Published in final edited form as:
Med Care. 2007 May ; 45(5 Suppl 1): S12–S21. doi:10.1097/01.mlr.0000254567.79743.e2.

Evaluation of Item Candidates: *The PROMIS Qualitative Item Review*

Darren A. DeWalt, MD, MPH[*], **Nan Rothrock, PhD**[†], **Susan Yount, PhD**[†], and **Arthur A. Stone, PhD**[‡] **on behalf of the PROMIS Cooperative Group**

[*]Division of General Internal Medicine and the Cecil G. Sheps Center for Health Services Research, University of North Carolina, Chapel Hill, North Carolina

[†]Center on Outcomes, Research, and Education, Evanston Northwestern Healthcare, Evanston, Illinois

[‡]Department of Psychiatry, Stony Brook University, Stony Brook, New York

Abstract

One of the PROMIS (Patient-Reported Outcome Measurement Information System) network's primary goals is the development of a comprehensive item bank for patient-reported outcomes of chronic diseases. For its first set of item banks, PROMIS chose to focus on pain, fatigue, emotional distress, physical function, and social function. An essential step for the development of an item pool is the identification, evaluation, and revision of extant questionnaire items for the core item pool. In this work, we also describe the systematic process wherein items are classified for subsequent statistical processing by the PROMIS investigators. Six phases of item development are documented: identification of extant items, item classification and selection, item review and revision, focus group input on domain coverage, cognitive interviews with individual items, and final revision before field testing. Identification of items refers to the systematic search for existing items in currently available scales. Expert item review and revision was conducted by trained professionals who reviewed the wording of each item and revised as appropriate for conventions adopted by the PROMIS network. Focus groups were used to confirm domain definitions and to identify new areas of item development for future PROMIS item banks. Cognitive interviews were used to examine individual items. Items successfully screened through this process were sent to field testing and will be subjected to innovative scale construction procedures.

Keywords

patient-reported outcomes; cognitive interviews; qualitative methods; questionnaire development

The PROMIS (Patient Reported Outcome Measurement Information System) project provides an opportunity to build on decades of work in the development of items to measure health. As such, the PROMIS investigators created a process for identifying and evaluating currently available items for consideration and optimization for the PROMIS item banks. Using both quantitative and qualitative methods, we hope to arrive at the most informative and efficient set of items for measuring health outcomes. This report describes the qualitative methods and framework for the item review. Aspects of the quantitative evaluation are described elsewhere in this supplement.[1]

Reprints: Darren A. DeWalt, MD, MPH, Assistant Professor of Medicine, Division of General Internal Medicine, 5039 Old Clinic Building. CB#7110, Chapel Hill, NC 27599-7110. dewaltd@med.unc.edu.

Most self-report items for health measurement in medical research and care were developed for scales using classic test theory and are administered as a set of items, regardless of respondent level of the latent trait. As a result, many items and scales are either highly precise and cover a small range of the latent trait or less precise and cover a larger range. For example, the SF-36 subscales cover a broad range of the latent trait (although still have important ceiling effects), but do not offer the precision for small differences within some samples.[2] Alternatively, the Headache Impact Test offers a relatively high degree of precision across a broad range of the continuum within a very specific patient population.[3] Many disease-specific questionnaires are designed for precision over a relatively small range or narrow sample.[4,5]

PROMIS aims to achieve both precision and range by using item response theory (IRT) and presenting items in a computerized adaptive testing (CAT) format based on the IRT results. IRT enables modeling of the latent trait and identification of individual item functioning. Using IRT, scale developers are able to draw information from a much larger selection of items to model the latent trait, and, subsequently, administer only those items that will offer the most precision for a given individual.[1,6–8] CAT is a system by which the item administered to the respondent is decided based on the response to previous items. CAT can decrease respondent burden while maintaining precision. To achieve these goals, PROMIS investigators needed to identify and develop items that cover the range of experience in the domains to be measured and items that can add precision to the final estimate of the level of the latent trait. The Qualitative Item Review (QIR) process was designed to do this.

The PROMIS investigators performed a step-wise QIR process that included: identification of extant items, item classification and selection or "binning and winnowing," item revision, focus group exploration of domain coverage, cognitive interviews on individual items, and final revision before field testing (Table 1). By following these steps, QIR aimed to arrive at an optimal set of items that would increase the likelihood of successful item bank development.

The QIR process described here was implemented by the PROMIS investigators between Spring 2005 and Summer 2006 for the following domains: pain, fatigue, emotional distress, physical function, and social function. This report summarizes the processes and some of the early findings as examples. Future work and technical reports will describe, in more detail, findings that influenced the PROMIS item banks.

Identification of Extant Items

Rather than develop all new items from scratch, PROMIS built on existing items that had undergone testing previously; in fact, many items we considered were from well-established instruments that had been extensively tested and had excellent track records. Additionally, the PROMIS investigators elected to perform a more inclusive search and evaluation of existing instruments to enrich the pool of domain-relevant items that would be potential candidates for the PROMIS item banks. Searches started with MEDLINE and Health and Psychosocial Instruments, but also included proprietary databases like Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID).[9] Each domain work-group constructed their own search strategy based upon the specific needs identified within the domain. For example, the Emotional Distress domain group identified 4 general areas (referred to as "subdomains") for starting bank development: depression, anxiety, anger, and substance misuse. They created search strategies to identify a breadth of items covering these topics as a starting point for PROMIS banks. Importantly, the process allowed manual searches of files by investigators to identify items that were not found through the database searches. For example, the Statistical Coordinating Center (SCC) had accumulated databases of items in a variety of domains in which they had researched previously. The SCC made these item lists available to PROMIS researchers. At this stage of the process, items were not filtered out if

they applied to a specific population. Rather, those items were kept for further qualitative analysis.

By performing these searches, PROMIS investigators identified thousands of items relevant to the domains PROMIS was trying to measure (Table 2). At that point, no judgment was made regarding the quality or redundancy of the items; they were only selected if they seemed relevant for the domain. All items were entered into a standardized item library at the SCC. Item characteristics recorded in the library included: (1) context: the instructions associated with answering the item; (2) stem: the part of the item that makes it unique from others in the same scale; (3) response options; (4) time frame: if stated, the period of time that the respondent was to consider in answering the question; and (5) instrument of origin.

Confronted with thousands of items, a method for sorting through the content and deciding on the most representative and informative items was needed. We called this next step “binning and winnowing.”

Binning and Winnowing

Binning

The PROMIS domain workgroups first selected those items from the item library that they believed represented their domain. This process was done in teams so that at least 2 people reviewed each item for inclusion. Upon completion of domain identification, domain workgroups proceeded with the task of binning items. Binning refers to a systematic process for grouping items according to meaning and specific latent construct. For example, “walking” became a bin within the physical function domain. The final goal was to have a bin from which a small number of items could be chosen to adequately represent the bin. We did not predetermine the number of items that would adequately represent a bin. Rather, the goal for this process was to identify enough items to capture the meaning of the bin and to eliminate unnecessary redundancy in the item pool. By grouping items systematically, the domain workgroups could observe redundancy among items and identify the best potential items based on qualitative characteristics.

PROMIS domain workgroups (including several investigators across the PROMIS research sites and the SCC) began by creating a set of bins based on a review of that domain's literature, including previous factor analytic studies of domain items, and theory-based studies of the domain.[10–19] This “top-down” approach began with a conceptual model of the facets of each domain.[20–22] However, each domain workgroup approached the process with the flexibility to add or subtract bins based on the content of items themselves. By taking this approach, we retained the organizational structure put forth by the domain experts, but took advantage of new ideas as expressed by the items written in the clinical literature. We believe this allowed for the most inclusive and open approach at this stage of item evaluation. Of course, there is a degree of arbitrariness to this process and we fully recognize that other investigators' review of items may yield a somewhat different set of bins. However, the purpose of binning was to enable the identification of redundant items. Thus, what is important is that the final set of items emerging from this process adequately represents the domains, and there are undoubtedly many different sets of bins that could yield such a set of items.

Winnowing

The goal of winnowing was to reduce the large item pool down to a representative set of items. The process of winnowing helped to identify item characteristics that would include or exclude them from the PROMIS item banks based on domain definitions. Ultimately, this process was based on the judgment of reviewers and was accomplished by a consensus process of 2 or 3 reviewers for each domain. We adopted a set of criteria for excluding entire bins or items within

bins because they were not applicable to the current domain activities for PROMIS. Many items excluded seem to measure important domains or subdomains that are not currently the focus of PROMIS item banks. PROMIS investigators used the following criteria to remove items from consideration: (1) item content was inconsistent with the domain definition; (2) an item was semantically redundant with a previous item; (3) the item content was too narrow to have universal applicability; (4) the stem of the item was disease specific, reducing general applicability of the item; and (5) the item was confusing. For example, items related to satisfaction with physical function were identified, binned and removed from the physical function item bank consideration, because satisfaction was not in the PROMIS definition of physical function. Across all domains, approximately 30% of the items were eliminated due to redundancy, and approximately 45% were eliminated because they did not fit within the domain definitions adopted by PROMIS investigators. Table 3 has examples of items that were eliminated and the reasons for doing so.

By carefully analyzing each item and comparing them to other items within a given bin, domain workgroups were better able to apply the several criteria to each item. As with all other aspects of the QIR process, all decisions about item winnowing were reviewed by multiple members of the domain workgroup and members of the SCC to ensure a high level of consensus and to impose some standardization of processes across domain groups. The process of binning and winnowing yielded a smaller set of items that were then subjected to editing to match PROMIS stylistic conventions in a process of item revision (Table 2).

Item Revision Process

After winnowing, each domain group had a set of items to carry forward for review by researchers and by potential respondents. This item set included items with a range of styles in phrasing, time frame of recall, response options, and literacy demands. Because of these variations, the items would be difficult to administer as a coherent test or on a CAT administration in their current form. PROMIS investigators made a substantial effort to create and use items that were accessible for a variety of literacy levels and that had little ambiguity or cognitive difficulty. As part of this effort, PROMIS favored uniformity in format when evidence did not suggest that diversity is better. The next phase of QIR involved item review and revision to provide consistency of style of questions, ease the literacy requirements of respondents, and apply a consistent set of response options and time frames. Network PRO experts worked to reach consensus on the item guidelines to be used across domains, including response options and time frame.

Response Options

Because hundreds of instruments were collected, dozens of response options were represented. In an attempt to minimize respondent confusion and reduce respondent burden, it was agreed that the number of available response options should be reduced. However, there is considerable debate regarding which response options would be most useful across all of the PROMIS domains. For example, response sets targeting frequency, intensity, and interference with functioning had all been used without consensus on what was most appropriate. As such, we used a consensus process to arrive at the PROMIS preferred response options. Rather than endorsing 1 set of response options as clearly superior to another, our consensus process acknowledged the need for some uniformity and the lack of empirical evidence that 1 set of options is clearly better than others.

The optimal number of response levels may vary for individual items, latent constructs, and context of item administration. Determining the optimal number of levels is an empirical exercise that can be accomplished by administering the same item with several different response sets. Such an empirical test was not a priority objective for PROMIS during the initial

item bank creation. Moreover, the ultimate goal of PROMIS, to develop item banks that can be administered via CAT, argues in favor of relative uniformity of response categories across banks of items. Learning a new response set for each item that appears on the computer screen would present an unnecessary cognitive burden on respondents and could yield less reliable data.

Based on experience with IRT analyses, PROMIS Investigators thought that a reasonable number of response levels would be between 4 and 6.[23] A greater number of response levels is likely to present more cognitive burden on respondents (having to parse out their symptoms to very fine levels) and smaller numbers of response levels do not adequately tap the item for all of its information. Frequently, when items have more than 6 response levels, 2 or more levels are collapsed together during IRT modeling to remove step disorder and to potentially improve the model fit.[23]

To select uniform response sets across domains, PROMIS domain groups analyzed frequency distributions of response sets for extant items. Although not a perfect strategy, it gave PROMIS investigators an idea of the preferences of scale developers of earlier instruments. We found that response options fell into specific categories depending on the intent of the item. The categories included intensity/severity, frequency, capability, and duration. Domain experts then proposed a smaller set of frequently used response sets that would be applicable across domains. These response sets were then reviewed by network psychometricians and language translation experts to aid in selecting sets most useful in IRT models and amenable to translation. Most of the PROMIS preferred response options (Table 4) include 2 sets within each category. About 90% of PROMIS items will use these options with the flexibility to use a different set if an important item cannot be satisfactorily reworded to fit one of the preferred sets. (For example, it is traditional for pain intensity items to be scored on a 0 to 10 point scale.)

Recall Time Frame

PROMIS also aimed to reach consensus on the time frame respondents will be instructed to refer to in answering questions (eg, “Thinking about the past week, please answer the following items.”) Understanding the optimal recall period over which the respondent should reflect while considering their experiences is a complex endeavor. There is a large literature from the fields of autobiographical memory, social-cognitive science, and survey research that highlights potential problems in questions that ask respondents to recall complex information over time.[24–26] First, memory is actually quite limited and selective in terms of what information from daily life is encoded and is subsequently available at recall.[27] Second, various systematic biases have been identified (also known as cognitive heuristics) in the manner that experiences are recollected. For example, more salient and/or intense events and the most recent events are highlighted in recall.[28,29] Third, when the information required by a question is not available to the respondent, other, less relevant information, may be used instead to answer the question.[30] For example, if the individual has difficulty remembering their pain intensity levels over a month, then current pain intensity may be used instead by the respondent.

For these reasons, we were concerned about selecting a recall period that would reduce the potential biases just described and yet be sufficient to capture a period of experience that was considered clinically relevant for outcome research. In fact, there is currently relatively little research that is available to inform this question, but our guiding principle was that relatively shorter reporting periods were to be preferred over longer ones to generate the most accurate data.

A 7-day reporting period was adopted as a general convention for PROMIS items. There is evidence that some symptoms, for example, pain intensity, exhibit a moderate correspondence between real-time reports of intensity and recalled reports, although there is also evidence that

when reported on the same scale, recalled pain levels are reliably higher.[31] Nevertheless, when we considered adopting an even shorter reporting period, such as a single day, we knew that the practical implications for outcome research would be significant. That is, investigators wishing coverage of outcomes for a week would have to administer the item bank each day and summarize the data (eg, by averaging) to have a single weekly score. At this point, we did not think that such a position was justified for this project. As with previous conventions, we recognized the importance of flexibility for those items that do not make sense with the 7-day period and that exceptions to the convention are acceptable. Certain infrequently occurring, yet highly salient symptoms (such as cardiovascular events), for instance, may be reliably reported with longer reporting periods, considering their highly salient nature, and we recognized that this might be necessary for some of the PROMIS items or domains.

One PROMIS domain, physical function, has chosen to not specify a time period, but to ask the question in the present tense. This decision reflects an important aspect of the domain definition adopted for physical function: that function will be measured by self-reported capability rather than self-reported performance. The distinction here can be appreciated by the difference between “are you able to run” versus “did you run.” In this context, using present tense and no time frame are reasonable.

Outside of the PROMIS banks, some investigators have taken an alternative position to time frame reporting and have adopted real-time approaches to outcome capture.[32] Although the jury is still out on the ultimate importance of differences between aggregated reports made in real-time and recalled reports, the growing literature summarized above suggests that there could be important differences. One of the PROMIS funded-projects is explicitly examining the accuracy of recall with various reporting periods and will be able to inform the PROMIS item development.

Item Revision

Items retained after the binning and winnowing process had numerous styles of language, instructions, recall periods, and response options. The Network recognized that most items would need some level of revision to adhere to the PROMIS format and to incorporate the PROMIS response option and recall period conventions. We also recognized that this was the opportunity to clarify vague or multibarreled questions before taking them to the field. Many questions also used language that was outdated, difficult to translate, or unnecessarily complex and could also be corrected during the item revision process.

Items were revised by Network experts in the specific domain. When revising the items, writers made the following assumptions: (1) items would need to stand alone, as only 1 item would be administered at a time on a computer screen; (2) all items would have similar context statements (eg, “In the past 7 days,”); (3) all else being equal, items should be as concise and simply worded as possible; and (4) items should be worded to use one of the preferred response options if possible. All writers targeted the sixth-grade reading level or less and attempted to choose words used commonly in English and tried to avoid slang. If items were multibarreled, the writer was encouraged to divide the item into at least 2 separate items. The derivative items were evaluated independently as to whether they still fit within the domain. After the initial revision, items were reviewed by at least 2 other members of the domain workgroup to achieve consensus about each item. Network translation experts also reviewed items for translatability.

Focus Groups

As the PROMIS project is committed to having both researcher (described above) and patient input in the development of item banks, both focus groups and cognitive interviews were included in the QIR process. Focus group interviews can help the researcher discover the

vocabulary and the thinking patterns of the target group to inform the development of questionnaire items.[33] More importantly for PROMIS, focus groups can help to identify important gaps in coverage of the current items and domain definitions. Although PROMIS has targeted domains extensively studied in the clinical literature, it was considered important to solicit feedback from potential respondents about the domains in question to make sure we were addressing topics that reflect how potential respondents experience the world. Therefore, the primary aim of the focus groups was to confirm the domain definitions and identify common language related to the domain. A secondary goal was to identify important measurement areas that are not currently covered by PROMIS item banks for consideration for future banks.

Because we are designing instruments to measure domains that cross multiple illnesses, ages, cultures, and lifestyles, we did not believe it was feasible to perform focus groups matched on each of those variables. Covering all important chronic illnesses alone would require hundreds of focus groups. For this reason, we adopted the strategy of selecting a sample of patients with and without chronic illness who had experienced a range of severity or limitation in the domain in question. A variety of ages and cultures were represented. As a group, we asked them to reflect on the various ways that their health affected their experience in a given domain.

Participants were recruited from a variety of settings including general medical clinics, arthritis registries, rehabilitation clinics, and outpatient psychiatric clinics (Table 5). Two to 4 focus groups were conducted for each domain, with the exception of Emotional Distress for which additional groups were conducted due to the number of subdomains (eg, Anger, Anxiety, Depression, Alcohol Abuse). After each focus group, PROMIS investigators conducted content analysis based on recall, notes taken by the cofacilitator, and transcripts from the session recordings. Specifically, we identified key words, phrases, and quotes regarding symptoms; additional emergent themes in each of the domains; and important issues not addressed by the initial 5 selected domains. Themes included in the final analysis were raised by more than 1 participant in a single group, and, ideally, by participants in more than 1 group. Overall, the focus groups confirmed the direction of the PROMIS domain definitions, but added important ideas for development into new item banks. More detailed results of the focus groups will be available in future reports.

Cognitive Interviews

We designed a cognitive interviewing process to elicit respondent feedback on all individual items considered for the PROMIS item banks. We queried individuals on the language, comprehensibility, ambiguity, and relevance of each item. Although PROMIS benefited from beginning with items that had already been used in clinical research, many of the extant items had not been subjected to formal cognitive interviewing. Subjecting potential items to cognitive interviewing has become a standard technique in the development of large-scale questionnaires, for example, by the National Center for Health Statistics.[34] Furthermore, through the item review process, most items' structure and response options were revised. As such, the PROMIS investigators consensus was that cognitive assessment with respondents could identify potentially problematic items and response scales and help to clarify items that were not easily understood and answered.

We based our cognitive interviewing protocol on the work of Willis.[35] The cognitive interviewing process ascertained: (1) comprehension of the question (ie, what does the respondent believe the question is asking; what do specific words and phrases in the question mean to the respondent); (2) the processes used by the respondent to retrieve relevant information from memory (ie, what does the respondent need to recall to be able to answer the question; what strategies does the respondent use to retrieve the information); (3) decision processes, such as motivation and social desirability (ie, is the respondent sufficiently

motivated to accurately and thoughtfully answer the question; is the respondent motivated by social desirability in answering the question); and (4) response processes (ie, can the respondent match his/her response to the question's response options).[36] Some of these processes may be "conscious," and others are outside the awareness of the respondent.[35]

The PROMIS cognitive interviews employed a "retrospective" verbal probing technique. In this technique, a participant completes a paper and pencil version of the questionnaire of interest. A trained interviewer then asks for other, specific information relevant to each question, or "probes further into the basis for the response."[35] This type of "retrospective" probing or debriefing is useful when a more "realistic" type of presentation of items is desirable, particularly at later stages of questionnaire development.[35] Additionally, this method reduces probing from biasing patients' responses to items later in the questionnaire. As the final PROMIS item banks will be self-administered and most items have been subjected to multiple research trials, a retrospective probing technique was considered most appropriate.

All PROMIS items underwent an initial set of 5 cognitive interviews. Proposed items were divided into sets of 30 items and each set of 30 was subjected to interviews with 5 individuals. Although items written de novo for questionnaires are often subjected to more cognitive interviews, items that are undergoing translation typically use a smaller number of interviews. The structure of cognitive interviews for PROMIS allowed for many more than 5 cognitive interviews on issues that cut across items such as context, response options, and time frame. For this reason, and because most items for PROMIS item banks are modifications of existing items rather than newly created items, the PROMIS investigators decided that 5 initial interviews was most appropriate. If, however, after 5 interviews the item underwent major revisions, the item was subjected to 3 to 5 additional interviews after the revisions.

Because cognitive interviewing uses small numbers of participants, representative sampling is difficult. Although many respondent characteristics may be associated with different interpretations of items, PROMIS investigators were most concerned about differences according to reading ability and racial group. To ensure that items were not evaluated by only white respondents, each item was reviewed by at least 1 nonwhite interviewee and at least 1 white interviewee. We recognize that this oversimplifies racial and ethnic categories, but enables some degree of diversity within a population of 5 interviewees. Additionally, each item was reviewed by at least 2 interviewees with one or more of the following criteria: (1) less than 12 years of education; (2) a measured reading level less than the ninth grade using the Wide Range Achievement Test-3 Reading subtest; or (3) a diagnosis associated with cognitive impairment (eg, traumatic brain injury or stroke). We recruited participants from clinical settings and from disease registries. This allowed us to target those who had completed less than 12 years of education. To mitigate differences according to geography, cognitive interviews were performed across the network representing the Northeast, South, Midwest, and Western parts of the United States. Table 6 presents the populations sampled and demographic characteristics of participants.

The cognitive interviewing process created a rich qualitative data set about items, and was remarkably efficient in that it used a small sample of individuals. We considered the possibility of using paper and pencil questionnaires to target the comprehensibility and relevance of items, because they would allow a larger sample to be used, increasing the likelihood that relatively infrequent, but important, responses would be captured. However, the downside of a paper-and-pencil method was that it would not allow the flexibility and richness of a cognitive interview. It also relied on the questionable ability of the respondent to self-reflect and communicate their thoughts in writing. Our compromise position was to have items in the physical functioning domain reviewed by respondents using the paper-and-pencil process and a small number of follow-up interviews, affording an opportunity to compare the 2 methods.

Participants were recruited through patient registries for arthritis, and aging (Table 7). With the paper-based survey method, each item was reviewed by 58–75 respondents who were asked to rate the item on clarity and importance. Those items that scored worst on clarity were selected for follow-up telephone cognitive interviews or deleted altogether. At this time, PROMIS has not decided on the best approach for future bank development. Both methods suffer from limitations and it is likely that PROMIS will weigh the benefits of each method depending on the banks developed in the future.

Final Item Revisions

On the basis of the cognitive interviewing results, final revisions will be completed before field testing. All items selected for testing will be subjected to testing with the Lexile Analyzer to assess readability.[37] The Lexile Analyzer gives an approximate reading level for the item based on the commonness of words in the item and the complexity of the syntax. This will create the opportunity to evaluate if more difficult to read items are problematic during field testing and serves as a final check on revising items to the easiest to read format. Additionally, item characteristics (for example, whether the item is referring to intensity, frequency, difficulty, or interference) can be classified for later analyses. By categorizing items, PROMIS can begin to understand the quantitative performance of items according to these subjective qualities. Once final revisions were completed, PROMIS items went to field testing with the aim of understanding the quantitative characteristics of the items.

Strength and Limitations

Our process for qualitative item review was designed to identify as many extant items as possible and optimize them based on expert review and respondent feedback. Through this process we aimed to have a cohesive set of items for field testing. Because we started with a review of thousands of items used in the clinical and research literatures, we capitalized on the decades of developmental work in each domain. Starting with a broad approach gave us a chance to review the breadth and depth of current instruments and allowed us to build on the expertise in the field.

The process for expert item review allowed us to focus our efforts on clarifying confusing items and simplifying language, where appropriate, to improve assessment in populations with low literacy. We also took the opportunity to make the items more uniform with regard to their instructions and response options. By unifying the item structure, we hope to reduce respondent burden and improve the accuracy of reporting.

Finally, we have designed a process to get feedback from potential respondents regarding the current conceptualization of each domain (focus groups) and on individual items (cognitive interviewing). By soliciting feedback from potential respondents, we can improve the likelihood that our items will be understood and interpreted as intended. We also improve the chance that our items reflect important patient experiences. This approach is consistent with the recent call by the US Food and Drug Administration's (FDA) preliminary guidance for the development of Patient Reported Outcomes (FDA Docket No. 2006D-0044), which calls for patient input in the development of self-report assessments.

An important limitation to the approach we have taken is that we are changing almost all existing items from their original format. Although many of these changes are minor (eg, changing the response categories or changing the instruction set), they are changes none-the-less, and the items may function differently. We have taken this approach in an effort to make the items more uniform as they are read and interpreted by the respondents. We believe that our methods will not worsen any good items and may improve those that do not currently

perform as well. As a check to this method, we will administer several well-known items in their current form to see if there is a difference in item function after the subtle changes.

Another limitation is the reliance on 5–10 cognitive interviews per item. To identify important conceptual difficulties with items, more cognitive interviews are occasionally recommended.[35] However, we felt a lesser number of interviews would be sufficient for our items because (1) we have performed extensive expert review and revision of existing items, (2) most items have had previous cognitive testing or field testing, and (3) the modifications we have made to existing items are more akin to a translation than to creation of entirely new items.

Summary

The PROMIS qualitative item review process is the result of a consensus process across 6 primary research sites, the statistical coordinating center, and PROMIS participants from the NIH. It reflects a diversity of views on questionnaire design and the role of qualitative methods. PROMIS investigators agreed that a rigorous and efficient approach toward qualitative review was an important and necessary step in producing the best items for use in the PROMIS item banks. This description of the process reflects that notion and our desire to unify and advance measurement of patient reported outcomes.

Acknowledgments

We acknowledge the efforts of coinvestigators at each of the sites that have contributed to the development of the qualitative item review process. Specifically, we would like to thank Deborah Irwin, Liana Castel, James Varni, Andrea Meier, Kelly Williams, Harry Guess (University of North Carolina); Joan Broderick, Chris Christodoulou, Doerte Junghaenel (Stony Brook University); Sue Eisen, Angela Stover, Paul Pilkonis (University of Pittsburgh); James Fries, Bonnie Bruce (Stanford University); Dagmar Amtmann, Karon Cook (University of Washington); Kevin Weinfurt (Duke University); Elizabeth Hahn, Sonia Eremenco, David Cella, Kimberly Webster, Benjamin Arnold (Evanston Northwestern Healthcare); Bryce Reeve, Gordon Willis (National Cancer Institute); William Riley (National Institute for Mental Health); Karon Coyne (Medtap Institute); and Liz Jansky, Lori Perez (Westat).

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grants 5U01AR052181, 5U01AR052177, 5U01AR052170, 5U01AR052158, 5U01AR052155, 5U01AR052171. Information on the Patient-Reported Outcomes Measurement Information System (PROMIS) can be found at http://nihroadmap.nih.gov/ and http://www.nihpromis.org.

References

1. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(Suppl 1):S22–S31. [PubMed: 17443115]
2. Ware, JE., Jr; Snow, K.; Kosinski, M., et al. SF-26 Health Survey: Manual and Interpretation Guide. Boston, MA: New England Medical Center, Health Institute; 1993.
3. Bjorner JB, Kosinski M, Ware JE Jr. Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. Qual Life Res 2003;12:981–1002. [PubMed: 14651417]
4. Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living with Heart Failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Pimobendan Multicenter Research Group. Am Heart J 1992;124:1017–1025. [PubMed: 1529875]
5. Juniper EF, Guyatt GH, Feeny DH, et al. Measuring quality of life in children with asthma. Qual Life Res 1996;5:35–46. [PubMed: 8901365]
6. Hill CD, Edwards MC, Thissen D, et al. Practical issues in the application of item response theory: a demonstration using items from the Pediatric Quality of Life Inventory (PedsQL) 4. 0 Generic Core Scales. Med Care 2007;45(Suppl 1):S39–S47. [PubMed: 17443118]

7. Hays RD, Liu H, Spritzer K, et al. Item response theory analyses of physical functioning items in the Medical Outcomes Study. Med Care 2007;45(Suppl 1):S32–S38. [PubMed: 17443117]
8. Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Mahwah: Lawrence Erlbaum Associates, Inc.; 2000.
9. Mapi Research Institute. Patient-Reported Outcome and Quality of Life Instruments Database. Mapi Research Institute; 2005.
10. Hahn EA, Cella D, Bode RK, et al. Social well-being: the forgotten health status measure [abstract]. Qual Life Res 2005;14:1991.
11. Stein KD, Martin SC, Hann DM, et al. A multidimensional measure of fatigue for use with cancer patients. Cancer Pract 1998;6:143–152. [PubMed: 9652245]
12. Smets EM, Garssen B, Bonke B, et al. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995;39:315–325. [PubMed: 7636775]
13. Brown TA, Chorpita BF, Barlow DH. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. J Abnorm Psychol 1998;107:179–192. [PubMed: 9604548]
14. Clark LA, Watson D. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. J Abnorm Psychol 1991;100:316–336. [PubMed: 1918611]
15. Krueger RF. The structure of common mental disorders. Arch Gen Psychiatry 1999;56:921–926. [PubMed: 10530634]
16. Watson D, Clark LA, Weber K, et al. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. J Abnorm Psychol 1995;104:15–25. [PubMed: 7897037]
17. Watson D, Weber K, Assenheimer JS, et al. Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. J Abnorm Psychol 1995;104:3–14. [PubMed: 7897050]
18. Krueger RF, McGue M, Iacono WG. The higher-order structure of common DSM mental disorders: internalization, externalization, and their connections to personality. Personality Individual Diff 2001;30:1245–1259.
19. Rose M, Bjorner JB, Becker J, et al. Evaluation of a preliminary physical function item bank supports the expected advantages of the Patient Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol.
20. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Road-map Cooperative Group during its first two years. Med Care 2007;45(Suppl 1):S3–S11. [PubMed: 17443116]
21. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005;23(5 Suppl 39):S53–S57. [PubMed: 16273785]
22. Fries JF. The promise of the future, updated: better outcome tools, greater relevance, more efficient study, lower research costs. Future Rheumatol 2006;1:415–421.
23. Bode RK, Lai JS, Cella D, et al. Issues in the development of an item bank. Arch Phys Med Rehabil 2003;84(4 Suppl 2):S52–S60. [PubMed: 12692772]
24. Bradburn NM, Rips LJ, Shevell SK. Answering autobiographical questions: the impact of memory and inference on surveys. Science 1987;236:157–161. [PubMed: 3563494]
25. Erskine A, Morley S, Pearce S. Memory for pain: a review. Pain 1990;41:255–265. [PubMed: 1697054]
26. Schwarz, N.; Sudman, S. Autobiographical Memory and the Validity of Retrospective Reports. New York, NY: Springer-Verlag; 1994.
27. Robinson MD, Clore GL. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychol Bull 2002;128:934–960. [PubMed: 12405138]
28. Gorin, A.; Stone, AA. Recall biases and cognitive errors in retrospective self-reports: a call for momentary assessments. In: Baum, A.; Revenson, TA.; Singer, JE., editors. Handbook of Health Psychology. Mahwah: Lawrence Erlbaum Associates; 2001. p. 405-413.

29. Redelmeier DA, Kahneman D. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain 1996;66:3–8. [PubMed: 8857625]

30. Menon, G.; Yorkston, E. The use of memory and contextual cues in the formation of behavioral frequency judgments. In: Stone, A.; Turkkan, J.; Bachrach, C., editors. The Science of Self-Report: Implications for Research and Practice. Mahwah: Lawrence Erlbaum Associates; 2000. p. 63-79.

31. Stone A, Schwartz J, Broderick J, et al. Variability of momentary pain predicts recall of weekly pain: a consequence of the peak memory heuristic. Personality Social Psychol Bull 2005;31:1340–1346.

32. Stone, A.; Shiffman, S.; Atienza, A., et al. The Science of Real-Time Data Capture. New York, NY: Oxford University Press; In press

33. Fowler, FJ. Survey Research Methods. 3rd. Thousand Oaks, CA: Sage Publications; 2002.

34. Johnson, C. Proposal guidelines for new content on the 2007-2008 National Health and Nutrition Examination Survey (NHANES). May 242004 [December 13, 2006]. Available at: http://www.cdc.gov/nchs/data/nhanes/proposal_guidelines_2007-8.pdf

35. Willis, GB. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage Publications; 2005.

36. Tourangeau, R. Cognitive sciences and survey methods. In: Jabine, T.; Straf, M.; Tanur, J., et al., editors. Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press; 1984. p. 730199

37. MetaMetrics. Lexile Analyzer. [December 13, 2006]. http://www.lexile.com/DesktopDefault.aspx?view=ed&tabindex=2&tabid=16&tabpageid=335

TABLE 1

Qualitative Item Review Steps and Timeline for Initial PROMIS Item Banks

	2005												2006					
Task	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Identification of items	X	X	X	X	X	X	X	X										
Item classification and selection					X	X	X	X	X	X	X	X						
Creation of items								X	X	X	X							
Item revision							X	X	X	X	X	X			X	X	X	X
Focus group studies													X	X	X			
Cognitive interviews															X	X	X	X

TABLE 2

Numbers of Items Identified by Searches for Extant Items

	Emotional Distress	Fatigue	Social Function	Physical Function	Pain
Number of items identified	2187	1066	1781	1860	644
Items for cognitive interviewing	299	135	129	247	191
Final item pool	224	112	112	224	168

TABLE 3

Examples of Items Eliminated at the Winnowing Stage

Item Stem	Domain	Reason for Removal
How many visits to mental health specialists have you made in the past 6 months	Depression (treatment)	Inconsistent with the domain definition
How much have you gotten fatigued easily	Depression (fatigue)	Semantic redundancy (24 in the bin)
I have difficulty sleeping	Depression (sleep problems)	Semantic redundancy (30 in the bin)
You are asked to place an “X” through these lines to indicate how you are feeling RIGHT NOW.	Fatigue	Item too narrow
Do you feel too much tiredness with normal or soft efforts?	Fatigue	Item confusing
Right now … do you feel such fatigue that you don't know what to do with yourself	Fatigue	Item vague
My motivation is lower when I am fatigued	Fatigue	Inconsistent with domain definition

TABLE 4

Initial PROMIS Preferred Response Options

Category	Preferred Response Options	
Frequency	Never	Never
	Rarely	Once a week or less
	Sometimes	Once every few days
	Often	Once a day
	Always	Every few hours
Duration	A few minutes	None
	Several minutes to an hour	1 day
	Several hours	2–3 days
	1–2 days	4–5 days
	>2 days	6–7 days
Intensity	None	Not at all
	Mild	A little bit
	Moderate	Somewhat
	Severe	Quite a bit
	Very severe	Very much
Capability	Without any difficulty	
	With a little difficulty	
	With some difficulty	
	With much difficulty	
	Unable to do	

TABLE 5

Focus Group Participants

	Emotional Distress	Fatigue	Social Function	Physical Function	Pain
No. groups	13	3	5	3	4
Total participants	104	17	31	15	24
Female, %	50	65	65	80	79
Ethnicity, %					
Hispanic	2	0	0	7	0
Race, %					
White	57	94	65	94	88
African American	38	6	29	0	8
Asian	3	0	3	7	0
American Indian/Alaska Native	2	0	0	0	0
Native Hawaiian/Pacific Islander	0	0	0	0	0
Multiple races	1	0	3	0	4
Age, mean (range)	50 (23–88)	48(26–65)	53 (23–83)	56(31–86)	60.7 (26–76)
Education, %					
≤8th grade	2	0	0	0	0
9th–11th grade	3	6	3	7	0
12th grade/GED	18	18	19	13	25
Some college	33	29	45	13	46
College degree	28	24	13	40	13
Advanced degree	16	24	19	27	17
Populations sampled	Outpatient psychiatric; mixed internal medicine outpatients	Rehabilitation; mixed internal medicine outpatients	Outpatient psychiatric; mixed internal medicine outpatients	Arthritis; rehabilitation	Arthritis; rehabilitation; mixed internal medicine outpatients

TABLE 6

Cognitive Interview Participants

	Emotional Distress	Fatigue	Social Function	Physical Function	Pain
No. interviews	34	29	22	18	52
Total participants	33	22	21	18	44
Female, %	64	55	40	67	59
Ethnicity, %					
Hispanic	0	0	0	6	9
Race, %					
White	76	50	71	67	82
African American	24	50	19	11	7
Asian	0	0	0	22	7
American Indian/Alaska Native	0	0	5	0	0
Native Hawaiian/Pacific Islander	0	0	0	0	2
Multiple races	0	0	5	0	2
Age, mean (range)	42 (20–60)	63 (38–83)	66 (39–86)	70 (48–93)	46 (18–83)
Education					
% ≤8th grade	0	5	19	6	2
% 9th–11th grade	0	27	14	6	2
% 12th grade/GED	21	18	5	22	9
% Some college	36	18	14	22	34
% College degree	15	14	14	28	27
% Advanced degree	27	18	33	17	25
Populations sampled	Outpatient psychiatric	Internal medicine outpatients; musculoskeletal disease registry	Internal medicine outpatients; musculoskeletal disease registry	Osteoarthritis; rheumatoid arthritis; aging cohort	Rehabilitation
Self-rated level of severity/impairment of domain, %					
None	12	9	14	Missing data	5
Mild	36	32	29		30
Moderate	48	26	33		50
Severe	3	18	24		16
WRAT-3 Reading Standard Score, mean (range)	47 (31–57)	47 (25–57)	46 (23–57)	Missing data	50 (35–57)

	Emotional Distress	Fatigue	Social Function	Physical Function	Pain
% <9th grade	30	24	29		7
Cognitive impairment, %					16

WRAT-3 indicates Wide Range Achievement Test 3. Only the reading subtest was used to identify approximate reading level of respondents. Cognitive impairment was caused by stroke (n = 3), traumatic brain injury (n = 3), and with brain damage secondary hypoxemia (n = 1).

TABLE 7

Participants in the Mailed Survey of Qualitative Rating of Items in Physical Function Domain

	Physical Function
Total participants	734
Female, %	53
Race, %	
White	87
African American	4
Asian	4
American Indian/Alaska Native	0
Hispanic	4
Age, years	
<30	0
30–39	0
40–49	3
50–59	12
60–69	20
70–79	28
80–89	34
90+	3
Education, %	
<7th grade	4
7th–12th grade	1
12th grade	10
>12th grade	85
Populations sampled, %	
Osteoarthritis	27
Rheumatoid arthritis	24
Aging cohort	50