Health Measurement, Industry, and Science

Leah McClimans

Patient-reported outcome measures (PROMs) are now common endpoints in clinical trials. PROMs measure latent variables such as mobility, health status, and quality of life typically by asking patients multiple questions using a Likert scale, e.g., “Does your health now limit you in lifting or carrying groceries?” Available answers might include yes, limited a lot; yes, limited a little; and no, not limited at all. Measures such as these gained traction in the 1970s when health measurement started to expand to include patient-reported outcomes alongside more traditional measures of mortality and morbidity. This addition was at least in part due to the development of health technology and improved standards of living, which led to a shift in attention from the cure of acute disease to the management of chronic illness (Cano and Hobart 2011). More recently, the popularity of PROMs can be traced to quality improvement initiatives that emphasize clinical effectiveness and patient centeredness. In the 2009 *FDA Guidance for Industry Patient-Reported Outcomes Measures: Use in Medical Product Development to Support Labeling Claims*, we see a marriage of these two motivations. PROMs provide patient-reported evidence of the effectiveness of drugs targeting the management of chronic illnesses and diseases, for example, in 2010 the FDA approved a patient-reported outcome claim for Actemra, a drug used to treat adults with rheumatoid arthritis (DeMuro et al. 2013).

Despite their popularity with industry and government agencies, PROMs face serious criticisms regarding their measurement properties, e.g., validity—the degree to which a questionnaire actually measures what it was intended to measure (Streiner and Norman 2008). The FDA *Guidance for Industry* goes some way toward recognizing the importance of PROM’s measurement properties. For instance, they emphasize that these properties are part of what should be evaluated when determining a measure’s suitability for use in medical product labeling. But this emphasis does not go far

L. McClimans (✉)
College of Arts and Sciences, University of South Carolina, Columbia, SC, USA
e-mail: mccliman@mailbox.sc.edu

D. Ho (ed.), *Philosophical Issues in Pharmaceutics*, Philosophy and Medicine 122,
DOI 10.1007/978-94-024-0979-6_6

enough. As critics, including this author, have argued, the methods that are typically used to establish the psychometric properties of these measures are themselves problematic, e.g., the methods to establish validity are themselves not valid (Hunt 1997; Hobart et al. 2007; McClimans 2010a, b). The result is that even when PROMs appear to meet the FDA criteria, they can still be unethical and unscientific.

Given this claim one might imagine that changes in our psychometric methods—what some refer to as new psychometric methods (Hobart et al. 2007)—would be welcome. But in this paper, I argue that we ought to proceed with caution. Science is not value-neutral and neither are the suggested changes that would make PROMs allegedly more scientific. Indeed, I will argue that by improving the ostensible scientific basis of these measures, we may make it more likely that PROMs will show a treatment effect. While this difficulty is in the best interests of industry, it may not be in the best interests of the public.

Problems with PROMs

Before discussing the kinds of changes and outlook that "new psychometric methods" suggest, it is helpful to understand the context in which they have arisen. Specifically, it is useful to examine some of the criticisms of PROMs that have led to the suggestion of new psychometric methods. The essence of the problem with PROMs can be summarized by a single deficiency that has multiple consequences; namely, PROMs lack a theory that provides a representation of the measurement interaction—the relationship between the construct and its instrument. This has been my argument throughout a number of papers, and, with regard to the general contours of my argument, I am not alone (McClimans 2010a, b, 2015; McClimans and Browne 2012). Consider Donna Lamping's 2008 Presidential address to the International Society of Quality of Life Research, where she identified the need for a theoretical framework as one of three challenges facing the future of PROMs. Or take Jeremy Hobart et al. (2007 *Lancet Neurology*) where they lament the lack of explicit construct theories in their article criticizing the current state of PROMs. In a final example, Sonja Hunt, in her 1997 editorial for *Quality of Life Research*, argues that the surfeit of poorly designed measures suggests that we do not know what quality of life *is* (Hunt 1997). In what follows, I provide a brief overview of three consequences that result from this lack of theory: problems with validity, interpretability, and responsiveness.

Validity

Without a robust theory, establishing the measurement properties of PROMs is often a viciously circular endeavor. Consider construct validity. Construct validity asks whether the questionnaire at hand performs in a way that the underlying theory suggests it should. An increasingly common criticism of construct validation in the context of classical test theory—the dominant psychometric paradigm—is its inability to determine if a measure represents its object of inquiry, i.e., its inability to provide evidence of validation (Hobart et al. 2007).

Construct validity is typically tested by assessing a measure's internal and external 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. But this process does not tell us anything about the construct itself, e.g., quality of life or mobility. It tells us only that certain questions tend to occupy the same conceptual space. External construct validity is examined via convergent and divergent validity testing. Here, multiple measurement scales deemed similar to and different from one another are applied to a sample population, and the scores derived from respondent answers are correlated. These correlations determine 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 actually assesses. It tells us only that some scales are correlated (or not) with other scales. But without some kind of theory, it is not clear what any of these measures actually measure.

Interpretability

Over the last 15 years, the discussion of how to interpret change in patient-reported outcomes has received considerable attention. Interpretability refers to the clinical significance of increases or decreases on a particular scale or measure over time. For instance, if I score a 30 on the Beck Depression Inventory (BDI-II), we know that I have scored in the middle of the scale—the BDI-II has 63 total points. But imagine two months later I score 42. What does this 12-point increase mean from a clinical point of view? Should my drug regime change? If so, how should it change? PROMs that have been developed using classical testing theory (CTT) only provide ordinal level information, i.e., we know that someone who scores 42 is more depressed than someone who scores 30, but we do not know the degree of that difference. PROMs are thus difficult to interpret.

This has led to the development of methods to enhance their interpretability. One popular method is the identification of a minimal important difference (MID). An MID is the smallest change in respondent scores that represent clinical, as opposed to merely statistical, significance and which would, ceteris paribus, warrant a change in a patient's care (Jaeschke et al. 1989). One popular method for determining a measure's MID is to map changes in respondent outcomes onto some kind of 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. The method asks the control group of patients 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 2007). 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.

This approach of acquiring an 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 suggests that this assumption is ill founded (Taminiau-Bloem et al. 2011; Wyrwich and Tardino 2006). Whether one understands the magnitude of change over the course of an illness as large or small is a matter of interpretation. As I have argued elsewhere, respondents' answers to TRI ought to be understood against the background of what makes for a good quality of life, e.g., the magnitude of change to which the answer "a little better" refers depends heavily on the significance that, say, worry has within the respondent's vision of the good (McClimans 2011). Thus, it is possible to have an outcome that indicates a large magnitude of change and to interpret this change as minimal.

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 in her understanding of herself and what she took to be a good life; ceteris paribus we would not expect a brown-eyed person to summarize their 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.

I suggest that what examples such as this illustrate is that our understanding of clinical significance ought to be closely linked to our understanding of the construct given the cohort of respondents for whom the measure is targeted. To put this point slightly differently, understanding change in PROMs requires that researchers have a grip on what quality of life or perceived health status means in the context of a particular PROM and the population it serves. In other words, we need a theory of the construct that the PROM aims to measure, i.e., a collection of sentences, propositions, statements, or beliefs and their logical consequences, and these can include statistical and general laws.

Responsiveness

As with validity and interpretability, responsiveness too needs a construct theory. Responsiveness refers to an instrument's sensitivity to change, although just what kind of change a responsive instrument should be able to detect is somewhat controversial, i.e., clinically important changes, changes due to treatment effects, or changes in the true value of the underlying construct (Terwee et al. 2003). But regardless of the kind of change an instrument is meant to identify, the development of a measure requires information about the appropriate distance between units of change. Take a simple example. In the USA, infants are measured in grams because measuring in kilograms is not sensitive enough—the extra grams that an infant weighs can make a difference to their prognosis. But older children and adults are typically weighed to the nearest kilogram because the extra grams are negligible to most of their health outcomes. This decision is in part theory driven including (1) our theoretical understanding of mass, (2) the role body mass plays in our understanding of health outcomes, and (3) the application of this theoretical understanding to different populations, e.g., infants and adults.

What level of sensitivity to a change in quality of life, health status, or mobility, for instance, should PROMs employ? The sensitivity of a scale in the context of PROMs is determined by the number and kinds of questions posed to respondents. For example, single item scales—scales that only ask one question—are limited in sensitivity since they must divide rich variables (e.g., spasticity) into only a few levels (Hobart et al. 2007). But just how finely *should* we divide a variable? In part, the answer to this question can be understood statistically. Questions that are considered "too close" to one another will have overlapping standard errors, but this statistic can be manipulated by increasing the sample size, i.e., the greater the sample size the smaller the standard error around the item estimates.[1] By increasing the sample size, one can increase the precision of the measure.

But as with questions about the appropriate sensitivity of measures of body mass, the responsiveness of a PROM requires a theory of the construct being measured, how that construct relates to other areas of interest, and how our theoretical under-

[1] Standard errors are a way of telling from a statistical perspective if x is significantly different from y. If standard errors overlap, this tells us that, in the case of PROMs, two items are similar enough to be indistinguishable.

standing relates to different populations. If we are trying to establish the effectiveness of a new drug using a PROM as one of the endpoints, then we need some theory that provides a representation of the measurement interaction in the context of the patient cohort as well as an understanding about how the construct in question relates to the condition or illness that is targeted by the new drug. These theoretical considerations cannot be achieved with statistics alone. Indeed, as I will argue below, determining the correct responsiveness of a scale must include considerations of value, in particular harms and benefits.

Solutions to PROMs

Cano and Hobart have been two of the most vocal and consistent critics of the use of traditional psychometric methods to develop PROMs. They have also been two of the most ardent supporters of the use of "new psychometric methods." In this section, I focus on Cano and Hobart's (2011) suggestion for correcting PROMs' current limitations.

While Hobart et al. agree that most of the PROMs in use lack theoretical development, they trace this error to CTT. The problem with CTT is that it does not provide the theoretical resources needed to model the measuring instrument, in this case a PROM. CTT theorizes that a person's observed score on the scale is the sum of the unobserved score to be estimated, i.e., the person's true score, plus measurement error (Hobart et al. 2007). Consider a physical functioning scale with a scoring range of 11–44, where higher scores indicate more limited functioning. Imagine that someone's observed score was 23. CTT tells us that this observed score is the result of their true score plus measurement error. Respondents' true scores are what we would like to know, but to get them, we need some idea of what, e.g., quality of life scores look like for this particular cohort of respondents (say, respondents with lung cancer). At the same time, we need some idea of what counts as instances of measurement error. For instance, does response shift count as an instance of measurement error or part of a person's true score? Response shift is defined as the change in the meaning of one's self-evaluation of a target construct (Schwartz and Sprangers 1999). A classic example of response shift is when a respondent becomes accustomed to her disease/illness/disability and recalibrates her internal standard of measurement. Is this recalibration best understood as measurement error—as much of the quality of life literature treats it—or is it best understood as a legitimate aspect of the quality of life construct, i.e., to be incorporated in our theory of quality of life? CTT leaves us unable to answer this question.[2]

The problem with CTT is that it does not give us a theoretical ideal for the true score as, say, the measurement of time has a theoretical ideal, i.e., the second is defined as the duration of exactly 9,192,631,770 periods of the radiation corre-

[2] For a longer discussion of the difficulties that CTT has with distinguishing between true scores and measurement error, see McClimans 2017.

sponding to a hyperfine transition of cesium-133 in the ground state (BIPM (Bureau International des Poids et Measures) 2006). Nor are CTT's target constructs sufficiently embedded within theories that would allow for approximations of measurement error as, for example, time is embedded within physical theory, e.g., the definition of the second assumes that cesium is in a flat space-time, but the cesium fountains (primary standards) that metrologists build are subject to gravitational redshift. Relativity theory helps us to estimate the error associated with these phenomena (Tal 2011). In other words, CTT does not provide a theoretical representation of the measurement interaction, i.e., the relationship between the construct of interest and its instruments (McClimans 2015).

In place of CTT, Hobart et al. argue for the use of new psychometric methods, particularly Rasch methodology. How is Rasch different from CTT? One important difference is that Rasch has an explicit mathematical model that provides a representation of the measurement interaction. Rasch measurement theory says that a person's response to an item is determined by the difference between a person's location on the ruler (i.e., how much ability they have) and an item's location on the ruler (i.e., how much ability an item requires). Thus it provides a representation of the measurement interaction. In particular, Rasch states that the higher a person's ability with respect to the difficulty of an item, the higher the probability that a person will answer "yes" to an item. The Rasch scale runs from plus to minus infinity with the zero point at the place where the difficulty of the items in the survey is equal to the ability of the sample population. Each item is located on the ruler relative to the point at which there is equal probability of respondents answering "yes" or "no" to that particular item. In Rasch, the probability of answering "yes" or "no" is modeled as a logistic function. The mathematical equation that governs this function is the model that represents Rasch measurement theory.

A second difference between Rasch and CTT is that Rasch prioritizes its mathematical model over the data. CTT defines what it is measuring, say, quality of life, by the items that are generated, usually from a qualitative sample. It is a descriptive approach. The Rasch model, however, operationally defines the construct one is trying to measure as a relationship between a person's ability and the probability they will answer "yes" to an item. When applied to a sample population, this model provides the characteristics and regression weights for selecting items and determining their difficulty, i.e., their place on the ruler (Stenner and Smith 1982). For example, if I ask respondents if they are balanced when seated and 95% say yes and if I ask if they can climb a flight of stairs and only 60% say yes, then balancing when seated is considered less difficult (i.e., requires less ability) than climbing stairs. Climbing stairs will appear farther down one end of the ruler than balanced when seated. Likewise, those who answer yes to more difficult items are considered to have more ability.

Within Rasch, the validity of a measure is determined by how well the data (i.e., respondent answers to survey questions) fit the predictions of the mathematical model applied to a sample (i.e., is balanced sitting less difficult than climbing stairs?) These predictions are made explicit in the construct specification equation in terms of the amount of variance that we should expect to find around the mean with respect to the balancing and climbing questions. If the model correctly describes the probability distributions of people responding to questions about walking and climbing, then we can say that the observed rating scale data satisfy the measurement model, i.e., this is a valid measure. If the model does not correctly describe the observed data, then so much the worse for the data. Within the confines of Rasch, data that do not fit the model cannot be measured.

For Hobart et al., it is not only the validity of PROMs that Rasch improves but also their interpretability. As I discussed above, PROMs developed using CTT are notoriously difficult to interpret. The clinical significance of a 10-point increase on a particular scale is unclear in part because CTT can only deliver ordinal level data. They are also difficult to interpret because they have dubious validity, and if we do not know what something measures, it is difficult to interpret the significance of changes in scale scores. Rasch instruments, on the other hand, provide each item a precise location on the scale. If a mobility scale is validated, then an improvement from −1.5 to .5 indicates an improvement from, say, having the ability to climb the stairs to having the ability to walk on uneven ground. Every increase or decrease in ability is tagged on a Rasch scale to items of various difficulties and thus provides estimates of clinical significance for each move up or down the ruler.

Industry

PROMs that are potentially invalid, difficult to interpret, and of questionable sensitivity can be, as Hobart et al. argue, an impediment to accurate estimates of effect size and detection of clinical change (Hobart et al. 2007). Indeed, they suggest that the failures of clinical trials to yield larger numbers of effective treatments may be due to the lack of scientific rigor of their measuring instruments. It is not surprising that pharmaceutical companies keen to demonstrate the effectiveness of their products while using measures that will satisfy the FDA guidelines are eager to explore methods that will improve their success. And it is not only industry that wants to see the acceleration of medical product development. The FDA also shares this goal.

Public-private partnerships, such as Critical Path Institute (C-Path) created under the auspices of the FDA's critical path initiative program, aim to create drug development tools (DDTs): new data, measurement and method standards to accelerate the pace and reduce the cost of medical product development, etc. (Critical Path Institute 2015). They do so by coordinating collaborations among scientists from the FDA, industry, and academia. Essentially, C-Path puts industry scientists and academics together to develop tools that will enhance the ability of industry to develop medical products. The FDA then provides iterative feedback on the tools they create hopefully ending in the approval of the DDT for use in specific product development. PROMs are one of the four types of clinical outcome assessments eligible to qualify as a DDT (Food and Drug Administration 2007).

The FDA's DDT qualification program along with C-Path is one way to build on the FDA guidelines for the use of PROMs in medical product labeling in order to streamline the process for an instrument's acceptance by the FDA. It is also an opportunity for industry and academics to work together to further their individual ends and, in doing so, flesh out the FDA guidelines, i.e., the FDA is not specific in its guidelines regarding what psychometric methods should be used to establish validity, interpretability, etc. DDTs provide industry and academics with the opportunity to develop new standards for measurement and methods, thus opening up room for new psychometric methods, such as Rasch. Indeed Hobart et al. explicitly call for such developments in their work.

Partly through the work of Sergio Sismondo, the philosophical and bioethics community has learned to have a healthy skepticism of industry/academic partnerships. Much of Sismondo's work focuses on violations of publication ethics through ghost-managed research (see Sismondo's chapter "Hegemony of Knowledge and Pharmaceutical Industry Strategy" in this volume; Sismondo and Doucet 2010; Sismondo and Nicholson 2009; Sismondo 2007). He identifies the entangled nature of ghost management as practically expedient, but ethically troublesome. It is practically expedient because at least at first gloss (almost) everyone involved wins: pharmaceutical companies get more market value out of their publications if well-respected academics put their names on the manuscripts; academics get publications in notable journals; and the journals get well-cited manuscripts, which if published will produce revenue in the form of offprints purchased by industry (Sismondo and Doucet 2010). But ghost management is ethically troublesome—we might even say corrupt—because it reveals how extensively clinical research is driven by market concerns, which in turn begs questions about (1) the justification of subjecting human subjects to research and (2) the integrity of that research. It also intimates a kind of sad desperation among academics for high impact publications and involvement in large clinical trials. As Sismondo points out, what they are doing is unethical, but ambitious academics may have few other options (Sismondo and Doucet 2010).

Although my objective in this last section is not to reveal the kind of widespread corruption that Sismondo does in his work, I do want to suggest that the collaboration of industry and academics to develop DDTs should be critically evaluated. In what follows, I suggest how a PROM developed using Rasch—for the sake of argument, a DDT—could be co-opted to provide evidence of clinical change. Thus, not only should we critically evaluate the collaborative partnerships that C-Path facilitates, but also the use of Rasch as a value-neutral improvement to the scientific rigor of PROMs.

As I discussed earlier, Rasch, unlike CCT, makes use of a more robust measurement theory. As such, it tells us what to make of respondent answers to survey questions, e.g., when respondents answer yes to more difficult items, then they have more ability than those who answer yes to easier items. It also provides us with a ruler with specific item locations. Recall that the Rasch scale runs from plus to minus infinity, with the zero point at the place where the difficulty of the items in the survey is equal to the ability of the sample population. Each item is located on the ruler relative to the point at which there is equal probability of respondents answering “yes” or “no” to that particular item. In sum, Rasch provides a formal theory that tells us where to locate items and where to locate people. But Rasch does not provide an attribute theory that guides us in choosing the content of the scale, i.e., the items or questions.

To be sure, there are constraints in the items that are chosen, the most obvious being that the data resulting from them must coincide with the Rasch model. And as I discussed earlier, if the standard error estimates of adjacent items overlap, then those items are taken to be too similar to one another. Although this latter constraint is not an absolute constraint, since increasing the sample size will decrease the standard error estimates and possibly preserve the questions under consideration. In any case, I want to put aside these two constraints and instead focus on the lack of an attribute theory within the Rasch model.

Rasch lacks a theory regarding the *content* of its target construct. Moreover, unlike the measurement of time, these target constructs are not enmeshed within a robust science such as physics. For example, Rasch does not tell us what is important about a particular construct (e.g., mobility) and neither does psychology. Thus, it is up to researchers who develop such scales to try out different questions if and until the survey data yields a fit with the Rasch model. But without theoretical guidance regarding the content of the construct of interest, how can we determine the adequate sensitivity of a scale? It seems that in this regard, Rasch is no better than CTT and possibly worse.

How might the use of Rasch be worse than CTT when it comes to the sensitivity of a scale? The problem is that Rasch makes it too easy to create a measure that is calibrated to detect clinical change. Consider the following example. It is possible to take survey data from questionnaires such as the European Quality of Life Five Dimensions (EQ-5D) and model it using Rasch. Imagine that when we do so, we find, not surprisingly, that the EQ-5D’s five questions are relatively insensitive to change because they divide wide variables into only a few levels, i.e., mobility, self-care, usual activities, pain discomfort, and depression/anxiety. Earlier we discussed a similar problem regarding sensitivity in the context of CTT. In Rasch language, the EQ-5D is too easy, i.e., even respondents without a lot of ability can answer all the questions positively. For instance, eye problems, sleep problems, sexual functioning, memory problems, problems communicating poststroke, and fatigue are a few of the deficits to which the EQ-5D is generally insensitive.

Now, suppose that you were looking at the EQ-5D data because you were interested in whether or not it was the appropriate measure to use in a clinical trial to establish the effectiveness of a drug. You have the mean pretreatment scores of your target population and you know that pretreatment they already have has more ability than the EQ-5D is able to measure. If you want to show a clinical improvement, then you need a measure that is more sensitive. In the language of Rasch, you need a measure that can target a higher-functioning population, i.e., respondents with more ability. Because you already know the mean pretreatment scores, you have an idea where on the ruler you need to develop the scale in order to measure the change you anticipate. Moreover, the more responsive the rulers (i.e., the closer together each step on the ruler), the more likely you will find a clinically significant change.

I want to be very clear: I am not suggesting that anyone is disingenuously using Rasch to demonstrate clinical change. What I am suggesting is that Rasch represents an opportunity to increase the likelihood of finding clinical benefit, while the choice to use Rasch is presented as a matter of scientific rigor. I am not alone in recognizing that Rasch represents this opportunity. Indeed Hobart et al. admit that one criticism of more sensitive measures is that they will increase type 1 errors (false positives) (Hobart et al. 2007). But while they more or less dismiss this concern since blunt instruments are equally problematic, I think it is worth taking seriously.

One reason to do that is because science is a value-laden enterprise. Indeed, as Heather Douglas writes in *Science, Policy, and the Value-Free Ideal*, social and ethical values are necessary to any science that has a public role, i.e., any science that has a role in policy, medicine, technology, etc., as health measurement certainly does. Douglas's argument is twofold. First, she reminds us that our evidence always underdetermines what we should believe (Douglas 2009). We can see her point, if we attend to Rasch measurement scales. Here we see that our knowledge of a construct, including respondent data from items thought to be related to the construct, underdetermines how many questions we ought to ask and at what difficulty level we should target our efforts, including how sensitive the scale should be and if certain areas of the scale should be more sensitive than others. Put another way, there is always an element of uncertainty in the use of scientific evidence. This uncertainty is overcome only when scientists use their judgment to determine which standard, characterization, claim, or theory is indicated (Douglas 2009).

For the second part of her argument, she claims that when science has a public role, when, for instance, a study has the potential to affect public policy or medical treatment options, then the use of expert judgment draws on social and ethical values. When science has the potential to affect others—and it clearly does in the context of health measurement—then the values employed in using one's judgment should be connected to an individual's perception of what is at stake should one make a mistake. Scientists ought to evaluate the social and ethical consequences of error (Douglas 2009: 87). In other words, when considering how sensitive a scale should be, researchers should contemplate the social and ethical consequences of creating an overly sensitive scale that increases the likelihood of type 1 errors. Some of the consequences might be loss of public trust, overmedication, rising healthcare costs, and industry (dis?)satisfaction. Equally, researchers should consider the consequences of creating less sensitive scales that increase the likelihood of type 2 errors (false negatives). Some of these consequences might be increased cost and time to medical product development and increased patient suffering due to the delay in medical product development.

For Douglas, the solution to scientific disagreements that stem from differences in social and ethical value orientations is to make the values on which decisions or judgments are based more transparent (Douglas 2009). In the context of health measurement, we might begin by simply acknowledging their existence. When Hobart et al. criticize CTT as unscientific and suggest Rasch as a replacement in the name of scientific rigor, we might soften the critique by recognizing that the choice to use CTT over Rasch is not only a lack of sophistication and knowledge as they sometimes seem to suggest, but also a value choice of prioritizing expediency and simplicity. Moreover, even if Rasch does provide the basis for more scientific measurement scales—as I believe it does—supporters need to recognize the value-laden decisions that still characterize these scales. Without recognition of values we employ under conditions of uncertainty, we cannot evaluate them. If we do not evaluate them, then I worry that similar to the case of ghost management, we might find ourselves building measures to tailor the marketing needs of pharmacy.

References

BIPM (Bureau International des Poids et Measures). 2006. The International System of Units (SI). Sèvres: BIPM. http://www.bipm.org/en/si/si_brochure/. Accessed 16 Apr 2015.

Cano, S.J., and J.C. Hobart. 2011. The problem with health measurement. *Patient Preference and Adherence* 5: 279–290.

Chauhan, C. 2007. Denouement: A Patient-Reported Observation. *Value in Health* 10: S146–S147.

Critical Path Institute. 2015. http://c-path.org/about/. Accessed 7 Apr 2015.

DeMuro, C., M. Clark, L. Doward, E. Evans, M. Mordin, and A. Gnanasakthy. 2013. Assessment of PRO Label Claims Granted by the FDA as Compared to the EMA (2006–2010). *Value in Health* 16(8): 1150–1155.

Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Fayers, P., and D. Machin. 2007. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*, 2nd ed. Chichester: Wiley.

Food and Drug Administration. 2007. Drug Development Tools Qualification Programs > Clinical Outcome Assessment Qualification Program. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm. Accessed 7 Apr 2015.

Hobart, J.C., S.J. Cano, J.P. Zajicek, and A.J. Thompson. 2007. Rating Scales as Outcome Measures for Clinical Trials in Neurology: Problems, Solutions, and Recommendations. *Lancet Neurology* 6(12): 1094–1105.

Hunt, S.M. 1997. The Problem of Quality of Life. *Quality of Life Research* 6(3): 205–212.

Jaeschke, R., J. Singer, and G.H. Guyatt. 1989. Measurement of Health Status. Ascertaining the Minimal Clinically Important Difference. *Contemporary Clinical Trials* 10(4): 407–415.

McClimans, L. 2010a. Towards Self-Determination in Quality of Life Research: A Dialogic Approach. *Medicine, Health Care, and Philosophy* 13(1): 67–76.

———. 2010b. A Theoretical Framework for Patient-Reported Outcome Measures. *Theoretical Medicine and Bioethics* 31(3): 225–240.

———. 2011. Interpretability, Validity, and the Minimum Important Difference. *Theoretical Medicine and Bioethics* 32(6): 389–401.

———. 2017. Measurement in Medicine and Beyond: Quality of Life, Blood Pressure and Time. In *Reasoning in Measurement*, ed. N. Mößner and A. Nordmann. London: Routledge.

McClimans, L., and J.P. Browne. 2012. Quality of Life Is a Process Not an Outcome. *Theoretical Medicine and Bioethics* 33(4): 279–292.

Schwartz, C.E., and M.A. Sprangers. 1999. Methodological Approaches for Assessing Response Shift in Longitudinal Health-Related Quality-of-Life Research. *Social Science & Medicine* 48(11): 1531–1548.

Sismondo, S. 2007. Ghost Management: How Much of the Medical Literature Is Shaped Behind the Scenes by the Pharmaceutical Industry? *PLoS Medicine* 4(9): e286.

Sismondo, S., and M. Doucet. 2010. Publication Ethics and the Ghost Management of Medical Publication. *Bioethics* 24(6): 273–283.

Sismondo, S., and S.H. Nicholson. 2009. Publication Planning 101. *Journal of Pharmacy & Pharmaceutical Sciences* 12(3): 273–279.

Stenner, A.J., and M. Smith. 1982. Testing Construct Theories. *Perceptual and Motor Skills* 55(2): 415–426.

Streiner, D.L., and G.R. Norman. 2008. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford: Oxford University Press.

Tal, E. 2011. How Accurate Is the Standard Second? *Philosophy in Science* 78(5): 1082–1096.

Taminiau-Bloem, E.F., F.J. Van Zuuren, M. Visser, C. Tishelman, C.E. Schwartz, M.A. Koeneman, C. Koning, and M. 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(11): 1414–1428.

Terwee, C.B., F.W. Dekker, W.M. Wiersinga, M.F. Prummel, and P. Bossuyt. 2003. On Assessing Responsiveness of Health-Related Quality of Life Instruments: Guidelines for Instrument Evaluation. *Quality of Life Research* 12(4): 349–362.

Wyrwich, K.W., and V.M. Tardino. 2006. Understanding Global Transition Assessments. *Quality of Life Research* 15(6): 995–1004.