2018

Epistemological Issues in Quality of Life Measurement and their Implications

Laura M. Cupples
University of South Carolina

EPISTEMOLOGICAL ISSUES IN QUALITY OF LIFE MEASUREMENT AND THEIR IMPLICATIONS

by

Laura M. Cupples

Bachelor of Science
Davidson College, 2002

Master of Arts
University of South Carolina, 2013

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Philosophy

College of Arts and Sciences

University of South Carolina

2018

Accepted by:

Leah McClimans, Major Professor

Michael Dickson, Committee Member

Tarja Knuuttila, Committee Member

A. Jackson Stenner, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

I am especially grateful to my major professor, Leah McClimans, and my three committee members, Michael Dickson, Tarja Knuuttila, and Jack Stenner for their generous contributions of time, encouragement, and intellectual guidance over the last eight years. I also wish to acknowledge the support I have received from my parents, Carol and Tommy Cupples, not only during the dissertation process, but also throughout my graduate career. I wish to thank the University of South Carolina Philosophy of Science Dissertation Group for reading and commenting on early drafts of many of these chapters, as well as anonymous reviewers for *Studies in History and Philosophy of Biological and Biomedical Sciences* and *Theoretical Medicine and Bioethics*. My thanks to Kevin Timpe, Marcel Boumans, and Anna Alexandrova, who read and commented on early drafts of individual chapters, and to Carolyn Schwartz, Nancy Cartwright, Eran Tal, and Hasok Chang for helpful correspondence. I am grateful to the anonymous individuals who agreed to be interviewed about their experience with disability, and to Emily Mann who offered useful guidance on the interview and IRB process. Finally, I wish to thank Rosa Fuller and Michelle Panchuk for their friendship and encouragement.

ABSTRACT

This project explores the epistemology of quality of life measurement. Quality of life measures face a dual epistemic burden both to serve as data for the evidence-based medicine movement and to give patients a greater voice (McClimans 2017). Much time, effort, and money has been invested in these measures over the last 50 years, and yet their theoretical foundations remain weak. It is not clear whether these instruments succeed in measuring what they purport to measure, or what precisely is meant by "quality of life" (Hunt 1997).

The epistemic challenges faced in quality of life measurement are not wholly unique. I argue for an analogy between quality of life measures and physical measures first in terms of the dynamics of their development and second in terms of their model dependence. I argue, following van Fraassen (2008) and McClimans (2010b) that the dynamics of measure development have a hermeneutic structure. And, following Tal (2012) I argue that judgments about the validity, accuracy, and comparability of quality of life measures are model dependent.

However, there are important contrasts between quality of life measures and archetypal physical measures as well. The concept of quality of life cannot be standardized the same way the concept of temperature can. Its meaning remains open to interpretation. In part, this is because quality of life is an imperfectly understood subject matter (McClimans 2010b). We might also argue that quality of life is inherently subjective (Schwartz and Rapkin 2004), in at least some sense, and that it is a socially

constructed Ballung concept (Cartwright and Runhardt 2014). Because the meaning of quality of life cannot be standardized, outcomes may not converge around a single determinate value.

In the final chapter, I extend my epistemological survey to the topic of epistemic justice in quality of life measurement. I argue that the chronically ill and disabled are better situated epistemically to evaluate their own health states than the general public (Barnes 2009; Harding 1993; Paul 2014), and that justice requires us to place them at the center of such deliberations (Fricker 2007). This is particularly true when these valuations affect policy.

TABLE OF CONTENTS

LITERATURE REVIEW

In recent years, philosophy of science has seen a turn toward more practice-oriented scholarship. Along with interest in both scientific modeling and experimentation, this new trend has brought about a resurgence in the study of philosophy of measurement, and measurement epistemology. While much of the initial work in epistemology of measurement focused on the physical sciences, philosophers soon turned their attention to measurement in the social sciences as well. A small subgroup of philosophers of social science turned their attention to the measurement of well-being in its various forms. This dissertation adds to that body of scholarship.

Since the early 1990s, patient-reported outcome measures—survey instruments targeting quality of life and subjective well-being—have played a growing role in medicine and health policy. These measures address a dual epistemic burden. First, they serve as sources of data for evidence-based medicine. Second, they give patients a voice in quality of life issues (McClimans 2017). There are now thousands of these measures in use all over the globe, and millions of dollars have been spent on their development. They are used to guide policy, to influence resource allocation decisions, to determine the efficacy of candidate treatments, and to facilitate conversations about quality of life issues between patients and providers. Yet both philosophers and thoughtful researchers have complained that these measures have only a weak grounding in theory. It is not clear that they really measure what researchers intend for them to measure. Inferences are being drawn from measurement data that may not be justified, and with patient well-

being at stake, this is no small problem (Hunt 1997). In this dissertation, I examine the epistemological challenges associated with quality of life measurement and their ethical implications.

What follows is a brief review of the recent philosophical literature on measurement. I draw on much of this literature in the chapters that follow.

1. Measurement in the Physical Sciences

Hasok Chang (2004), Bas van Fraassen (2008), and Eran Tal (2012) all address epistemic challenges associated with measure development in the physical sciences, including the problem of coordination. Coordination involves both defining the quantity to be measured, and finding a principled way to assign numeric values to that quantity.

1.1. Hasok Chang

In 2004, Hasok Chang, one of the founding members of the Society for Philosophy of Science in Practice, published his book, *Inventing Temperature: Measurement and Scientific Progress*. Chang's monograph was a philosophical and historical exploration of the development of the modern temperature standard. Chang's (2004) work exposed what he calls the problem of nomic measurement. The problem of nomic measurement is not limited to temperature measurement, but is common to all measurement that rests on empirical law. In order to develop a system of measurement based in empirical law, we are reliant on experimental data, but that data can only be derived through measurement. To negotiate this epistemic circularity, Chang (2004) proposes a progressive, coherentist framework for measure development and scientific progress—a framework he observes at work in the decisions of historical actors responsible for developing the modern temperature standard. Chang (2004) dubs this methodological framework "epistemic

iteration". Epistemic iteration involves simultaneously obeying both the principle of respect, a directive to work within existing research paradigms as far as possible, and the imperative of progress, which pushes researchers to iteratively advance the epistemic virtues (scope, precision, etc.) of their measures (Chang 2004).

1.2. Bas van Fraassen

Like Chang, van Fraassen is concerned with problems of measure development, though he frames his discussion not in terms of the problem of nomic measurement but in terms of coordination. How do we go about systematically tying numerical values to a physical quantity? When developing a system of measurement, we ask both "what counts as a measurement of (physical quantity) X?" and "what is (that physical quantity) X?" (van Fraassen 2008, 116). According to van Fraassen (2008, 116), these two questions cannot be answered independently of one another. Like Chang, van Fraassen identifies an epistemic circularity that must be addressed if we are to move forward with measure development. Measurement relies on the identification of empirical regularities, but we cannot identify those regularities outside some framework for measurement. The method van Fraassen proposes for addressing that circularity, and for answering the interdependent questions posed above, is hermeneutical. We cannot take a view from nowhere, but must instead "presuppose an understanding both of the measurement procedure and of what is measured" (van Fraassen 2008, 121). For van Fraassen, these pre-understandings may be retrospective (we may come to understand what we once measured by looking back with a mature theory to aid us), or we may instead rely on and build from historical understandings of measurement procedure for a given quantity (van

Fraassen 2008). When we choose to build on historical understandings, this hermeneutical approach mirrors Chang's framework of epistemic iteration.

1.3. Eran Tal

Eran Tal has composed a model-based account of measurement epistemology that builds on the coherentist and historically contextualist approaches advocated by Chang and van Fraassen. In Tal's case, context is supplied by an idealized model of the measurement process. His (2012) doctoral thesis addresses three questions:

1. How is it possible to tell whether an instrument measures the quantity it is intended to?
2. What do claims to measurement accuracy amount to, and how might such claims be justified?
3. When is disagreement among instruments a sign of error, and when does it imply that instruments measure different quantities? (Tal 2012, ii).

Tal argues that while these questions are conceptually distinct, they are also epistemically entangled. They cannot be answered independently of one another. To address these questions, Tal examines the role idealized models of the measurement process play for metrologists studying the measurement of time. Tal argues that until a measurement process is subsumed under an idealized model, we cannot justify knowledge claims about measure validity, accuracy, or comparability. That is, we cannot solve the problems of coordination, accuracy, or quantity individuation (Tal 2012).

2. Measurement in the Social Sciences

Marcel Boumans observes that "because the traditional concept of measurement is based on measurement in physics, the scientific content of field science, as well as social science, was and still is contested" (2015, 28). Scholars studying measurement in the social sciences have worked to develop new theories of measurement that are better matched to their subject matter. Both philosophically minded psychometricians and

philosophers of science have contributed to recent philosophical work on measurement in the social sciences. I briefly examine contributions by Denny Borsboom (2005), Marcel Boumans (2015), and Nancy Cartwright (Cartwright and Runhardt 2014; Bradburn, Cartwright, and Fuller 2017).

2.1. Denny Borsboom

In his (2005) monograph, *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*, psychometrician Denny Borsboom evaluates traditional and modern psychometric models based on their philosophical implications. A central goal of his project is to answer the question, do psychological tests really measure something, and if so what do they measure? He describes this as the problem of validity and argues that, of the three types of measurement models he discusses, latent variable theory is best equipped to answer it. This is because latent variable theory, unlike classical test theory and the representational or axiomatic theory of measurement, is realist in its orientation. That is, it posits the existence of a latent attribute that operates as the cause of test outcomes (Borsboom 2005). Philosophically minded psychometricians tend to be far more concerned with the ontology of their target constructs than do scholars of physical measurement (personal communication with Leah McClimans 2016), and Borsboom is not the only psychometrician who insists that only a realist account of psychological constructs provides a sound basis for valid measurement (see Michell 1999).

2.2. Marcel Boumans

Scientific research, especially research in the physical sciences, prizes objectivity. Yet in Boumans's (2015) book *Science Outside the Laboratory: Measurement in Field Science and Economics*, he argues that measurement outcomes are a result of both mechanical

objectivity and expert judgment. He notes that, "When pure objectivity is impossible, we need to accept subjectivity to complement the incomplete objective knowledge. The question is not how to exclude subjective judgment, but rather where do we allow it, how much, and in what sense?" (120). This is especially true in the social sciences. One of the loci for expert judgment in the measurement process is in the way measurement is modeled. Modeling requires imagination, and is thus a subjective endeavor (Boumans 2015).

But how do we ensure that our subjective reasoning is rational? How do we avoid bias and common pitfalls of fallacious reasoning? The evidence based medicine movement, for instance, considers expert clinical judgment a poor form of evidence (Boumans 2015), and Tversky and Kahneman (1974) have demonstrated that even scientific experts are prone to predictable errors in reasoning, especially probabilistic reasoning. Boumans works to rehabilitate expert judgment by showing that the rationality of a conclusion is dependent on the modeling assumptions made by the reasoner, and by the way the reasoner interprets the problem presented to him or her. When evaluating the rationality of expert judgment, we should be cognizant of what Boumans (2015) calls the "two-model problem". Experimenters and their subjects (in this case clinical experts) often interpret research problems differently from one another; they operate with different modeling assumptions. Those modeling assumptions come together to determine which solutions count as rational for a given problem. What looks like an error in reasoning from one perspective may be perfectly rational from another (Boumans 2015).

2.3. Nancy Cartwright

While the physical sciences trade in objectivity, natural kinds, and law-like regularities, many of the concepts we are interested in measuring as social scientists will be socially constructed and value laden. The way we characterize and represent concepts in the social sciences will depend on our purposes, and the way we frame our research problems will depend in part on our values, both epistemic and social. Additionally, many of the concepts social scientists traffic in will be what Otto Neurath (1936), and later Cartwright and Runhardt (2014), and Bradburn, Cartwright, and Fuller (2017) call Ballung concepts. Rather than being defined by necessary and sufficient conditions, and thus supporting the law-like regularities we expect to see in the physical sciences, Ballung concepts are often characterized instead by family resemblance. The boundaries of Ballung concepts, such as disability, well-being, poverty, and civil war, tend to be unclear. Ballung concepts are usually best represented by a table of indicators corresponding to their properties rather than by a single value. For instance, a measure of poverty might specify the goods (nutrition, education, housing) a person or population's income provides them access too. It might also specify both absolute and relative poverty levels for a person or population (Cartwright and Runhardt 2014; Bradburn, Cartwright, and Fuller 2017).

3. Measurement of Quality of Life and Well-being

A small cohort of scholars studying measurement in the social sciences focus their attention on the measurement of happiness, quality of life, or subjective well-being. While the measurement of well-being shares many characteristics of social science measurement writ large, the philosophers I discuss below also bring to light concerns specific to these constructs.

3.1. Leah McClimans

Following van Fraassen's (2008) lead, McClimans brings her expertise on philosophical hermeneutics to bear on the epistemology of quality of life measurement. Specifically, McClimans (2010b) looks to Gadamer's (1991) logic of question and answer to illuminate the way respondents interpret questions about quality of life and to show researchers what they stand to learn from respondents' non-uniform interpretations. McClimans (2010b) argues that because quality of life is imperfectly understood, we should be asking genuine rather than merely apparent questions about how quality of life and other patient-reported outcomes behave. Researchers should put their understandings of quality of life at risk and allow them to be reshaped by what they learn from respondents (McClimans (2010b).

For McClimans (2010a), and for Gadamer (1991), this attitude of openness does not mean anything goes when it comes to the way we conceptualize quality of life. Not every interpretation will bear fruit. According to McClimans (2010a), fruitful conceptualizations of quality of life will be coherent. The whole will make sense of the parts, and vice versa. Furthermore, we should see our quality of life measures as seeking to uncover something true about the world (McClimans 2010a). Texts and text analogues for Gadamer (1991) are vehicles for discovering truth, not merely works of creative expression.

3.2. Dan Hausman

While McClimans focuses on the meaning of quality of life for both researchers and respondents, Hausman focuses instead on the relationship between the value of generic health states, health policy, and health economics. Hausman (2016) is skeptical that

researchers can measure health, and instead suggests that what they are really measuring is the value of certain health states. Conceptual tools such as the quality adjusted life year, or QALY, have been pressed in to service in recent decades by national health services and insurers seeking a way to value health states. QALYs and other utility measures of health typically rely on preferences of the general public or of individuals who are ill or disabled. These preferences are often elicited through time trade off or standard gamble tasks. While Hausman notes a number of problems with trying to value health states in this way, he admits that there does not seem to be a good alternative. When certain safeguards are in place—namely, preferences must be self-interested, consistent, and based on true belief—Hausman argues that preferences can serve as evidence for well-being (Hausman 2016).

3.3. Anna Alexandrova

Alexandrova (2017) approaches the study of well-being from a science and values perspective. If we are to measure well-being, we must have some idea what it is. We must have some theory about what makes for a good life. This means our science, and our measures, must be value-apt. Yet the three dominant theories of well-being advanced by philosophers—hedonism, preference satisfaction theories, and objective list theories—are of little help to scientists aspiring to study well-being. These theories are too general and abstract to assist practicing scientists, who need a means of making predictions and facilitating measurement (Alexandrova 2017). Alexandrova (2017) calls for more contextual, mid-level theories to bridge the gap between high level philosophical theorizing and empirical science. Well-being will look different for Syrian refugees than it does for British children in foster care, and it will differ again for patients with colon

cancer. If we are to measure well-being responsibly, and in a value-apt fashion, scientists and philosophers must work together to supply the conceptual foundations for these measures, and they must be open and transparent about the values they deploy (Alexandrova 2017).

4. Synopsis

This dissertation focuses on the epistemological challenges of measuring quality of life in clinical and health research settings. I examine important analogies and disanalogies between physical measures and measures in the human sciences, using quality of life as a case study. The first chapter of this dissertation examines parallels between the epistemic challenges of measure development in the physical and human sciences and examines the role hermeneutic methodology has to play in each case. Hasok Chang (2004) presents a progressive, coherentist strategy for measure development that he calls epistemic iteration that I argue, following Bas van Fraassen (2008), is hermeneutical in nature. Leah McClimans (2010b) describes a hermeneutic method for developing measures of quality of life and for achieving a better understanding of their target construct. Despite intriguing parallels between these two hermeneutic methods, an important difference remains. Unlike physical constructs, such as temperature, whose operationalizations and meanings are fixed (Chang 2004), the operationalization of well-being is not fixed, because the meaning of quality of life, and the meanings of the questions that target it, are imperfectly understood and thus open to interpretation (McClimans 2010b).

In Chapter 2 I argue, following work in the physical sciences by Eran Tal (2012), that models play a key role in supporting judgments about measure validity, accuracy, and comparability for measures of quality of life. While patient reported outcome

measures, including quality of life measures, typically have well developed statistical models, and some progress is being made in developing their qualitative models, theoretical models of quality of life are generally underdeveloped. There is no consensus in the field about what constitutes quality of life, as it relates to health, and yet, as Sonja Hunt (1997) has complained, there has been a rush to measurement, fueled by the needs of researchers and health policy makers for quality of life data. Mature statistical models, such as the Rasch measurement model, support judgments about measure comparability, but without good qualitative and theoretical models of quality of life, we should be skeptical of claims about measurement accuracy and validity.

In the third chapter, I ask what sort of concept quality of life is and what consequences this has for its measurement. Unlike physical measures, measures of quality of life are inherently subjective (Schwartz and Rapkin 2004), at least in some sense of the term. The way respondents interact with questionnaires that measure quality of life will depend on their cognitive processes of appraisal. For this reason, I argue that their outcomes are context dependent and imprecise. Quality of life is operationalized differently for different patients; individuals' judgments about their own or about hypothetical health states depend on their frame of reference, their sampling strategy, their standards for comparison, and their personal values and priorities (Schwartz and Rapkin 2004).

Quality of life is both socially constructed and value laden. Furthermore, according to Cartwright and Runhardt (2014) as well as Bradburn, Cartwright, and Fuller (2017), quality of life is a Ballung concept. Ballung concepts are not defined by necessary and sufficient conditions for concept membership, but instead rely on less

precise criteria such as family resemblance. The way we measure Ballung concepts will depend on our research purposes, as they can be instantiated—and thus operationalized—in a variety of ways. Whether we conceive of quality of life as inherently subjective or as a Ballung concept, its measurement outcomes will be context dependent. Rather than thinking of quality of life as a single valued construct, it is more realistic to conceive of it as a pluralistic quantity.

My final dissertation chapter focuses on the role of utility measures of health states in guiding resource allocation decisions in health care. Members of the general public are asked to evaluate the utility of hypothetical health states, most of which they have never personally experienced. Policy makers argue that self-evaluations of these health states by the disabled and chronically ill are unreliable because they are distorted by adaptive preference. Drawing on the resources of feminist standpoint epistemology, I argue that this assessment is unwarranted and that failure to take the evaluations of the disabled and chronically ill seriously is an epistemic injustice. Furthermore, because members of the general public routinely undervalue these health states relative to the disabled and chronically ill, distributive justice for health care resources is dependent on epistemic justice. Following Dan Hausman (2017), I suggest that health values might better be uncovered by deliberative focus groups. To be both just and epistemically sound, these groups should place the disabled and chronically ill at the center of their deliberations.

CHAPTER 1

MEASURE DEVELOPMENT AND THE HERMENEUTIC TASK[1]

1. Introduction

I examine the dynamics of measure development using two case studies: temperature, and health-related quality of life. I argue, following Bas van Fraassen (2008) and Leah McClimans (2010b) that in each case these dynamics have a hermeneutic structure. Furthermore, I show that Hans Georg Gadamer's (1991) philosophical hermeneutics in particular are an effective lens through which to examine the development of the temperature standard as described by Hasok Chang (2004). Despite similar grounding in hermeneutics, I note an important difference between measure development for temperature and for health-related quality of life. Namely, while the meaning of temperature can be standardized, the meaning of health-related quality of life cannot. A strategy of progressive coherentism leads to a determinate value for temperature, but the same cannot be said for health-related quality of life. This standardization of meaning for the temperature concept represents a limit to the analogy with hermeneutics. Finally, I argue that the indeterminacy we find in quality of life measurement is a result not only of analogy with the hermeneutic task, but of full-fledged participation in it.

[1] Cupples, Laura. 2018. "Measure Development and the Hermeneutic Task." To be submitted.

2. Setting the Stage

2.1. Van Fraassen, Coordination, and the Hermeneutic Task

In his 2008 monograph, *Scientific Representation: Paradoxes of Perspective*, Bas van Fraassen explains that solving measurement's problem of coordination is a hermeneutic task. When scientists work to develop new measures, they must seek out principles of coordination that link the construct to be measured to a specific value on a measurement scale. As they do so, they must try to answer two entangled questions: "(1) What counts as a measurement of (physical quantity) X? and (2) What is (that physical quantity) X?" (van Fraassen 2008, 116). These questions cannot be answered separately from one another, according to van Fraassen, and approaching them together seems to mire us in epistemic circularity. How do we determine the value of our target construct without an established measurement procedure, and how do we establish a measurement procedure without knowing what function relates the quantity we are able to observe to the quantity we are trying to uncover? Hermeneutics famously addresses itself to a similar form of circularity. We cannot uncover the meaning of a text or text analogue without presupposing some as yet unjustified, and possibly erroneous meaning coming in to the task (Warnke 1987). Somehow we must use this provisional assumption as a basis for revising our understanding moving forward—we must be able to correct our own understanding. How do we address ourselves to this apparent circularity, and what can measure developers learn from hermeneutics?

2.2. Chang's Progressive Coherentism

Like van Fraassen, Chang (2004) recognizes the epistemic circularity inherent in measure development. Chang describes what he calls the "problem of nomic measurement" as follows:

1. We want to measure a quantity X.
2. Quantity X is not directly observable, so we infer it from another quantity Y, which is directly observable.
3. For this inference, we need a law that expresses X as a function of Y, as follows: $X = f(Y)$.
4. The form of this function f cannot be discovered or tested empirically, because that would involve knowing the values of both Y and X, and X is the unknown variable that we are trying to measure (Chang 2004, 59).

The problem of nomic measurement is illustrated by researchers' attempts to progress from ordinal thermoscopes to numeric thermometers. In this case, X is temperature and Y is the volume of the thermometric fluid. Once fixed points at 0° and 100° Celsius had been added to thermoscopes, researchers needed to find a way to fill in the rest of the temperature scale. The question is, were they justified in assuming that temperature increased linearly with volume. It seems impossible to answer this question without knowing temperature values independently (Chang 2004). And yet, as Chang shows, by using a progressive, coherentist strategy, researchers were able to develop and justify the modern temperature scale.

Chang (2004) describes researchers as reliant on two methodological commitments: first, "the principle of respect", and second, "the imperative of progress" (2004, 44). The principle of respect involves a commitment to "respect the prior standard as far as it is plausible to do so" (Chang 2004, 44). We must recognize that our attempts at measure development are historically situated. We cannot, as van Fraassen observes, take a "view from nowhere", but must work within the tradition we find ourselves in

(2008, 122). Chang (2004) demonstrates this principle by describing the way early researchers relied on coherence with physical sensation to develop their temperature measures. They chose to respect sense data that taught them that thermometric fluids tend to expand with the addition of heat. While they were not fully justified in trusting sense data, choosing to treat it with too much skepticism would have made the advancement of thermometry impossible. Thus, sense data was treated with at least provisional trust. This trust was not indefeasible, as researchers left open the possibility that sensation might at times be mistaken. Indeed the hope was that thermoscopes built on the foundation of sense data would, when complete, be able to challenge that same sense data (Chang 2004).

Chang's second methodological commitment is to the imperative of progress. We want each new iteration of measure development to improve upon the epistemic virtues of its predecessors. Thus, we may look for improvements in the consistency, precision, or scope of our measures (Chang 2004, 44). Regnault's painstaking work testing the comparability of measurement values obtained using various thermometric fluids can be seen as an advance in both precision and consistency. Wedgwood's attempt to expand the standard into the high temperature range through pyrometry can be seen as an advance in scope. Progress is achieved through "creative evolution" (Chang 2004, 46), with each new step in measure development building on its predecessor, but not straightforwardly derived from it. Chang calls this process of creative evolution "epistemic iteration" (2004, 44-46), emphasizing that each step brings us closer to our epistemic goals.

Chang (2004) argues that a measurement system is not complete until researchers bridge the gap between the abstract and the concrete, between theoretical concepts and practice. Theoretical standards must be operationalized if they are to have empirical content. Researchers had established a working temperature standard long before Lord Kelvin tied temperature measurement first to Carnot's heat engine and then to the ideal gas law, but that working standard had not been married to a mature theory (Chang 2004). Chang (2004) argues that bringing the two together involved taking a theoretical system and creating an image of it—an image that would function as an idealization of a matching physical system. A series of iterative corrections were then made to bring the values generated by the two systems closer together. The end result was a determinate temperature value in each context (Chang 2004).

2.3 The Hermeneutic Task

The hermeneutic task is the task of interpreting the meaning of a text or text analogue. This task is worked out differently for different hermeneutic scholars. Romantic era scholars such as Schleiermacher saw their task as uncovering the creative intention behind a text. Because misunderstanding was the natural result of our attempts to interpret historical works, Schleiermacher emphasized the need for a rigorous hermeneutic method. This method was both grammatical and psychological: the grammatical method carefully examined dialect, sentence structure, and genre while the psychological method called for the reader to transpose himself into the author's position in an attempt to recreate the conditions and mindset under which the work was originally created. Finally, coherence between grammatical and psychological elements constrained possible interpretations (Warnke 1987).

Dilthey, like Schleiermacher, sought to uncover the original vantage point from which creative works were made. A member of the historical school, he sought a methodology that would both separate the human sciences from the natural sciences, and place the human sciences on a secure footing. He emphasized the need for a rigorous hermeneutic method that would make objective understanding in the human sciences possible. Dilthey's hermeneutic philosophy distinguished between two types of experience—scientific experience, which was repeatable and thus verifiable, and life experience, which could not be repeated in the same way. For Dilthey, life experience was the basis upon which the human sciences were built. Individually, we learn from life experience and begin to see the world in new ways. Collectively, life experience forms a kind of *Geist* or spirit that infuses a society's understanding of their shared history. Despite his insight that understanding is conditioned by historical experience, Dilthy, like Schleiermacher, believed that it was necessary to transpose ourselves into the position of historical actors if we wished to understand them and their original intentions. According to Gadamer, Dilthey's concern that human sciences be made objective undermined his more valuable insight into the historical situatedness of understanding (Warnke 1987).

Hans Georg Gadamer (1991) broke with both Schleiermacher and Dilthey in important ways. First, he distinguished between two types of understanding: understanding of an author's creative intentions, and substantive understanding of a truth claim. For Gadamer, the purpose of hermeneutics was primarily the second. He eschewed calls for a rigorous or defined hermeneutic method, and instead worked to describe the conditions under which understanding was made possible. For this reason, Gadamer did not believe that we should attempt to transpose ourselves into the position

of the original author of a text. Instead, he believed that the perspective brought to bear on a text by the interpreter was invaluable (Gadamer 1991; Warnke 1987).

Gadamer (1991) adopted Dilthey's insight that as interpreters, our understanding is inevitably historically situated. He emphasized the essential role of tradition in our coming to understand a text. Tradition, Gadamer argues, strongly influences our prejudices about the meaning of a text. Beginning our encounter with a text with some pre-understanding or prejudice, Gadamer argues, is a necessary condition for coming to understand its meaning more fully. In seeking to understand a text, we place our pre-understandings at risk and allow them to be challenged by the text itself. We engage in dialogue with the text, posing questions that lead to a greater understanding of its subject matter and allowing the text to answer those questions. In this way, we achieve what Gadamer calls a "fusion of horizons" (1991, 306-307)—a new, shared understanding of the subject matter (Gadamer 1991; Warnke 1987). This fusion of horizons, however, is not the final word when it comes to the meaning of a text. Tradition and the subjectivity of interpreters are ever evolving, and so are the pre-understandings that may be brought to bear on a historical text. When trying to learn the truth about a subject matter, we continue to address new questions to the text that treats it. Our horizons remain open, as does our interpretation of the text (Gadamer 1991). Thus, according to Gadamer (1991), we must learn to accept a certain degree of indeterminacy with regard to meaning.

3. Measure Development as a Hermeneutic Task

3.1. Whose Hermeneutics?

I argue that among the hermeneutic philosophies available to us, Gadamer's (1991) is best suited to illuminate the process of measure development. It is important that

Gadamer sees texts as truth claims, and not merely as aesthetic objects.[2] The goal of Gadamer's hermeneutics is reach agreement about the meaning of the subject matter at hand through truth-seeking dialogue, and not merely to uncover the author's creative intentions (Gadamer 1991; Warnke 1987). Science, like Gadamer's hermeneutics, is typically seen as a progressive enterprise. It seeks a nearer approximation, if not to truth, then at least to epistemic virtue. A hermeneutics that merely seeks to uncover the creative intentions of an author or artist, and defines that interpretation as the only correct one, seems therefore, to be an inadequate model of the scientific enterprise. Such a hermeneutics would mirror the naïve conventionalism that both Chang (2004) and van Fraassen (2008) reject as a solution to the epistemic circularity encountered in measure development. A conventionalist chooses one instrument and its outcomes as definitive of the standard, and defines all other instruments as being in error. This choice of standard may be entirely arbitrary, or it may be made to maximize the simplicity of the measurement model. Van Fraassen argues that conventionalism fails because of its ahistorical approach to the problem of coordination, while Chang observes that choosing an arbitrary standard is unlikely to satisfy scientific researchers given their traditionally realist attitudes toward measurement.

Gadamer's (1991) claim that tradition carries normative, though not indefeasible force in our attempts to make sense of a text is also important. Tradition informs our pre-understandings of a text, constraining subjectivity in interpretation. Those provisional pre-understandings are a condition for the possibility of uncovering meaning (Warnke

[2] This truth Gadamer seeks is not the singular Truth with a capital T of realist philosophy of science. Instead, Gadamer believed that a text could express multiple truths about a subject matter. A plurality of meanings are possible.

1987). Theory carries a similar normative force in measure development (Chang 2004). To deny the normative role of theory would be to advocate for operationalism. Each instrument would define its own concept, and there would be no need to seek agreement between instruments (Chang 2004). The hermeneutic equivalent would be unconstrained subjectivism of meaning. Any interpretation would be legitimate. But according to Chang (2004) and van Fraassen (2008), the problem of measure development is no more solved by operationalism than it is by conventionalism. Chang (2004) worries, first of all, that operationalism fruitlessly multiplies measurement concepts. Scientists' purposes are better served by unified concepts.

If van Fraassen (2008) and Chang (2004) reject conventionalism and operationalism, what is their solution to the problem of epistemic circularity faced by measure developers? Both propose a process of iterative refinement and correction of candidate standards that brings measurement values into agreement with one another and with theory. This process is dialogical, and truth seeking, as are Gadamer's hermeneutics (Chang 2004). We create a new iteration of our standard, or gain a new understanding of a text, by putting the old one at risk and allowing ourselves to learn from the encounter. In so doing, we come to an agreement about the content of the text, or the value of our measure.

3.2. Gadamer and Chang

In Section 3.1, I explained at a general level why Gadamer's (1991) approach to hermeneutics is a better analogue to the dynamics of measure development than other historical approaches to hermeneutics. In Section 3.2, I explain in greater detail why the

development of the temperature standard in particular, as described by Chang (2004), is hermeneutical in Gadamer's sense.

For Gadamer (1991), the subjectivity of our interpretations is constrained in a number of ways. First, it is limited by the "anticipation of completeness" (Warnke 1987, 82-91). The anticipation that the text has something true to tell us and that it forms a unified and coherent whole informs and regulates our provisional interpretation of the text. That anticipation is also the basis upon which we test those provisional interpretations against the text itself. We assume that the parts of the text will form a self-consistent whole when interpreted correctly, and if they fail to do so, we have reason to revise our interpretation (Gadamer 1991; Warnke 1987). Warnke (1987) for instance, gives the example of provisionally taking a book to be a detective story, and then finding that under that interpretation, the elements of the plot fail to form a coherent whole. Our encounter with the text, along with our anticipation of its completeness, lead us to revise our initial interpretation (Gadamer 1991; Warnke 1987).

Second, the subjectivity of our interpretation is limited by the need for coherence between the historic text and its present day application. The application of a text to our own lives is an essential part of the interpretive act (Gadamer 1991; Warnke 1987). Gadamer and Warnke illustrate the need for both the normative force of tradition and for present day application through an analogy with Aristotelian ethics. For Aristotle, we must not only possess theoretical knowledge about general ethical norms, but we must also translate those norms into concrete and situationally appropriate action (Aristotle 1999; Gadamer 1991; Warnke 1987). If we are to understand the virtue of generosity or justice or courage writ large, we must know how to act it out in a specific situation.

How do these dynamics align with Chang's (2004) progressive coherentism? Recall that measure development for Chang involves commitments to both the principle of respect and the imperative of progress. Like Gadamer, Chang recognizes the indispensability of tradition—in his case, scientific tradition. Whether relying on sense data or ordinal thermoscopes, Chang's researchers provisionally accepted the authority of scientific tradition. Doing so was a precondition of scientific progress (Chang 2004). Having done so, they tested experimental findings based on those traditional methods against an anticipation of completeness. These constraints guided researchers in establishing more coherent measurement systems by allowing them to refine and correct the systems they had provisionally endorsed (Chang 2004).

Furthermore, recall that for Chang (2004) a measurement system is incomplete until the gap between the abstract and the concrete has been bridged. Abstract concepts, such as temperature, must ultimately be married to operation, as it was when Lord Kelvin linked temperature measurement first to Carnot's heat engine and then to the ideal gas law (Chang 2004). Abstract concepts carry with them the normativity of tradition, or if you'd rather, of a general ethical principle, while their operationalization is the acting out of that norm or tradition in a concrete, empirical context (Gadamer 1991; Warnke 1987). In this way, the norm, or the abstract concept, takes on empirical content (Chang 2004; van Fraassen 2008). Temperature becomes more than a mere mathematical term; it becomes a target of measurement.

4. Hermeneutics and Measure Development in the Human Sciences

In this section, I extend the analogy between measure development and the hermeneutic task to encompass measures in the human sciences as well, relying Leah McClimans's (2010b) philosophical work on the measurement of quality of life.

4.1. McClimans, PROMs, and Gadamer

McClimans addresses the relationship between measure development and Gadamer's philosophical hermeneutics in her (2010b) paper "A Theoretical Framework for Patient-Reported Outcome Measures." She sees patient-reported outcome measures (PROMs) as texts, the meaning of whose subject matter is to be uncovered. PROMs are survey instruments used to measure health-related quality of life and subjective well-being. They pose questions to respondents about pain, mobility, functional status, fatigue, emotions, and social connectedness. When respondents answer these questions, they give researchers access to phenomena that would not otherwise be observable (McClimans 2010b).

Like temperature measurement, the development of PROMs is plagued by epistemic circularity. Recall Chang's problem of nomic measurement. Here quality of life is the unobservable phenomenon we wish to measure, and the factors that contribute to it—factors that respondents report on such as self-care, mobility, and emotional well-being—are made observable for researchers through those reports. Presumably quality of life is a function of these factors, but exactly how they come together to make up quality of life is unknown. There is no gold standard for quality of life—no independent means of measuring it apart from discerning its relationship to these factors (McClimans 2010b). At present, different researchers, and different measures, posit different relationships

between quality of life and mobility, pain, fatigue, etc. Apart from a general consensus that quality of life is multidimensional, encompassing physical, emotional, and social well-being, there is no widely accepted theory of quality of life to act as a norm or constraint on its many measures (Hunt 1997). McClimans's purpose in her (2010b) paper is to provide a framework for greater theoretical development for PROMs.

McClimans (2010b) sees PROM development as dialogical and truth seeking. She looks to Gadamer's (1991) logic of question and answer to help make sense of the conversation taking place between researchers and respondents. Importantly, the subject matter of their conversation, quality of life, is imperfectly understood. This imperfect understanding has consequences for the type of questions it's appropriate for us to ask about the subject matter (McClimans 2010b). McClimans (2010b) argues, following Gadamer (1991), that we should be asking genuine questions about quality of life instead of merely apparent questions. We ask genuine questions when we do not fully understand a subject matter and wish to learn more about it. Genuine questions are open to interpretation and do not have pre-determined answers. We ask apparent questions on the other hand when a subject matter is well known to us and when there are definite criteria for a correct response (McClimans 2010b).

Unfortunately, because PROM questions are typically standardized, they function as apparent questions and close the door for us to learn more about the meaning of quality of life from the respondents we are in dialogue with (McClimans 2010b). McClimans (2010b) suggests that we can reopen that door by supplementing standardized PROMs with think aloud studies and qualitative interviews. Initially, when researchers are trying to define the concept to be measured, respondents may be queried about what aspects of

quality of life are most relevant to them given their particular illness or disability. For instance, breast cancer survivors undergoing reconstructive surgery nominated the following outcomes as important to their surgical experience and recovery: satisfaction with breasts, overall outcome, process of care, as well as physical, psychosocial, and sexual well-being (Pusic et al. 2009, 345).

When researchers begin to test a preliminary measure, respondents are often asked to speculate about the meaning of the questions they are faced with and to share their reasons for responding to those questions the way they do. While typically these interviews are conducted to ensure that respondents are interpreting questions as researchers intend, they need not serve that role (McClimans 2010b). McClimans (2010b) argues that we should take respondents' answers and interpretations seriously, even when they are unexpected and challenge our preconceived ideas about quality of life. For instance, does a disability that limits mobility necessarily affect one's quality of life, and if it does, how does it affect it? Does adaptation to the circumstances of one's illness play a legitimate role in quality of life, or is that role illusory? When researchers are willing to put their own understandings at risk and to learn from respondents, a new and shared understanding—what Gadamer calls a fusion of horizons—becomes possible (Gadamer 1991; McClimans 2010b; Warnke 1987). In this way, researchers can better understand the meaning of quality of life data and the inferences they draw from those data are more likely to be sound (McClimans 2010b; Schwartz and Rapkin 2004).

Importantly, McClimans (2010b) rejects the arguments of some researchers (see e.g., Schwartz and Rapkin 2004) that the meaning of quality of life is inherently subjective, and perhaps even idiosyncratic. Like Gadamer (1991), she sees certain

constraints on interpretation at work in the hermeneutics of quality of life. By now, those constraints should be familiar—coherence, the anticipation of completeness, and the indispensability of application as a part of the interpretive act. While respondents' interpretations should be taken seriously, and we should grant them the authority to teach us more about quality of life, they are not indefeasible. We cannot say that anything goes when it comes to quality of life (McClimans 2010a). McClimans (2010a) offers the example of a woman who has been culturally conditioned to accept female circumcision as a normal social practice. While this woman may claim that female circumcision is not qualitatively different from male circumcision and is perfectly compatible with a good quality of life, if we consider the coherence of her claim there are legitimate reasons to be skeptical. Unlike male circumcision, female circumcision tends to be practiced in the context of oppressively patriarchal cultures that limit the well-being of women in a number of significant ways (McClimans 2010a).

5. Commonalities and Differences Between Temperature and Quality of Life

Using case studies in temperature and quality of life, I have tried to show that the dynamics of measure development in both the physical sciences and the human sciences are hermeneutical. Furthermore, I have argued that Hans Georg Gadamer's (1991) hermeneutics are a better model for the dynamics of measure development than earlier hermeneutic philosophies. In this section, I will demonstrate an important difference between measure development in the physical sciences and the human sciences, and how that difference is rooted in Gadamer's hermeneutics. Namely, I will show that the meaning of temperature is standardizable, while the meaning of quality of life is not (McClimans 2010b). As a consequence, temperature outcomes in a given situation can

be made to converge around a single, determinate value, while quality of life outcomes cannot. This difference is rooted in an incomplete analogy between temperature measurement and the hermeneutic task.

5.1. Standardization of Meaning

In order to see measurement outcomes converge around a single value, we must standardize the meaning of the concept being measured. For measurement concepts, that meaning is jointly determined by theory and operation through a process of progressive coherentism. Thus the meaning of temperature is determined in part by norms set forth by the ideal gas law and the theory of the heat engine, and also in part by functioning thermometers that operationalize the temperature standard (Chang 2004). For a text, meaning is determined hermeneutically through genuine, truth-seeking dialogue resulting in a fusion of horizons. Tradition, subjectivity, and the "thing in itself" are brought together for the interpreter, who must make sense of the text before her (Gadamer 1991; Warnke 1987).

It is common for immature measures to lack grounding in theory. Measurement practices fall into place and are widely agreed upon long before those practices are married with theory (Hacking 1983). Chang's (2004) narrative about the development of the temperature standard illustrates as much, as the numeric thermometer had been established well before Lord Kelvin anchored the temperature scale to the ideal gas law and thereby establish abstract norms for its meaning. While the temperature standard has slowly come to maturity over the course of centuries, PROMs are, by contrast, relative newcomers having appeared on the scene in the 1960s and 70s both as part of the social indicators movement, and risen significantly in popularity of use around 1990. Sonja

Hunt (1997) has complained about the widespread proliferation of quality of life measures in the absence of solid theoretical grounding, and her complaint remains valid twenty years later.

As things stand now, the meaning of quality of life I primarily defined operationally. As such, it varies from measure to measure. What's more, if we adhere to McClimans's (2010b) commitment to take unexpected interpretations of PROM questions seriously, we open the door to still further operational variance as different respondents may, in effect, be answering different questions from one another. This variance is not without bound, but is still problematic for those who would try to nail down a standardized meaning for quality of life (McClimans 2010a). McClimans (2010b) sees this variance in interpretation and conceptualization as a perennial factor in quality of life measurement. While researchers and certain cohorts of respondents may reach a fusion of horizons regarding the meaning of quality of life, that shared understanding will not be universal, and it will not be finally determinate any more than the meaning of a historical text. Instead, as Anna Alexandrova (2017) has argued, we should see the meaning of quality of life as varying with context. Indeed, this is why PROMs must be newly validated for different populations of respondents or when they are put to work in different contexts (Food and Drug Administration 2009).

5.2. Convergence of Measurement Values

According to Chang (2004), temperature measurement is regulated by the ontological principle of single value. That is, researchers are committed to finding or constructing a single value for temperature each time it is measured. Historically, however, different thermometers often gave different temperature readings from one another when used to

measure the same body. These disparate outcomes were brought into accord with each other, or made to converge, by iteratively correcting each outcome and bringing different operationalizations into alignment with theory (Chang 2004).

Convergence around a single value is a hard won achievement, and not one that is guaranteed (Chang 2004). Without a widely accepted theoretical model of the measurement system, without a determinate meaning for the measurement concept, it is difficult, though not impossible, to make the case for choosing a single outcome over others. Regnault, for instance, was able to endorse the outcomes provided by the air thermometer over those of the mercury and spirit thermometer by testing for consistency. He found that while different air thermometers typically agreed well with one another, different mercury and spirit thermometers did not. However, there was no guarantee that Regnault's strategy would be successful. The choice of thermometric fluid might well have been underdetermined by his findings (Chang 2004). For this reason, theory has an important role to play in providing a normative standard.

For instance, consider Eran Tal's (2012) account of the construction of the standard second. The standard is an idealized one whose meaning is defined by a theoretical model of a cesium fountain clock. Because (1) that standard can never be perfectly realized by an actual clock, and (2) an abstract idealization alone does not have empirical content, metrologists bridge the gap between the two by de-idealizing the standard. De-idealization involves identifying sources of error and uncertainty and taking them into account. It allows researchers to correct the outcomes given by real cesium fountain clocks and to bring them into alignment with the single value endorsed by the ideal standard (Tal 2012).

Yet as I have argued above, there is no widely accepted theoretical definition for quality of life (Hunt 1997; McClimans 2010b). There is no ideal model to provide a norm for the correction of disparate measurement values. While not indefeasible, respondents' subjective conceptualizations of quality of life and their individual interpretations of PROM questions play an important role in shaping the meaning of quality of life, and thus in its measurement (McClimans 2010b). When a measurement concept does not have a determinate meaning, or a theoretical definition, different operationalizations and different measurement values often remain permissible. I argue that this indeterminacy is a feature of quality of life measurement (McClimans 2010b).

Insofar as a single theoretical definition, or a single conceptual meaning for target constructs is possible in the physical sciences, measure development in the physical sciences seems, in fact, to diverge from the dynamics of the hermeneutic task. For Gadamer (1991), there are always new questions to ask of a text, and new pre-understandings to bring to bear on that text as tradition evolves. For this reason, new and different understandings are always possible. The perspective brought to bear on a text by the historically and socially situated interpreter is an essential part of the fusion of horizons that determines the meaning of a text. Because that perspective is ever changing, that meaning is not determinate (Gadamer 1991; Warnke 1987).

While measure development in the physical sciences is in many ways analogous to the hermeneutic task, that analogy has limitations. The most important of these limitations is the determinacy of meaning for the measurement concept and the resultant convergence of measurement outcomes around a single value. PROM development represents a more complete analogy with the hermeneutic task, as the meaning of its

target concept remains indeterminate. Indeed, the relationship goes beyond mere analogy. PROMs, according to McClimans (2010b), are texts whose subject matter is imperfectly understood and whose meaning is to be uncovered. These instruments are hermeneutic objects, and the researchers and respondents who interact with them are interpreters of meaning.

6. Conclusion

In this chapter I have shown that the dynamics of measure development are analogous to the hermeneutic task as described by Hans Georg Gadamer (1991). This is true not only for archetypal measures in the physical sciences, such as temperature measures (van Fraassen 2008), but also for measures in the human sciences, such as PROMs (McClimans 2010b). Chang's (2004) work on the development of the temperature standard, and McClimans's (2010b; 2010a) work on quality of life measurement, illustrate that measure development is dialogical, is shaped by tradition, and is constrained by coherence. Yet this analogy has certain limitations, at least for physical measures. Insofar as a determinate and standardized meaning for measurement concepts is achievable, the analogy with Gadamer's hermeneutics is broken. In this way, the development of quality of life measures bears a closer resemblance to the hermeneutic task than the development of the temperature standard.

CHAPTER 2

THE EPISTEMOLOGICAL ROLES OF MODELS IN HEALTH SCIENCE MEASUREMENT[3]

1. Introduction

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 (Tal 2012). 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.

[3] Cupples, Laura. 2017. "Epistemological Roles of Models in Health Science Measurement." In Leah McClimans (Ed.) *Measurement in Medicine: Essays in Assessment and Evalutation*. London: Rowman Littlefield International. Reprinted with permission, March 6, 2018.

While other authors have discussed the roles of qualitative models (McClimans 2010b), statistical models (Streiner, Norman, and Cairney 2015; Bond and Fox 2007), and theoretical models (Rapkin and Schwartz 2004; Stenner et al. 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.

2. 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.

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. Finally, 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).

3. 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 2010b; 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 2008 and McClimans 2011)), patients may see a broad range of abilities as indicative of relatively good mobility (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 prefer their understandings of health status and health-related quality of life. While patients' interpretations are not indefeasible, researchers should generally work to build qualitative models that describe patients' actual 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 2010b).

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 et al. 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; 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 better qualitative models of the measurement process.

The good news is that many 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 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.

4. 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, Norman, and Cairney 2015) and the models used by Rasch represent how patients interact with test items to produce an outcome or test score (Stenner et al. 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.

4.1. 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, expressed in Equation 2.1, posits three variables: a true score (T_T), an observed score (T_O), and a random error term (E).

$$T_O = T_T + E \qquad \text{(Equation 2.1)}$$

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, Norman, and Cairney 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.[4] Instead of providing

[4] 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.

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, Norman, and Cairney 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 2010b; 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.[5] 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

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

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.

4.2. 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.[6] 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 et al. 2009).

[6] 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.

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?"[7] 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 et al. 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. Equation 2.2 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. Equation 2.2 is graphically represented in Figure 2.1.

$$P(x_i = 1) = \frac{e^{(b - di)}}{1 + e^{(b - di)}} \qquad \text{(Equation 2.2)}$$

[7] PROMIS measures are actually built and analyzed using item response theory's two parameter model for measurement rather than using Rasch. However, they involve the same unidimensionality and gradations of difficulty as Rasch measures.

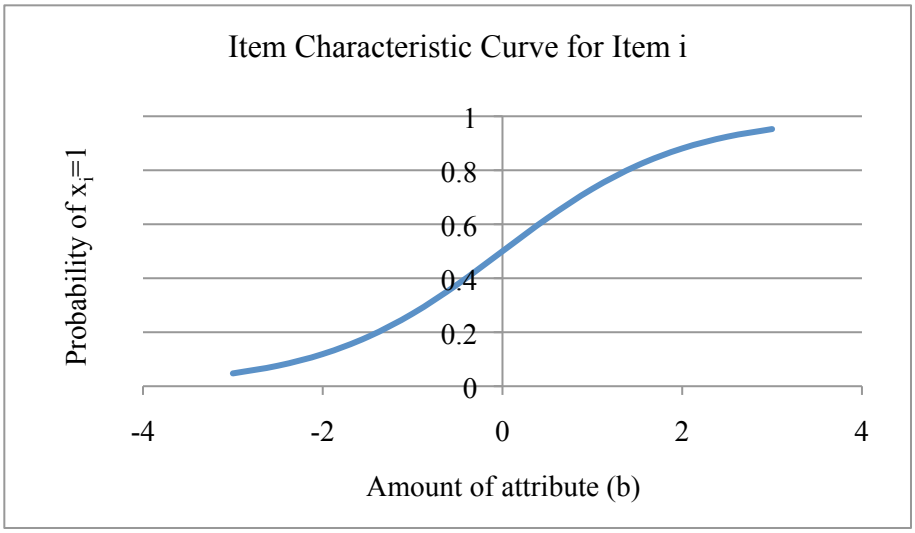


Figure 2.1. Item characteristic curve showing the probability of the respondent choosing the answer $x_i = 1$. The difficulty of item i is set to 0 logits.

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, estimations of item difficulty do 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 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.

5. 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 2010b).

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 he claimed he was only a little bit limited, yet by all accounts he was more physically limited than when he responded the first time (McClimans 2010b; Westerman et al. 2008, 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 2010b, 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 2010b). 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.[8] When testing these theoretical models, we can compare our empirical estimations of item difficulty

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

(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 would allow 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.[9] But 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 2010b; 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

[9] 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.

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.

6. 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, if we can develop them, would help us distinguish legitimate variation in the target attribute from patient error, and would help us establish the operational accuracy of Rasch measures.

CHAPTER 3

QUALITY OF LIFE: A CONTEXTUALLY VALUED, PLURALISTIC CONSTRUCT[10]

1. Introduction.

When measures of physical constructs are operationalized in different ways, we take pains to ensure that the outcomes of those operations converge around a single value. In the case of physical measures, we iteratively correct disparate outcomes until they come to agree with the outcomes predicted by an idealized model (Chang 2004). As we do this, we may also adjust the model itself to bring its predictions in line with empirical data. I argue that we should not expect this convergence of values for patient reported outcome measures (PROMs) of quality of life, which is contextually valued and pluralistic. I examine two complementary characterizations of the concept of quality of life. Together these characterizations imply that quality of life measures may have multiple legitimate outcomes—outcomes that are not in need of correction and that need not agree.

First, according both Cartwright and Runhardt (2014) and Bradburn, Cartwright, and Fuller (2017), quality of life is a Ballung concept—a fuzzy bordered cluster concept whose instances are unified by family resemblance rather than by necessary and sufficient conditions for concept membership. Because of their ubiquity in the social

[10] Cupples, Laura. 2018. "Quality of Life: A Contextually Valued, Pluralistic Construct." Submitted to *Studies in History and Philosophy of Biological and Biomedical Sciences*.

sciences, Ballung concepts are often socially constructed and value laden, and they are measured in different ways depending on the purposes of the investigator. Quality of life meets both these criteria. For instance, some PROMs are used as secondary endpoints in randomized controlled trials to determine the efficacy of candidate treatments, others are used as guides to resource allocation, and still others are used to facilitate conversations between patients and providers. As with many Ballung concepts, outcomes for quality of life measures often take the form of a multi-dimensional profile. Yet the content of that profile varies from instrument to instrument depending on researchers' purposes and their subsequent conceptualizations of quality of life (Cartwright and Runhardt 2014).

Second, Schwartz and Rapkin (2004) have argued that the measure of quality of life is inherently subjective. I flesh out what is meant by "subjectivity" in this context. According to this subjective model, they way we evaluate quality of life depends on individual cognitive processes, as well as personal experiences of health and illness, and the respondent's values and priorities. Individual patients and their proxies will evaluate quality of life differently both from one another and from themselves over time depending on their processes of appraisal. While respondents' interpretations are not indefeasible, and are subject to certain constraints, Schwartz and Rapkin (2004) argue that there are many legitimate conceptualizations of quality of life, and thus many legitimate interpretations of measurement items. Thus, in addition to the variance we see due to instrumentation, we will also see variance within the same instrument due to differences in subjective interpretation and the cognitive process of appraisal.

Thus, quality of life is contextually valued in more than one sense of the word. First, the constituents of quality of life will vary depending on the purpose of

measurement and the way researchers conceptualize the construct. But this does not mean that patient reported outcomes are straightforwardly determined by the way researchers' questions interact with clinical facts. We must also take into account the conceptualizations and interpretations of respondents. Just as researchers may legitimately have many purposes for measuring quality of life and may thus choose to conceptualize it in many ways, respondents may legitimately understand and interact with these measures in a variety of ways depending on the cognitive processes they deploy. In sum, this collision of contextualities suggests that, unlike the outcomes of physical measures, quality of life outcomes need not be single-valued, or even defined by a unique profile of indicators, and the outcomes associated with different operationalizations of the construct should not be expected to converge.

2. Physical Measures and the Ontological Principle of Single-Valuedness

When physical and humanistic constructs are operationalized in different ways, those separate operationalizations often produce distinct measurement outcomes. An air thermometer and a mercury thermometer may, for instance, give different temperature readings for the same beaker of water. When our measurement outcomes disagree, we normally suppose that one or both of those outcomes must be in error, or at least that it is appropriate to take steps to bring them into agreement. We assume that once our thermometers are properly calibrated, and random and systemic sources of error are brought under control, there is some single, determinate temperature value to be found or constructed.

Hasok Chang (2004) argues that measures of physical constructs, such as temperature, are governed by the ontological principle of single-value. If temperature is

to be a meaningful physical concept, it is assumed that it must carry a single value in any given instance. For this reason, even when temperature is operationalized in multiple ways, measurement outcomes are intentionally brought into agreement with one another. When values disagree, they must each be iteratively corrected in light of the predictions of an idealized measurement model. Sources of error and uncertainty must be identified, and measurement outcomes must be adjusted accordingly (Chang 2004; Tal 2012).

Eran Tal (2012) describes how this modeling process works for time measurement and the standardization of the second. No physical clock defines the duration of a standard second; instead, the definition of the second is based on the radiation period of a cesium atom at ground state at a temperature of absolute zero (Tal 2012, 34). This idealized definition is unrealizable in practice. Thirteen cesium fountain clocks serve as primary, though imperfect, realizations of the standard second for metrologists, who model sources of systematic error and uncertainty in their operation. By accounting for these sources of error, metrologists are able to "de-idealize" their model of the standard second and measure the accuracy of their primary realizations (Tal 2012). Thus, they are able to take an unruly collection of thirteen distinct outcomes, corresponding to thirteen imperfect realizations, and correct those outcomes to bring them into alignment and construct an authoritative time standard.

While the ontological principle of single-value is an appropriate constraint in the domain of physical measurement (e.g., time and temperature), and is indeed a precondition for meaningful measurement for those quantities, I will argue that it is not an appropriate constraint for at least some humanistic measures, and specifically that it is not appropriate for the PROMs. There is no single idealized model that standardizes

quality of life measurement. As such, there are multiple acceptable operationalizations of quality of life, each producing their own legitimate outcomes. While these differences are not indefeasible, and are subject to some constraints, differences between these outcomes should not, in general, be conceived of as measurement error, and therefore should not be corrected. I describe two complementary characterizations of the concept that together imply that it carries a plurality of contextual values rather than a determinate single value.

3. First Conceptualization: Quality of Life is a Ballung Concept

Wittgenstein argued in his *Philosophical Investigations* that just as members of families resemble one another without necessarily sharing any one defining trait, games also share a family resemblance. Some games are played with teams, others one on one, and still others are solitary pursuits. Some involve careful strategy, while others are primarily a source of amusement. Because there are no necessary or sufficient conditions for concept or category membership, family resemblance concepts will not have clearly defined borders. But Wittgenstein argues that these concepts are still useful and significant. It is still meaningful to tell someone that a certain practice is a game, just as it is meaningful to tell someone that a woman resembles both her aunt and her brother (Wittgenstein 1973).

Nancy Cartwright and Rosa Runhardt (2014), and later Bradburn, Cartwright, and Fuller (2017), make use of a similar framework for conceptual classification in their recent revitalization of the Ballung concept. Ballung concepts are fuzzy bordered cluster concepts. While many Ballung concepts are united by family resemblance, instances need not resemble one another (personal communication with Cartwright, 2017). The

commonality between Ballung concepts and family resemblance concepts lies in the lack of necessary and sufficient conditions for concept membership. Ballung concepts include exemplars such as well-being, disability, poverty, and civil war. Ballung concepts are different from natural kinds or homeostatic property clusters in that the relationship between properties is not, for the most part, causal. Good psychological well-being, for instance, is not significantly more likely to be present for a patient with good mobility than for one with poor mobility. Furthermore, because of their frequent appearance in the social sciences, Ballung concepts are often socially constructed rather than natural (Cartwright and Runhardt 2014; Rubin 2008; Bird and Tobin 2017).

Otto Neurath was concerned by the ubiquity of Ballung concepts in the social sciences (Cartwright and Runhardt 2014). He worried that, because of their fuzzy criteria for membership, these concepts would not support scientific laws the way more clearly defined physical concepts do (Bradburn, Cartwright, and Fuller 2017). And indeed, even today we find it more difficult to make predictions based on social scientific theory than physical theory. Additionally, Ballung concepts deployed by the social sciences are often normative, defying science's value free ideal. Both poverty and disability are value-laden concepts, and the way we characterize them depends on those values.

The measurement of Ballung concepts presents a particular challenge. First we must deal with problems of classification and make decisions about what particulars do and do not fall within a given category. These decisions, and others about how we will quantitatively represent a Ballung concept, are dependent in large part on our research purposes (Cartwright and Runhardt 2014). Consider disability, for instance. Whether we classify a given condition as a disability will depend on whether we are determining a

person's eligibility for social security benefits, their need for accommodation at school or in the workplace, their need for medical intervention (if we subscribe to a medical model of disability), or their status as someone properly advocated for by the disability rights movement.

Decisions about how to represent and measure Ballung concepts are also dependent on how we define other subsidiary concepts. Whether a conflict is deadly enough to be classified as a civil war, and whether that conflict is genuinely two sided or merely genocidal, may depend on who we classify as a combatant and who we classify as a civilian, and on which deaths are attributed directly to the conflict in question, and which deaths are instead attributed to lack of medical resources or to famine (Cartwright and Runhardt 2014).

Ballung concepts can be operationalized in different ways, and they will carry different measurement values depending on how they are constituted. Cartwright and Runhardt (2014) argue that Ballung concepts are typically better represented as a table of indicators than as a single value. Poverty measures might include indicators telling us a person or group's net annual income, what statistical quartile they fall into relative to other earners in their society, and more importantly, what resources—including education, daily nutrition, and housing—that income allows. Depending on our purposes in measuring poverty, different sets of indicators may be chosen in different contexts.

Bradburn, Cartwright, and Fuller (2017) argue that quality of life is an archetypal Ballung concept. There is no consensus on necessary and sufficient conditions for the good life, and there are now thousands of PROMs, each operationalizing quality of life in a different way (Hunt 1997). Measurement outcomes, whether they take the form of a

table of indicators or a single aggregate outcome, vary with specific measurement purposes as encapsulated in particular instruments, and measures that aggregate relevant factors into a single outcome will weight or prioritize those factors in different ways.

The Short Form-36, for instance, was developed as a generic measure of health status to be used in population surveys and in the evaluation of health policy (McDowell 2006). It breaks quality of life down into eight dimensions: physical functioning, role limitations due to physical health problems, bodily pain, social functioning, general mental health, role limitations due to emotional problems, vitality, fatigue, and general health perceptions. The measure generates both an eight dimensional profile score, and two sum scores—one for physical well-being and one for mental well-being (McDowell 2006). The World Health Organization (WHO) Quality of Life Scale, on the other hand, was designed to measure the effect of health problems in patient's lives, facilitate communication between patient and clinician, and to determine the efficacy of candidate treatments. It was also meant to be a tool to guide health policy. Its developers envisioned an instrument that could be modified to be appropriate across cultures (McDowell 2006, 619). The instrument measures quality of life across six dimensions: physical health, psychological health, level of independence, social relationships, environment, and spirituality. In addition to a six dimensional profile, it also generates a general quality of life or health perception score (McDowell 2006, 620).

Like many Ballung concepts in the social sciences, quality of life is both socially constructed and value laden. Its makeup is contingent on social, cultural, and personal values as well as on the professional interests of researchers. It is for this reason that the WHO Quality of Life Scale is modified to be culturally specific rather than merely

translated into other languages. Not all cultures will value independence in the same way, for instance. It is also the impetus behind idiographic measures of quality of life—that is, individualized measures that allow respondents to identify and weight important quality of life domains according to their own values and priorities—such as the SEIQoL.

But if quality of life outcomes are contingent on our purposes for measurement, are there any purposes that are more legitimate than others? Is there any way to privilege one operationalization and corresponding quality of life outcome over others, given that its measure is contingent on various values and purposes? And are there any values or conceptualizations of the good life that are more defensible than others? While it makes sense that some operationalizations of quality of life—some instruments—would be better suited to particular purposes than to others, there are a wide variety of legitimate purposes for quality of life measurement (see McClimans 2010b for further development of this discussion). Admittedly, few researchers present a clear account of why their instruments target the content they do, or for how and why quality of life should vary with the factors they choose to focus on (Gill and Feinstein 1994). Established instruments are often repurposed, sometimes against the objections of their creators (Hunt 1997). This may be a good reason to question the validity of quality of life measures—to question whether they truly measure what they set out to measure—but it does not change the fact that many purposes and corresponding operationalizations are permissible.

This does not mean that anything goes when it comes to quality of life measurement. Just as the concept of a game is still significant, despite a lack of necessary and sufficient conditions for inclusion in the category, so too is the concept of quality of life. We may have good reasons to reject some conceptualizations of the good

life. For instance, the good life may be incompatible with unremitting pain or with a lack of access to necessary medical care (McClimans 2010a; McClimans 2010b). We have reasons to be skeptical of measures that do not take these values into account. McClimans (2010a) argues that our conceptualization of quality of life should be bound by concerns for coherence and that those conceptualizations should be truth seeking. The way we take account of individual factors believed to contribute to quality of life (e.g., pain, mobility, fatigue) should cohere with our sense of the good life as a whole. If, for instance, we believe that well-being is contingent on a respondent having access to a certain set of goods or capabilities (see Sen 1993; Nussbaum 2001)—capabilities that facilitate a life of meaningful choice—we should be skeptical of a measure of quality of life whose data are inconsistent with that view.

4. Second Conceptualization: Quality of Life Measurement as Inherently Subjective

Social scientists Carolyn Schwartz and Bruce Rapkin (2004) argue that quality of life measurement is also inherently subjective. The plausibility of this argument depends, in large part, on what exactly is meant by "subjective" and on what sense of "objective" it is contrasted with.

For some quality of life researchers, subjectivity and objectivity are related to particular theories of well-being implicit in the measure. Both Derek Parfit (1984) and James Griffin (1989) describe three broad theories of well-being that we might choose from. The first two, experiential theories and desire satisfaction theories, are subjective. Experiential theories of well-being hold that quality of life depends on our personal mental states. For example, hedonistic theories that equate well-being with happiness would fit into this category. Desire satisfaction theories posit that our well-being depends

on the extent to which our individual goals and desires are met. Such theories may or may not require that our goals be rational or well-informed. This constraint pushes desire satisfaction theories in the direction of objectivity (Brock 1993). A third group of theories, sometimes called objective list theories, claim that there is some fact of the matter about what basic goods well-being requires (Parfit 1984; Griffin 1989; Brock 1993). Sen's capabilities approach to well-being is one such theory (1993; see also Nussbaum 2001).

While some PROMs claim to be based on particular models of well-being, such as the capabilities approach, many other take no stand on the nature of well-being writ large. The majority of measure developers consider such questions to be the province of philosophers rather than scientists (Alexandrova 2017). Havi Morreim (1986) contrasts objective measures of quality of life, which incorporate both individual clinical data and societal norms about well-being, to subjective measures, in which clinical data are contextualized by individual values and conceptualizations of the good life. Objective measures of well-being rest on a community consensus about requirements for well-being, while subjective measures depend instead on individual perspectives (Morreim 1986). Differences in values and priorities, as well as differences in the way individuals conceptualize the good life, lead respondents to interpret the questions in quality of life measures in non-uniform ways. That is, they lead respondents to operationalize quality of life differently both from one another (a patient and her spouse may evaluate that patient's quality of life differently from one another) and intrapersonally over time (say at time of diagnosis and after months of debilitating chemotherapy).

Given that Parfit (1984), Griffin (1989), and Morreim (1986) leave room for both subjective and objective conceptualizations of quality of life, as well as both subjective and objective measurement, how do Schwartz and Rapkin (2004) defend their claim that quality of life measurement is inherently subjective? While Schwartz and Rapkin's sense of subjectivity is similar in some ways to Morreim's (it acknowledges variability in patient values and priorities), it is less philosophical. As social scientists, Schwartz and Rapkin (2004) are focused more on the cognitive processes involved in appraising quality of life. They appeal to the science of survey design and administration to make their case that individuals interpret survey items in non-uniform ways.

The cognitive process a respondent goes through when formulating an answer to an evaluative question is called his or her method of appraisal. Appraisal comprises four factors: a frame of reference, a sampling strategy, a standard of comparison, and a combinatory algorithm (Rapkin and Schwartz 2004). These factors, which are shaped by a person's individual concept of quality of life, determine how an individual will interpret the measure's questions. First, a respondent's frame of reference is the range of experiences he or she considers relevant to the question at hand. For instance, a patient undergoing treatment for cancer might consider their overall experience with illness since they became symptomatic, or they might focus more on the recent side effects of a particular treatment (Schwartz and Rapkin 2004). They may consider only concerns they associate with the illness at issue, and not those associated with aging. Within an individual's frame of reference, he or she uses a sampling strategy to access particular memories or experiences (Rapkin and Schwartz 2004). For instance, a patient's most recent or most painful experiences may be easier to access than others. Or a patient may

recall an occasion when their health prevented them from engaging in an activity that was especially meaningful to them.

When asked about their quality of life or about limitations in functioning respondents also employ a particular contrast class or standard of comparison. They may compare their level of well-being to what it was prior to becoming ill, or to some personal ideal, or they may compare their level of function to how well they perceive others with the same diagnosis or of the same age to be doing (Rapkin and Schwartz 2004). Depending on who or what a patient compares herself to, she may assess her quality of life or her limitations differently.

Finally, when making the judgments necessary to answer questions about quality of life, respondents will prioritize some health concerns over others. A respondent's values and priorities will be shaped by culture, personality, life experience, and the ways he or she has adapted over time to illness or disability (Schwartz and Rapkin 2004; Schwartz et al. 2007). For instance, a respondent who is used to an active lifestyle may be more concerned about chronic fatigue than about pain control, while an individual with a permanent physical disability may have learned to prize social and psychological well-being over mobility. Patients who have grown up in a culture that values autonomy and self-reliance may evaluate their well-being differently from patients who have been raised with a strong sense of duty to family or community.

Schwartz and Rapkin (2004) contend that even questions that seem to be straightforward, or that are written to constrain one or more of these factors, inevitably leave room for variance in interpretation:

> Common phrases like "bodily pain" or "some help" are highly subject to interpretation. Efforts to introduce more precise terms may reduce

> variance in appraisal parameters, but narrower concepts like "headache" or "unaided every time" can still connote different meanings. Even ratings of very specific functions ("difficulty lifting your arm over your head") may be affected by individual differences in standards of comparison ("compared with how I used to be?" or "compared to my mom after she had the same surgery?") and salience ("how often do I really need to lift my arm like that?") (Schwartz and Rapkin 2004, 4).

How bad does a headache have to be to affect one's quality of life, for instance? And if a patient is able to get around independently using only his cane, does that count as unaided? If the respondent can lift her arm over her head but experiences pain or stiffness each time she so, can she really say that she is able to lift her arm in this way?

Schwartz and Rapkin (2004) argue that because quality of life measurement is inherently subjective, there are multiple legitimate operationalizations of quality of life. When respondents appraise quality of life in individualized ways—accessing different memories or deploying different standards for comparison—they are not, for the most part, acting in error—even if their interpretations do not match those intended by researchers. There is no single ideal model of quality of life measurement whose outcomes others ought to be assimilated to. This is true not only across instruments, as entailed by the fact that quality of life is a Ballung concept, but also within the same instrument, as Schwartz and Rapkin's (2004) work demonstrates.

Here again we could ask whether there are any subjective interpretations or methods of appraisal that should be excluded—is any perspective permissible? McClimans (2010a) uses the example of a respondent who, due to socio-cultural influences, claims that female genital mutilation is compatible with a good quality of life. Perhaps this respondent sees the practice as no different from male circumcision as

widely practiced in the West. Should we be skeptical of such a claim? McClimans argues that we should indeed be skeptical, since female genital mutilation is usually accompanied by other forms of gendered oppression that we would not see as consistent with the good life for women (McClimans 2010a). Once again, concerns for coherence constrain plausible conceptualizations of quality of life.

Researchers may attempt to constrain processes of appraisal with explicit item instructions—instructions to focus on the recent side effects of treatment for instance. They may specify the contrast class or standard of comparison to be used by a respondent: compare your current health to how you were feeling a month ago. Yet McClimans (2011) argues that even such specification fails to fully constrain respondents' interpretations, since there are many ways that my quality of life now may differ from what it was a month ago, and it is not always clear where symptoms end and side-effects begin. Furthermore, many quality of life researchers intentionally leave their questions ambiguous in these respects in order to increase their bandwidth—that is, the range of interpretations and quality of life experiences they can target (Schwartz and Rapkin 2004). Indeed, Schwartz and Rapkin (2004) argue open-endedness may actual be more desirable than greater specificity. Even if we could constrain all appraisal factors and force respondents to interpret questions in standardized ways, they suggest that doing so would subvert one of the primary purposes of quality of life measurement—i.e., it would undermine efforts to give patient perspectives a voice on quality of life issues (Schwartz and Rapkin 2004). As the authors say, "We would not hand an individual an inkblot and ask, 'What does this butterfly look like to you?'" (Schwartz and Rapkin 2004, 5). Leaving room for a variety of interpretations allows respondents to project their own

values and perceptions onto the measure. And if we are willing to listen to patients, and to attempt to uncover how they appraise quality of life and why, we will learn more about their experiences of the good life.

5. Implications for Quality of Life Measurement

I have shown that quality of life can be operationalized in many ways. There are a range of constituent factors that could come together to make up quality of life, from physical and emotional functioning, to pain and fatigue, and symptom load to spirituality, environment, and independence. These factors will vary from instrument to instrument according to the purposes of the researchers involved. Furthermore, within the same instrument, respondents will often interpret survey items in non-uniform ways based on their processes of appraisal. The range of memories they access will vary, as will their standards for comparison. If we agree that quality of life is a Ballung concept, and that it is inherently subjective (at least in Schwartz and Rapkin's sense of the word), it looks as if multiple operationalizations of quality of life are permissible and legitimate, as are the values they generate. Quality of life is pluralistic and its value is context dependent.

It might seem, given my characterization of quality of life as a Ballung concept, that the intentions of researchers should be the primary determinants of any model of quality of life. While conceptualizations of quality of life vary, they would be determined in each case by the interaction between researchers' purposes and clinical facts. Yet Schwartz and Rapkin's argument that quality of life is inherently subjective complicate that picture, as do best practices in quality of life measurement that call for patient input in instrument design. Quality of life outcomes are also shaped by the way respondents understand and interact with researchers' instruments. Not only do patients'

cognitive processes of appraisal vary, so too do their values and priorities. All of these factors come together to make quality of life outcomes context dependent and pluralistic.

The measurement of quality of life will not be defined by a single idealized model, and its many operationalizations are not, for the most part, imperfect approximations in need of correction or alignment. Instead, each stands as a legitimate realization of the construct. Our goal should not be the construction of a single, convergent value for quality of life outcomes, as is the goal with physical constructs. We should instead aim for a better understanding of how quality of life varies with context, whether that context is provided by the purpose of research, respondents' cognitive processes, or the values and priorities that have arisen out of their personal experiences with health and illness.

CHAPTER 4

EPISTEMIC JUSTICE, HEALTH STATE VALUATIONS, AND THE QUALITY ADJUSTED LIFE YEAR[11]

1. Introduction

The quality adjusted life year (QALY) is a generic measure of disease burden used in health economic analysis. By estimating the degree of improvement in a patient's quality of life associated with a given intervention, and also taking into account how many years the patient is projected to enjoy that improvement, policy makers attempt to gauge the relative worth of that intervention—i.e., they can gauge the number of QALYs that will be gained by it—and weigh that against the cost of the intervention. From a utilitarian standpoint, it is argued that because resources are limited, society and the third party payers that function as its surrogates should try to maximize the good they are able to do for patients with available health care dollars.

Yet who, from an epistemic standpoint, should provide values for the health states in question? At present, values are typically solicited from a representative sample of the general public rather than directly from disabled and chronically ill persons. This is the case despite the fact that, arguably, only this latter group has the necessary experience to place an informed value on these states. Nonetheless, a number of arguments, both ethical and epistemic, are typically advanced in favor of soliciting values from an

[11] Cupples, Laura. 2018. "Epistemic Justice, Health State Valuations, and the Quality Adjusted Life Year." To be submitted to the *International Journal of Feminist Approaches to Bioethics*.

unrestricted sample. Some of these arguments are ethical in nature—(1) that, in the interest of fairness to all, the insurance principle prohibits parties who are already experiencing the states to be covered by a policy from making resource allocation decisions about those states, and (2) that because members of the general public are each stakeholders in resource allocation decisions, their views should be solicited in making health state valuations that will determine the outcome of those decisions. The third argument is epistemological, and holds that disabled and chronically ill people are not credible reporters because they exhibit an adaptive preference for their own health states. In this paper, I focus on this epistemological argument, and argue that it does not constitute an adequate reason for soliciting values primarily from healthy and able-bodied individuals.

I draw on Elizabeth Barnes's work on disability and adaptive preference, on feminist standpoint theory, and on Laurie Paul's work on transformative experience to explain why, from an epistemic standpoint, the perspectives and valuations of the disabled and chronically ill should be valued. I also draw on interviews with disabled and chronically ill individuals to flesh out and complicate the story told by philosophers. In spring of 2018, I solicited volunteers who self-identified as disabled or chronically ill via social media. Those volunteers were asked:

1. First, it would help me to know in broad terms what sort of illness or disability you have.
2. If you were not born with a disability or chronic illness but acquired or developed it at some point, what was the experience of becoming ill or disabled like?
3. What sorts of misconceptions do people tend to have about your life as a disabled or chronically ill person?
4. Do people who are familiar with your illness or disability tend to over or underestimate its impact on your life?

5. Do you think it's possible for someone who hasn't experienced your illness or disability to accurately imagine what it is like to live with it?
6. Are there things you value about your experience of illness or disability?

Approval for these interviews was secured through the University of South Carolina's internal review board.

2. Background: Definitions and Current Practices

Disability and chronic illness are not coextensive. It is possible to be disabled without being ill, just as chronic illness need not be disabling. But in the context of health utility measurement, we are primarily interested in those chronic illnesses that are in some way disabling. The EuroQol (EQ-5D) for instance, targets multidimensional health states, which include: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (Williams 1995). Similarly, the Health Utilities Index (HUI(3)) targets: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain (Horsman et al. 2003). Neither of these indices exhaustively captures all possible elements of chronic illness or disability, but they do aim to capture disabling elements of paradigm cases.

Researchers and policy makers currently elicit the health state valuations they use to calculate QALYs by surveying a representative sample of the general public rather than directly targeting the chronically ill or disabled. They ask members of the public about the values they would assign to a number of hypothetical health states and then aggregate the results. Different health states of varying severity are described in a few sentences, and then respondents are presented with either a time trade off or standard gamble task. In the time trade off task, respondents are asked if they would prefer, for instance, to live 10 years in a state of moderate disability, or to live 7 years in a state of full health. Time intervals are adjusted until the respondent is indifferent between the

two outcomes. In the standard gamble task, respondents might be asked, for instance, whether they would prefer to undergo a treatment that had a 70% chance of restoring them from a state of moderate disability to full health and a 30% chance of causing immediate death, or whether they would prefer a second treatment that had a 100% chance of restoring them to a state of only mild disability. These methods assume that years lived in full health are more valuable than years lived in a state of disability or illness, and thus expect respondents to be willing to trade off years of life for the opportunity to live in a healthier or more able-bodied state. Once utilities have been assigned to the various health states presented in the task, those values are correlated with the health state profiles of actual illnesses and disabilities as described by persons affected by those conditions.

Health state valuations range in magnitude from zero (the purported value of immediate death) to one (full health and functionality). Life with a colostomy, for instance, is valued at approximately 0.8 (Ubel, Loewenstein, and Jepson 2003). This means that members of the general public would typically trade ten years of life with a colostomy for eight years without one. A single year lived with a colostomy is worth 0.8 QALYs, and ten years lived in such a state would be worth eight QALYs. If a medical intervention could repair a patient's gastrointestinal tract, that patient would gain 0.2 QALYs (the difference between 0.8 and 1.0) for each year gained without a colostomy, assuming the patient was otherwise healthy and able-bodied. Resource allocation decisions are based on how many QALYs can be gained per unit cost. The idea is to maximize the number of QALYs that can be gained by the population given limited financial resources.

3. Chronic Illness, Disability, and Epistemic Justice

For a variety of reasons, researchers and policy makers do not target disabled or chronically ill individuals when they solicit values for a given health state. Researchers argue that the values members of this community place on their own health states may be skewed by adaptation, so that their purported valuations are unreasonably high (Salomon et al. 2012). Others appeal to "the insurance principle", asserting that we cannot justifiably determine what treatments should be covered by an insurance plan or health service after we've already become ill (Hadorn 1991). They argue that allowing patients to assign values to pre- and post-intervention states would violate this principle. Still others argue that since the financial resources being mobilized to pay for treatments belong to the general public, members of the public should be the ones consulted about how those resources should be used (Dolan 1999).

The choice to use valuations from the general public is controversial, primarily because the healthy and able-bodied tend to systematically undervalue the quality of life of persons with chronic illnesses and disabilities compared to patients' own valuations. The deflated valuing of disabled and chronically ill health states is ethically problematic because it can indicate to policy makers that it is less worthwhile to save the life of a disabled person than a non-disabled person (Ubel, Loewenstein, and Jepson 2003). Thus, for both ethical and epistemic reasons, some (Carel 2014; Nord et al. 1999) have argued that it would be better to seek valuations from patients who have actual experience living in these states.

Drawing first on Miranda Fricker (2007), and later on Havi Carel and Ian James Kidd (2014) and Kidd and Carel (2017), I argue that epistemic justice requires us to seek health state valuations from persons with disability or chronic illness in greater numbers than are included in a representative sample of the general public. According to Fricker, epistemic injustice is "a wrong done to someone specifically in their capacity as a knower" (Fricker 2007, 1). One form of epistemic injustice is testimonial injustice. This injustice occurs when an individual's credibility is systematically underestimated due to identity prejudice—i.e., prejudice based on the social group to which a person belongs. Fricker illustrates testimonial injustice by examining the plight of Tom Robinson from Harper Lee's classic novel, *To Kill a Mockingbird*. Tom, a black man living in 1930s Alabama, is falsely accused of rape by a young white woman and her father. His courtroom testimony is systematically discounted by the all white jury considering his case, despite the evidence corroborating his account. Fricker's explanation for the jury's verdict is that, given the racist prejudices of the day, they see black men as untrustworthy sources of testimony.

While Fricker (2007) focuses on marginalized groups such as women and persons of color, Carel and Kidd (2014 and 2017) extend the discussion to include patients. I argue that not only patients (i.e., the chronically ill), but also the disabled can be subject to systematic epistemic injustice—especially when disability is viewed under a medical model. Health utility indices and policy-makers' responses to these indices seem to assume a medical model of disability, i.e. a model in which disabilities are medical problems that need to be treated or cured (Amundson 2005). I am not committed to such a model. Indeed, I argue that such a model is inconsistent with the way many disabled

persons see themselves. In many cases, I agree with disability rights advocates that what is disabling about a given condition is not the condition itself but the lack of accommodation that person encounters in his or her social environment.

Carel and Kidd argue that patients are vulnerable to epistemic injustice, since medical professionals and policy makers often underestimate both their credibility and the relevance of their medical testimony. Patients are, for instance, said to be overly emotional (and hence irrational) and prone to include irrelevant and unhelpful information in their accounts of their experiences with illness (Carel and Kidd 2014 and Kidd and Carel 2017). Scientifically objective, clinical accounts of illness given by medical professionals are epistemically privileged over the more phenomenological accounts rendered by patients. Yet if we are to heed the call of the disability rights movement for "nothing about us without us", we must be attentive to the testimony of the disabled and chronically ill and grant it the credibility it deserves.

I suggest that an example of testimonial injustice toward the disabled and chronically ill is the assumption that their valuations of health states are not credible because they tend to adapt to their conditions. "Adaptive preference" is a preference for a suboptimal state brought about through a constricted set of options (Elster 1983; Nussbaum 2001). Elizabeth Barnes explains adaptive preferences by way of John Elster's illustration of a hungry fox who, unable to reach the grapes he wants, decides that grapes are too sour for foxes anyway (Barnes 2009; Elster 1983). Martha Nussbaum (2001) and Daniel Brock (1995; 2005) have argued that disability positive testimony, i.e. testimony from disabled persons that they value or even prefer being disabled to being

able-bodied, is evidence of adaptive preference, since, clearly, disabled states are suboptimal. [12]

Barnes disagrees with this position and (2009) argues, that, ceteris paribus, we should view individuals' assessments of their own well-being as more accurate than those of third parties. Indeed, there should be a high epistemic bar for discounting this assumption. This position is consistent with the arguments presented in feminist standpoint theory, that marginalized persons often have access to knowledge not available to more socially privileged groups (Harding 1993; Medina 2013). I asked one woman with obsessive compulsive disorder if she thought others were able to understand her experience: "I think it takes a lot of extra effort to understand what OCD feels like if you don't have it, and even then, some parts are so hard to describe, I don't know how other people could understand it when I barely do." An autistic man with right arm paralysis notes: "I find that people observing my life (but not closely connected) tend to underestimate the impact of my disability, as if my life and disability were incompatible. That statement applies to both my physical disability and autism."

Barnes (2009) further notes that we cannot use the popular assumption that disability is suboptimal to epistemically undermine disabled persons' positive assessments of their own well-being, since the fact that disability is suboptimal is exactly what is in question. The credibility of disabled and chronically ill persons' testimony

[12] Martha Nussbaum adheres to a capabilities approach to well-being, which claims that a good quality of life depends on our access to certain essential goods—goods that give us opportunities to flourish. One of those essential goods, on Nussbaum's view, is good health (2001). Brock holds a similar view. He believes that well-being requires a wide range of opportunity, and that since disability, by definition, involves a limitation in one or more species-typical functional abilities, it limits the opportunities of those who experience it (2005).

about their own well-being is typically devalued precisely because they live with disabling medical conditions, and it is thought that the unfortunate necessity of coping with those conditions makes them incapable of rationally and accurately valuing their own health states. And while there may be instances in which the testimony of the disabled or chronically ill is less than credible, a systematic prejudice against accepting their testimony is epistemically unwarranted.

While some may claim that third person testimony about quality of life is more objective than first person testimony (see for instance Brock 1995 and 2005), Amundson (2005) warns that this perspective is often a manifestation of societal prejudice or of the stigma surrounding disability. For instance, third parties may assign low values to disabled states because of the societal prejudice that living with a disability is inherently a misfortune and is worthy of pity. Rather than replacing subjective assessments by the disabled with objective assessments of quality of life by the healthy and able-bodied, Amundson suspects that we are simply replacing one subjective assessment with another (2005).

Sandra Harding, on the other hand, suggests that knowledge built from the perspective of marginalized lives can in fact be more objective—i.e., less biased and distorted—than knowledge built from a dominant perspective (1993). In this case, working from the perspective of persons with disabilities by engaging in the sorts of activities they do on a daily basis would generate less biased or distorted research problems and hypotheses. Harding calls the objectivity we achieve by taking this perspective "strong objectivity". Furthermore, she emphasizes that this perspective will not be monolithic, but will encompass heterogeneous and sometimes even contradictory

points of view. This heterogeneity is exactly what we should expect and we should use it as a resource for scientific discovery (Harding 1993).

Currently, the injustice perpetrated against the chronically ill and disabled goes beyond soliciting their valuations and then dismissing them as incredible. At present, researchers soliciting health state valuations more or less ignore the disabled and chronically ill. While this population is not excluded from unrestricted samples of the general public, their testimony is far outweighed by healthy and able-bodied persons who make up the majority of the sample. Thus their values are likely to be swamped when the sample's values are aggregated. When the valuations provided by healthy and able-bodied individuals overwhelm those provided by the disabled and chronically ill, those valuations can seem to provide evidence for ableist assumptions about the quality of life of the disabled and chronically ill; it can seem to serve as evidence for a medical model of disability.

Research consistently shows that healthy and able-bodied persons, including family members and caretakers, and even otherwise knowledgeable medical professionals, systematically underestimate the quality of life of those living with disabilities and chronic illnesses (see for instance Ubel, Loewenstein, and Jepson 2003; Carel 2014). Carel notes that Dolan (1997) identifies eighty-three health states evaluated as worse than death by members of the general public. However, Carel observes that people who actually live in these health states typically report levels of well-being similar to those of healthy individuals (2014, 253).

Dolan's finding has profound implications for justice and health policy.[13] While in one sense, assigning lower values to disabled or chronically ill health states makes treating those conditions seem more advantageous—for instance, cochlear implants are associated with a large gain in QALYs per unit cost—in general, it makes little sense, if one wishes to maximize QALYs gained per unit cost, to prioritize care for conditions that are less treatable. For instance, if two disabled persons both have health state values of 0.4, and we have the ability to restore one person to a health state valued at 0.6, but could, with the same resources, restore the other to a state valued at 0.8, it seems to make more sense to invest in helping the second individual. We would gain an extra 0.2 QALYs per year by helping the second individual that we would not gain by helping the first. Or consider whether it makes more sense, from the standpoint of QALY maximization, to give a scarce donor organ to an individual who can be restored to full health, or to an individual who will remain in one of Dolan's 83 worse than death states. Such individuals would gain far fewer QALYs from an organ transplant than otherwise able-bodied individuals. For this reason, cost-benefit analysis seems to tell us that the organ should go to the individual who can be restored to full health.

Yet fairness seems to dictate that whether you are disabled or not, you have an equal claim on resources that would preserve your life or improve its quality (Nord 1999;

[13] This finding has a second potentially problematic implication. If we believe the disabled and chronically ill individuals who value their own health states just as highly as those of healthy individuals, it makes little sense to invest in care that would prevent such disabilities. It is difficult to accept such an implication. Elizabeth Barnes discusses the ethical permissibility of causing disability or of failing to prevent it in Chapter 5 of her 2016 monograph, *The Minority Body*. Since she is committed to a mere difference view of disability, she also must deal with this objection. Barnes concludes that while living with a disability once one as gotten used to it is not a significant harm in many cases, the transition to becoming disabled is traumatic for most people and may indeed be conceived of as harmful.

Brock 2005). Indeed, some scholars, such as Martha Nussbaum and Amyrta Sen, who are proponents of a capabilities approach to quality of life, would argue that disabled persons have a greater claim on society's resources than those who are able-bodied (Nussbaum 2001). And John Rawls has argued that when we make decisions from behind a veil of ignorance about the most ethical way to distribute societal resources, we should look for ways to maximize benefit to the worst off (1999). Furthermore, Amundson (2005) argues that we tolerate this bias against providing health care to those with a lower baseline quality of life (a quality of life that cannot be much improved) due to disability in a way that we would not if that lower baseline quality of life were due to, for instance, poverty. For instance, we do not believe that the poor should automatically receive less priority in the distribution of health care resources because living in poverty ostensibly lowers a person's baseline quality of life.

4. Chronic Illness and Disability as Transformative Experiences

In this section, I extend Elizabeth Barnes's assertion that the testimony of the disabled and chronically ill regarding their own well-being is more credible than that of third parties by arguing, following L.A. Paul (2014), that becoming disabled or chronically ill is an epistemically and personally transformative experience. Paul notes that, "People with different skin colors, genders, or histories will have very different experiences in their day-to-day interactions. If you are a man who has grown up and has always lived in a rich Western country, you cannot know what it is like to be a woman living in Ethiopia, and if she has never left her village, she cannot know what it is like to be a man like you" (Paul 2014, 7). I argue that the same is true of the experience of living with a disability or chronic illness. Indeed, Paul admits as much in her book, particularly in her discussion

of hearing vs. deaf parents and the decision each party must make about whether to pursue cochlear implant surgery for their deaf children (2014, 56-70). Barnes also references Paul's work on transformative experience in her recent book, *The Minority Body* (2016, 107), claiming that becoming disabled is a transformative experience.

An epistemically transformative experience is a subjective experience that you cannot accurately imagine prior to living through it. Indeed, you cannot imagine what it would be like to be you having undergone such a transformation. Because you cannot imagine what a transformative experience will be like, you cannot make rational decisions regarding its relative value or desirability. For instance, a hearing individual cannot accurately envision the subjective experience of life as a member of the deaf community or rationally decide to become deaf by imaginatively projecting himself into that state in order to ascertain its value. Nor can a person born deaf accurately assess, at least through projective imagination, the value of acquiring a foreign fifth sense, according to Paul (2014).

Often, experiences that are epistemically transformative are also personally transformative, which compounds difficulties in assigning values to undergoing those transformations. A personally transformative experience alters your values and preferences in ways that may be unpredictable (Paul 2014). At best, we may be able to guess what our new preferences might be by asking those who have already undergone the transformation in question. Research on response shift in quality of life measurement shows that individuals who become chronically ill or disabled adapt to their conditions over time in ways that change their values and preferences (Schwartz and Sprangers 1999). Persons with recently acquired chronic illnesses and disabilities reconceptualize

what it means to be limited and what it means to be doing well, and they reprioritize their values in response to their new lives.

Because, on the view I have a been discussing, acquiring a chronic illness or disability is an epistemically and personally transformative experience, healthy and able-bodied individuals cannot accurately project themselves into the subjective experience of disabled or chronically ill persons and thereby assign values to their health states. And yet, this is exactly what quality of life researchers ask of healthy and able-bodied individuals when soliciting valuations for QALYs. Researchers describe hypothetical health states in a few sentences, ask respondents to imagine life in those states, and then confront them with time trade off or standard gamble tasks involving those health states. But if becoming disabled or chronically ill is epistemically and personally transformative, individuals cannot reliably complete these tasks for hypothetical health states that they have never experienced. For this reason, I argue that the healthy and able-bodied are, at least by themselves, epistemically ill-equipped to make judgments about the quality of life of disabled and chronically ill individuals—i.e., to place values on those health states—in the ways that researchers ask them to.

5. Should Members of the General Public Have a Voice in Health State Valuations?

Up to this point, I have tried to show that the epistemic argument from adaptive preference against soliciting health state values directly from the chronically ill or disabled fails, but I have not addressed arguments in favor of soliciting values from the general public. While standpoint epistemology argues that marginalized groups can have greater access to some types of knowledge than more dominant groups, Patricia Hill Collins notes that Black women, for instance, can produce only an attenuated version of

Black feminist thought when separated from other groups (1990). Similarly, Sandra Harding advocates that men and women work together across their differences to produce feminist thought (1993). In the same way, it may be necessary to include able-bodied persons alongside the disabled and chronically ill in socially conscious discussions of the value of disability and illness. This may be the case despite the fact that healthy and able-bodied individuals are, by themselves, epistemically ill-equipped to make such valuations.

Paul (2014) resists the conclusion that those of us who have not yet had a particular experience cannot, under any circumstances, rationally make decisions about whether we would want to find ourselves in that state. She argues, for instance, that it would be disastrous to conclude that because a currently childless couple cannot accurately imagine, and therefore place an informed value on, the subjective experience of having a child of their own, their present preferences should have no bearing on their decision about whether to start a family. We are not automata, relying solely on scientific evidence and third person testimony to make the decisions that affect us most intimately (Paul 2014, 87). The ability to rationally make decisions about transformative experience is relevant because health state values are not only used to determine the worth of treating existing illnesses and disabilities, they are also used to determine the worth of preventing chronic illness and disability. For instance, we can surgically prevent a cataract patient from becoming blind. By successfully managing diabetes we can prevent disabling complications such as blindness and neuropathy. A more controversial example is the value of preventing genetically transmitted disabilities through the use of genetic testing. While many disability rights advocates object to the

selective abortion of disabled fetuses (Brock 2005; personal communication with Kevin Timpe 2017), there is still broad social and financial support for such technologies.

By extending Paul's argument about rational decision-making regarding the transformative experience, I suggest that it may be possible to rationally decide not to become chronically ill or disabled—even if we cannot accurately place a value on those subjective states. For Paul, this possibility is not rooted in our ability to project ourselves forward into those states and thereby attach a value to them—she believes this is impossible. Instead, she believes that we can rationally put a value on the revelatory nature of transformation. That is, we may value undergoing a transformative experience merely for the sake of experiencing something new. Thus, the fact that we do not know what it would be like for us to have a child, for instance, ceases to be an impediment to rational decision making and becomes instead the very thing that makes a rational decision possible. Most of us do not value revelation highly enough to take the risk of experiencing a lower quality of life as a disabled or chronically ill individual, even if disability positive testimony tells us that many disabled individuals value or even prefer life with a disability.

One problem with Paul's argument is that the choice not to take the risk of experiencing a lower quality of life for the sake of revelation is rooted in what Barnes (2016) calls a "bad difference" model of disability. I.e., it is rooted in the popular assumption, supported by the medical model of disability and contested by many disability rights activists, that being disabled inherently makes you worse off. So even decisions that seem rational are grounded in a controversial conception of disability that we may not be justified in taking for granted. These supposedly rational decisions may

be influenced more by prejudice than reason. If we are to include the voices of the healthy and able-bodied when valuing health states, we must somehow address the problem of systematic stigmatization of disabled states that occurs when healthy and able-bodied people adhere to a medical, or bad difference, model of disability. I discuss one potential solution to these problems in the section below.

6. On the Role of Deliberative Focus Groups in Health State Valuation

Healthy and able-bodied individuals are stakeholders in decisions about resource allocation for the prevention of chronic illness and disability, and they are potential stakeholders in decisions about resource allocation for treatment of existing illnesses and disabilities. I have presented an (imperfect) argument, following Paul (2014) suggesting it may be possible for the healthy and able bodied to rationally choose not to become ill or disabled—an argument that is vulnerable to charges that it relies on a bad difference model of disability. Given their position as stakeholders, and given that it may be possible for them to rationally choose not to become disabled, it might seem that it is appropriate to include the healthy and able-bodied in the process of valuing health states for the purposes of resource allocation.

In this section, I present a proposal for soliciting health state values from both the disabled and chronically ill and the healthy and able-bodied. I argue that if we are to include input from healthy and able-bodied members of the general public in the process of determining health state valuations, the values they express should be informed by testimony from the disabled and chronically ill. I propose, following Daniel Hausman (2015), that one way to do this is by soliciting values not through individual surveys but from deliberative focus groups, and I further argue that those groups should give greater

representation to the chronically ill and disabled than would be found in a random sample of the general public. There are a number of reasons to think that values determined through an inclusive, deliberative process are superior to those determined by individuals, especially when some of those individuals are epistemically under-equipped.

Hausman (2015) argues that one problem with the individual surveys currently used to solicit health state valuations is that they encourage their users to respond with gut reactions rather than carefully considered values and preferences. When respondents are asked to value fifteen health states in as many minutes, it is impossible to do otherwise. Few healthy and able-bodied individuals have thought carefully about whether they would truly prefer to spend 10 years in a state of moderate disability or chronic illness or 7 years in a state of perfect health, nor have they thought about how living in a state of moderate disability would affect their lives or those of their families. Even if they have thought about living in such a state, they typically don't have the requisite knowledge of what is involved in living with a particular disability to have informed preferences, and the short descriptions of these health states offered by most utility measures are of little help in stirring the imagination. What's more, by instructing respondents that there is no right or wrong answer to the questions being posed, these measures further discourage carefully considered deliberation (Hausman 2015).

When important decisions affecting the well-being of patients depend on the valuations solicited from respondents, it is important that those valuations be both well-informed and well thought out. Joining members of the general public with the chronically ill and disabled in mixed focus groups can help healthy individuals to take the human capacity to adapt to illness and disability into account when making their

valuations, and it can help them to overcome the focusing illusion. The focusing illusion is the tendency of healthy and able-bodied individuals to focus the ways disability or illness will limit their capabilities rather than recognizing the many ways their lives will remain the same or might even be enhanced. Bringing together members of the general public and the chronically ill and disabled allows healthy and able-bodied individuals to ask questions about what is involved in living, for instance, with chronic kidney disease, or what it is like to be blind or to live with depression. How do individuals with these conditions value their own lives? How have they learned to adapt to their conditions, and how are their experiences of chronic illness and disability different from what they expected when they first became disabled or ill? One man explains, for instance, how his initial expectations of life with a physical disability were proven wrong. "I used to fear I'd be a horrible dad because of what I can't do, but now, I'd argue that my daughter had a richer childhood because of it. Unlike most dads, I actually NEED my kid's help. This has brought us closer and made her more independent in all the right ways." Life with a disability or chronic illness moves from being an abstract problem that happens to strangers to being a more familiar condition that someone they know experiences every day. Thus, in addition to increasing the general public's familiarity with disability and chronic illness, deliberating about the values of those states as a member of a mixed focus group gives healthy and able-bodied individuals the opportunity to develop a greater understanding of these conditions and the people who live with them.

Discussion among group members spurs more careful deliberation because proponents of a particular valuation are challenged to defend that valuation to others who may disagree with them. They are forced to give reasons for their preferences. Gut

reactions are no longer acceptable as responses to researchers' queries. Respondents must listen to and consider one another's values, preferences, and reasons—including the disability positive testimony of many individuals with disabilities—in ways that they are not required to when values from individual surveys are simply aggregated and averaged by researchers. Respondents must work to understand one another's points of view by expanding their own interpretive horizons.

But wouldn't it be sufficient for these deliberative focus groups to share the same makeup as an unrestricted representative sample of the general public? I don't believe that it would. First, it is important that a broad range of disabilities and chronic illnesses be represented in each group. The life experience of a person with chronic kidney disease is very different from that of a person with multiple sclerosis, and so may the values and preferences expressed by persons with these conditions. Second, it is important to note that even within a particular illness or disability, values and preferences are not monolithic. The autistic man I spoke with observed that given the choice, he would prefer to be autistic rather than neurotypical. The young woman with OCD noted that while others had told her that her illness made her more observant than most people, she didn't see her experience as valuable in any way. A second young woman with a motor disability in her hands stated that while everyday tasks were more difficult for her than they were for most of her peers, she also feels a greater sense of accomplishment when she is successful.

The mere presence of the disabled and chronically ill in deliberative focus groups is not enough to ensure that those groups are inclusive in nature. Their testimony must be solicited and valued by other members of the group, and those group members must

exhibit the epistemic virtues of humility, open-mindedness, and curiosity/diligence (Medina 2013). That is, they must be willing to consider their own viewpoints defeasible, to actively seek out and listen to others' perspectives. Harding argues that dominant groups must be challenged to collaborate with marginalized people (1993). "Such a project requires learning to listen attentively to marginalized people; it requires educating oneself about their histories, achievements, preferred social relations, and hopes for the future; it requires putting one's body on the line for "their" causes until they feel like "our" causes; it requires critical examination of the dominant institutional beliefs and practices that disadvantage them; it requires critical self-examination to discover how one unwittingly participates in generating disadvantage to them … and more" (1993).

One practical way to facilitate inclusivity and epistemic justice would be to open up our conception of deliberation to include a broad range of communicative styles, thus allowing group members to engage not only in rational argument but also in telling their own stories. Loosening the requirement that participants be able to engage in formal reason giving leads to greater inclusivity by ensuring that groups need not be composed only of highly-educated or formally trained members. Since less educated, but still experientially expert members of the disabled and chronically ill community are often subject to even greater marginalization, this affordance is an important one.

Finally, allowing protest to inform the results of deliberation addresses concerns expressed by Iris Marion Young (2001) that democratic deliberation, when performed against a backdrop of systematic inequality, can perpetuate that injustice rather than remedying it. She worries not only about the inclusiveness of deliberative groups, but also that they often buy into hegemonic formulations of social problems. For instance, it

is not at all clear that the questions we have been presenting to respondents up till this point make sense when addressed to the disabled and chronically ill—or to respondents that do not buy into the medical model of disability. This is just one illustration of the way different standpoints can generate different research questions (Harding 1993).

Studies show that the chronically ill and disabled are reluctant to give up even a small portion of their lives in order to live as healthy or able-bodied individuals (Fowler et al. 1995; O'Leary et al. 1995). They simply do not buy into the otherwise common assumption that healthy and able-bodied life years are more valuable than disabled life years. Given the prevalence of disability positive testimony—testimony from the disabled that they value their lives highly and sometimes even prefer being disabled to being able-bodied—this is not surprising. Thus, the typical time trade off task used to solicit health values becomes both ethically and epistemically problematic—even offensive—when addressed to the disabled and chronically ill. What's more, in many cases, restoring the disabled or chronically ill to full health or functionality is impossible, making a non-sense of the task. There is no potential real world choice to be made between living some years in disability or some other number of years as an able-bodied individual.

It may be that because it is rooted in a medical model of disability, the entire model of assigning relative values or utilities to health states based on time trade off or standard gamble preferences is fatally flawed, or at least irremediably biased toward dominant perspectives. Unfortunately, considerable social, intellectual, and financial capital has, at this point, been invested in the QALY framework, and it is unlikely that policy makers will abandon it unless a better tool for guiding resource allocation presents

itself. If one accepts the premise that health care resources are limited and that difficult resource allocation decisions must be made, it seems some system must be used to guide the process. It is up to those of us who value epistemic and distributive justice to see to it that that system is as good as it can be. The development of such a system remains an important task for policy makers and stakeholders.

7. Conclusion

This chapter argues that policy makers are not epistemically justified in soliciting health state valuations for resource allocation decisions from the general public while neglecting the perspectives of the disabled and chronically ill. Their testimony cannot be dismissed as the product of adaptive preference. Instead, feminist standpoint theory and Laurie Paul's work on transformative experience both suggest that the disabled and chronically ill occupy a privileged epistemic position when it comes to valuing their own health states. Furthermore, it is epistemically unjust to systematically discount the testimony of the disabled and chronically ill. Yet rather than soliciting health state valuations exclusively from this epistemically privileged group, I argue that valuations should be sought, if they are sought at all, from mixed focus groups. While the disabled and chronically ill should hold a central place in deliberations about health care priority setting, without the participation of other groups, they can form only an attenuated standpoint.

CONCLUSION

Health-related quality of life measures have become a major fixture in evidence-based medicine and serve a key role in health policy discussions. Given their power to impact the well-being of patients, it is important that they be interpreted correctly and that they be epistemically sound. It is important to know what judgments they do and do not license us to make. This dissertation has tried to make some headway in examining the epistemic issues surrounding quality of life measurement in the health sciences and their ethical implications.

Some of the epistemic challenges faced by measure developers in the human sciences are unique—for instance, quality of life researchers must contend with the subjectivity of their measures as well as with their Ballung nature. These measures target and affect moral agents—agents with their own values and preferences. These values and preferences impact the way agents evaluate their own quality of life, and that of others. They affect the way respondents conceive of quality of life and the way they interpret the questions posed to them by researchers. Other epistemic challenges are remarkably similar to those faced by scientists working on archetypal physical measures. Both physical and humanistic measure development are hermeneutic enterprises, and both types of measures depend on models for contextualization.

My work has explored commonalities and differences between measurement in the human sciences, as exemplified by PROMs, and archetypal physical measurement, as

exemplified by time and temperature. Often work that seeks to bridge the gap between these two areas sees physical measurement as normative. My work does not. I believe that there are important senses in which we should not expect humanistic measures to follow the lead of physical measures. I advance an ontological thesis about quality of life as a target of measurement in several chapters, suggesting that it is both socially constructed and contextually valued.

Asking for quality of life measures to be epistemically sound is not the same thing as asking them to behave like physical measures. Meeting the latter request seems to be one of the goals of Rasch measurement. Forcing humanistic measures to fit a prescriptive, statistical model carries both advantages and disadvantages. I discuss some of these advantages in Chapter 2: the Rasch model facilitates judgments about measurement comparability, for instance, and it supports interval level measurement. However, I worry that in its quest to mirror physical measurement, Rasch problematically narrows the construct to be measured, and fails to take into account variation that naturally arises from respondents' interpretations and conceptualizations.

Physical measures are typically single valued, and I have argued that quality of life measures are instead contextually valued. A perhaps an unavoidable lack of theoretical underpinnings leaves these measures to be defined operationally. Both their Ballung nature and their subjectivity lead to variation in the ways quality of life measurement is operationalized. Instead of trying to bring the multiple values produced by these operationalizations into agreement, as we would if they were in error, we should see these multiple values as reflective of legitimate differences in the ways quality of life may be conceptualized. The question of whether, given enough time, a theoretical model

for quality of life can be developed is a challenging one. While I was initially optimistic about the possibility of developing a theoretical model for quality of life, as evidenced by my work in Chapter 2, I later began to worry that it simply was not possible to arrive at such a model given the subjectivity and Ballung nature of quality of life.

Rather than viewing physical measures as normative, and taking human measures as second class measures, I would argue that researchers in the physical sciences can in fact learn from measures in the human sciences. For instance, while measures in the human sciences are more straightforwardly texts to be interpreted, measure development in the human sciences is also a hermeneutical enterprise, as van Fraassen has observed. My work elaborates on this claim by examining Hasok Chang's historical treatment of temperature measurement. While physical measures depart from hermeneutics insofar as their meanings can be standardized, it is important to realize that the meanings and values we choose for these measures are contingent. They are the result of often underdetermined choices by researchers. I argue, following McClimans et al. (2013), that quality of life measures are socially constructed. But so too are physical measures. The single value that characterizes outcomes in the physical sciences is not a straightforward discovery, but a social achievement born of the iterative correction of disparate outcomes.

Finally, I examine the ways in which self-evaluations of quality of life may differ from the evaluations of others, and argue that *ceteris paribus*, we should trust self-evaluations more. Epistemic injustice involves unfairly discounting the testimony of certain knowers based on social identity prejudice. While that prejudice is often based on gender or race, it may also be based on disability. Moving forward, I hope to explore not just testimonial injustice in the context of quality of life measurement but also

hermeneutic injustice. Does the attempted standardization of quality of life measures problematically constrain interpretive resources available to respondents?

One of the strengths of this dissertation is the breadth of philosophical approaches it employs. I examine the epistemology of quality of life measures using tools from philosophical hermeneutics, metaphysics, the modeling literature, and feminist philosophy. This amalgam of approaches shows the philosophical richness of the quality of life measurement as a topic for investigation. On the surface, it represents a very practical problem affecting health policy and resource allocation. But at its core, it requires us to explore deep philosophical questions about the meaning of well-being for individuals and for society. Much work remains to be done.

REFERENCES

Aikin, Scott F. and J. Caleb Clanton. 2010. "Developing Group-Deliberative Virtues." *Journal of Applied Philosophy* 27(4): 409-424.

Alexandrova, Anna. 2017. *A Philosophy for the Science of Well-being*. Oxford: Oxford University Press.

Amundson, Ron. 2005. „Disability, Ideology, and Quality of Life: A Bias in Biomedical Ethics." In *Quality of Life and Human Difference: Genetic Testing, Health Care, and Disability*, edited by David Wasserman, Jerome Bickenbach, and Robert Wachbroit, 101-124. Cambridge: Cambridge University Press.

Anatchkova, Milena D., John E. Ware, and Jakob B. Bjorner. 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.

Aristotle. 1999. *Nicomachean Ethics.* 2nd ed. Edited by Terence Irwin. Indianapolis: Hackett Publishing Company.

Bachtiger, Andre, Simon Niemeyer, Michael Neblo, Marco R. Steenbergen, and Jurg Steiner. 2010. "Disentangling Diversity in Deliberative Democracy: Competing Theories, Their Blind Spots and Complementarities." *Journal of Political Philosophy* 18(1): 32-63.

Barnes, Elizabeth. 2009. "Disability and Adaptive Preference." *Philosophical Perspectives* 23: 1-22.

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.

Bird, Alexander and Tobin, Emma. Spring 2017. "Natural Kinds." *Stanford Encyclopedia of Philosophy.* Edited by Edward N. Zalta. Accessed October 12, 2017. https://plato.stanford.edu/archives/spr2017/entries/natural-kinds/.

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.

Bradburn, Norman, Nancy Cartwright & Jonathan Fuller. 2017. "A Theory of Measurement." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, edited by Leah McClimans, 73-88. New York: Rowman and Littlefield International.

Brock, Dan. 1993. "Quality of Life Measurement in Health Care and Medical Ethics. In *The Quality of Life,* edited by Martha Nussbaum and Amartya Sen, 95-132. Oxford: Oxford University Press.

Brock, Dan. 1995. "Justice and the ADA: Does Prioritizing and Rationing Health Care Discriminate Against the Disabled?" *Social Philosophy and Policy* 12(2): 159-185.

Brock, Dan. 2005. "Preventing Genetically Transmitted Disabilities while Respecting Persons with Disabilities." In *Quality of Life and Human Difference: Genetic Testing, Health Care, and Disability,* edited by David Wasserman, Jerome Bickenbach, and Robert Wachbroit, 67-100. Cambridge: Cambridge University Press.

Brooks, Richard and the EuroQol Group. 1996. "EuroQol: The Current State of Play." *Health Policy* 37: 53-72.

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

Browne, John, Hannah M. McGee & Ciaran A. O'Boyle. 1997. "Conceptual Approaches to the Assessment of Quality of Life." *Psychology & Health* 12: 737-751.

Cano, S. J., L. E. Barrett, J. P. Zajicek, and J. C. Hobart. 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.

Carel, Havi. 2014. “Ill, But Well: A Phenomenology of Well-Being in Chronic Illness.” In *Disability and the Good Human Life,* edited by Jerome Bickenbach, Franziska Felder, and Barbara Schmitz, 243-270. Cambridge: Cambridge University Press.

Carel, Havi and Ian James Kidd. 2014. “Epistemic Injustice in Healthcare: A Philosophical Analysis.” *Medicine, Health Care, and Philosophy* 17: 529-540.

Cartwright, Nancy and Rosa Runhardt. 2014. “Measurement.” In *Philosophy of Social Science: A New Introduction,* edited by Nancy Cartwright and Eleonora Montuschi, 265-287. Oxford: Oxford University Press.

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*, edited by Judith M. Tanur, 15-48. New York: Sage Foundation.

Craig, Edward. 1990. *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Clarendon Press.

Dolan, Paul. 1996. “The Effect of Experience of Illness on Health State Valuations.” *Journal of Clinical Epidemiology* 49(5): 551-564.

Dolan, Paul. 1997. “Modeling Valuations for EuroQol Health States.” *Medical Care* 35(11): 1095-1108.

Dolan, Paul. 1999. “Whose Preferences Count?” *Medical Decision Making* 19: 482-486.

Donovan, Jenny, Stephen J. Frankel & John D. Eyles. 1993. “Assessing the Need for Health Status Measures.” *Journal of Epidemiology and Community Health* 47:158-162.

Elster, J. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

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.

Fowler Jr., Floyd J., Paul D. Cleary, Michael P. Massagli, Joel Weissman, and Arnold Epstein. 1995. « Describing and Measuring the Values of Health States : The Rôle

of Reluctance to Give up Life in the Measurement of Values of Health States. » *Medical Decision Making* 15 : 195-200.

Frank, Lori, Ethan Basch, and Joe V. Selby. 2014. "The PCORI Perspective on Patient-Centered Outcomes Research." *Journal of the American Medical Association* 312(15): 1513-1514.

Fricker, Miranda. 2007. *Epistemic Injustice: Power & the Ethics of Knowing*. Oxford: Oxford University Press.

Gadamer, Hans Georg. 1991. *Truth and Method*, 2nd Rev. ed., translated by Joel Weinsheimer and Donald G. Marshall. New York: Crossroad Publishing.

Gill, Thomas M. and Alvan R. Feinstein. 1994. "A Critical Appraisal of Quality of Life Measurements." *Journal of the American Medical Association* 272(8): 619-626.

Griffin, James. 1989. *Well-being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hadorn, David C. 1991. "The Role of Public Values in Setting Health Care Priorities." *Social Science & Medicine* 32: 773-781.

Harding, Sandra. 1993. „Rethinking Standpoint Epistemology: What is Strong Objectivity?" In *Feminist Epistemologies*, edited by Linda Alcoff and Elizabeth Potter, 49-82. New York: Routledge.

Hausman, Daniel. 2015. *Valuing Health: Well-being, Freedom, and Suffering*. Oxford: Oxford University Press.

Hill Collins, Patricia. 1990. *Black Feminist Thought : Knowledge, Consciousness, and the Politics of Empowerment*. Boston : Unwin Hyman.

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, Solutions, and Recommendations." *Lancet Neurology* 6: 1094-1105.

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).

Hobart, Jeremy, Stefan Cano, Rachel Baron, Alan Thompson, Steven Schwid, John Zajicek, and David Andrich. 2013. "Achieving Valid Patient-Reported Outcomes Measurement: A Lesson from Fatigue in Multiple Sclerosis." *Multiple Sclerosis Journal* 0(0): 1-11.

Horsman, John, William Furlong, David Feeny, and George Torrance. 2003. "The Health Utilities Index (HUI): Concepts, Measurement Properties and Applications." *Health and Quality of Life Outcomes* 1:54: 1-13.

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

Joyce, C.R.B., A. Hickey, H.M. McGee & C.A. O'Boyle. 2003. "A Theory Based Method for the Evaluation of Individual Quality of Life: The SEIQoL." *Quality of Life Research* 12:275-280.

Kane, Michael T. 1982. "A Sampling Framework for Validity." *Applied Psychological Measurement* 6(2): 125-160.

Kidd, Ian James and Havi Carel. 2017. "Epistemic Injustice and Illness." *Journal of Applied Philosophy* 34(2): 172-190.

Klassen, Anne, Andrea Pusic, Amie Scott, Jennifer Klok, and Stefan J. Cano. 2009. "Satisfaction and Quality of Life in Women Who Undergo Breast Surgery: A Qualitative Study." *BioMed Central Women's Health*, 9(11): 1-8.

McClimans, Leah. 2010a. "Towards Self-Determination in Quality of Life Research: A Dialogic Approach." *Medical Health Care and Philosophy* 13: 67-76.

McClimans, Leah. 2010b. "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.

McClimans, Leah, Jerome Bickenbach, Marjan Westerman, Licia Carlson, David Wasserman & Carolyn Schwartz. 2013. "Philosophical Perspectives on Response Shift." *Quality of Life Research* 22: 1871-1878.

McClimans, Leah. 2017. "Measurement in Medicine and Beyond: Quality of Life, Blood Pressure and Time" In *Reasoning in Measurement,* edited by A. Nordmann and N. Mößner, N., 133-146. New York: Routledge.

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

Medina, Jose. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. 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.

Morreim, Haavi. 1986. "Computing the Quality of Life." In *The Price of Health,* edited by G.J. Agich and C.E. Begley, 45-69. Dordrecht, Holland: D. Reidel Publishing Company.

Murray, D.W., R. Fitzpatrick, K. Rogers, H. Pandit, D. J. Beard, A. J. Carr, and J. Dawson. 2007. "The Use of the Oxford Hip and Knee Scores." *The Journal of Bone and Joint Surgery* 89B: 1010-1014.

Neurath, Otto. 1983. « Physicalism and Investigation of Knowledge. » In *Philosophical Papers 1913-1946,* edited and translated by Robert Cohen and Marie Neurath, 159-167. Dordrecht : Reidel.

Nord, Erik, Jose Luis Pinto, Jeff Richardson, Paul Menzel, and Peter Ubel. 1999. "Incorporating Societal Concerns for Fairness in Numerical Valuations of Health Programmes." *Health Economics* 8: 25-39.

Nussbaum, Martha C. 2001. *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge, MA: The Belknap Press of Harvard University Press.

O'Leary, June F., Diane L. Fairclough, M. Kay Jankowski, and Jane C. Weeks. 1995. "Comparison of Time-Tradeoff Utilities and Rating Scale Values of Câncer Patients and Their Relatives: Evidence for a Possible Plateau Relationship." *Medical Decision Making* 15: 132-137.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Patrick, Donald L., Laurie B. Burke, Chad J. Gwaltney, Nancy Kline Leidy, Mona Martin, Elizabeth Molsen, and Lena Ring. 2011a. "Content Validity—Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1—Eliciting Concepts for a New PRO Instrument." *Value in Health* 14: 967-977.

Patrick, Donald L., Laurie B. Burke, Chad J. Gwaltney, Nancy Kline Leidy, Mona Martin, Elizabeth Molsen, and Lena Ring. 2011b. "Content Validity—Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: part 2—Assessing Respondent Understanding." *Value in Health* 14: 978-988.

Paul, L.A. 2014. *Transformative Experience*. Oxford: Oxford University Press.

Pusic, Andrea, L., Anne F. Klassen, Amie M. Scott, Jennifer A. Klok, and Stefan J. Cano. 2009. "Development of a New Patient-Reported Outcome Measure for Breast Surgery: The Breast-Q." *Plastic and Reconstructive Surgery* 124: 345-353.

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.

Rawls, John. 1999. *A Theory of Justice*, Rev. ed. Cambridge, MA: Harvard University Press.

Rubin, Michael. 2008. "Is Goodness a Homeostatic Property Cluster?" *Ethics* 118: 496-528.

Salomon, Joshua A., Theo Vos, Daniel R Hogan, Michael Gagnon, Mohsen Naghavi, Ali Mokdad, Nazma Begum, Razibuzzaman Shah, Muhammad Karyana, Soewarta Kosen, Mario Reyna Farje, Gilberto Moncada, Arup Dutta, Sunil Sazawal, Andrew Dyer, Jason Seiler, Victor Aboyans, Lesley Baker, Amanda Baxter, Emelia J Benjamin, Kavi Bhalla, Aref Bin Abdulhak, Fiona Blyth, Rupert Bourne, Tasanee Braithwaite, Peter Brooks, Traolach S Brugha, Claire Bryan-Hancock, Rachelle Buchbinder, Peter Burney, Bianca Calabria, Honglei Chen, Sumeet S Chugh, Rebecca Cooley, Michael H Criqui, Marita Cross, Kaustubh C Dabhadkar, Nabila Dahodwala, Adrian Davis, Louisa Degenhardt, Cesar Díaz-Torné, E Ray Dorsey, Tim Driscoll, Karen Edmond, Alexis Elbaz, Majid Ezzati, Valery Feigin, Cleusa P Ferri, Abraham D Flaxman, Louise Flood, Marlene Fransen, Kana Fuse, Belinda J Gabbe, Richard F Gillum, Juanita Haagsma, James E Harrison, Rasmus Havmoeller, Roderick J Hay, Abdullah Hel-Baqui, Hans W Hoek, Howard Hoffman, Emily Hogeland, Damian Hoy, Deborah Jarvis, Jost B Jonas, Ganesan Karthikeyan, Lisa Marie Knowlton, Tim Lathlean, Janet L Leasher, Stephen S Lim, Steven E Lipshultz, Alan D Lopez, Rafael Lozano, Ronan Lyons, Reza Malekzadeh, Wagner Marcenes, Lyn March, David J Margolis, Neil McGill, John McGrath, George A Mensah, Ana-Claire Meyer, Catherine Michaud, Andrew Moran, Rintaro Mori, Michele E Murdoch, Luigi Naldi, Charles R Newton, Rosana Norman, Saad B Omer, Richard Osborne, Neil Pearce, Fernando Perez-Ruiz, Norberto Perico, Konrad Pesudovs, David Phillips, Farshad Pourmalek, Martin Prince, Jürgen T Rehm, Guiseppe Remuzzi, Kathryn

Richardson, Robin Room, Sukanta Saha, Uchechukwu Sampson, Lidia Sanchez-Riera, Maria Segui-Gomez, Saeid Shahraz, Kenji Shibuya, David Singh, Karen Sliwa, Emma Smith, Isabelle Soerjomataram, Timothy Steiner, Wilma A Stolk, Lars Jacob Stovner, Christopher Sudfeld, Hugh R Taylor, Imad M Tleyjeh, Marieke J van der Werf, Wendy L Watson, David J Weatherall, Robert Weintraub, Marc G Weisskopf, Harvey Whiteford, James D Wilkinson, Anthony D Woolf, Zhi-Jie Zheng, Christopher J L Murray. 2012. "Common values in assessing health outcomes from disease and injury: Disability weights measurement study for the Global Burden of Disease Study 2010." *Lancet* 380: 2129-2143.

Schwartz, Carolyn, Elena Andresen, Margaret A. Nosek, Gloria L. Krahn & the RRTC Expert Panel on Health Status Measurement. 2007. "Response Shift Theory: Important Implications for Measuring Quality of Life in People with Disability." *Archives of Physical Medicine and Rehabilitation* 88:529-536.

Schwartz, Carolyn E. and Mirjam A.G. Sprangers 1999. "Methodological Approaches for Assessing Response Shift in Longitudinal Health-Related Quality of Life Research." *Social Science & Medicine* 48: 1531-1548.

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.

Sen, Amartya. 1993. "Capability and Well-being." In *The Quality of Life*, edited by Martha Nussbaum and Amartya Sen, 30-53. Oxford: Oxford University Press.

Stenner, A. Jackson, William P. Fisher Jr., Mark H. Stone, and Donald S. Burdick. 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 Geoffrey R. Norman, and John Cairney. 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.

Tversky, Amos and Daniel Khaneman. 1974. « Judgment Under Uncertainty : Heuristics and Biases. » *Science*, New Series 185(4157) : 1124-1131.

Ubel, Peter A., George Loewenstein, and Christopher Jepson. 2003. “Whose Quality of Life? A Commentary Exploring Discrepancies Between Health State Evaluations of Patients and the General Public.” *Quality of Life Research* 12: 599-607.

Van Fraassen, Bas. 2008. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

Warnke, Georgia. 1987. *Gadamer: Hermeneutics, Tradition and Reason*. Stanford: Stanford University Press.

Westerman, Marjan J., Tony Hak, Mirjam A. G. Sprangers, Harry J. M. Groen, Gerrit van der Wal, Anne-Mei The. 2008. “Listen to Their Answers! Response Behavior in the Measurement of Physical and Role Functioning.” *Quality of Life Research*, 17: 549-558.

Williams, Alan. 1995. *The Role of the EuroQol Instrument in QALY Calculations*. University of York Centre for Health Economics.

Wittgenstein, Ludwig. 1973. *Philosophical Investigations*, 3rd ed., translated by G.E.M. Anscombe. New York: The McMillan Company.

Young, Iris Marion. 2001. “Activist Challenges to Deliberative Democracy.” *Political Theory* 29(5) : 670-690.

APPENDIX A

LETTER OF INVITATION FOR EXEMPT RESEARCH

Dear ___,

My name is Laura Cupples. I am a doctoral candidate in the Philosophy Department at the University of South Carolina. I am conducting a research study as part of the requirements of my degree in Philosophy, and I would like to invite you to participate.

You are being asked to participate in this study because you self-identify as disabled or chronically ill. If you decide to participate, you will be asked to complete a brief questionnaire about your experience with chronic illness or disability. In particular, you will be asked questions about what kind of disability or illness you have, what peoples attitudes and misconceptions are about your life, and whether or in what ways you value being ill or disabled. You do not have to answer any questions that you do not wish to. I will share the questions with you electronically, and you can respond at a convenient time. The questionnaire should take about 15 minutes to complete. After completion of the study, your responses will be deleted.

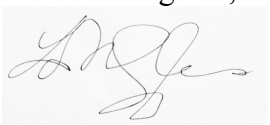
Participation is confidential. Study information will be kept on my personal computer, which is password protected. I will anonymize your interview responses once I have received them. The results of the study may be published or presented at professional meetings, but your identity will not be revealed.

Taking part in the study is your decision. You do not have to be in this study if you do not want to. You may also quit being in the study at any time or decide not to answer any question you are not comfortable answering.

I will be happy to answer any questions you have about the study. You may contact me at 803-528-7111 or cupples@email.sc.edu or my faculty advisor, Emily Mann (emann@mailbox.sc.edu) if you have study related questions or problems. If you have any questions about your rights as a research participant, you may contact the Office of Research Compliance at the University of South Carolina at 803-777-7095.

Thank you for your consideration. If you would like to participate, please contact me at cupples@email.sc.edu. .

With kind regards,



Laura M. Cupples
University of South Carolina
Byrnes Building Room 444
Columbia, SC 29208
803-528-7111
cupples@email.sc.edu

APPENDIX B

INTERVIEW QUESTIONS

1. First, it would help me to know in broad terms what sort of illness or disability you have.

2. If you were not born with a disability or chronic illness but acquired or developed it at some point, what was the experience of becoming ill or disabled like?

3. What sorts of misconceptions do people tend to have about your life as a disabled or chronically ill person?

4. Do people who are familiar with your illness or disability tend to over or underestimate its impact on your life?

5. Do you think it's possible for someone who hasn't experienced your illness or disability to accurately imagine what it is like to live with it?

6. Are there things you value about your experience of illness or disability?

APPENDIX C

PERMISSION TO REPRINT

From: **Sarah Campbell** scampbell@rowman.com
Subject: RE: assistance re Measurement in Medicine chapter
Date: March 6, 2018 at 4:48 AM
To: Isobel Cowper-Coles icowpercoles@rowman.com, CUPPLES, LAURA cupples@email.sc.edu

SC

Hi Laura

Thanks for your note. I am happy to grant you permission to reuse your chapter from *Measurement in Medicine* (ed. McClimans) in your forthcoming doctoral dissertation. We would only ask that you acknowledge our publication appropriately. Please note you will need to reapply for permission should you wish to publish the material commercially in the future.

All best wishes

Sarah

From: Isobel Cowper-Coles
Sent: 01 March 2018 20:40
To: CUPPLES, LAURA
Cc: Sarah Campbell
Subject: RE: assistance re Measurement in Medicine chapter

Hi Laura

Thanks for your message. I am copying my colleague Sarah Campbell, the publisher of the anthology, who will be able to assist.

Best wishes
Isobel

From: CUPPLES, LAURA [mailto:cupples@email.sc.edu]
Sent: 01 March 2018 12:42
To: Isobel Cowper-Coles <icowpercoles@rowman.com>
Subject: assistance re Measurement in Medicine chapter

Isobel -

I contributed a chapter to a recent anthology - Measurement in Medicine - edited by Leah McClimans and published by Rowman and Littlefield Intl in 2017. I need to get permission to reproduce that chapter in my doctoral dissertation. The title of the chapter was Epistemic Roles of Models in Health Science Measurement. Who do I need to speak to about that, or can you help me?

Thanks,
Laura Cupples
University of South Carolina