueness theorems that identify what kind of measurement scale a specific measurement procedure will produce.

RMT is concerned with determining the relations that must manifestin the empirical data to construct a measurement scale. As such, RMT is often criticized for its overly rational orientation and inability to engage with everyday issues in applied measurement (e.g., measurement error and uncertainty, calibration, reliability, and so on) (Borsboom 2005; Heilmann 2015; Tal 2013).

In the same vein, it is criticized for being overly abstract and thus not engaged with concrete scientific inquiry. As Borsboom (2005) notes, RMT does not hypothesize theoretical constructs or latent variables; the measurement scales created using RMT require only basic empirical relations, logic, and mathematics. It implies an antirealist ontology closely related to logical positivism by linking empirical (observed) relations to numerical (theoretical) relations via axioms and theorems, which for Borsboom resemble correspondence rules (Borsboom 2006; Michell 2005).

Classical Test Theory

Although RMT does not serve to inspire many contemporary psychological measuring instruments, CTT does. In fact, CTT is the dominant measurement paradigm within much of psychological measurement including clinical outcome assessments (COAs) (Borsboom 2006; Cano and Hobart 2011). CTT turns on a simple model where an observed score (O) (i.e., the empirical data acquired after respondents complete a test or questionnaire) is equal to a person's true score (T) plus uncertainty, commonly termed random error (E); thus $O = T+E$.

When using CTT the value of the true score is taken to be a theoretically unknown value, which is assumed to be constant, and the observed score is assumed to be a random variable, which produces a bell-shaped curve around the true score. The error score is taken to have an expectation value of zero. The idea here is that as the number of observations (i.e., administrations of the test or questionnaire) increases, the random errors will tend to cancel one another out; thus, the mean of the observations is taken as an estimate of the true score. To acquire an empirical value for T in the context of COAs, a person must be measured repeatedly on a scale (fill out the items of a questionnaire), and each observation (individual items or repeated administration of the same questionnaire) must be independent of the others (Hobart and Cano 2009).

The problem with these requirements is that in much of the behavioral sciences, they are not met. For instance, repeated administrations of a COA are not independent of one another. Respondents remember the questions from previous administrations and reevaluate their answers considering them. Moreover, COAs often do not function as a successive series; rather they function as measurements taken at a single point in time (e.g., to determine one's physical functioning three months post-operation). Third, the interpretation of the observed score as an estimate of the true score significantly depends on the assumption of a continuous variable (e.g., distance) with a normal probability distribution. But many of the variables in the context of COAs (e.g., physical functioning) are categorical, not continuous, as the

responses elicited from respondents to individual questions take a limited number of values (e.g., strongly agree, agree, disagree, strongly disagree).

These difficulties, as well as others, are well known (Borsboom 2006; Cano and Hobart 2011). Typically, a thought experiment is given to manage the first two: imagine the person being measured is brainwashed in between a series of administrations (Lord and Novick 2008). This thought experiment renders, by definition, a series of administrations, and the brainwashing renders those administrations independent of one another. The third difficulty is often dealt with by simply ignoring the categorical nature of the data elicited from individual questions and assuming that the variable approximates continuity given a large enough number of possible values that can be derived from combinations of responses to different questions.

One drawback of this thought experiment is that it renders CTT unfalsifiable (Borsboom 2005; Hobart et al. 2007). Borsboom (2005) goes so far as to call it a "tautology." This is because CTT is rooted in the theory of errors. Within this theory, the idea that random errors will cancel one another out in the long run (i.e., the error score will have an expectation value of zero) is an empirical assumption (Borsboom 2005). That an observed score estimates the true score is a hypothesis contingent on this empirical assumption. But in the context of psychological measurement, there are no empirical grounds for making these assumptions since (1) measurements are not taken in a series and (2) even if they were, they are not independent of one another. To put CTT into practice, the error score must be assumed to have a zero expectation value, and as a result the true score is simply defined in terms of the mean observed score. In practice, the model reduces to O = T.

The claim that CTT is a form of antirealism derives from this simplified model, and at least for Borsboom (2006), it evokes an antirealist version of operationalism.[2] Operationalism is best known through the work of Percy Bridgman (1927, 5), who wrote, "In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations." CTT is open to an operationalist interpretation since the observed score, which is the result of a particular instrument's "operations" essentially defines the meaning of the attribute in question through its equivalence with the true score. Indeed, we can see operationalism at play in the context of COAs where we find a proliferation of measures targeting quality of life or subjective health status.

Epistemic Risk and Psychological Measurement

CTT is widely criticized as a theory for psychological measurement, but in this section I want to focus specifically on the kind of criticisms that those who advocate for realism in measurement make to illustrate the kind of risk they understand CTT to proliferate and thus the values they take it to represent.

In his article "Attack of the Psychometricians" Borsboom (2006) analyzes (and laments) various factors that have stalled the integration of psychometrics with psychology. One might imagine that because psychological measurement is a large part of psychology, psychometrics should be an integral part of it. Yet the advances in modeling that psychometrics has achieved in the last century have not been taken up by psychologists who develop and use psychological measures. In this paper, Borsboom is interested in why this integration has not occurred. He cites the prevalence of CTT in psychology textbooks, operationalism as a popular theory in psychology, and a lack of interest and training in the math necessary to understand and use the developments in psychometric modeling. Of interest to this chapter is, why does Borsboom think these advances in psychometrics *should* be integrated into psychology? He provides two different explanations. First, an integration is necessary for the progress of psychology as a science. Second, current instruments are not fit for purpose, and because they directly affect individuals' lives, we have a social obligation to improve them.

Jeremy Hobart and colleagues (Hobart et al. 2007) elaborate on these explanations in their paper "Rating Scales as Outcomes Measures for Clinical Trials in Neurology: Problems, Solutions and Recommendations." They argue that using new psychometric methods (i.e., latent trait theories) improves measuring instruments' scientific rigor and thus the chances of coming to a correct conclusion about the effect of a disease and the efficacy of a treatment given clinical change. They examine two methodological limitations of CTT to make these points.

Ordinal vs. Interval Measurement Scales

CTT yields measurement scales at the ordinal level. It can be difficult to interpret the significance of (e.g., clinical change) from ordinal-level measurement scores. For instance, if a group scores 15 on the Beck Depression Inventory and then scores 19, their depression has gotten worse, but how much worse? Does this change indicate a need for clinical intervention? Or is the change nominal? If another group changes from a score of 41 to 45, their depression has also gotten worse, but is this increase comparable to the first group? Because ordinal scales lack a consistent unit, these questions are difficult to answer with much precision, and this lack of precision can be problematic for clinical and evidence-based purposes. Clinicians use measuring instruments at least in part to provide information regarding whether patients need to change treatment regimens, but to do so they need measurements that provide

clear information regarding magnitudes of change. Moreover, one popular use of measuring instruments is to determine efficacy, but this requires the ability to compare magnitudes of change (e.g., between different arms of a study) (see McClimans 2011).

Some of the new psychometric methods, such as those that use the Rasch family of models, claim to generate interval-level measurement. Interval-level measurement allows us to say that a change score of, for example, four, is of the same magnitude no matter where on the scale this change occurs (e.g., whether the change is from 15 to 19 or from 41 to 45). Thus, interval-level measurement allows us to compare change scores across the scale. It also allows us to use mathematical operations, such as the addition or subtraction of a constant, to the measurement values without changing the form of the scale. We can also use the arithmetic mean as a measure of average value. The practical result is that clinicians need not wonder whether a four-point change at one point on the scale represents a different magnitude than the same change at another point in the scale, and researchers can compare magnitudes of change across study populations.

No doubt interval-level scales have practical advantages over ordinal-level scales.[3] But those who advocate for interval-level scales tend to go beyond these practical concerns. One assumption that motivates much of the emphasis on interval-level scales is that ordinal scales are not *scientific measurements*. Measurement, in this view, requires a quantitative variable. This position echoes that of Michell, which I discussed above. It is not an uncommon position. In the paper by Hobart and colleagues (2007), the subheading that leads the section arguing for interval-level measurement is titled "The Requirement for Rating Scales to Generate Rigorous Measurements" and then later, "Ordered Scores [from CTT] Are Not Scientific Measurements."

This emphasis matters. If you see interval-level scales in psychological measurement as having mainly, practical benefits, then the choice of what instrument to use (i.e., instruments with interval-level or ordinal-level scales) will depend on context. But if you understand interval-level measurement as the only form of legitimate scientific measurement, then there is no decision to make: all measuring instruments should be interval-level scales. The risk one wages when using CTT is an epistemic risk to scientific credibility. Proponents of realism in measurement understand their commitment as a commitment to truth and see CTT as committed to, at best, expediency.

To be sure, Borsboom does not hold the same realist view in that he is not committed to interval-level scales. Instead, he sees the commitment to realism as a way of understanding the relation between an attribute and the scores generated by a measuring instrument. One furthers this understanding by choosing a measurement model that explains this relationship. In choosing a measurement model, one must specify the structure of the attribute and the

function that relates it to the measurement scores. This process furthers scientific progress at least in part because it provides an argument for the validity of a measuring instrument, and for Borsboom (Borsboom and Zand Scholten 2008), knowing what one is measuring is essential to measuring it.

Validity

Determining validity is the second methodological limitation of CTT that Hobart et al. (2007) discuss in their article. Construct validity, determining whether a measuring instrument measures what it aims to measure, is CTT's primary validation method. It is typically tested by assessing a measure's internal and external construct validity. Internal construct validity is tested by examining the extent to which the questions or items within a measurement scale are statistically related to one another based on the responses given by a sample population. The criticism is that this process does not tell us anything about the construct itself (e.g., quality of life); it tells us only that certain questions tend to behave similarly in the same conceptual space. External construct validity is examined via convergent and divergent validity testing. Multiple measurement scales deemed like and different from one another are applied to a sample population, and the scores derived from respondent answers are correlated. These correlations provide information about whether the scale being validated correlates higher with scales that measure similar constructs than with those measuring dissimilar constructs. Once again, this process does not tell us what construct a measure assesses. It tells us only that some scales are correlated (or not) with other scales.

Construct validity is accused of circularity, vacuity and meaninglessness (Borsboom, Mellenbergh, and van Heerden 2004; Hobart et al. 2007). Hobart and colleagues (2007) argue similarly to Borsboom et al. (2004) that the solution to construct validity is the development of models (Hobart and colleagues refer to them as "construct specification equations") that predict the variation of the data from the measuring instrument. Borsboom and colleagues (2004) add that validity is achieved if and only if (1) the attribute exists and (2) variations in the attribute causally produce variations in the data. For Hobart and colleagues, the idea that the attribute is real is presupposed given that they require interval-level measurement of quantitative variables.

The risk of using CTT in the context of validity is primarily a form of inductive risk. The worry is that claiming validity using tests of internal and external construct validation does not provide sufficient evidence that one is measuring what one intends to measure. Treating statistical correlations as though they provide evidence of an attribute is potentially a bad inference because it is possible that such correlations speak only to the questionnaire as an artifact (i.e., the inference is uncertain). What does the aversion to

taking correlations as evidence of an attribute tell us about the values realists prioritize?

At least in part, this aversion speaks to Michell's (2005) and Borsboom's (2005) interest in measurement as furthering scientific discovery. Construct validation as used by CTT purports to provide evidence that an instrument is measuring the attribute it is intended to measure. Insofar as the instrument is validated it also provides evidence of that attribute through its measurement outcomes. If the inference from correlations to an attribute is invalid, then the outcome information that the instrument provides is questionable. If we, nonetheless, take these outcomes as evidence of an attribute, we adulterate the integrity of science and stall progress. To be sure, Borsboom also argues that the use of CTT-based instruments violates our social obligation to the public. But this obligation rests on the assumption that measures that are fit for purpose must be anchored to real attributes. The use of CTT measures is a violation of our social obligation because they are scientifically inadequate, and thus the instruments are materially dishonest.

From a realist perspective, CTT could hardly be worse. On one hand, it purports to measure attributes, and on the other hand, its methods glide across the surface of reality mutually reinforcing the legitimacy of what is a constructed space. Because the attributes are never modeled, nor defined through a robust theory, it is unclear what CTT instruments measure. Yet volumes of articles are devoted to their respective validity. Similarly, because they make use of ordinal scales, it is often unclear how to interpret changes over time, and yet they are used in clinical practice and health policy as evidence of change.

NONEPISTEMIC VALUES AND PSYCHOLOGICAL MEASUREMENT

CTT measuring instruments are easy targets for critics, and I have also criticized them, particularly those CTT instruments used in health care. I (McClimans 2010, 2011) have argued like those above that these measures are invalid and difficult to interpret, and like Borsboom (Borsboom and Zand Scholten 2008), I have argued that these problems ultimately stem from a lack of theorizing about the target attribute.

But unlike the authors above, my concern with these instruments is not limited to a concern with scientific integrity and progress; I am not simply concerned with epistemic values of rigor and truth. Rather, I situate COAs as bearing a double burden of epistemic and ethical credibility. Moreover, because the ethical risk involved in COAs is intertwined with the epistemic choices of measurement methodology and measuring instrument, it is not

a matter of one interest taking priority over the other. Thus, we cannot say that a measuring instrument fulfills our social obligation simply because it embodies a certain degree of scientific rigor. Ethical risks must be considered and balanced alongside epistemic risks. Ethical considerations are not new to medical research, but they are relatively absent from discussions of measurement.

Although I think ethical risk is part of many aspects of epistemic practice in measurement, patient-reported forms of COAs present a particularly salient case for taking ethical risk and ethical values seriously. Patient-reported outcome measures (PROMs) measure attributes such as mobility, health status, and quality of life by asking patients questions (e.g., "Does your health now limit you in lifting or carrying groceries?"). PROMs are very popular with health policy makers both in the United States and abroad because they incorporate the patients' point of view into assessments of effectiveness, thus bringing together patient centeredness and clinical effectiveness in one instrument (Black 2013; Department of Health 2008; Speight and Barendse 2010; Washington and Lipstein 2011).

Given their dual purpose, PROMs embody certain ethical-epistemic imperatives. For instance, validity, ostensibly an epistemic concern of measurement, takes on a distinctive ethical dimension when we consider that PROMs are intended to assess the subjective experience of patients' health and well-being by asking them questions about it (Schwartz and Rapkin 2004). If our knowledge of the target attribute is underdeveloped or systematically excludes certain legitimate kinds of subjective experience, then it is not simply that we come to know less about health and well-being, but also we do not live up to what we owe the patients to whom we pose our questions. We may, for example, attribute to them a level of subjective experience they do not have, or we may attribute an experience that is not theirs while making the claim that this is what patients report (and giving ourselves credit for asking them). The nature of PROMs (i.e., that they speak on behalf of patients) means that more is at stake than scientific integrity when we develop and use them (McClimans et al. 2017).

To make my point, I provide two examples. Both examples take up again the concern regarding interpretability and scale development.

Classical Test Theory Reconsidered

Earlier I discussed how CTT measuring instruments are difficult to interpret because they yield ordinal-level scales. Not surprisingly, this difficulty has led to the development of methods to enhance their interpretability. One popular method is the identification of a minimal important difference (MID). A MID is the smallest change in respondent scores that represents clinical

significance and which would ceteris paribus warrant a change in a patient's care (Jaeschke, Singer, and Guyatt 1989). One popular method for determining a measure's MID is to map changes in respondent outcomes onto a control. These are referred to as "anchor-based" approaches. The idea is to determine the minimal amount of change that is noticeable to patients and to use this unit of change as the MID.

Here is how it works. A control group of patients are asked to rate the extent of their symptom change over the course of an illness or intervention on a transition-rating index (TRI). TRIs are standardized questionnaires that ask patients questions such as "Do you have more or less pain since your first radiotherapy treatment?" Typically, patients are given seven possible answers ranging from "no change" to "a great deal better" (Fayers and Machin 2015). Those who indicate minimal change (i.e., those who rate themselves as just "a little better" than before the intervention) become the patient control group. The mean-change score of this group is used as the MID for the PROM.

The approach of acquiring a MID via a patient control group assumes that respondents who rate their symptom change as "a little better" on a transition question should ceteris paribus also have comparable change scores from the PROM. Put differently, similarities in respondent answers to transition questions ought to underwrite similarities in respondents' magnitude of change over the course of an intervention or illness. But qualitative data from interviews with patients suggest that this assumption is ill founded (Taminiau-Bloem et al. 2011; Wyrwich and Tardino 2006). To take a concrete example, consider Cynthia Chauhan, a patient advocate during the deliberations on the FDA guidelines for the use of PROMs in labeling claims. In response to the deliberations, Chauhan cautioned those present "not to lose the whole person in your quest to give patient-reported outcomes free-standing autonomy" (Chauhan 2007). To make her point, she discussed the side effects of a drug called bimatoprost, which she uses to forestall blindness from glaucoma. One of the side effects of bimatoprost is to turn blue eyes brown. Chauhan has "sapphire blue" eyes, in which, she says, she has taken some pride. As she speaks of her decision to take the drug despite its consequences, she notes that doing so will affect her identity in that she will soon no longer be the sort of person she has always enjoyed being (i.e., she will no longer have blue eyes). Moreover, she points out that although the meaning that taking this drug has for her is not quantified on any outcome measure, it nonetheless affects her quality of life (Chauhan 2007).

We can imagine that even if the bimatoprost is only minimally successful and Chauhan's resulting change score from the PROM is low, she will nonetheless have experienced a significant change—she will not be the same person she was before. But this significance is tied to the place that her blue eyes had to her identity and what she took to be a good life; ceteris paribus we would not expect a brown-eyed person to summarize his or her experience in the same way. Thus, it would not be surprising if Chauhan's answer to the transition question was "quite a bit" while the magnitude of her change score was minimal.

This example illustrates how a popular and widely used attempt to improve the epistemic quality of CTT instruments is infused with epistemic-ethical concerns. The assumption that motivates the use of a MID is ill founded. Patients' assessments of change are influenced by the way that change interfaces with their identity and their accounts of what makes for a good quality life. Using a MID as the interpretive key for a PROM threatens to lose the "whole person," as Chauhan put it, while only providing the appearance of epistemic improvement. Epistemically, MIDs do not necessarily underwrite any particular magnitude of change from PROMs; thus, they do not necessarily represent the smallest clinically significant change. Using a MID as though it represents clinical significance threatens to ignore or exacerbate the harms or benefits patients experience depending on the particulars of the intervention in question.

Latent Trait Theory Reconsidered

Earlier I discussed that, the Rasch family of models claim to be able to establish interval-level measurement. For my purposes here, I'm going to assume that in principle this claim is true (for a debate over this claim see Borsboom and Mellenbergh 2004; Borsboom and Zand Scholten 2008; Michell 2000).

To provide some background, the Rasch model requires (1) that a person with greater ability should have a greater probability of answering a question correctly and (2) that given two questions, one of which is more difficult than the other, a person has a greater probability of answering the easier question. Given these relationships, Rasch models can establish measurement values for person ability and question difficulty. The function used in Rasch models is a logit, and the formula for one variant of this model for questions with dichotomous response options is:

$$Pr\{x_{ni}|\beta_n, \delta_i\} = \frac{e^{x_{ni}(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

where $x_{ni} \in [0,1]$; β_n and δ_i are the measurements of the ability of person n and the difficulty of item i, respectively, upon the same latent trait, and e is the natural logarithm constant (2.718). In COA applications, these models often concern levels of severity or frequency on a rating scale interpreted as indicating less or more health, disease burden, functionality, engagement in decision making, and so on.

The extent to which observed data (patients' responses to questions) "fit" the predictions of those responses from a Rasch model, within acceptable uncertainty, indicates the extent to which "measurement" is achieved. This is because, if ordered items fit the Rasch model's predictions, we can infer interval-level measurement of a latent variable. But we know that observed data never perfectly fit the predictions of a model. Thus, where do we set the threshold? To be sure, we are looking for a convergence between the empirical data and the Rasch model that is able to support useful inferences, but whether the convergence is sufficient for such inferences is open to legitimate disagreement.

This area of legitimate disagreement is a place where nonepistemic values come into play. Consider a questionnaire whose data imperfectly fit the Rasch model, but the way they imperfectly fit the model has an easy "fix": you can get rid of some of the questions in the original questionnaire.[4] Should you remove these questions to get a better fit? The qualitative research during the development of the questionnaire suggested these questions were important—research that involved patients and clinical experts in the field—but if we get rid of them, we can claim that this attribute is a continuous quantity. Doing so might mean that we can meet the demands of the government agency or pharmaceutical company funding us; it also might make us look good, and/or we might be able to sell the instrument and make a profit. But we might also consider how removing these items could affect how people fare when the attribute in question is measured by the instrument. Ethically speaking, might removing the items harm people; could it benefit them? Moreover, although we can claim that the attribute is a continuous quantity, removing the items weakens this claim and thus weakens the argument of an epistemic advance of Rasch over CTT.

CONCLUSION

Contemporary debate on psychological measuring instruments tends to focus on the importance of a realist ontology. I have argued that proponents of this position understand realism to support epistemic values of truth and rigor in measurement, values they take to be undermined by using antirealist approaches to measurement, namely CTT. I have suggested that whatever the inadequacies of CTT, these problems cannot be considered purely ontological or epistemic failings, and neither can the measures from latent trait theory be understood as simply advancing science. Psychological measures, particularly PROMs, are characterized by ethical-epistemic entanglements. In some cases, these entanglements mitigate against realist claims as they do in the example of Rasch, or they can complicate the antirealist picture psychometricians attribute to CTT as they do in the example of MIDs.

What, I think, these entanglements illustrate is that the emphasis on a realist ontology as a threshold over which serious psychologists and legitimate measures must cross is overly simplistic. Nonepistemic, often ethical values, enter into a large proportion of questions during measurement development and use—both for CTT and latent trait theory. These questions are not divisible from the epistemic questions whose values realists seek to uphold and whose answers are often used as evidence of a realist ontology. Ignoring them falsely suggests that latent trait theory is inherently superior to CTT both scientifically and in terms of our obligations to respondents.

NOTES

1. Borsboom (2005) argues for a between-subject causal connection between latent variables and item responses (opposed to within-subject causal account, which he rejects).

2. To be sure, operationalism need not imply antirealism. As Eran Tal (2016) points out, the methods chosen could underwrite an attribute that refers to a mind-independent reality. But the way CTT is used tends to render this interpretation irrelevant since the point Borsboom and Michell are making is that psychological measurement requires realism to be a form of measurement.

3. I have argued that in some cases they have some epistemic advantages as well; see McClimans, Browne, and Cano, 2017.

4. This is not simply a thought experiment. See, for example, the development of the ABILHAND questionnaire, where items that failed the statistical tests of fit to the Rasch model were discarded (Durez et al. 2007).

REFERENCES

Black, Nick. 2013. "Patient Reported Outcome Measures Could Help Transform Healthcare." *The British Medical Journal* 346: f167. doi:10.1136/bmj.f167.

Borsboom, Denny. 2005. *Measuring the Mind*. Cambridge: Cambridge University Press.

Borsboom, Denny. 2006. "The Attack of the Psychometricians." *Psychometrika* 71: 425–40. doi:10.1007/s11336-006-1447-6.

Borsboom, Denny, and Gideon J. Mellenbergh. 2004. "Why Psychometrics Is Not Pathological: A Comment on Michell." *Theory & Psychology* 14: 105–20. doi:10.1177/0959354304040200.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111: 1061–71. doi:10.1037/0033-295X.111.4.1061.

Borsboom, Denny, and Annemarie Zand Scholten. 2008. "The Rasch Model and Conjoint Measurement Theory from the Perspective of Psychometrics." *Theory & Psychology* 18: 111–17. doi:10.1177/0959354307086925.

Bridgman, Percy. 1927. *The Logic of Modern Physics*. London: The Macmillan Company.

Cano, Stefan J., and Jeremy C. Hobart. 2011. "The Problem with Health Measurement." *Patient Preference and Adherence* 5: 279–90. doi:10.2147/PPA.S14399.

Chauhan, Cynthia. 2007. "Denouement: A Patient-Reported Observation." *Value in Health* 10: S146–S47. doi:10.1111/j.1524-4733.2007.00276.x.

Department of Health. 2008. *High Quality Care for All: NHS Next Stage Review Final Report*. London: The Stationary Office.

Durez, Patrick, Virginie Fraselle, Frédéric Houssiau, Jean-Louis Thonnard, Henri Nielens, and Massimo Penta. 2007. "Validation of the ABILHAND Questionnaire as a Measure of Manual Ability in Patients with Rheumatoid Arthritis." *Annals of the Rheumatic Diseases* 66: 1098–105. doi:10.1136/ard.2006.056150.

Fayers, Peter, and David Machin. 2015. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. 2nd Edition. Hoboken, NJ: John Wiley & Sons.

Heilmann, Conrad. 2015. "A New Interpretation of the Representational Theory of Measurement." *Philosophy of Science* 82: 787–97. doi:10.1086/683280.

Hobart, Jeremy, and Stefan Cano. 2009. "Improving the Evaluation of Therapeutic Interventions in Multiple Sclerosis: The Role of New Psychometric Methods." *Health Technology Assessment* 13: 1–177. doi:10.3310/hta13120.

Hobart, Jeremy C., Stefan J. Cano, John P. Zajicek, and Alan J. Thompson. 2007. "Rating Scales as Outcome Measures for Clinical Trials in Neurology: Problems, Solntions, and Recommendations." *Lancet Neurology* 6: 1094–105. doi:10.1016/S1474-4422(07)70290-9.

Jaeschke, Roman, Joel Singer, and Gordan H. Guyatt. 1989. "Measurement of Health Status: Ascertaining the Minimal Clinically Important Difference." *Controlled Clinical Trials* 10: 407–15.

Lord, Frederic M., and Melvin R. Novick. 2008. *Statistical Theories of Mental Test Scores*. Charlotte: Information Age Publishing.

Maul, Andrew. 2013. "On the Ontology of Psychological Attributes." *Theory & Psychology* 23: 752–69.

McClimans, Leah. 2010. A Theoretical Framework for Patient-Reported Outcome Measures. *Theoretical Medicine and Bioethics* 32: 47–60.

McClimans, Leah. 2011. "Interpretability, Validity, and the Minimum Important Difference." *Theoreticol Medicine and Bioethics* 32: 389–401. doi:10.1007/s11017-011-9186-9.

McClimans, L, Browne, J and Stefan, C. 2017. "Clinical Outcome Measurement: Models, Theory, Psychometrics and Practice", *Studies in the History and Philosophy of Science* (DOI) 10.1016/j.shpsa.2017.06.004 available at https://doi.org/10.1016/j.shpsa.2017.06.004

Michell, Joel. 1997. "Quantitative Science and the Definition of Measnrement in Psychology." *British Journal of Psychology* 88: 355–83.

Michell, Joel. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Cambridge: Cambridge University Press.

Michell, Joel. 2000. "Normal Science, Pathological Science and Psychometrics." *Theory & Psychology* 10: 639–67. doi:10.1177/0959354300105004.

Michell, Joel. 2005. "The Logic of Measurement: A Realist Overview." *Measurement* 38: 285–94.

Schwartz, Carolyn, and Bruce Rapkin. 2004. "Reconsidering the Psychometrics of Quality of Life Assessment in Light of Response Shift and Appraisal." *Health and Quality of Life Outcomes* 2: 16.

Speight, Jane, and Shalleen M. Barendse. 2010. "FDA Guidance on Patient Reported Outcomes." *British Medical Journal* 340: c2921. doi:10.1136/bmj.c2921.

Tal, Eran. 2013. "Old and New Problems in Philosophy of Measurement." *Philosophy Compass* 8: 1159–73. doi:10.1111/phc3.12089.

Tal, Eran. 2016. "Measurement in Science." In *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2016/entries/measurement-science/.

Taminiau-Bloem, Elsbeth F., Florence J. Van Zuuren, Mechteld R. M. Visser, Carol Tishelman, Carolyn E. Schwartz, Margot A. Koeneman, Caro C. E. Koning, and Mirjam A. G. Sprangers. 2011. "Opening the Black Box of Cancer Patients' Quality-of-Life Change Assessments: A Think-Aloud Study Examining the Cognitive Processes Underlying Responses to Transition Items." *Psychology & Health* 26: 1414–28. doi:10.1080/08870446.2011.596203.

Washington, A. Eugene, and Steven H. Lipstein. 2011. "The Patient-Centered Outcomes Research Institute: Better Information, Decisions, and Health." *New England Journal of Medicine* 365: e31.

Wyrwich, Kathleen W., and Vicki M. Tardino. 2006. "Understanding Global Transition Assessments." *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 15: 995–1004. doi:10.1007/s11136-006-0050-8.

Chapter 7

The Epistemological Roles of Models in Health Science Measurement

Laura M. Cupples

Patient-reported outcome measures are survey instruments used by health researchers and clinicians to quantify health-related quality of life or health status.[1] These measures are epistemically sound only when they can be shown to be valid, comparable to other measures of the same attribute, and accurate. In this paper, I introduce three different kinds of models that I argue are essential for supporting judgments of validity, comparability, and accuracy, respectively. The first types of models are qualitative models. These models represent patients' and researchers' interpretations of test items and their conceptualizations of target attributes. Second, I examine statistical models; they are models that give an account of how patients interact with questionnaire items. The third kind of models I discuss are theoretical models. These models tell a story about the composition of the attribute, its behavior over time and across patient groups, and the relationship between patients' raw scores and the level of the attribute they possess.

While other authors have discussed the roles of qualitative models (McClimans 2010), statistical models (Bond and Fox 2007; Streiner, Norman, and Cairney 2015), and theoretical models (Rapkin and Schwartz 2004; Stenner et al. 2013), in many cases they have not tied these models to their epistemic roles. That is, they have not necessarily associated them with judgments about content validity, comparability, and accuracy.

BACKGROUND

In what follows, I discuss the relationship between patient-reported outcome measures and the models that I contend ought to be used to support them. Patient-reported outcome measures are survey instruments used by medical