Theor Med Bioeth (2010) 31:225–240
DOI 10.1007/s11017-010-9142-0

A theoretical framework for patient-reported outcome measures

Leah McClimans

Published online: 5 June 2010

Abstract Patient-reported outcome measures (PROMs) are increasingly used to assess multiple facets of healthcare, including effectiveness, side effects of treatment, symptoms, health care needs, quality of care, and the evaluation of health care options. There are thousands of these measures and yet there is very little discussion of their theoretical underpinnings. In her 2008 Presidential address to the Society for Quality of Life Research (ISOQoL), Professor Donna Lamping challenged researchers to grapple with the theoretical issues that arise from these measures. In this paper, I attempt to do so by arguing for an analogy between PROMs and Hans-Georg Gadamer's logic of question and answer. While researchers readily admit that the constructs involved in PROMs are imperfectly understood and lack a gold standard, they often ignore the consequences of this fact. Gadamer's work on questions and their importance to philosophical hermeneutics helps to show that the questions researchers ask about such constructs are also imperfectly understood. I argue that these questions should not be standardized, and I instead propose a theoretical framework that understands PROMs as posing genuine questions to respondents—questions that are open to reinterpretation.

Keywords Patient-reported outcomes measures (PROMs) · Quality of life · Perceived health status · Hermeneutics · Interpretation · Theoretical

In her 2008 Presidential address to the International Society for Quality of Life Research (ISOQoL), Professor Donna Lamping identified three challenges facing

L. McClimans (✉)
Department of Philosophy, University of South Carolina, Columbia, SC, USA
e-mail: mccliman@mailbox.sc.edu

this field.[1] One of these challenges was the need for a theoretical framework to support patient-reported outcome measures (PROMs).[2] Lamenting the lack of theoretical discussions in leading journals she suggested that it is time for some researchers to step back from the creation of new measures and reflect on more conceptual issues.[3] In this paper, I begin to meet this challenge by arguing for an analogy between Hans-Georg Gadamer's theoretical account of the logic of questions and answers in *Truth and Method* and the logic of questions and answers in PROMs.

Gadamer: The logic of question and answer

For Gadamer, the hermeneutical task is one of understanding and interpreting meaning.[4] We achieve this task by asking questions about a subject matter of interest, for instance, by asking questions about quality of life or the nature of courage or Barack Obama's rise to the Presidency [6]. But we use questions in at least two different ways, and for Gadamer, only one of these is associated with understanding meaning. For example, we sometimes ask questions when we already know their answer, and in such cases, our questions serve as a kind of test. Instances of pedagogy often take the form of these questions. Sometimes, however, we ask questions precisely because we do not know their answer. In these cases, we ask questions in order to better understand something we do not know already. In a Gadamerian context, this situation represents the hermeneutical task.

The difference between asking questions when one does not know the answer and asking questions when one does respectively marks for Gadamer the difference between genuine and merely apparent questions. Genuine questions are 'open' [6]. This state of openness refers both to the way in which genuine questions open inquiry into a subject matter and also the way in which genuine questions are themselves open to reinterpretation. Take, for instance, the question, 'Why is the *Twilight Saga* so compelling to young girls?' This question is not merely a request for information about a subject matter; it also opens up the possibility of understanding the meaning of these novels in new ways. As Gadamer puts this

[1] Although ISOQoL refers explicitly to 'quality of life research', in practice, the interests of the society extend beyond these specific measures. In this paper, I use the term 'patient-reported outcome measures' to cover these more general interests.

[2] It is important to note that some philosophers *have* explored the theoretical basis of the constructs that these measures are meant to assess. These discussions tend to focus on questions about the nature of quality of life or welfare, e.g., are these constructs 'subjective' or 'objective', or are these constructs best characterized in terms of happiness or capabilities or preferences. For examples, see [1–4]. These discussions, however, often fail to engage with the practical grounding of PROMs as questionnaires or psychometric instruments. Thus in many ways they fail to meet the theoretical needs of researchers in this field. Antaki and Rapley make a similar point; see [5].

[3] Professor Lamping's Presidential Address was made at the 2008 ISOQoL annual conference in Montevideo, Uruguay on 25 October 2009.

[4] Following Gadamer, I will use the terms 'interpretation' and 'understanding' synonymously.

point, 'The significance of questioning consists in revealing the questionability of what is questioned' [6]. In pursuing my question about *Twilight*, i.e., in revealing the subject matter's questionability, I may come to understand the novels as propaganda for abusive relationships as opposed to the harmless romance I once took it to be. Moreover, as I come to understand the novel differently, I may also come to see that my original question was ill formed; perhaps the question is not so much why *Twilight* is compelling, but rather, why we are fascinated by relationships characterized by abuse.

Apparent questions, however, are not open. When, for instance, we assess students via an exam, we claim to know something of the subject matter being tested, and thus, we can claim to know the correct answers to the questions we ask. Apparent questions do not reveal the 'questionability' of a subject matter because the subject matter is allegedly already known in these cases. This lack of 'questionability' means that we can assess students based on criteria determined in advance. Moreover, while we may judge that students have answered our questions incorrectly, we are not equally susceptible to student claims that we have asked the wrong question. Apparent questions are not equally open to reinterpretation since, for Gadamer, we can claim to know an answer only when we understand it as the answer to a *particular* question—a question whose meaning is not in doubt [6]. This particularity is in turn a consequence of our knowledge of the subject matter.[5]

The study of philosophical hermeneutics asks how we come to understand the meaning of *Twilight* or a PROM. For Gadamer, this understanding is achieved by asking a text or its analogue genuine questions about its subject matter. How does a genuine dialogue uncover the meaning of a text or text analogue? It does so by revealing the text as something from which we can learn. As Gadamer conceives of it, a text or its analogue is given to us as an answer to some yet-to-be-determined question. The text's meaning is properly conveyed when we come to understand the text as an answer to a particular question. Our job then is to articulate the question for which the text or its analogue is the answer [6].[6] To complete this task we must enter into a dialogue with the text. For by presenting itself to us as an answer to some question, the text itself poses a question to us: How is this text or text-analogue related to the subject matter of our interest? [6]

In attempting to answer this question we bring our own limited understanding of the subject matter to bear on the text. By trying to make sense of the subject matter in terms of what the text says about it, we may find that our understanding of the subject matter has changed or come to disagree with what the text says (we may understand the insight that the text purports to provide and yet still think it is mistaken). The important point, however, is that understanding the meaning of a text or its analogue requires us to ask genuine questions about its subject matter, and in doing so, we open ourselves up to different ways of understanding the subject matter and the questions we ask.

[5] To be sure, we may have misunderstood the subject matter and then our questions may still be wrong.

[6] Collingwood [7] makes a similar point.

PROMs: Questions and answers

In their simplest form, PROMs are a series of questions designed to elicit information about a target construct, i.e., quality of life or perceived health status. Researchers use questions as their probe because information about these constructs cannot be acquired otherwise, for instance, via needles or blood pressure machines.[7] As researchers often say, these constructs lack a 'gold standard'; we lack evidence for the validity of these constructs independent of the measures that assess them—independent, that is, of the questions we ask about them. Quality of life and perceived health status, we might say, are subject matters that researchers imperfectly understand.

According to Gadamer, when we have an imperfect understanding of our subject matter, the questions we ask about it are genuine—questions that open us up to the possibility of understanding the meaning of our subject matter and our questions differently. Yet most PROMs are standardized, thus the questions and their respective meanings are determined in advance. Such questions function as apparent questions and thereby suggest that our subject matter is already understood. To be sure, as the guidelines recently published by the FDA and the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) stipulate, when developing a PROM, researchers should seek respondent input into the meaning of the questions and answers posed [9, 10]. Nonetheless, the point of this qualitative research is to develop questions with a standardized meaning. Thus, the FDA and ISPOR's guidelines suggest that our understanding of quality of life and perceived health status is only imperfect prior to the completion of the qualitative research necessary for measurement development.[8]

Some have disagreed with this conceptualization of PROMs. For instance, proponents of an individualized conception of quality of life, and more recently, of those working on response shift—the phenomena in which respondents understand the same questions differently over time—argue for the on-going legitimacy of different individual judgments about the meaning of quality of life or perceived health status. They argue that these differences are legitimate because the constructs of interest are inherently subjective and thus, on their reading, open to idiosyncratic interpretations [11, 12]. These researchers typically conclude both that we cannot standardize PROMs—no matter how much qualitative work is done—and that we should not attempt to do so [11, 12].

Although I tend to agree with these general conclusions, I have argued elsewhere that it is misleading to hang them on the idea of the 'subjective' or idiosyncratic

[7] To be sure, some have argued that there are 'objective' indicators for measuring quality of life, and thus we need not ask respondents questions. But there is now a strong consensus amongst PROMs researchers that quality of life and perceived health status are 'subjective' constructs, i.e., we need to ask patients questions; see [8].

[8] Although I am suggesting here that the FDA and ISPOR's conceptualization of quality of life and perceived health status is inadequate, it is important to note that these new guidelines represent a *significant* step forward. Most existing measures have been developed with little if any patient input or qualitative research; see [10].

nature of these constructs [13]. In what follows, I offer a different kind of argument, one that hangs on the logic of questions and imperfectly understood constructs.

An imperfectly understood construct

In the June 2005 issue of the *Canadian Journal of Ophthalmology*, Lorne Bellan attempts to answer the following question: 'Why are patients with no visual symptoms on cataract waiting lists?' This question was motivated by the Regional Evaluation of Surgical Indication and Outcomes (RESIO) study in British Columbia where about 30% of patients placed on waiting lists for cataract surgery had a score on the Visual Function Index (VF-14) of 91 or more out of 100 [14].

A score of 100 on the VF-14 indicates no visual complaints and scores of 90 or more suggest virtually no visual problems. Moreover, there are multiple studies that indicate that subjective measures of functional impairment, such as the VF-14, are the best indicators of the degree of impairment and the potential gain from surgery [15, 16]. But if these studies are correct, then why do 30% of the patients placed on cataract waiting lists in British Columbia have scores that indicate that they do not need the surgery [17]? Researchers from the RESIO study answered this question by suggesting that the threshold indications for cataract surgery were now very low, and the media suggested to tax payers that these thresholds were *too* low [17, 18]. Bellan's study probes the accuracy of these suggestions by taking a closer look at the validity of the VF-14 using data from the Manitoba Cataract Waiting List Program [17].[9]

All individuals who were scheduled to undergo cataract surgery in Winnipeg were asked to answer the questions on the VF-14. It asked respondents about the degree of difficulty they perceived themselves to be having while performing common visual tasks, such as watching TV or driving a car. Between January and May 2002, prospective cataract patients who completed the VF-14 and whose scores reported *no* functional impairment (VF-14 scores of 100) were asked to participate in Bellan's study. The 149 individuals who agreed to participate were asked three questions: (1) Are you experiencing any problems with your vision that you were not asked about during the VF-14 questionnaire? (2) Please tell me the reason, as you understand it, for why you have been scheduled to have cataract surgery? (3) What activities do you think will be easier for you after your surgery? Of the 149 patients, 108 were having surgery because of symptoms not specified on the questionnaire, 28 were having it purely based on the doctor's advice, and 13 were asymptomatic [17].

In January 2003, the same patients were contacted after their surgery was completed. They were asked four questions: (1) How satisfied are you with your vision in the eye that was operated on? (2) Did you find that your vision was more impaired than you thought before surgery? (3) Did you feel that your vision

[9] This program is used to monitor and prioritize patients waiting for cataract surgery using the VF-14. This program makes it particularly easy to identify and follow-up on patients with high VF-14 scores; see [17].

improved after cataract surgery? (4) Would you be willing to repeat this type of surgery again [17]? Out of the 149 original participants, 105 had completed their surgery at the time of this second round of questions. Of these participants, 85% were very or extremely satisfied with their surgery, 75% felt that their vision had markedly improved, and 94% were willing to repeat the procedure. Only 9% reported being unsatisfied, saying that their vision had not improved or that they would not repeat the surgery [17].[10]

Bellan concludes from these results that most of the patients who had no perceived impairment according to the VF-14, in fact, did have a large enough perceived impairment to significantly benefit from surgical correction. Bellan concludes that, 'the VF-14 cannot reliably and accurately identify all patients who are likely to benefit from cataract surgery' [17].[11] But notably, he goes on to say that this conclusion is consistent with the findings of the original study in which the VF-14 was developed and validated: the measure was not designed to arbitrate in these kinds of situations [17].

Indeed, Bellan does more than suggest that the use of the VF-14 was inappropriate. He suggests that those familiar with the original design of the VF-14 should have known in advance that the VF-14 would not be sufficiently responsive to determine the need for cataract surgery. In other words, the questions asked in the VF-14 are not sufficiently sensitive to capture those problems that matter to surgical candidates and that cataract surgery can correct. But knowing in advance that the VF-14 is not sufficiently responsive requires us to know quite a lot about the construct of perceived functioning. For one thing, we would need to know that what is important to the perceived functioning of possible candidates for cataract surgery is significantly different from what is important to the perceived functioning of other cataract patients. *Ceteris paribus*, is it reasonable to expect researchers to know this information in advance, i.e., before asking respondents the questions in the VF-14?

I suggest that it is not reasonable. To be sure, PROMs are designed to be specific to a population. So when the target population differs from a PROM's developmental population, the FDA and ISPOR recommend additional qualitative research to ensure external validity [9, 10]. Yet I suggest that it is not always clear in advance when a population is and is not significantly different from the one in which a PROM has been previously validated. When does a difference in population make a difference in the construction of the construct? The FDA's guidelines require that populations be similar with respect to age, sex, ethnic identity, cognitive ability, and, depending on the type of measure, disease state [9]. But the RESIO and

[10] These findings compare favorably with a study in the UK in which the mean VF-14 score was 82.7. In this study 93% of patients described the results of their operation as good or better; 82% found their visual problems much better; 3.4% felt that the surgery was of no benefit or worse [19].

[11] Indeed in the UK, Black et al. found similar results in their 2009 study. In this study between 30 and 50% of patients had pre-operative VH-14 scores that indicated they could achieve little to no improvement. Yet after surgery 93.5% of these patients reported that their problem was better and 93.1% reported their outcome as good to excellent. Black et al. have suggested that a new, improved instrument for assessing the impact of cataract surgery on perceived visual functioning and quality of life is needed; see [19].

Manitoba populations were not significantly different in regards to these variables from the population in which the VF-14 was originally validated [14, 17, 20].

I agree with the spirit of the FDA and ISPOR recommendations, namely, that what constitutes quality of life and perceived health status differs depending on the concrete situation in which they are applied. Additionally, however, I propose that we cannot always determine in advance the features of a situation that might affect the measure. We can be surprised by the meanings that respondents attach to different functionings or experiences depending on the context at hand. I suggest that we *need* studies such as Bellan's precisely because they have the ability to bring these features of a situation into sharper relief.

I suggest that these studies of the VF-14, which were applied to candidates for cataract surgery, reinforce the fact that PROMs' constructs are imperfectly understood. Before applying the VF-14 to the cohorts in British Columbia and Winnipeg, we did not know as much about the perceived functioning of possible candidates for cataract surgery as we did afterward. A charitable interpretation would suggest that researchers did not make a mistake in using the VF-14, i.e., that Bellan's study was not unnecessary. Rather, it was in applying the VF-14 to this context that we were able to learn both about the limitations of the VF-14 and more about the features of perceived functional impairment. To be sure, the researchers in British Columbia *may* have had reason to be cautious with the VF-14 since they were using it in a slightly different context than it had been used previously. But they *were* cautious: 30% of people who did not show any perceived impairment still received surgery.

Our imperfect understanding of PROMs' constructs suggests that we can always learn more about them, and that we can do so by applying them in new situations. Our imperfect understanding of these constructs should not be particularly surprising, after all, PROMs do lack a gold standard. Moreover, the very idea of construct validity appreciates our ability to learn more about constructs since it requires us to continually test and reformulate our theoretical understanding of a construct by measuring it in different situations. Nonetheless, the consequences of the fact that PROMs' constructs are imperfectly understood are often overlooked and, as my discussion of Gadamer indicates, they are far reaching.

Imperfectly understood questions

One of these consequences, I suggest, is that the questions in PROMs ought to be construed as genuine questions; their meaning cannot be standardized. Yet most PROMs are designed as standardized measures. The expectation built into properly developed standardized PROMs is that respondents will understand the questions and answers consistently and uniformly; if they do not, then something has gone wrong. But is this a valid expectation? To answer this question we must first ask another question: what are the conditions required for a consistent and uniform understanding of the questions and answers in PROMs? If PROMs cannot meet these conditions, then we cannot expect respondents to understand them in a standardized fashion.

In their paper, 'Asking Questions and Influencing Answers', Herbert Clark and Michael Schober discuss the importance of context for the consistent and uniform understanding of survey questions. They begin with a discussion of what they refer to as 'common ground'. Here they show how words become meaningful against a shared context; when one changes the context, one often changes what people understand a word to mean [21]. Making a similar claim, Larry Wright asks us to consider the sentence, 'The cat is on the mat.' Typically we conjure ideas of an animal on a rug. The context that provides this interpretation is perhaps a domestic setting. But we can also imagine a different context in which this statement might arise—for instance, a construction site. In this new context the sentence above takes on a new meaning, namely that a piece of machinery is parked on a blasting mat [22].

Common ground, Clark and Schober write, '…is essential in interpreting everything people say' [21]. But simply understanding the contextual setting that is implied by a statement is not enough to disambiguate its meaning; the context should also give us insight into the purpose or aim someone has in uttering a sentence. For example, imagine that one is in the kitchen, and one's partner suddenly says 'the cat is on the mat'. One may know that he is referring to the pet Cooper, but one may not understand the point of what he is saying. Was he looking for the cat? Is the mat off-limits, and therefore, is he expressing frustration? Is his statement a code for, 'now is a good time to give Cooper his meds'? Indeed, in their attempt to make more precise what is involved in their understanding of common ground, Clark and Schober stress the importance of a 'common purpose' [21]. They maintain that purposes are important to understanding because they shape what people mean by what they say [21]. Sharing a common purpose or developing one avoids the confusion that inevitably arises when conversation partners are at *cross*-purposes. But if Clark and Schober are correct, then the consistent and uniform understanding of the questions in PROMs requires, at least in part, that respondents and researchers share an understanding of their purpose.

Echoing elements of Gadamer, Wright suggests that we understand the purpose of a conversation or statement when we understand the *question* to which the discussion or statement is the *answer* [22]. For instance, we understand the significance of 'the cat is on the mat' when we understand it as the answer to a particular question, such as, 'Where is the cat?' or 'What is Cooper doing?' or 'When should we give him his meds?' But Wright's suggestion is misleading if we then think that we can understand these questions without further contextual support; without contextual clues, these questions are also unclear. For instance, in their study, 'Assessing the need for health status measures', Donovan et al. ask respondents to answer yes or no to the question, 'I take tablets to help me sleep'. One respondent answers, 'I take tablets at night for the cramp and they help me sleep. What do I put there?' [23]. Without any further information the respondent is unsure what the question is asking her: is the question a causal one—does she take tablets *because* they help her sleep—or is it interested in *whether* she takes them, for whatever reason?

In another example, highlighting the physical functioning dimension of the SF-36, Sarah Mallinson asks respondents whether their current health limits them in

walking half a mile. They are asked to answer in terms of 'Yes, limited a lot', 'Yes, limited a little', or 'No, not limited at all'. But again respondents were unsure of the meaning of the question: does the question seek to illuminate their fitness level over different terrain or does it seek to illuminate their mere capacity to walk such a distance? Consider, for example, what one respondent had to say:

> I can walk down to the garden centre but there's no way I could get back because it's up-hill, and as soon as I, I can't walk up that hill so it depends which, if you're talking about on the flat, slowly, not talking or carrying anything… I can walk around the shopping precinct and round the supermarket because you're going slowly and you're stopping and looking at things and you're not talking to anybody. [24]

In these examples and others like them the answers provided are not made clear even in light of their respective questions. For these answers to make sense, respondents need to better understand the *question* itself. If we apply Wright's question/answer structure to these examples, then in order to understand what the answers signify, the respondent needs to understand the question about taking tablets to sleep or walking a half-mile as the *answer* to a further question. But what might this question be?

The idea that we understand the purpose of individual questions only when we understand these questions as answers to another question follows the cascading design of PROMS nicely. PROMs set out to answer specific questions, for instance, 'What is the quality of life of those suffering from depression?' But researchers do not ask respondents *this* question. Instead, using a mixture of qualitative and quantitative methods, the researchers break down the research question into a series of further questions, for instance, the question from earlier about taking tablets to sleep. It is *these* questions that respondents are asked to answer. Taken together respondent answers to these questions are intended to provide an answer to the broader research question.

I suggest that it is this research question that provides the purpose with respect to which the individual questions in a measure should be understood. But as I have already discussed, our understanding of questions such as these—our understanding of PROMs constructs—is imperfectly understood. Consequently, if the purpose of the individual questions is imperfectly understood, then the questions themselves are too. We cannot expect respondents to consistently and uniformly understand the questions and answers in these measures because the constructs that these measures assess are themselves not well understood. When respondents understand the questions and answers in PROMs differently than expected, it is not because something has gone wrong; instead, these differences are the natural consequences of asking questions about an imperfectly understood subject matter.[12] Respondents, in attempting to understand the meaning of a set of questions, bring their current understanding of the subject matter to bear on them. It should be no surprise that they often understand them in numerous ways.

[12] This is not to say that every interpretation is legitimate, see [13].

But differences in interpretation need not obstruct research into patient-reported outcomes. In the next section, I examine two studies of PROMs that ask patients to think out loud as they answer questions. Using these kinds of studies in conjunction with quantitative PROMs measures can help us to better understand our measures, but only if we take seriously the insights provided by these patients, and in so doing, treat the questions we ask as genuine ones.

Nottingham Health Profile

Qualitative studies of PROMs that ask respondents to think out loud as they answer the questions in a measure are often used to show the various ways that respondents interpret and respond to questions. For the most part these differences are conceptualized as a problem that research into PROMs should overcome or at least tame.[13] I suggest instead that we should capitalize on these differences, using them to raise questions that will in turn help us to better understand our measures and their constructs.

In their article, 'Assessing the need of health status measures', Donovan, Frankel, and Eyles examine the propriety of using PROMs to determine the health needs of local populations. Their study proceeds by comparing the differences between respondent answers to questions taken from interviews with their answers to the questions in the Nottingham Health Profile (NHP) [23].

During the interview section of the study, Donovan et al. find discrepancies between respondents' answers on the NHP and their interview responses. Donovan and her team, in an effort to overcome these discrepancies, propose that they are due to the NHP's categorical response options; the yes/no alternatives, they claim, are too limiting for respondent answers [23]. Donovan et al. offer the following examples to support their diagnosis:

Things are getting me down: yes/no
I have pain at night: yes/no
I have unbearable pain: yes/no
I take tablets to help me sleep: yes/no

To the first question, one respondent answers, 'I won't let them if I can. Can I put sometimes?' [23] Here the categorical scaling of the question does appear to be limiting. The respondent's answer perhaps reflects the need to implement a continuous judgment scale in the NHP, which would allow for a larger range of response options.

But changing the response scale will not solve all of the NHP's problems. In the other three questions, respondents have rather different sorts of difficulties—difficulties which reflect more than a problem with the yes/no answers. For instance, to the question about pain at night, one respondent replies that she doesn't so much have pain as discomfort [23]. Here her confusion seems to be whether discomfort counts as pain. Similarly, with respect to the question about unbearable pain one

[13] In terms of overcoming, see [25, 26]; in terms of taming, see [27].

respondent replies that the pain is only unbearable when she has a backache [23]. We might say here that the phrase 'unbearable pain' is unclear—is an occasional backache sufficient to count as unbearable pain in general? Lastly, as I discussed in the previous section, to the question about taking tablets to help with sleep, one respondent answers, 'I take tablets at night for the cramp and they help me sleep. What do I put there?' [23]. Here the respondent is unsure what the question is asking her: is it a causal question or a whether question?

While I agree with Donovan and her team that categorical scaling constrains responses, I disagree with their conclusion, namely, that a solution to the respondents' confusion requires a simple adjustment to the scaling options. For if it was clear how we should understand, say, unbearable pain, then the yes/no format of the NHP would not generate the uncertainty that Donovan et al. discovered. As I suggested in the previous section the real difficulty is over how to understand the meaning of these questions. *Should* discomfort count as pain? *Is* occasional pain unbearable pain? *How* do sleeping pills embody quality of life?

One advantage of qualitative studies such as this one is that they bring these sorts of questions to light. I suggest that we should take such questions seriously and attempt to answer them. Attempting to answer such questions allows us to acknowledge that we have something to learn from respondents and reveals our questions as genuine ones. In doing so, we can consider new aspects of our constructs, aspects that are particularly relevant to the circumstances of the cohort of respondents who raise them. Moreover, reflecting on the construct also illuminates the meaning of the questions themselves. For instance, if we decide that discomfort at night is an infringement on the perceived health status of this cohort, then we might rephrase the question, 'I have pain at night' to make this point clearer.

But when respondents are asked to 'think aloud' as they answer the questions in a PROM, they do not always make comments about the questions. Sometimes they simply interpret the questions posed to them and answer accordingly. If their interpretation is unexpected then their answers may be confusing. In such cases, how might we use respondents' answers to improve our understanding of PROMs?

Consider the following question and response from the study by Donovan et al.: 'I find it hard to bend'; 'I do find it hard to bend, but I'm not ticking yes there.' Donovan et al. interpret this response as being contradictory [23]. This result, they suggest, is due to the pressure that respondents feel to give socially acceptable responses [23]. Their interpretation allows them to disregard the respondent's answer thus overcoming the contradiction. But while this respondent's answer *may* represent a contradiction it does not necessarily do so. To see why, consider another question and response: 'Worry is keeping me awake at night', 'Well yes, but it's only stupid things. I lie awake thinking. I'll put no because I'm just being stupid' [23]. Unlike the bending example, this respondent explains why she has decided to mark 'no'; yes she worries, but she only worries about stupid things because she's 'being stupid'.

Although Donovan and her team interpret this second example as yet another instance of contradiction, we might understand it differently. We might take this respondent to understand her worry as different from the sort of worry in which the NHP is interested. The kind of worry the respondent experiences is petty and

'stupid', but the kind of worry that she understands the measure to be interested in involves, say, existential doubt and acute torment.

In these cases, the respondents' answers are contradictory only if we cannot imagine ways of understanding the questions other than our own. I suggest instead that we take all respondent answers seriously. In this case, this means that we begin by assuming that these answers are not contradictory and try to learn something about the respondents' perceived health status by trying to figure out what questions they take themselves to be answering. These different interpretations would then allow us to think about perceived health status as I suggested we should do when asking questions about pills and pain. For example, we might ask, 'Does finding it hard to bend *always* affect quality of life?' 'Is the affect on quality of life *different* when one lies awake worrying about stupid things instead of existential issues?'

EORTC QLQ-C30

It is not only generic measures such as the NHP that provide learning opportunities. In their article, 'Listen to their answers! Response behaviour in the measurement of physical and role functioning,' Westerman et al. listen to small-cell lung cancer patients as they answer the questions on the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30) [28]. In their study, they also find instances similar to those in the previous study: cases where respondents question the meaning of a term and cases where respondents interpret questions in unexpected ways [28]. But additionally, the authors illustrate ways in which the same respondents understand the same questions differently over time. Thus, they explicitly address the issue of response shift: the phenomenon whereby the meaning of one's self-evaluation of quality of life changes over time [29].

For example, one question asks respondents, 'Were you limited in doing either your work or other daily activities?' The possible answers include not at all, a little, quite a bit, and very much. During the second interview, four weeks into chemotherapy, one respondent answers, 'A little, it depends how I'm feeling. If I have a good day, I can take on the whole world. Vacuum cleaning, my motorbike, my car'. At the fourth interview, six weeks after the completion of treatment the same respondent answers, 'A little, the first day back at work again, the tension having to tell everyone the same story over and over again, but of course I feel much better than I did 6 weeks ago.' In this case the respondent answers 'a little' to both questions—according to the data there is no improvement. But Westerman et al. argue that things *did* improve for this patient—he just changed his perspective on what he considered normal [28].

This adjustment can also go in the other direction. For instance, consider the question, 'Were you limited in pursuing your hobbies or other leisure time activities?' During the first interview at the start of chemotherapy one respondent answers, 'My hobby is working in the garden, that's very difficult, quite a bit'. But at the second interview four weeks later the same respondent answers, 'I'm reading at the moment. Gardening is not possible anymore, a little' [28]. This respondent

first answers 'quite a bit' and then 'a little'; according to the data, he appears to be improving. But given the interview portion of his answers, the authors conclude that there is no real improvement. They explain his so-called progress as merely an adjustment to his new situation [28].

In both of the examples above, Westerman et al. conceive of the respondents' second answers as some form of adjustment, either to improved or deteriorating health. In light of this fact, Westerman et al. deny that the second answers signify genuine accounts of quality of life. The real account of quality of life is instead placed in the one respondent's acknowledgement that he feels better than he did six weeks ago; and it is placed in the other respondent's acknowledgement that he can no longer garden. Moreover, in summarizing the different adjustments that respondents made when answering the EORTC questions, Westerman et al. write, 'Through this behaviour it seems that, at a subconscious level, the patients are distancing themselves from the meaning behind the question, i.e., measuring the impact of treatment and disease on their functioning' [28].

Yet, as I have argued already, the fact that the EORTC's questions are attempting to measure the impact of treatment and disease on functioning does not render their meaning unambiguous. I suggest that the respondent 'adjustments', instead of signifying that the respondents are 'distancing themselves from the meaning behind the question', may tell us something important about their quality of life. Indeed, if our questions and constructs are imperfectly understood, then we ignore respondent adjustments at the cost of our own ignorance. I suggest that the phenomenon of response shift is of a piece with other instances where respondents understand questions differently than expected. As in the cases from earlier, we ought to take these respondent answers seriously; we might do so by trying to imagine how their answers could provide us with a coherent account of their quality of life.

Take, for instance, the respondent who consistently answers that he is only a little limited in doing work or other daily activities. Here we can easily imagine a situation in which an individual continues to feel a little limited despite the fact that he can do more activities. This is not unusual. For instance, very ambitious people may report feeling somewhat less than fulfilled in spite of achieving the goals they set for themselves—there is always more to accomplish. These adjustments to one's new situation are common and in many cases admirable. Moreover, we would expect such striving to be part of one's subjective experience and to affect one's quality of life. In some circumstances, striving for more may improve quality of life, for instance, if it helps one to remain optimistic; but it may also adversely affect quality of life, for instance, if it leads to discontent. In any case we cannot assume from this respondent's answer that his quality of life is 'really' improving. His adjustment to his situation is instead part of his quality of life and should serve as a point of departure for further questions.

Similarly, with regard to the gentleman whose limitations appear to improve despite the fact that he can no longer garden, his adjustment may represent an appreciation of new hobbies, hobbies which he had previously overlooked when gardening was possible. Certainly such an adjustment could improve quality of life. Indeed under certain circumstances we would say that such an adjustment was admirable, brave even. To be sure, this respondent may have adapted too quickly or

he may be fooling himself in thinking that reading can replace his garden—his quality of life may have decreased. But we cannot draw this conclusion simply from the fact that he now reads instead of gardens. Instead we must ask further questions to better understand how such adjustments might legitimately and illegitimately relate to quality of life in the context of small cell lung cancer patients undergoing chemotherapy.

By suggesting that these respondents' adjustments are part of their quality of life, I am reiterating the point I made earlier, namely, that what constitutes the meaning of PROMs' constructs changes in different concrete contexts. As patients' health change, we should *expect* changes in their understanding of quality of life, just as it did for patients with different limitations. Indeed these differences follow from the nature of interpretation: if we want to understand a text or text analogue, we must apply it to our present situation [6]. As Gadamer has put the point, '…we understand in a *different* way, *if we understand at all*' [6].

Conclusion

Qualitative studies such as those above are already used in tandem with quantitative studies during the development of PROMs, and there are calls from the FDA, ISPOR, and elsewhere for the qualitative studies to be used more often [9, 10, 30]. My suggestion that we turn to such studies for insight into PROMs is thus not a radical one, and yet the logic and significance of my suggestion is different from other similar appeals. If quality of life and perceived health status are imperfectly understood, context dependent constructs, then so too are the questions that make up PROMs. Our use of quantitative measures must take these facts into account. One suggestion is that we use qualitative studies to survey a sub-section of respondents at different periods throughout the duration of a study—not just during the measure's development—to help provide insight and raise questions regarding the meaning of our constructs and questions. As I have shown, these insights and questions should be taken seriously and will often lead to further questions, questions which may require conceptual analysis, a reassessment of the aims of a study, a need for more respondent input, and so on.

The answers we give to these questions should in turn affect our understanding of the data we collect and when applicable should also affect the measures themselves—we may end up eliminating, reformulating, or adding questions. These improvements, however, are not steps towards stand-alone quantitative measures because PROMs' constructs and questions are imperfectly understood and context dependent; qualitative studies should be an inherent, on-going part of assessing quality of life and perceived health status. On the other hand, qualitative studies are not sufficient. Quantitative PROMs are more efficient than qualitative studies, allow for larger sample sizes, and complement the uses to which other outcome measures are put. Moreover, the data from quantitative PROMs may themselves serve as a point of departure for further questions, as it did in Bellan's cataract study as well as Westerman et al.'s study on lung cancer patients.

Jean Grondin has said, 'To understand something means to have related it to ourselves in such a way that we discover in it an answer to our own questions…' [31]. When respondents attempt to understand the questions in these measures we have seen how they relate it to themselves—how they use these questions to articulate their own questions. It is no wonder that respondents understand the questions in these measures differently. We need to acknowledge this condition and use it to create better measures. An on-going infusion of qualitative work in our quantitative measures is, I believe, one way to achieve this goal; how this process would be operationalised is, however, beyond the scope of this paper. But in time perhaps we will learn to see PROMs more as tools to enhance genuine communication about quality of life and perceived health status, and less as apparent questions used to make determinate assessments.

References

1. Griffin, James. 1986. *Well-being*. Oxford: Clarendon Press.
2. Sen, Amartya. 1993. Capability and well being. In *The quality of life*, ed. Martha C. Nussbaum, and Amartya Sen, 30–53. Oxford: Clarendon Press.
3. Nordenfelt, Lennart. 1993. *Quality of life, health and happiness*. Aldershot: Ashgate Publishing.
4. Sumner, Leonard W. 1996. *Welfare, happiness and ethics*. Oxford: Clarendon Press.
5. Antaki, Charles, and Mark Rapley. 1996. Quality of life talk: The liberal paradox of psychological testing. *Discourse and Society* 7: 293–316.
6. Gadamer, Hans-Georg. 2003. *Truth and method*, 2nd ed. Trans. Joel Weinsheimer and Donald G. Marshall. New York: Continuum Press.
7. Collingwood, Robin G. 1939. *An autobiography*. London: Oxford University Press.
8. Stenner, P. H., D. Cooper, and S.M. Skevington. 2003. Putting the Q into quality of life; the identification of subjective constructions of health-related quality of life using Q methodology. *Social Science and Medicine* 57: 2161–2173.
9. 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.
10. Rotherman, M., L. Burke, P. Erickson, N.K. Leidy, D.L. Patrick, and C.D. Petrie. 2009. Use of existing patient-reported outcome (PRO) instruments and their modification: The ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO Task Force Report. *Value in Health* 12: 1075–1083.
11. Hickey, A., C.A. O'Boyle, H.M. McGee, and C.R.B. Joyce. 1999. The schedule for the evaluation of individual quality of life. In *Individual quality of life: Approaches to conceptualisation and assessment*, ed. C.R.B. Joyce, C.A. O'Boyle, and H. McGee, 119–134. Australia: Harwood Academic Publishers.
12. Schwartz, Carolyn E., and Bruce D. Rapkin. 2004. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health and Quality of Life Outcomes* 2: 14.
13. McClimans, Leah. 2010. Towards self-determination in quality of life research: A dialogic approach. *Medicine, Health Care and Philosophy* 13: 612.
14. Wright, C.J., and Y. Robens-Paradise. 2001. Evaluation of indications and outcomes in elective surgery: A feasibility study in the acute care hospitals of the Vancouver/Richmond health region. Results from the Regional Evaluation of Surgical Indications and Outcomes (RESIO) project. http://www.chspr.ubc.ca/files/publications/2001/hpru01-06R.pdf. Accessed 4 February 2010.
15. Desai, P., A. Reidy, D.C. Minassian, G. Vafidis, and J. Bolger. 1996. Gains from cataract surgery: Visual function and quality of life. *British Journal of Ophthalmology* 80: 868–873.
16. Schein, O.D., E.P. Steinberg, S.D. Cassard, J.M. Tielsch, J.C. Javitt, and A. Sommer. 1995. Predictors of outcomes in patients who underwent cataract surgery. *Ophthalmology* 102: 817–823.

17. Bellan, Lorne. 2005. Why are patients with no visual symptoms on cataract waiting lists? *Canadian Journal of Ophthalmology* 40: 433–438.
18. Wright, C.J., G.K. Chambers, and Y. Robens-Paradise. 2002. Evaluation of indications for outcomes of elective surgery. *Canadian Medical Association Journal* 167: 461–466.
19. Black, N., J. Browne, J. van der Meulen, L. Jamieson, L. Copely, and J. Lewsey. 2009. Is there overutilization of cataract surgery in England? *British Journal of Ophthalmology* 93: 13–17.
20. Steinberg, E.P., J.M. Tielsch, O.D. Schein, J.C. Javitt, P. Sharkey, S.D. Cassard, M.W. Legro, et al. 1994. The VF-14. An index of functional impairment in patients with cataracts. *Archives of Ophthalmology* 112: 630–638.
21. Clark, Herbert H., and Michael F. Schober. 1992. Asking questions and influencing answers. In *Questions about questions: Inquiries into the cognitive bases of surveys*, ed. Judith M. Tanur, 15–48. New York: Sage Foundation.
22. Wright, L. 2001. *Critical thinking: An introduction to analytical reading and reasoning*. Oxford: Oxford University Press.
23. Donovan, Jennifer L., S.J. Frankel, and J.D. Eyles. 1993. Assessing the need for health status measures. *Journal of Epidemiology and Community Health* 47: 158–162.
24. Mallinson, Sarah. 2002. Listening to respondents: A qualitative assessment of the short-form 36 health status questionnaire. *Social Science and Medicine* 54: 11–21.
25. Krabbe, Paul F.M. 2008. Thurstone scaling as a measurement method to quantify subjective health outcomes. *Medical Care* 46: 357–365.
26. Krabbe, Paul, F.M., Noor Tromp, Theo J.M. Ruers, and Piet van Riel. 2008. Better measurement methods may notably reduce response-shifts. 2008 International Society for Quality of Life Research meeting abstracts. *QLR Journal*: A-22, Abstract #1214. http://isoqol.org/pdfs/AbstractsForBooklet_2008v3.pdf. Accessed 4 April 2010.
27. Rapkin, Bruce D., and Carolyn E. 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.
28. Westerman, Marjan J., Tony Hak, Mirjam A. Sprangers, Harry J.M. Groen, Gerrit van der Wal, and Anne-Mei The. 2008. Listen to their answers! Response behaviour in the measurement of physical and role functioning. *Quality of Life Research* 17: 549–558.
29. Goddard, Ruth-Barclay, Joshua D. Epstein, and Nancy E. Mayo. 2009. Response shift: A brief overview and proposed research priorities. *Quality of Life Research* 18: 335–346.
30. Pool, Jan J.M., Sharon R. Hiralal, Raymond W.J.G. Ostelo, Kees van der Veer, and Henrica C.W. de Vet. 2010. Added value of qualitative studies in the development of health related patient reported outcomes such as the pain coping and cognition list in patients with sub-acute neck pain. *Manual Therapy* 15: 43–47.
31. Grondin, Jean. 1994. *Introduction to philosophical hermeneutics*. Trans. Joel Weinsheimer. New Haven: Yale University Press.