The Epistemological Roles of Models in Health Science Measurement

Laura M. Cupples

University of South Carolina

Patient-reported outcome measures are survey instruments used by health researchers and clinicians to quantify health-related quality of life or health status. These measures are only epistemically sound 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; these are models that give an account of how patients interact with questionnaire items. The third kinds 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 that they possess.

While other authors have discussed the roles of qualitative models (McClimans 2010), statistical models (Streiner et al. 2015 and Bond and Fox 2007), and theoretical models (Rapkin and Schwartz 2004 and Stenner 2013), in many cases they have not tied these models to their particular 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 researchers and clinicians to quantify patients' health status or health-related quality of life. These measures rely on self-report to make patients' private experiences public and accessible to clinicians and researchers. They typically ask respondents questions about, for example, physical and psychological functioning, mobility, social connectedness, pain levels, or other factors that researchers believe contribute to health status and health-related quality of life.

While some of the instruments used to measure health status and health-related quality of life are generic (e.g. the Short Form-36 (Stewart and Ware 1992) and the Nottingham Health Profile (McDowell 2006)) and thus supposed to quantify well-being for patients with a wide range of ailments and health statuses, other instruments are disease specific and are designed to be used only with patients who have particular illnesses (e.g. asthma, arthritis, or cancer). Still other instruments are site specific, and focus on the effect of injury to, deterioration of, or intervention upon certain body parts (e.g. the Oxford Hip Score (Murray et al. 2007) and the BREAST-Q (Klassen et al. 2009)).

Measurement of health-related quality of life and health status involve complex processes, which include patient understandings and interpretations of survey questions, the cognitive abilities of patients, their powers of memory, and the values that shape their appraisal of quality of life. Moreover, patient interactions with survey items also depend on the statistical properties of those items and their intended conceptual content. Furthermore, measurement involves the numeric representation of outcomes and the management of error. Models of the measurement process are holistic representations that take into account some subset of these

factors. I will argue below that in order to obtain a full picture of the measurement process, and to facilitate judgments about an instrument's validity, accuracy, and comparability, three different models of the measurement process must be deployed, namely qualitative models, statistical models, and theoretical models.

Throughout this paper I will understand models to be abstract and idealized representations of dynamic systems that are constructed based on theoretical, statistical, and pragmatic assumptions about those systems. While models are based in part on abstract theory, they also function separately from that theory because they often incorporate material constraints and affordances, assume background conditions specific to the local system in question, and reflect the limits of our mathematical capabilities (Morgan and Morrison 1999).

Qualitative Models and Content Validity

In this section, I argue that qualitative models of the measurement process have an important role to play in supporting judgments about the content validity of measures in the health sciences. I take a qualitative model of the measurement process to be an explication of patient or researcher interpretations of test items. These interpretations help to determine the actual conceptual content of the measure, since they affect the operationalization of the measure. Yet we also hope that the intended conceptual content of the measure matches up with patient conceptualizations and interpretations. Unfortunately, patients and researchers often understand test items, and target attributes such as health status and health-related quality of life, differently from one another and differently over time (McClimans 2010; Rapkin and Schwartz 2004). Varying understandings of the attribute in question mean that patients can interpret the meanings of test questions in different ways. Thus patients may, in effect, be answering different questions from

the ones researchers believe themselves to be asking. As I will explain below, when this happens, the content validity of our measures suffers.

A measure with good content validity comprehensively covers all domains that are part of the target attribute. All and only those domains that are part of the target attribute are captured by such a measure (Food and Drug Administration 2009). Content validity is important because it helps to secure inferences from a measure's outcomes to an attribute of interest, i.e. that the quantitative representation given by the measure's outcome is representative of some portion or level of the attribute. If a measure is intended to assess quality of life after mastectomy and breast reconstruction, but the items focus on physical functioning and neglect aesthetic appearance, then we might reasonably lack confidence in the inference that the measure's outcome represents quality of life after these interventions. Our lack of confidence is due to the fact that the measure has poor content validity, i.e. it neglects aspects of the attribute that are relevant to making inferences from the outcomes.

But how is content validity diminished by a mismatch in patients' and researchers' interpretations of test items? When patients' interpretations of test items fail to coincide with the interpretations envisioned by quality of life researchers, the operationalization of the measure when applied to patient populations may differ from the operationalization intended by researchers. Test items will carry different meanings, and thus different conceptual content, from what was envisioned. This difference results in diminished content validity because the inference from the measure's outcomes to the intended attribute is invalid. The instrument does not, in fact, measure what researchers meant for it to measure.

Why do patients and researchers sometimes disagree in their understandings and interpretations of items? Moreover, what might such disagreement look like? Imagine we are

trying to get a sense of how limited patients are in their mobility. To determine this, we ask several groups of patients how difficult it is for them to engage in strenuous exercise. Healthy patients may envision a run of several kilometers, while for patients with a chronic illness or disability, a walk of a few hundred meters may be considered strenuous. For very elderly patients, or patients with significant disability, even a walk across the house may be challenging. Because of the different contexts informing their interpretations, these patients are answering different test items from one another and perhaps different test items from those researchers may have intended them to answer. Depending on the contrast class they apply to the question (for instance, how limited in their mobility they were a month ago, how limited they perceive other patients with the same illness or injury might be, or how limited they were when in full health (see van Fraassen 1980 and McClimans 2011), patients may see a broad range of abilities as indicative of relatively good mobility for them (Rapkin and Schwartz 2004). When they talk about how limited they are in their mobility, even patients who cannot engage in very strenuous activity may feel less limited than we might imagine. On the flip side, patients who are still relatively mobile may feel more limited than we might see them as being.

If we want our measures to demonstrate good content validity, we need to find a way to bring the qualitative models of the measurement process—the models that specify patients' and researchers' interpretations of test items and therefore the conceptual content of the measures we are interested in—into agreement with one another. How can we best accomplish this goal? Because patient-reported outcome measures were created to give patients a voice with regard to their own subjective health status, it seems that we should privilege their understandings of health status and health-related quality of life. This means that researchers should concentrate on building qualitative models that describe patients' true interpretations of test items. Researchers

cannot simply assume that mismatches between their interpretations and those of patients constitute error on the part of patients. They must re-examine their own interpretations in light of those held by patients (McClimans 2010).

How can researchers discover the content of patients' conceptualizations of health status and health-related quality of life, and how can they learn about patients' interpretations of test items? This is done through qualitative research during the instrument development process. Patient focus groups are asked about the domains they feel are most important to their health-related quality of life or health status. They can be asked which symptoms make the biggest impact on their lives, and which capabilities are most important for them to maintain. This sort of information, along with input from clinical experts, helps researchers write items that are relevant to patients' experiences with health and illness (Klassen 2009). Once a draft of the instrument is completed, patients can be interviewed individually as part of a think aloud study (Westerman et al. 2008 and Bellan 2005). Patients can be queried about the relevance and clarity of test items, i.e. about how they interpret the items they are presented with and why. When researchers have access to these interpretations, and when they are able to write instruments that cover the conceptual content that patients feel is most relevant to the attributes to be measured, they will be able to build relatively accurate qualitative models of the measurement process.

The good news is that most quality of life researchers do now rely on patient input during measure development. The practice of interviewing patients about their experiences with illness and treatment has become much more common since the 2009 publication of a new FDA guidance on the development of patient-reported outcome measures. This recent change in practice is an important first step in establishing sound qualitative models.

Statistical Models and Comparability

In this section I discuss two statistical models used to represent the process of health status and health-related quality of life measurement. Specifically I will examine the model(s) used in classical test theory (CTT) and those used by Rasch measurement theory. In general the models used by CTT give an account of how observed scores relate to true scores (Streiner et al. 2015) and the models used by Rasch represent how patients interact with test items to produce an outcome or test score (Stenner 2013). In what follows I examine the ways these statistical models epistemically support or fail to support judgments about comparability among measures of the same attribute.

Classical Test Theory

Most patient-reported outcome measures are designed and analyzed using classical test theory. While modern psychometric methodologies such as Rasch measurement theory boast greater utility in many respects (e.g., CTT produces ordinal level measures while Rasch produces interval level measures), CTT is still very popular due to its flexibility and ease of use. CTT employs a thinner statistical model than modern psychometric theories such as Rasch. Because of the way it models measurement, it gives us little information about the mechanics of the measurement process, or about the ways patients interact with individual test items (Borsboom 2005). Moreover, as I will show, the model employed by CTT does not easily facilitate the creation of comparable measures of the same target attribute (Bond and Fox 2007).

The CTT model posits three variables: a true score (T_T), an observed score (T_O), and a random error term (E).

$$T_O = T_T + E$$

We can think of a respondent's true score as the expected value of the observed score (the actual score achieved on the measure) over a universe of possible observations of the same construct. As shown above, the observed score is the sum of the true score plus the random error term. The expected value of the random error term over many test administrations is zero (Borsboom 2005).

In CTT, the individual items are taken to be members of a random sample drawn from a population of possible items (Kane 1982). Answers to each item contribute equally to the final raw score, and what matters is how items perform en masse rather than individually (Streiner et al. 2015). This is because the unit of analysis in CTT is the test as a whole rather than, say, individual items in the questionnaire. The result is that CTT gives us little or no insight into how respondents interact with individual test items. For example, CTT does not specify the difficulty of each test item, nor does it tell us how likely it is that a respondent with a certain level of the target attribute will answer an item in a particular way.[i] Instead of providing information at the item level, CTT helps us understand how groups of respondents interact with the test as a whole. In what is called norm-referenced measurement, patients' scores on CTT tests are compared with the performance of norm groups in order to place outcomes in context.

Because CTT focuses on how groups of respondents interact with the test as a whole, it is difficult to achieve comparability of measuring instruments. In other words, it is difficult to develop parallel measures of the same attribute for which the same scores carry the same meaning (i.e., signify the same level of quality of life or health status). In order to say that two instruments measure the same attribute, we must ensure that test items cover exactly the same range of content. But this coverage is difficult to ensure with CTT at least in part because attributes measured by CTT instruments are often multi-dimensional (Borsboom 2005), e.g.

health status and health-related quality of life are usually taken to include physical, functional, emotional, and social dimensions (Cella 1994). In a CTT measure, the content of the attribute is determined by the specific content of the totality of the questions (Streiner et al. 2015). It is tricky to perfectly replicate the conceptual content of a CTT test as a whole, even if you try to match questions by conceptual content item by item. (For instance, do turning a key and fastening a button require the same type of capability? Or do questions about these two tasks in fact cover different conceptual content?) Nonetheless with good qualitative and theoretical models of the measurement process, it may be possible to create tests that measure the same attribute. When good conceptual definitions are used to inform the content of test items, there is a better chance that those items will cover the same conceptual content as comparable tests. This is because good conceptual definitions can help answer exactly such questions as whether or not fastening a button and turning a key require the same sort of capability. However, a common criticism of patient-reported outcome measures is that their conceptual and theoretical foundations are usually rather weak (McClimans 2010; Hunt 1997; Hobart et al. 2007). This means that these sorts of questions are usually left unanswered.

In addition to ensuring that two instruments measure the same attribute, comparability requires that the same scores carry the same meanings on those instruments. CTT tests differ in the manner in which their scores are determined. Most tests simply sum responses to questions to arrive at a raw score, but once this is done they often rescale that raw score in some way to arrive at a final outcome. The algorithm used to rescale outcomes differs from instrument to instrument (see for example the Disabilities of the Arm, Shoulder and Hand in Cano et al. 2010; and the non-normed physical function scale for the Short Form-36 in Stewart and Ware 1992). For this reason, identical scores on two different instruments may signify different levels of the

same attribute. Similarly, questions may be posed either positively or negatively—targeting, for instance, either mobility or its inverse.[ii] For this reason, even the directionality of scales may differ.

Lately, efforts have been made by quality of life researchers to place scores on normed scales. By placing outcomes on a scale from 0 to 100, calibrating mean score values to 50, and scaling standard deviations to 10 for a number of quality of life instruments, researchers have been able to facilitate comparability among measures of the same attribute (e.g. the Short Form-36 in Stewart and Ware 1992). Unfortunately, these efforts can also be misleading. Placing outcomes on the same scale does not ensure that measures cover the same conceptual content, and thus does not ensure that they target the same attribute. As I have argued, two requirements must be met for comparability between measures. Measures must target the same attribute and like scores must carry like meanings.

Rasch Measurement Theory

Recently health researchers have begun to take advantage of the resources offered by modern testing methodologies such as Rasch measurement theory and item response theory (IRT) (Hobart et al. 2007). The Rasch model is often considered to be a subset of IRT models so for the sake of simplicity, I will focus on the Rasch model in this section.[iii] Rasch measurement theory deploys a thicker statistical model than CTT, primarily because it tells a more complete story about how patients with a certain level of the measured attribute interact with individual test items of varying difficulty. Rasch locates an instrument's items on a continuum according to difficulty so that successively ranked items should each be more difficult for patients to answer (Bond and Fox 2007). The more items a patient can endorse on the BREAST-Q, for instance, the

more favorable her surgical outcome is estimated to be in terms of satisfaction with surgical results and care as well as resultant quality of life (Klassen 2009).

Unlike CTT, instruments designed and analyzed using Rasch measurement theory are intended to measure unidimensional attributes. The level of attribute possessed by the patient counters the difficulty of individual test items, so that when the level of attribute exceeds the difficulty of a test item, a patient is more likely to answer a question in the affirmative (Bond and Fox 2007). For instance, the Patient-Reported Outcome Measures Information System (PROMIS) physical function instrument asks questions such as "Are you able to sit at the edge of your bed?" and "Are you able to carry a laundry basket up a flight of stairs?" Intuitively, a patient must possess more mobility to be able to answer the second question in the affirmative than the first. Rasch measurement theory tells us the probability (P) that an item (x_i) will be answered in a particular way ($P(x_i = 1)$)—for instance, that an item will be endorsed (1) rather than rejected (0), is determined solely by the relationship between item difficulty (d_i) and the amount of the attribute possessed by the patient (b) (Stenner 2013). So, for example, the probability that a patient will agree that she is able to dress herself depends on the relationship between the difficulty of the task and her amount of functional ability. The equation below describes what is called an item response curve, for a given item (x_i). An item response curve describes the probability that an item of a given difficulty will be endorsed, or that a particular answer will be given, based on the level of attribute possessed by the respondent.

$$P(x_i = 1) = \frac{e^{(b - di)}}{1 + e^{(b - di)}}$$

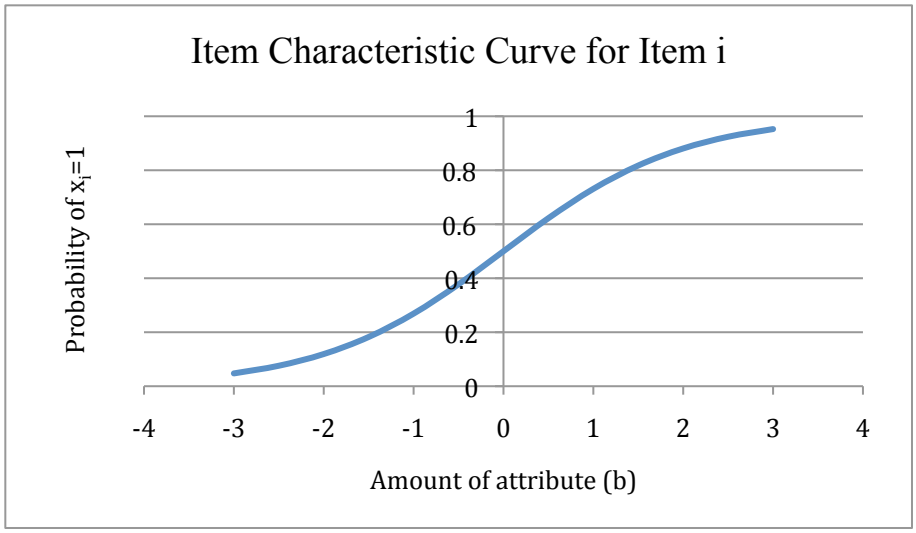

Figure 1. Item characteristic curve showing the probability of the respondent choosing the answer $x_i = 1$. The difficulty of item i is set to 0 logits.

The Rasch model boasts a number of advantages over CTT. For instance, as a result of the mathematical separability of item difficulty and level of attribute in the Rasch model, these two factors are invariant across patient populations and with respect to the subset of test items employed, respectively. That is, the difficulty of items does not depend on who is responding to them or on how much of the measured attribute they possess. Likewise, estimates of a patient's level of the measured attribute do not depend on the specific items employed by the measure. The function of the measuring instrument does not depend on the context in which it is employed (i.e., whether a meter stick is used to measure a table or a rug), and measurement outcomes do not depend on the specific instrument used, as long as that instrument is properly calibrated. Together, these qualities are often referred to as specific objectivity (Stenner and Burdick 1997).

The specific objectivity of these types of measures is an extremely useful trait because it makes it possible to compose comparable tests of the same attribute using the method of item banking. With item banking a large bank of items is created, with all items measuring the same unidimensional attribute, and subsets of items from that bank are combined to form tests of

various lengths (often with the goal of minimizing the burden placed on patients) (Bond and Fox 2007). Tests can also be created that are targeted to patients with a particular amount of the measured attribute, so that the instrument provides greater more precise measurement at that attribute level.

Nevertheless, it is still important when composing comparable instruments using Rasch to base those measures on qualitative models of the target attribute. Though the mathematical characteristics of Rasch measurement can ensure that these measures are both unidimensional and specifically objective, and hence that the various instruments composed from associated item banks all measure the same attribute, it is still important to know what the conceptual content of that attribute is, i.e., to ensure good content validity. For instance, it is important to know whether a measure targets depression or anxiety. These two attributes are often comorbid, and similar questions can be used to assess them, so a good qualitative model is necessary to separate measures of one from the other.

Theoretical Models and Accuracy

In this section, I discuss the epistemic role of theoretical models of the measurement process and argue that they facilitate judgments about measurement accuracy. Theoretical models of the measurement process are models informed by a theory of the attribute. Like qualitative models, they tell us about the conceptual content of the measure. They might tell us, for instance, when it is permissible to drop a statistically ill-fitting item from a Rasch measure, and when that item is essential to the instrument's conceptual content. But they also tell us how the attribute behaves—how it changes over time and across circumstances or patient groups. So a theoretical model would tell us what kind of change in quality of life we might expect over the course of a

patient's illness or treatment, and whether an unexpected change should be classed as legitimate variation in the target attribute or an instance of error (McClimans 2010).

Marjan Westerman and her colleagues (2008) studied a phenomenon called response shift among a group of cancer patients receiving chemotherapy. Response shift is an unexpected change in a patient's measured level of quality of life, or some other target attribute, due to adaptation to illness or treatment. For instance, a patient may change her frame of reference—the standard to which she compares her current condition—and this may alter her appraisal of her quality of life. Or a patient may reconceptualize what it means to be limited in his pursuit of leisure activities. One of Westerman's cancer patients claimed at the beginning of treatment that he was very limited in pursuing his leisure activities. He was an avid gardener but found his hobby difficult to maintain once he became ill. Several weeks later claimed he was only a little bit limited, yet by all accounts he was more physically limited than when he responded the first time (Westerman et al. 2008, p. 555). Most quality of life researchers hold the pre-theoretic assumption that quality of life, and the domains that make it up, are standardizable. That is, they take the meaning of quality of life, or of limitation in this case, to be constant from one case to the next. Thus, when these concepts shift in their meaning, as they did for Westerman's patient, they assume it must be due to measurement error. But a theory of the measured attribute might suggest that quality of life and its constituent domains cannot be standardized in this way. It may suggest that meanings shift according to patients' circumstances. If so, this patient's reconceptualization of what it means to be limited might not be an instance of measurement error at all, but instead an example of legitimate qualitative variation in the target attribute (McClimans 2010, Rapkin and Schwartz 2004).

A theoretical model helps us make judgments about measurement accuracy in part by allowing us to distinguish between legitimate changes in a patient's level of quality of life and responses that should be considered errors. Without a theory of quality of life, it is premature to make the judgment that the patient whose quality of life appeared to improve—due to his adaptive change in leisure activities—was in error about his quality of life (McClimans 2010). Quality of life may in fact change based on subjective assessments of limitation rather than due to objective improvement or deterioration. Many people with acquired disabilities, after initially rating their quality of life as lower than when they were able-bodied, later claim to value their new lives just as highly as their previous, able-bodied lives (Barnes 2016). Taking their testimony seriously may require us to see changes in quality of life due to response shift as legitimate variation. Some proponents of Rasch measurement (Stenner et al. 2013, Hobart et al. 2007) see a somewhat different role for theoretical models of the measurement process. They see these models essentially as helpers to the statistical model. According to Jack Stenner and his colleagues (2013) theoretical models help to predict the difficulty of test items based on certain causal factors that explain their position on the measurement scale. For instance, the difficulty of mobility items might vary in terms of the strength and range of motion required to complete the relevant mobility tasks. [iv]When testing these theoretical models, we can compare our empirical estimations of item difficulty (estimations based on the probability distribution of actual patient responses) with our calculated theoretical values for mobility in order to determine how closely empirical values match theoretically calculated values. Once our theoretical model has been well confirmed, we can use it to make judgments about item fit to the model.

I suggest that in the case of Rasch measures, having this sort of theoretical model of the measurement process allows us to make judgments about what Eran Tal calls operational

accuracy – or accuracy relative to some standard (2012). We use measurement standards to calibrate individual instruments, or to ensure that they conform to an idealized scale and thus measure their target attributes accurately. In this case, the standard against which empirical measures of item difficulty are being calibrated is the idealization of item difficulty predicted by the theoretical model.[v] Without a theory of the attribute to facilitate calculation of theoretical values for item difficulty, we do not have a standard for comparison with empirical values, and we cannot make judgments about the operational accuracy of our measures. That is, we cannot make judgments about the accuracy of our measures relative to the standard set by theory.

Unfortunately, patient-reported outcome measures have notoriously weak conceptual grounding, and in most cases lack a theory of the attribute, and a forteriori, a theoretical model of the measurement process (McClimans 2010; Hunt 1997; Hobart et al. 2007). This frequent lack of theoretical model has consequences for our ability to make judgments about the accuracy of patient-reported outcome measures for both CTT and Rasch measures. Thus, one suggestion for future research is to develop good theoretical models for these measures. Not only will uncovering these models aid us in making judgments about measurement accuracy, but since theoretical models also incorporate certain roles of qualitative models, they will also aid us in making judgments about content validity.

Conclusion

The model-based account of measurement epistemology developed by Eran Tal (2012) for use in physical measurement argues that in order for us to make legitimate inferences about measure validity, comparability, and accuracy, our measures must be epistemically supported by abstract and idealized models of the measurement process. I have discussed three broad types of models

of the measurement process for patient-reported outcome measures of health-related quality of life and health status. Qualitative models reflect patients' understandings and interpretations of the construct in question and the associated test items. These models facilitate judgments about content validity. Statistical models give an account of how patients interact with test items in the Rasch framework and how observed scores relate to true scores in the CTT framework. The statistical model that a measure is rooted in helps to determine how comparable measures of the same attribute can be constructed. Finally, theoretical models are models that are derived from a theory of the measured attribute. Not only do they tell us about the conceptual content of the measure, they also tell us about the behavior of the target attribute over time and across patient populations. Theoretical models help us distinguish legitimate variation in the target attribute from patient error, and they help us establish the operational accuracy of Rasch measures.

References.

Anatchkova, Milena D. et al. 2011. "Assessing the factor structure of a role functioning item bank." *Quality of Life Research* 20: 745-758.

Andrich, David. 2004. "Controversy and the Rasch model." *Medical Care* 42(1): I-7-I-16.

Barnes, Elizabeth. 2016. *The Minority Body: A Theory of Disability*. Oxford: Oxford University Press.

Bellan, Lorne. 2005. "Why are patients with no visual symptoms on cataract waiting lists?" *Canadian Journal of Opthalmology* 40: 433-438.

Bond, Trevor G. and Christine M. Fox. 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* 2nd Ed. New York: Routledge, Taylor & Francis Group.

Borsboom, Denny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.

Boumans, Marcel. 2015. *Science Outside the Laboratory: Measurement in Field Science and Economics*. Oxford: Oxford University Press.

Cano, Stefan et al. 2011. "Beyond the reach of traditional analyses: Using Rasch to evaluate the DASH in people with multiple sclerosis." *Multiple Sclerosis Journal* 17(2): 214-222.

Cano, Stefan and Jeremy Hobart. 2011. "The problem with health measurement." Patient Preference and Adherence 5: 279-290.

Cella, David F. 1994. "Quality of life: Concepts and definition." *Journal of Pain and Symptom Management* 9(3): 186-192.

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

Clark, Herbert H. and Michael F. Schober. 1992. "Asking questions and influencing answers." In *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, ed. Judith M. Tanur, 15-48. New York: Sage Foundation.

Food and Drug Administration. 2009. *Guidance for industry on patient-reported outcome measures: Use in medicinal product development to support labeling claims*. Federal Register 74: 1–43.

Hobart, Jeremy et al. 2007. "Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations." *Lancet Neurology* 6: 1094-1105.

Hobart, Jeremy et al. 2013. "Achieving valid patient-reported outcomes measurement: A lesson from fatigue in multiple sclerosis." Multiple Sclerosis Journal 0(0): 1-11.

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(12).

Hunt, S.M. 1997. "The problem of quality of life." *Quality of Life Research*, 6: 205-212.

Kane, Michael T. 1982. "A sampling framework for validity." *Applied Psychological Measurement* 6(2): 125-160.

Klassen, Anne. 2009. "Satisfaction and quality of life in women who undergo breast surgery: A qualitative study." *BioMed Central Women's Health*, 9(11).

McClimans, Leah. 2010. "A theoretical framework for patient-reported outcome measures," *Theoretical Medicine and Bioethics* 31: 225-240.

McClimans, Leah. 2011. "The art of asking questions." *International Journal of Philosophical Studies* 19(4): 521-538.

McDowell, Ian. 2006. *Measuring Health* 3rd Ed. Oxford: Oxford University Press.

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

Morgan, Mary, and Margaret Morrison, eds. 1999. *Models as Mediators*. Cambridge: Cambridge University Press.

Murray, D.W. et al. 2007. "The use of the Oxford Hip and Knee Scores." *The Journal of Bone and Joint Surgery* 89B: 1010-1014.

Rapkin, Bruce, and Carolyn Schwartz. 2004. "Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift." *Health and Quality of Life Outcomes* 2: 16.

Stenner, A. Jackson et al. 2013. "Causal Rasch models." *Frontiers in Psychology* 4:1-14.

Stenner, A. Jackson and Donald S. Burdick. 1997. *The Objective Measurement of Reading Comprehension: In Response to Technical Questions Raised by the California Department of Education Technical Study Group*. Durham, NC: Metametrics, Inc.

Stewart, Anita and John Ware. 1992. *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Durham: Duke University Press.

Streiner, David et al. 2015. *Health Measurement Scales: A Practical Guide to their Development and Use, 5th Ed*. Oxford: Oxford University Press.

Tal, Eran. 2012. "The epistemology of measurement: A model based account" (PhD diss., University of Toronto)

Van Fraassen, Bas. 1980. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

Westerman, Marjan, et al. 2008. "Listen to their answers! Response behavior in the measurement of physical and role functioning." *Quality of Life Research*, 17: 549-558.

[i] Indeed, though I use the language of attributes for the sake of consistency, the CTT framework (unlike the Rasch framework) need not even hypothesize the existence of an underlying causal attribute. In general, CTT speaks of constructs rather than attributes.

[ii] It is debatable whether positively and negatively worded questions about, say, mobility even measure the same attribute. See for instance (Anatchkova et al. 2010).

[iii] The mathematical model deployed by Rasch measurement theory is identical to the one-parameter item response theory model. The only distinction is that the Rasch model is prescriptive, while the item response model aims only to be descriptively accurate (Andrich 2004). Two and three parameter IRT models incorporate additional variables in an attempt to better describe measurement data, but in doing so, they forfeit certain functional advantages shared by Rasch and the one-parameter model.

[iv] A. Jackson Stenner, telephone conversation with author, June 1, 2016.

[v] In his 2012 "How Accurate is the Standard Second", Tal notes that the duration of the standard second is defined and determined by idealized models, not by the ticks of physical clocks. This is because the duration of the tick of even the best physical clock is disrupted by a number of outside forces that carry it away from the duration described by the definition.