Quality of Life Research (2019) 28:3395–3401
https://doi.org/10.1007/s11136-019-02285-8

COMMENTARY

Viewing assessments of patient-reported heath status as conversations: Implications for developing and evaluating patient-reported outcome measures

Kevin P. Weinfurt[1]

Accepted: 23 August 2019 / Published online: 4 September 2019

Abstract
Patient-reported outcome measures (PROMs) are frequently used in research to reflect the patient's perspective. In this commentary, I argue that further improvements can be made in how we develop and evaluate PROMs by viewing assessment as a type of conversation. Philosophically speaking, a PROM assessment can be conceptualized as a formal conversation that serves as a model of an informal, longer, and more nuanced conversation with a research participant about their health experience. Psychologically speaking, evidence from research in survey methodology and discursive psychology shows that respondents to self-report measures behave in ways consistent with the idea that they are doing their best to participate in a conversation, albeit an unusual one. Several suggestions are offered for creating a better conversational context through study materials and PROM instructions, and by improving the yield of cognitive interviews. It is hoped that this commentary can stimulate further discussions in our field regarding how to integrate insights about the conversational nature of assessment from survey research and discursive psychology to better reflect the patient's voice in research.

Keywords Patient-reported outcomes · Theory · Qualitative methods · Survey methods · Discursive psychology · Philosophy · Cognitive interviews

The past decade has witnessed concerted efforts to improve the ability of researchers to reflect the "patient's voice" in clinical research. Toward that end, researchers who develop and use patient-reported outcome measures (PROMs) have tried to improve the quality of these measures through more rigorous development processes that incorporate significant input from key stakeholders, including patients. As someone who has participated in this field for some years, I see even greater opportunities to convey patients' experiences effectively by focusing on the assessment of patient-reported outcomes (PROs) as a form of conversation, or discourse. This idea is not new [1] but seems absent from much of the research that our field is producing. In this commentary, I present a philosophical argument for the essentially discursive nature of PRO assessment, followed by a review of psychological research showing that respondents to standardized self-report questions act in ways consistent with conversational conventions. Finally, I suggest several implications for improving the quality of PRO assessment. (Note that for the purposes of this paper, I use the term *assessment* to mean the act of collecting information using a standardized set of items. This term avoids some of the debate concerning whether an activity is or is not considered a *measurement* [2].)

Conversation and the epistemology of PRO assessment

Upon hearing that PRO assessment is essentially a conversation, one might be concerned that this connotes something unscientific, or at least less scientific than "objective" measures of the body's structures and processes. It is therefore useful to locate PRO assessment within a philosophy of scientific experimentation to understand what type of scientific activity takes place when using PROMs in research. Many of us think of PRO assessment as analogous to using

✉ Kevin P. Weinfurt
kevin.weinfurt@duke.edu

[1] Department of Population Health Sciences, Center for Health Measurement, Duke University Medical Center, 215 Morris Street, Suite 210, Durham, NC 27701, USA

instruments in the physical sciences to measure quantities such as mass, temperature, or velocity. But I agree with Harré [3] that the more appropriate analogy to the physical sciences is between a self-report measure and an experimental *apparatus*. An apparatus is a creation of the researcher designed to serve as a *model* of some phenomenon out in the world—in essence, to create what Harré calls a "mini-world" [4]. For example, a researcher studying osteomyelitis (bacterial infection of bone) might wish to understand how certain bacteria attach themselves to human bones. So, the researcher grows a culture of *Staphylococcus aureus* and places it in a sealed test tube containing a sample of bone matrix from a human cadaver. This experimental setup is a model of the exposure of *S. aureus* in the body to the bones of a living human being (Fig. 1). The researcher makes an observation of the mechanism by which the *S. aureus* in the test tube binds to the bone matrix and, because of the similarity between the test tube's "mini-world" and aspects of real, living human bodies, the researcher infers that this is the same mechanism of binding that occurs in the bones of a living human being. How well the mechanism inferred from observing the test tube setup matches the actual mechanism in living humans is unknown; it is hoped that it is accurate based on the similarity between key aspects of the laboratory setup and real human bodies.

In the case of health status assessment, the natural world consists of a living person who might experience different symptoms or functional impairments, which we come to know about through informal, nuanced conversations with the person. For example, I might learn a great deal about a person's ability to participate in their social roles if I spoke to them for an hour over coffee. Such a conversation would yield a rich understanding of the person's social role participation but would be time consuming and difficult to summarize along with other such interviews. And so, we can create a PROM as a special kind of formalized conversation that serves as a *model* for those longer, richer conversations (Fig. 2). The PROM's formal conversation is much more manageable than informal conversations with research participants, just as working with a test tube in the lab is more manageable than trying to make observations in a living human being. After a person responds to the items on a PROM, the researcher observes those responses to get an impression of the person's social role participation. Mathematical tools (e.g., summary score) can be used to summarize (i.e., "score") the information in the responses, reducing their complexity. The researcher infers that the impression of this person's social role participation is essentially the same as what would be gleaned from having a conversation with the person "out in the wild." Stated differently, the researcher infers that the PROM responses are a reasonable substitute for what would be learned from an extended conversation with the person. As with the laboratory example (Fig. 1), the accuracy of the inference based on the PROM responses depends upon how well the PROM assessment approximates what would be learned by conversing with a person about his or her social role participation under natural conditions.

One of the clearest examples of how PROM assessment can serve as a model for conversations in real life is through the use of computerized adaptive testing (CAT) based on Item Response Theory. Specifically, a CAT algorithm and accompanying item bank can serve as a model of a questioner who is trying to form an impression about a person's health status through a conversation with that person. The questioner forms follow-up questions based on answers to

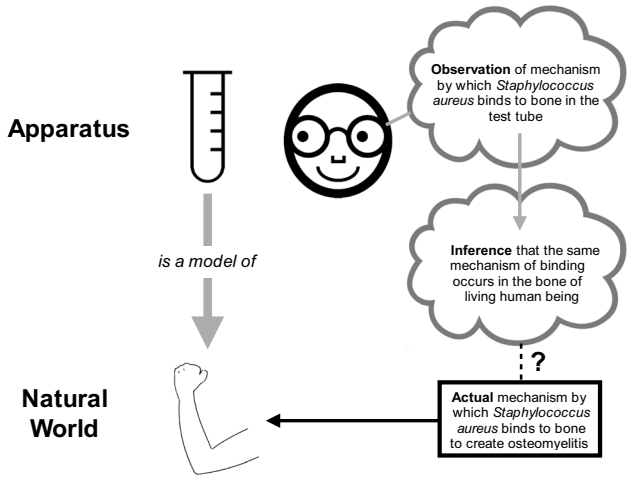

Fig. 1 Rationale for the use of a laboratory apparatus as a model to understand how *S. aureus* binds to bone in osteomyelitis

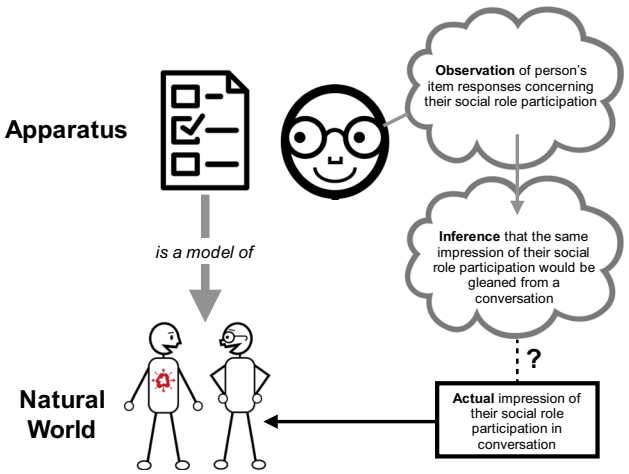

Fig. 2 Rationale for the use of a patient-reported outcome measure as a model to understand a person's experience of social role participation. (Pair of interlocutors pictured in the bottom left include the person with a health condition on the left and the researcher on the right.)

previous ones. The CAT "questioner" uses a similar strategy. To assess someone's mobility it might begin with a general question, such as "How much difficulty are you having getting around during the day?" If the person indicates "Much difficulty," then the next question would be selected to provide the most information about the person's degree of mobility. This would continue and after each question and response, the CAT "questioner" revises its impression of the person's mobility and decides whether it is (a) confident enough in its impression (i.e., the prespecified minimum degree of precision is reached) or (b) that asking more questions would be inconsiderate or create burden for the respondent (i.e., the prespecified maximum number of items has been reached). If the answer to either is "yes," then the questioning stops. If the answer is "no," then the CAT "questioner" continues asking questions, choosing each item from the item bank likely to provide the most information in the region of latent mobility corresponding to the current estimate of the person's mobility. CAT assessment is not a perfect analogue of how such a conversation would unfold in the real world. For example, in a real conversation, if the respondent answered a question in a way that strongly contradicted the emerging impression of mobility, a real questioner would seek to understand the contradiction by probing the respondent.

Respondents to PROMs act as participants in a conversation

The previous section addressed the philosophical status of PROMs as models of informal conversation. In this section, I briefly review psychological work that demonstrates how respondents to standardized questions behave in ways consistent with the view that they are trying to participate in a conversation. Most of this work comes from decades of scholarship in survey methodology, for which there are several excellent summaries [5–7], as well as work in discursive psychology [8, 9].

Respondents' search for meaning in standardized conversations

The dominant body of research in this area makes use of Grice's [10] idea that conversations are governed by tacit conventions that are subsumed by the *cooperative principle of communication*: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (p. 26). People who are responding to self-report items assume that the cooperative principle of conversation applies to the questions being asked and the answers they should give. As Schwarz [7] observes, however, the researchers are typically not acting in accordance with such principles when they design the self-report measure and other study materials. Rather, the researchers are motivated to reduce unwanted variation from the assessment by standardizing the content and presentation of questions to the participants. But this standardization also removes the typical contextual cues and opportunities for negotiation of meaning that occur in day to day life. Absent such cues, the participants do their best to glean meaning from all aspects of the questionnaire. This has been documented in many studies [6, 7], a few examples of which I describe here.

Researcher's interests

In normal conversation, we consider the position, interest, and knowledge of our interlocutor in crafting our responses. Most standardized surveys lack such information, and so participants might look for clues in the study materials. For example, in one study [11] adults were queried about the degree to which 16 different workplace behaviors reflected sexual harassment. Half of the sample of participants were told that the sponsor of the study was a feminist organization called Women Against Sexual Harassment and the questions appeared under the heading "Sexual Harassment Survey." The other half were told that the sponsor was a research organization called Work Environment Institute and the questions appeared under the heading "Work Atmosphere Survey." Respondents in the first half rated the behaviors as slightly more indicative of sexual harassment than those in the second half. This is consistent with the idea that respondents in the first group used the affiliation of the researchers to infer that the researchers had an interest in detecting sexual harassment, and thus the behaviors the researchers included in the survey were likely to be potential cases of harassment.

Numeric values of rating scales

Were we to have a conversation about how successful you thought you were, our conversation would likely include some back and forth about how we are each interpreting "success." In a standardized questionnaire, such discussion is not possible and so participants may look for clues about the intended meaning of "success" based on formal features of the survey, such as the numbers corresponding to the response options. This is exactly what Schwarz et al. [12] found when they randomly presented their sample with response options from either 0 = "not at all successful" to 10 = "extremely successful" or −5 = "not at all successful" to +5 = "extremely successful." Many more participants receiving the first numeric scale rated themselves as less successful relative to those receiving the second scale. Follow-on studies showed that the first scale connoted to the participants that success is a unipolar concept (i.e., degrees

of success), whereas the second connoted a bipolar concept (i.e., unsuccessful to successful). This illustrates how respondents might look to formal features of the questionnaire for clues about the researcher's understanding and intentions, even though the researcher's selection of those features may have been made for reasons other than the meaning of the question (e.g., to keep response scales consistent across items).

Frequency scale response options

When asked to indicate the frequency of some experience, such as physical symptoms, respondents assume that the researchers are engaging in a conversation in good faith and thus chose response scales based on their expert knowledge about the distribution of that experience in the population. Thus, when their memory for the experiences is less than perfect and/or the symptom is relatively ambiguous, respondents may use the distribution of response options as clues as to what constitutes a typical frequency (corresponding to the middle response option) versus extremely low or high frequencies (corresponding to the lowest and highest options, respectively). The result is that reports of the frequency of a symptom might differ substantially depending upon the response options provided [13]. People reading frequency responses given by others can be similarly affected by the framing of the response options, such that even experienced doctors interpreted a symptom with a given frequency as being more severe when that frequency was at the higher end of a set of response options than the lower end [14].

Respondent's use of item responses to do things

Another important result of viewing PRO assessment as conversation is that it encourages us to focus on the pragmatic character of the linguistic expressions involved. Verbal expressions have a semantic meaning—the literal meaning of the words—but in normal conversation, every expression is also some act committed by speaker to pursue an agenda. For example, if a patient were to say to her doctor "My sleep has gotten a lot worse since the last visit," the semantic meaning of the expression is about change in sleep quality. But the patient's intention is to trigger the doctor to do something to address the worsened sleep. The acts that we perform through some expression—requesting help, reassuring family members, describing how well we can do something—are known as *speech acts*, a term coined by Austin [15].

There are two varieties of speech acts. One is the speaker's *intended act* and the other is the *actual effect* that it has on the listener. (Austin used the more technical terms *illocutionary act* and *perlocutionary act* to refer to what I am calling the *intended act* and *actual effect*, respectively.) For example, a patient in a primary care clinic might be asked to complete a pain intensity measure on a 0–10 scale. The doctor's intended act in administering the item is to solicit the patient's experienced level of pain intensity to inform a treatment decision. The patient circles a "9," and the doctor interprets the patient's response as having been done to describe the patient's pain intensity. However, the patient's intended act in circling the "9" might be to get a prescription for a stronger analgesic. The actual effect is the doctor's treatment decision. Thus, it is less appropriate to describe this episode as "a measurement of pain intensity," and more appropriate to describe it as discursive episode, rife with agendas and assumptions.

As another example, consider Cella et al.'s study [16] that assessed the relationship between patients' global ratings of their change with changes in their scores on the Functional Assessment of Cancer Therapy—General (FACT-G). The mean increase in FACT-G scores for people who said they were "minimally better" was smaller than the mean decrement in FACT-G scores for people who said they were "minimally worse." Viewing the patients' global ratings of change as the patients' perceptions of their change would seem to suggest a true disconnect between one method of assessment (i.e., the global rating of change) and another (i.e., difference in FACT-G scores between two time points). In contrast, viewing the person's responses to the global rating of change questions as purposeful expressions suggests alternative interpretations of this asymmetry. For example, studies have documented that, among patients with cancer, there is strong cultural support for expressing a positive attitude and the belief that positive expressions can improve one's outcomes [17]. The intended act of projecting a positive outlook could strongly moderate patients' responses to a question about overall change. A small increase in health might be heralded as a real *improvement*, but patients might be reluctant to endorse the same amount of change in a negative direction as indicating real *worsening*. Larger levels of decrement would be required to overcome culturally based reluctance to express negative results. This illustrates how failure to consider the speech acts increases the risk of misinterpreting the meaning of a health status assessment. (For a rich and related discussion, see McClimans [1, 18].)

Some implications for developing and evaluating PRO assessments

While there are many implications of viewing PRO assessments as discursive activities, I will highlight a few that have the most direct implications for how we develop and evaluate PROMs.

Explicitly communicate the context for the PRO assessment to the participant

To create a more accurate model of an informal conversation, the respondents need to understand the social and motivational context of the questions: Who is really asking me these questions? Why are they asking me these questions? What, if anything, will happen if I give a particular answer? When respondents do not know the answers to these questions, the risk increases that the respondent's intention in providing their answer will not match the researcher's intention in asking the question [7]. Recalling an earlier example, a research participant might overreport their pain severity, because they believe that their doctor will view their answer and make a decision about prescribing an opioid. (Note that an excellent and more elaborated treatment of the role of context in the use of language in PROMs can be found in McClimans [18].)

Those wishing to assess PROs should intentionally craft various communications to participants that can help to "set the stage" for the conversation that is the PROM. Those communications could include conversations with an investigator or study coordinator, informed consent documents, and materials—including PROM instructions—that precede the actual PROM questions. In addition to the typical description of the cognitive tasks involved in completing the PROM, these communications should also provide information about the main actors and their motivations as discussed earlier. What should be included and how it should be expressed could be fruitful subjects of study in more expanded cognitive interviews. Plans to communicate the context to participants should be mindful that many participants have become habituated to boring and/or uninformative documents and instructions, so might be inclined to skip over materials if allowed. Creative solutions should be explored, such as quizzing participants about answers to key questions (e.g., "Why do the researchers want to know about your symptoms today?"), using more attention-grabbing titles, or using brief and engaging videos instead of written instructions. While all of these require additional resources, there might be a significant payoff in the reduction of unhelpful response variation and better overall engagement with participants.

Get more out of cognitive interviews

Cognitive interviews are a standard part of measure development for many researchers [19]. In light of the preceding discussion, there are at least two things that we could do (or do more often) to improve the usefulness of cognitive interviews during the development process for PROMs.

First, cognitive interviews might be enriched by querying people about what they believe their responses mean and what they think are the implications of responding a certain way. This is important because some of the "measurement error" we observe in our PROMs could be due to a mismatch between the researcher's intentions in crafting a question and the respondent's intentions in answering the question. One example of this comes from our work on how participants in phase 1 oncology trials understood and responded to questions about expectations of benefit [20]. These questions were created by researchers with the intention of eliciting the participant's understanding of their prognosis. We used a cognitive interview strategy that uncovered some interesting insights into the speech acts people accomplished with their responses to these questions about expectation of benefit. When someone selected an answer at one end of the scale (e.g., indicating a high expectation that they would benefit from participating in a phase 1 trial), we asked them to imagine someone who selected an answer at the opposite end of the scale and imagine what that person's situation might look like. We found that people regarded a low expectation as morally unacceptable—something a cancer patient should not say. This helped us to understand both the discursive conventions used by patients when talking about expectations and their agendas in crafting their answers—to rally themselves and promote a positive attitude, not to provide their best guess at a prognosis [17]. Thus, the question writers' intentions to solicit estimates of prognosis and the respondents' intentions were fundamentally mismatched. Failure to appreciate this would lead to faulty conclusions made on the basis of patients' responses to these items.

Because intentional speech is usually intended *for* some real or imagined person, it might also be interesting to ask respondents—if it is not already clear from the PROM's instructions or other study communications—who they think will be looking at their responses (e.g., their doctor, an insurance company employee, their spouse), why that person wants to know about their health status, and what they might do with the answer. Additionally, we might ask participants how their responses might change under different contexts, e.g., a researcher trying to understand the health of a large sample of patients, their doctor using the answers to inform treatment decisions, or a hospital trying to improve the quality of care delivered. (Thanks to my colleague, Bryce Reeve, PhD, for this last suggestion.) All of these beliefs might inform a research participant's response to the items. Also, these beliefs might also be amenable to change with a modification to the PROM instructions, items, or both.

Second, cognitive interview data can also help us to appreciate the places where the PROM assessment is a good model of a richer conversation and where it is lacking. This helps us to know what we can expect from a PROM. For example, when developing an earlier version of the Patient-Reported Outcomes Measurement Information System®

(PROMIS®) Sex Functioning and Satisfaction (SexFS) measurement system for sexual function, we found that, despite repeated rounds of cognitive interviews and revisions of items and instructions, there was no way to eliminate variability in how women who used lubricants (e.g., K-Y Jelly) responded to items about their vaginal dryness during sexual activity [21]. In other words, a standardized, formal conversation (i.e., a PROM assessment) is a relatively poor model for the type of conversation needed to really understand the functioning and experience of a woman who is using lubricants. And so, we know that there will be more ambiguity about responses from women who use lubricants. This is similar to other situations in science when we identify types of cases for which a model tends to generate more or less accurate predictions.

Conclusion

Drawing on prior work in survey methodology and discursive psychology, I have argued that viewing PRO assessment as a kind of conversation can focus efforts toward creating a more effective communication. I have offered some recommendations for incorporating this insight into the construction and evaluation of richer conversational contexts for assessment, as well as for increasing the yield of cognitive interviews. I hope that these thoughts stimulate the much-needed conversations within our field about how best to incorporate insights from research in survey methodology and discursive psychology—fields that, to date, have gone unnoticed by the majority of PRO researchers, myself included. (I am grateful to an anonymous reviewer for directing me to the large volume of work done in survey methodology.) Understanding the conversational nature of PRO assessment has the potential to unpack and perhaps reduce what we now label as "measurement error" and, in the process, make more space for the patient's voice in our assessment activities.

Acknowledgements I am grateful to my colleagues who provided invaluable feedback on earlier drafts of this paper: Theresa Coles, PhD; Karon Cook, PhD; Kathryn Flynn, PhD; Bryce Reeve, PhD; Christy Zigler, PhD; and Nancy Zucker, PhD.

Author contribution KW: Conceptualization and drafting of final manuscript.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

Research involving human and animal participants This article does not contain any studies with human participants performed by any of the authors.

References

1. McClimans, L. (2010). Towards self-determination in quality of life research: A dialogic approach. *Medicine, Health Care and Philosophy, 13*(1), 67–76. https://doi.org/10.1007/s11019-009-9195-x.
2. Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory*. New York: Routledge.
3. Harré, R. (2002). *Cognitive Science*. Thousand Oaks, CA: Sage.
4. Harré, R. (1998). Recovering the experiment. *Philosophy, 73*(285), 353–377. https://doi.org/10.2307/3751988.
5. Gobo, G., & Mauceri, S. (2014). *Constructing survey data*. Thousand Oaks: Sage.
6. Schwarz, N., & International Statistical Review/Revue Internationale de Statistique. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. *JSTOR, 63*(2), 153. https://doi.org/10.2307/1403610.
7. Schwarz, N. (2010). Measurement as cooperative communication: What research participants learn from questionnaires. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 43–61). Thousand Oaks, CA: Sage.
8. Harré, R., & Stearns, P. (Eds.). (1995). *Discursive psychology in practice*. Thousand Oaks: Sage.
9. Harré, R., & Gillett, G. (1994). *The discursive mind*. Thousand Oaks, CA: Sage.
10. Grice, P. (1989). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 22–57). Cambridge, MA: Harvard University Press.
11. Galesic, M., & Tourangeau, R. (2007). What is sexual harassment? It depends on who asks! Framing effects on survey responses. *Applied Cognitive Psychology, 21*(2), 189–202. https://doi.org/10.1002/acp.1336.
12. Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55,* 570–582.
13. Schwarz, N. (1999). Frequency reports of physical symptoms and health behaviors: How the questionnaire determines the results. In D. C. Park, R. W. Morrell, & K. Shifren (Eds.), *Processing of medical information in aging patients* (pp. 93–108). New York: Academic Press.
14. Schwarz, N., Bless, H., Bohner, G., Harlacher, U., & Kellenbenz, M. (1991). Response scales as frames of reference: The impact of frequency range on diagnostic judgements. *Applied Cognitive Psychology, 5*(1), 37–49. https://doi.org/10.1002/acp.2350050104.
15. Austin, J. L. (1962). *How to do things with words* (2nd ed.). Cambridge, MA: Harvard University Press.
16. Cella, D., Hahn, E. A., & Dineen, K. (2002). Meaningful change in cancer-specific quality of life scores: Differences between improvement and worsening. *Quality of Life Research, 11*(3), 207–221.
17. Sulmasy, D. P., Astrow, A. B., He, M. K., Seils, D. M., Meropol, N. J., Micco, E., et al. (2010). The culture of faith and hope: Patients' justifications for their high estimations of expected therapeutic benefit when enrolling in early phase oncology trials. *Cancer, 116*(15), 3702–3711. https://doi.org/10.1002/cncr.25201.
18. McClimans, L. (2010). A theoretical framework for patient-reported outcome measures. *Theoretical Medicine and Bioethics, 31*(3), 225–240. https://doi.org/10.1007/s11017-010-9142-0.
19. DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(Suppl 1), S12–S21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2.
20. Weinfurt, K. P. (2013). Understanding what participants in empirical bioethical studies mean: Historical cautions from William

James and Ludwig Wittgenstein. *AJOB Primary Research, 4*(3), 49–54. https://doi.org/10.1080/21507716.2013.807893.
21. Fortune-Greeley, A. K., Flynn, K. E., Jeffery, D. D., Williams, M. S., Keefe, F. J., Reeve, B. B., et al. (2009). Using cognitive interviews to evaluate items for measuring sexual functioning across cancer populations: Improvements and remaining challenges. *Quality of Life Research, 18*(8), 1085–1093. https://doi.org/10.1007/s11136-009-9523-x.