Qual Life Res (2008) 17:1239–1246
DOI 10.1007/s11136-008-9402-x

Cognitive interviewing in the evaluation of fatigue items: Results from the patient-reported outcomes measurement information system (PROMIS)

Christopher Christodoulou · Doerte U. Junghaenel · Darren A. DeWalt · Nan Rothrock · Arthur A. Stone

Accepted: 18 September 2008 / Published online: 12 October 2008

Abstract

Objectives Cognitive Interviewing (CI) is a technique increasingly used to obtain respondent feedback on potential items during questionnaire development. No standard guidelines exist by which to incorporate CI feedback in deciding to retain, revise, or eliminate potential items. We used CI in developing fatigue items for the National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System (PROMIS) Roadmap initiative. Our aims were to describe the CI process, formally evaluate the utility of decisions made on the basis of CI, and offer suggestions for future research.

Methods Participants were 22 patients with a diverse range of chronic health conditions. During CI, each participant provided feedback on a series of items. We then reviewed the CI data and decided whether to retain, revise, or eliminate each potential item. Following this, we developed or adopted three quantitative methods to compare retained versus eliminated items.

Results Retained items raised fewer serious concerns, were less likely to be viewed as non-applicable, and were less likely to display problems with clarity or to make incorrect assumptions about respondents.

Conclusions CI was useful in developing the PROMIS fatigue items and the methods used to judge CI for the present item set may be useful for future investigations.

Keywords Cognitive interviewing · Outcomes assessment · Qualitative methods · Quality of life · Questionnaire development

Introduction

The Patient-Reported Outcome Measurement Information System (PROMIS) is a multi-center, collaborative project funded under the NIH Roadmap for Medical Research Initiative to improve the measurement of clinically important symptoms and outcomes. PROMIS aims to optimize the accuracy and efficiency by which patient-reported outcomes (PROs) are assessed and employed in research and clinical practice. Its goal is to develop and standardize a set of item banks that allow the assessment of key symptoms and health concepts relevant to a wide range of patient-reported chronic disease outcomes [1]. The first set of PROMIS item banks focuses on the domains of emotional distress, social function, physical function, pain, and fatigue.

The present investigation focused on fatigue, an experience familiar to almost all people and applicable to a variety of situations (e.g., occupational, academic,

C. Christodoulou (✉)
Department of Neurology, HSC T12-028, Stony Brook
University, Stony Brook, NY 11794-8121, USA
e-mail: christopher.christodoulou@sunysb.edu

D. U. Junghaenel · A. A. Stone
Department of Psychiatry and Behavioral Sciences, Applied
Behavioral Medicine Research Institute, Stony Brook
University, 125 Putnam Hall, South Campus, Stony Brook,
NY 11794-8790, USA
e-mail: arthur.stone@sunysb.edu

D. A. DeWalt
Division of General Internal Medicine, University of North
Carolina at Chapel Hill, Chapel Hill, NC, USA

N. Rothrock
Center on Outcomes, Research, and Education (CORE),
Evanston Northwestern Healthcare and Northwestern University
Feinberg School of Medicine, Chicago, IL, USA

athletic, and medical). Researchers have struggled for years to define and measure fatigue in a broadly acceptable manner [2–4]. Fatigue can be measured as a decline in behavioral performance over time (e.g., when the number of pounds a weightlifter can bench press lessens with repetition), but it is most commonly assessed as a subjective feeling by means of self-report questionnaires [3, 5]. The provisional definition used for PROMIS reflects an interest in medically relevant pathological fatigue. Fatigue is defined as an overwhelming, debilitating, and sustained sense of exhaustion that decreases one's ability to carry out daily activities, including the ability to work effectively and to function at one's usual level in family or social roles [6–8].

During the PROMIS project, a variety of qualitative and quantitative methods have been applied to potential fatigue items in an effort to create item banks characterized by items with a high degree of both precision and range [9, 10]. One of these techniques is cognitive interviewing (CI) [11]. During CI, respondents are probed for their interpretation of question content and response options to determine potential problems or concerns associated with each item. This feedback is used by test developers to refine and improve their questionnaires. Methods are available to help interviewers elicit and categorize types of respondent feedback, for example, in terms of ambiguity, language, comprehensibility, and relevance of items [11, 12].

Despite the increasing use of CI in questionnaire development [13–15], little has been written to quantify its benefits [16, 17]. This is most likely due to the qualitative nature of CI and the feedback it generates. While such data are not easily quantified, we sought to determine if there was an approximate correspondence that could serve to quantitatively corroborate its benefit to questionnaire design. We recognized that such an approach would never capture the full richness and complexity of the CI process, but wanted to provide an initial quantitative assessment of its utility. This was important for the PROMIS project because the initial list of potential items came from existing questionnaires that were already in use.

Focusing on fatigue, the goal of the present investigation was to demonstrate that the items chosen for retention during CI were quantifiably better than those that were eliminated. We evaluated the performance of retained versus eliminated items in terms of: (1) the number of serious concerns raised by the respondents about the item, (2) the number of items respondents viewed as non-applicable, and (3) the number of specific types of problems reported (e.g., clarity), following the scheme described by Willis and Lessler in the Question Appraisal System (QAS-99) [18].

Methods

General PROMIS methods

The process began with a step-wise qualitative item review that included: (1) identification of items from existing fatigue scales, (2) item classification and selection, (3) item revision, (4) focus group exploration of domain coverage, (5) CI on individual items, and (6) final revision before field testing [9]. More than 80 fatigue questionnaires were initially reviewed, resulting in a list of over 1,000 potential fatigue items, though many of these were quite similar to one another. By the time of the CI step, a total of 136 potential items remained (examples in Table 1). These items were grouped into four non-overlapping sets of 34 items each, with one set administered to each subject in the first round of CI (some subjects were administered additional revised items during a second round of CI). We allowed similar questions to be grouped together so that respondents could consider and comment on the similarities and differences between wording choices if that seemed important to them.

Recruitment of CI participants

The participant sample was intended to represent a diverse range of chronic health conditions (e.g., diabetes, chronic pulmonary disease, cardiovascular disease, musculoskeletal disease, chronic pain, and chronic gastrointestinal conditions) and socio-demographic characteristics. Fatigue is common in many of these conditions, but there is also substantial variability between individuals. The aim was to include subjects with mild, moderate, and severe levels of fatigue in the review of each item. Participants were interviewed at the University of North Carolina (UNC),

Table 1 Examples of PROMIS fatigue items

PROMIS fatigue item examples
1. How often did your fatigue make it difficult to plan activities ahead of time?
(1) Never
(2) Rarely
(3) Sometimes
(4) Often
(5) Always
2. How bushed were you on average?
(1) Not at all
(2) A little bit
(3) Somewhat
(4) Quite a bit
(5) Very much

Chapel Hill, Medical School. Potential participants were recruited from two sources: the North Carolina Musculoskeletal Health Project and the UNC General Internal Medicine Practice.

The North Carolina Musculoskeletal Health Project is a collaborative database established by researchers and clinicians at the Thurston Arthritis Research Center and Department of Orthopedics at the UNC Medical School. The database contains a list of consecutive patients from the rheumatology and orthopedics clinics seen at UNC who consented to participate in future studies. Potential cognitive interview participants were mailed an invitation letter that provided an overview of the purpose and nature of the cognitive interviews and asked if they would be willing to participate. Interested patients could contact the study personnel by email or phone. The research staff also followed up with phone calls to assess interest in participating in the study and to determine eligibility. In addition, patients were directly approached and screened for eligibility at the UNC General Internal Medicine Practice with the permission of the treating physician. This study was previously approved by the UNC Institutional Review Board and is protocol # 05-2571.

Inclusion criteria

Patients were eligible to participate in the cognitive interviews if: (1) they were at least 18 years of age, (2) had seen a physician for a chronic health condition within the past 5 years, (3) were able to speak and read English, (4) were willing to provide written informed consent prior to study entry, (5) had no concurrent medical or psychiatric condition that, in the investigator's opinion, may preclude participation in this study, and (6) had no cognitive or other impairment (e.g., visual) that would interfere with completing an interview.

Conducting the cognitive interviews

Each CI was conducted face-to-face and lasted approximately 45–60 min. Patients completed paper and pencil questionnaires consisting of 34 items (from the total of 136), and then were debriefed by the interviewers. Going item by item, the interviewer asked a series of open-ended questions, following a script, seeking comments with regard to the item stem (body of the question), the response options, and the time frame (the period covered by the questions, which was uniformly set at 7 days). The interviewer asked summary questions at the end of the interview (Table 2).

Table 2 Probes for cognitive interview

Cognitive interview probes
(1) How would you say this question in your own words?
(2) How easy or hard was this question to answer? (If interviewee finds it hard to answer) How would you reword the question to make it easier to answer?
(3) How easy or hard was it to tell the difference between each choice?
(4) You chose (quote their choice) as your answer. What does (quote their choice) mean to you?
(5) If you could change the answers to this question to make them easier to understand, what would you do—if anything?
(6) When you read the words "past 7 days" which days did you think of? (e.g. From which day to which day?
(7) Did you think mostly about your experience on specific days or what was typical for you over the last 7 days (If specific days, "can you tell me more about what made you think about those specific days?")
(8) Would you have responded to this statement differently if we asked you about what had happened over the last 30 days instead of only the last 7 days?
(9) Finally, what could we do, if anything, to improve these questions when we use them in the future with other people like you?

All 136 items were reviewed by five to six participants during the first interview round, and the 19 items subjected to a second round of CI were reviewed by a minimum of three more participants. While this is not a large number of CI per item, it should be noted that most of the PROMIS items were taken and slightly modified from existing questionnaires that had already been used in large numbers of subjects. In addition, CI was only one in a series of techniques used to refine the questionnaire items.

CI data was collected by trained interviewers at UNC, Chapel Hill Medical School. They were faculty or graduate students in public health or social work who underwent two CI training sessions for 4 h each, including methods, protocol review, and practicing with feedback. All interviews were conducted by two staff. One conducted the interview while the other took detailed notes and recorded the interview. Recordings were only used to fill in gaps in the notes and were not transcribed verbatim. After the interview, one staff took the notes and organized them into a cohesive report along with the comments from the other cognitive interviews for the given item.

Modification of items on the basis of CI

After completion of the first round of cognitive interviews, on an item-by-item basis, we decided if each item needed revision based on feedback from cognitive debriefing. As mentioned in the introduction, there is no standard method for using CI feedback to modify items. The summary of CI feedback for each item was reviewed by a group of five individuals at Stony Brook University including persons with expertise in the study of fatigue and the development of self-report measures [the present authors (A.A.S., D.J., and C.C.) and two other members of the research team]. The group decided on a consensus basis whether to retain, revise, or eliminate each item. In arriving at a decision, the group placed particular weight on comments that arose in the feedback of more than one respondent. However, a single negative remark was occasionally enough to lead to a decision to revise or eliminate an item (e.g., a remark signaling a serious misunderstanding of the item stem). For items judged as requiring substantial revision, a second round of CI was undertaken with three to five participants reviewing each item.

Evaluation of decisions made in response to CI

After the CI process was complete, we decided to formally evaluate whether the items we accepted fared better than the eliminated items in terms of the concerns raised by subjects during CI. Revised items that were sent back for re-evaluation after Round 1 were not re-rated, since we were most interested in the *final disposition* of an item (retained versus eliminated). Items revised after Round 1 were only rated after completion of CI Round 2 when their final disposition (i.e., retained versus eliminated) was known. (There were two items that were revised after Round 2 without another CI round, and we decided to exclude these items from analysis.)

We developed and adopted methods to assess the following questions regarding the accepted versus eliminated items: (1) Did the retained items have fewer serious CI concerns than eliminated items? (2) Were eliminated items more likely to be seen as non-applicable to respondents' lives? (3) What types of concerns were raised for eliminated versus retained items using the QAS-99 [18]? Below, we describe the methods used to address each of the questions raised. Two of the present authors (D.J. and C.C.) employed these methods approximately 4 months after the initial decisions had been made. In an effort to minimize the influence of the earlier decisions on the more quantitative formal evaluations, information on item disposition was removed from item spreadsheets that were used during the formal evaluations.

Metric for evaluating if the retained items raise fewer serious concerns during CI than eliminated items

We categorized the number of concerns that were raised for each item into *mild* concerns and *serious* concerns. Concerns were defined as follows.

Mild concern

We considered a concern *mild* when a subject *suggested* alternate wording without specifically stating that the current wording was bad. Words like "preferred", "offered", or "suggested" were considered a "mild" concern.

Serious concern

We considered a concern *serious* if one or more of the following conditions were met: (1) the respondent *insisted* on a wording change, using expressions like "should", "needs to", "must", etc.; (2) the respondent specifically said something negative about the existing item, regardless of whether the respondent provided alternate wording or not; or (3) the comments of a respondent reflected a misunderstanding of either the item stem or the response options.

Each item was reviewed by three to six CI participants (five to six in the first round of CI, and at least three in the second round) who could indicate whether they had problems with the stem and/or response categories. Because the total number of concerns raised (either mild or serious) could differ based on the number of participants, the number of concerns for each item was divided by the number of participants who viewed that item. We calculated this separately for mild and for serious concerns. Our primary focus was on the serious concerns. This procedure allowed us to get a quantitative picture of the number of concerns raised for each of the eliminated as well as retained items.

We determined the degree to which we were able to reliably assess the severity of concerns raised by participants by having two raters jointly review a subset of items ($n = 19$). The raters coded and then discussed the items one at a time in an effort to increase inter-rater coding consistency. The remaining items were then rated independently. The two raters classified participants' concerns as mild and/or serious in an identical fashion for 93% of the items (111/119). As an alternative measure of reliability, we obtained the intraclass correlation, which yielded a value of 0.91 ($P < 0.001$) for *mild concerns* and 0.93 ($P < 0.001$) for *serious concerns* (using a two-way mixed model). The few remaining differences between raters were resolved by joint discussion of the items, so that all ratings reported were the consensus opinion of both raters.

Metric for evaluating if eliminated items were more likely to be seen as non-applicable to respondents' lives

We counted the number of subjects who stated that an item was not applicable to their lives during the past 7 days. For example, respondents commented that the particular experience or particular event mentioned in the item did not occur for them during that time period. For example, subjects not working rated the following item as non-applicable: “how often did you feel used up at the end of the workday?”

As with the severity ratings, two raters jointly reviewed a subset of items ($n = 19$). The raters coded and then discussed the items one at a time in an effort to increase inter-rater coding consistency. The remaining items were then rated independently. A 99% agreement rate (118/119) was achieved for the applicability ratings and the intraclass correlation for applicability ratings was 0.86 ($P < 0.001$). The difference between raters on the single item was resolved by joint discussion of the item, and all ratings reported were the consensus opinion of both raters.

Metric for evaluating the types of concerns raised for eliminated versus retained items using the QAS-99

We used the QAS-99 [18] as a method of categorizing the item problems identified during the CI process. The QAS-99 consists of eight major categories that address item problems (Table 3). Most of the QAS-99 categories (categories 3–8) identify types of problems that are associated with each item from the respondent’s perspective, but category 1 (Reading) pertains to difficulties reading items from the *interviewer’s perspective* and category 2 (Instructions) pertains to difficulties respondents have with the *overall instructional set* rather than to any individual item. Because the focus of the CI in PROMIS was to obtain *item-by-item analysis from the respondent’s perspective*, we excluded categories 1 and 2. Therefore, the major categories we assessed were: Clarity, Assumptions, Knowledge/Memory, Sensitivity/Bias, Response Categories, and Other Problems.

To ensure that the complexity of the rating task was captured, we chose to establish inter-rater reliability for the QAS-99 classifications on the items with the highest likelihood of exhibiting problems; that is, the 55 items that were revised or eliminated in each round of CI. For training purposes, a subset of the items ($n = 8$) that would undergo QAS-99 classification was jointly reviewed and discussed by the two raters. The raters coded and then discussed the items one at a time in an effort to increase inter-rater coding consistency. Inter-rater reliability was established on the 47 remaining items.

Table 3 QAS-99 categories [18]

Step 1: *Reading*: Determine if it is difficult for the interviewers to read the question uniformly to all respondents.

Step 2: *Instructions*: Look for problems with any introductions, instructions, or explanations from the *respondent’s* point of view.

Step 3: *Clarity*: Identify problems related to communicating the *intent* or *meaning* of the question to the respondent.

Step 4: *Assumptions*: Determine if there are problems with assumptions made or the underlying logic.

Step 5: *Knowledge/Memory*: Check whether respondents are likely to *not* know or have trouble remembering information.

Step 6: *Sensitivity/Bias*: Assess questions for sensitive nature or wording, and for bias.

Step 7: *Response Categories*: Assess the adequacy of the range of responses to be recorded.

Step 8: *Other*: Look for problems not identified in Steps 1–7.

Establishing inter-rater reliability for QAS-99 ratings was more complicated than for the severity and non-applicability ratings, because it was possible to assign more than one QAS-99 problem category to each item [18]. Thus, the raters could agree on some but not all of the same categories. For example on item X, Rater 1 could assign problems with Clarity and Assumptions; for the same item, Rater 2 could assign problems with Clarity and Response Options. As a result, inter-rater agreement could be determined in multiple ways and we defined two levels of agreement: identical and partial. *Identical agreement* required that the choices of the two raters were exactly the same. Identical agreement was obtained for 79% (37/47) of the items. *Partial agreement* is a more lenient standard. It required that at least one of the choices of the two raters (and possibly more) was the same. Partial agreement was obtained for 91% (43/47) of the items. All differences on the QAS-99 were resolved by the two raters, following discussion of the items, and all ratings reported can be considered the consensus opinion of both raters.

Results

Cognitive interview participants

The sample of cognitive interview participants consisted of 22 patients (Table 4). Patients reported a wide array of medical diseases, including diabetes, high blood pressure, depression, liver disease, and inflammatory bowel syndrome. The most frequent diagnoses were arthritis, heart disease, and chronic pain. A median of three medical diagnoses was reported.

The two CI rounds

A total of 136 fatigue items were submitted for the *first round* of CI. Of those, 33 (24%) were eliminated and 19 (14%) were sent back for reevaluation in a *second CI round*. Not all of the 19 items sent back for re-evaluation were revised; three were reviewed again in their original form, because we felt that the original CI comments were not definitive and additional CI feedback was desired. In the second round of testing, only three (16%) items were eliminated ($n = 1$) or revised ($n = 2$). At the completion of both rounds of CI, 102 items (75%) were deemed acceptable and 34 items (25%) were eliminated.

Table 4 Characteristics of cognitive interview participants ($n = 22$)

Characteristic	
Mean age in years	63.5 (SD = 11.6)
% Female	55%
Reported race	50% White 50% Black
Education	
≤High school graduate	50% ($n = 11$)
Some college/technical degree	18% ($n = 4$)
College graduate (BA, BS)	14% ($n = 3$)
Advanced degree (MA, MD, PhD)	18% ($n = 4$)
Reported ethnicity	0% Latino/Hispanic
Median number of reported medical diagnoses	3 (range 1–9)
Most frequently reported medical diagnoses	Arthritis (n = 16, 73%) Heart disease (n = 10, 45%) Chronic pain (n = 9, 41%)

Did the retained items raise fewer serious concerns during CI than eliminated items?

Non-normal distributions were found for each of the variables measuring concern: mild concerns per participant (median = 0.0, range 0–1.0), serious concerns per participant (median 0.0, range, 0–2.2). Hence, analyses comparing the number of concerns and non-applicability ratings between retained and eliminated fatigue items were conducted using the non-parametric Mann Whitney test (Table 5). Results showed higher mean ranks of serious concerns for eliminated items compared to retained items ($P < 0.001$). No differences were found for mild concerns.

Were eliminated items more likely to be seen as non-applicable to respondents' lives?

Deviations from normality were also found for non-applicability ratings per participant (median = 0.0, range 0–0.7), so the comparisons of non-applicability ratings between retained and eliminated fatigue items were conducted via the non-parametric Mann Whitney test (Table 5). Eliminated items had a higher mean rank of non-applicability ratings compared to retained items ($P < 0.01$).

Table 5 Higher numbers indicate more concerns/non-applicability

	Mean rank		z	P
	Retained items	Eliminated items		
Mild concerns	70.3	63.2	−1.28	0.20
Serious concerns	57.0	103.2	−6.58	<0.001
Non-applicability	65.9	76.3	−3.08	<0.01

What types of concerns were raised for eliminated versus retained items using the QAS-99?

Concerns raised about items were classified into QAS categories (71% had one QAS concern, and 27% had two QAS concerns). Differences were examined between the types of concerns raised for retained and eliminated items for each of the QAS-99 categories. Concerns in the category of Clarity were more common for eliminated items (70%; 24/34) than for retained items (30%; 31/102) ($P < 0.01$). Likewise, concerns pertaining to Assumptions (i.e., items that were non-applicable to some people, items that assumed constant behavior, or items that were double-barreled) were raised for 29% (10/34) of the eliminated items compared to 11% of the retained items (11/102) ($P = 0.01$). No differences were found for Knowledge/Memory-related concerns, and concerns regarding Response Categories. Sensitivity/Bias-related concerns or Other Problems were not coded for any of the items.

Discussion

Cognitive interviewing (CI) is a technique increasingly used to elicit respondent feedback to aid in the development of questionnaires [11, 19]. The process of CI has been described, and the kinds of problems that items can display have been categorized [11, 20]. However, despite its increased use, there is little quantitative evidence demonstrating the benefits of the qualitative feedback it generates [16, 17]. Our goal was to develop measures to quantify the impact of CI on the development of fatigue items for PROMIS, and we present our experience as a suggested way of synthesizing the vast amount of input one can acquire from even a small number of interviews.

Our decisions on how to utilize respondents' feedback during CI were originally based on informal group consensus following a review of the transcripts; that is, they were qualitative in nature. In an a posteriori effort, we developed quantitative strategies to measure CI feedback in terms of severity and non-applicability and adopted methods to determine specific problem types. This allowed us to evaluate the quality of the initial group decisions.

Results from each of the strategies provided consistent support that the retained fatigue items were better received by CI respondents than eliminated items and supported our initial qualitative decisions. First, we found that respondents raised fewer serious concerns for retained versus eliminated items. Second, respondents were more likely to view the eliminated items as non-applicable to their lives during the recall period of 7 days. The low average number of serious concerns raised per item was likely due to the relatively high quality of the original questions, which were adapted from existing fatigue questionnaires.

With regard to QAS categories of concern, retained items were less likely than those eliminated to raise concerns regarding their clarity or to make incorrect assumptions regarding the respondent. Some categories of QAS concerns were not mentioned at all by participants (e.g., Sensitivity/Bias), which may reflect the absence of probes addressing those issues or may have resulted from idiosyncratic properties of these fatigue items.

There may be a concern that individuals who made the initial qualitative CI decisions were also involved in the quantification of CI feedback after the fact, and that this may have increased the level of consistency found between qualitative and quantitative methods. The influence of the earlier qualitative decisions on the later quantitative measurements was minimized to some extent by the waiting period of approximately 4 months before quantitative measurements were taken. In addition, we removed items' final disposition (i.e., retained versus eliminated) resulting from our informal method from the relevant spreadsheets. We decided after the qualitative decisions had been made to determine whether it was possible to apply quantitative methods to this type of qualitative information. Refinement of CI procedures is warranted given their increased use in questionnaire development [12]. Future studies attempting to compare the two approaches to measurement should ensure that they are made independently.

CI is an inherently qualitative process that is not easily quantified [19]. Interpreting CI feedback is not a simple exercise where a majority rules. If one person in a small sample has a different interpretation of an item, that can be enough for test developers to consider revision or elimination [19]. Nonetheless, a quantitative approach can help bring data across items into a standard metric, making it easier to identify common concerns among item candidates (e.g., inappropriate assumptions about the lives of respondents). The use of a quantitative approach can make the process of CI interpretation more transparent, consistent, and reproducible. We recognize that our approach to its quantification is imperfect. However, there did appear to be a high enough correspondence to validate the utility of CI for this set of items. It can be useful to point to this type of quantitative data in demonstrating the benefits of CI, particularly to persons less familiar with the technique.

Acknowledgements The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, and Susana Serrate-Sztein, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the web site at http://www.nihpromis.org for additional information on the PROMIS cooperative group. We thank the team at the University of North Carolina, Chapel Hill that conducted the cognitive interviews. We also thank Pamela Calvanese and Maria Bolger for helping to review the CI feedback on the fatigue items. A.A.S. is the chair of the scientific advisory board of invivodata, inc. and a senior scientist at the Gallup Organization.

References

1. Cella, D., Yount, S., Rothrock, N., et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*, S3–S11. doi:10.1097/01.mlr.0000258615.42478.55.
2. Bartley, S. H., & Chute, E. (1947). *Fatigue and impairment in man*. New York: McGraw-Hill.
3. Christodoulou, C. (2005). The assessment and measurement of fatigue. In J. DeLuca (Ed.), *Fatigue as a window to the brain* (pp. 19–35). New York: MIT Press.
4. Yellen, S. B., Cella, D. F., Webster, K., et al. (1997). Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *Journal of Pain and Symptom Management, 13*, 63–74. doi:10.1016/S0885-3924(96)00274-6.
5. Wessely, S., Hotopf, M., & Sharpe, D. (1998). *Chronic fatigue and its syndromes*. New York: Oxford University Press.
6. Stewart, A. L., Hays, R. D., & Ware, J. E. (1992). Health perceptions, energy/fatigue, and health distress measures. In A. L. Stewart & J. E. Ware (Eds.), *Measuring functional status and well-being: The medical outcomes study approach* (pp. 143–172). Durham, NC: Duke University Press.
7. North American Nursing Diagnosis Association. (1996). *Nursing diagnoses: Definition and classification, 1997–1998*. Philadelphia, PA: McGraw-Hill.
8. Glaus, A. (1998). *Fatigue in patients with cancer: Analysis and assessment*. Heidelberg, Germany: Springer.
9. DeWalt, D. A., Rothrock, N., Yount, S., et al. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*, S12–S21. doi:10.1097/01.mlr.0000254567.79743.e2.
10. Reeve, B. B., Hays, R. D., Bjorner, J. B., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*, S22–S31. doi:10.1097/01.mlr.0000250483.85507.04.

11. Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage.
12. Bartenfeld, T. A. (2003). Department of health and human services: Center for disease control and prevention: Proposed data collections submitted for public comment and recommendations. *Federal Register, 68*, 35227.
13. Wu, H. S., & McSweeney, M. (2004). Assessing fatigue in persons with cancer—An instrument development and testing study. *Cancer, 101*, 1685–1695. doi:10.1002/cncr.20540.
14. Hyde, M., Wiggins, R. D., Higgs, P., et al. (2003). A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP-19). *Aging & Mental Health, 7*, 186–194. doi:10.1080/1360786031000101157.
15. Carbone, E. T., Campbell, M. K., & Honess-Morreale, L. (2002). Use of cognitive interview techniques in the development of nutrition surveys and interactive nutrition messages for low-income populations. *Journal of the American Dietetic Association, 102*, 690–696. doi:10.1016/S0002-8223(02)90156-2.
16. Murtagh, F. E. M., Addington-Hall, J. M., & Higginson, I. J. (2007). The value of cognitive interviewing techniques in palliative care research. *Palliative Medicine, 21*, 87–93. doi:10.1177/0269216306075367.
17. Napoles-Springer, A. M., Santoyo-Olsson, J., O'Brien, H., et al. (2006). Using cognitive interviews to develop surveys in diverse populations. *Medical Care, 44*, S21–S30. doi:10.1097/01.mlr.0000245425.65905.1d.
18. Willis, G. B., & Lessler, J. T. (1999). *Question appraisal system: QAS-99*. National Cancer Institute.
19. Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287–311. doi:10.1093/poq/nfm006.
20. Conrad, F., & Blair, J. (1996). From impressions to data: Increasing the objectivity of cognitive interviews. In *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, pp. 1–9.