Published in final edited form as:
Clin Trials. 2008 ; 5(6): 575–586. doi:10.1177/1740774508098414.

Using Item Banks to Construct Measures of Patient Reported Outcomes in Clinical Trials: Investigator Perceptions

Kathryn E. Flynn[a,b], **Carrie B. Dombeck**[a], **Esi Morgan DeWitt**[c], **Kevin A. Schulman**[c,d], and **Kevin P. Weinfurt**[a,b]

a*Center for Clinical and Genetic Economics, Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC*

b*Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, NC*

c*Department of Pediatrics, Duke University School of Medicine, Durham, NC*

d*Department of Medicine, Duke University School of Medicine, Durham, NC*

INTRODUCTION

Evaluating how a patient experiences a health care intervention has become increasingly important in a wide range of clinical studies, including pharmaceutical, behavioral, device, and procedural trials. Such experiences are typically captured via "patient-reported outcomes" (PROs). Although PROs have played an important role in major clinical trials and in labeling decisions by the Food and Drug Administration (FDA) [1], the assessment of PROs has been problematic. For example, many PRO measurement instruments are long and burdensome for patients and research staff. Even rigorously designed instruments can miss the full range of patient experiences or be insensitive to change over time because of floor or ceiling effects. Furthermore, there is a lack of standardization in most therapeutic areas, with many similar PRO measures but no standard metric, making it difficult to compare or combine scores across studies. For these reasons, the National Institutes of Health (NIH), under the NIH Roadmap theme of reengineering the clinical research enterprise (http://nihroadmap.nih.gov/), has identified better assessment of PROs as a pressing need. Accordingly, the NIH has funded the Patient-Reported Outcomes Measurement Information System Network (PROMIS, http://www.nihpromis.org/) to develop better measures of PROs for chronic diseases [2]. The primary focus of this study is to assess clinical investigators' perceptions of the utility of PRO measures in clinical trials and to anticipate perceived barriers to the adoption of new PRO measures and methods.

PROMIS, along with other recent initiatives (eg, [3-5]), is exploring a new method for improving PRO measurement based on item response theory (IRT). IRT is a psychometric framework that has become the standard in educational testing and is gaining popularity in health assessment because of its ability to generate shorter measures without compromising reliability or sensitivity. IRT models are appropriate when one is interested in measuring an unobserved trait (or *construct*) that presumably exists along a continuum, such as fatigue or physical functioning. For questions (or *items*) with multiple ordinal responses (eg, *Not at all* to *Very much*), IRT models describe the likelihood of selecting each of the response options as a function of the person's continuous underlying trait. Once IRT models have been estimated for all of the items, one is able to do two things: (1) identify which items will provide the most

Author for correspondence: Dr. Kathryn Flynn, PO Box 17969, Durham, North Carolina, United States 27715. E-mail: kathryn.flynn@duke.edu.

information about different parts of the underlying continuum, and (2) translate a person's responses to items into an estimate of that person's status on the underlying continuum [6]. Additional details about IRT have been described elsewhere [7-9].

IRT models allow individual questions to be linked, using psychometric principles, in an "item bank." An item bank is a comprehensive collection of questions (and their response options) designed to measure an underlying construct (eg, fatigue) across its entire continuum. Creation of an item bank requires rigorous qualitative (e.g., expert review, patient focus groups, and cognitive interviews) and quantitative (eg, large-scale field testing of candidate items) evaluation to ensure the items are comprehensible to patients, valid, and precise. Items are calibrated to establish item statistics and properties. With an item bank, researchers have multiple options for instrument development, including creating a customized short form made up of a fixed set of items or administering a tailored test through computerized adaptive testing (CAT). Customized short forms are created by selecting the items that are most sensitive to the distribution of a given trait in the population under study. For example, investigators interested in measuring fatigue in patients with severe heart failure could select the items that are most sensitive to high levels of fatigue. Investigators studying milder heart failure might opt for a short form that includes items more sensitive to lower levels of fatigue. Because all of the items in the two short forms are from the same item bank, the fatigue scores estimated in both clinical studies will be on the same metric.

CAT is a computer algorithm that selects successive items based on an individual's responses to previously administered items. For example, if an initial item asked how difficult it was for the respondent to walk up a flight of stairs and the respondent's answer was "extremely difficult," this would suggest that the respondent's physical functioning is on the low end. The CAT algorithm would then select a second question that asks about difficulty doing an easier activity, such as walking on flat ground. On the other hand, if the response to the initial item was "not at all difficult," a second question might ask about a more difficult activity, such as running a mile. In this way, CAT is always searching for the next item that will provide the most unique information about the person [8]. CAT is already used for a variety of standardized educational tests, including the Graduate Record Examination [10], the North American Pharmacist Licensure Examination [11], and the United States Medical Licensing Examination [12]. Although CAT is the more sophisticated option than short forms, any option that uses items from the same item bank will produce comparable scores, regardless of whether the same items were asked of respondents [8]. This is another key advantage of measures based on IRT.

Widespread use of IRT item banks could be advantageous to health outcome assessment. However, an important dimension of the effectiveness of this technology is adoption by the end user community—clinical trialists. The purpose of this study was twofold: (1) to evaluate a brief tutorial designed as a basic introduction to IRT-based item banks, and (2) to elicit investigators' questions and concerns about both current and IRT-based PRO measurement strategies in clinical trial settings. Understanding these concerns is valuable to promote use of item banks and CAT in the clinical trials setting. This information will also be helpful for sponsors who may want to learn more about the issues and challenges of using PRO item banks as well as for developers of item banks who bear the burden of educating both potential users and relevant stakeholders (such as funding agencies and regulatory boards) on the merits of this new technology.

METHODS

Data Collection

We planned to conduct a total of 40 interviews with investigators conducting clinical research. We targeted 10 interviews in each of 4 areas—cardiovascular outcomes, oncology outcomes,

pediatrics, and mixed outcomes—in order to hear from investigators from different backgrounds and therapeutic areas as well as to achieve data saturation. We e-mailed all lead authors affiliated with US academic medical centers or research institutions who published results from clinical trials between July 2005 and April 2006 in the *Journal of the American Medical Association* (n=32) or the *New England Journal of Medicine* (n=50). To recruit additional participants in cardiovascular outcomes, oncology outcomes, and pediatrics, we contacted lead authors of papers describing phase 3 or 4 clinical trial results published between July 2005 and April 2006 in *Circulation* (n=12), the *Journal of the American College of Cardiology* (n=4), the *Journal of Clinical Oncology* (n=15), and *Pediatrics* (n=6). Recruitment was not dependent on an investigator having experience with PROs, but 2 investigators in pediatrics excluded themselves from participating because they worked solely in neonatology, where PROs are not used. After one week, we sent a follow-up e-mail to nonresponders. Two investigators in pediatrics were recruited via recommendations from a PROMIS project investigator before we decided to recruit participants primarily through literature search. These 2 investigators were not affiliated with PROMIS. In all, 42 investigators participated (10 each in cardiology and mixed outcomes, 11 each in oncology outcomes and pediatrics).

Once an interview with an investigator was scheduled, we e-mailed 8 questions (the parent questions in Appendix A) about previous experience with PROs and provided the 6-slide Item Bank Tutorial as an attachment (Appendix B). We advised participants that the questions would be discussed at the start of the interview. We also included the following working definition of PROs, which was repeated at the beginning of the interview: “For the purposes of these questions, PRO refers to any endpoint derived from patient reports, including single-item outcome measures, event logs, symptom reports, formal instruments to measure health-related quality of life (HRQoL), health status, adherence, and satisfaction with treatment.”

We obtained verbal consent from participants during semistructured telephone interviews (and in-person interviews, when possible [n = 3]) conducted by a trained interviewer. Interviews were audio-recorded and lasted 30 minutes on average. Investigators received $200 for participating. The institutional review board of the Duke University Health System approved this study.

Measures

Following the parent questions and appropriate probes outlined in the Interview Guide, which is available in Appendix A, the interviewer asked participants about their past experiences with PRO instruments, their opinions about the value and usefulness of PROs, and their views on the quality of current instruments. Two additional questions were asked about the value of PROs generally and whether the participant was familiar with the FDA draft guidance on PROs [13]. The interviewer then guided participants through the 6-slide tutorial, which described common problems with current PRO instruments (2 slides), a simplified definition of an item bank and a general process for developing item banks (2 slides), and the novel products of IRT item banks, including customized short forms and CAT (1 slide each). Most of the slides in the tutorial were adapted from a presentation from the National Cancer Institute (B. B. Reeve, personal communication). The slides and accompanying interview notes are available in Appendix B. We added some information to the slides based on the analysis of interviewee responses that indicated need for clarification of tutorial material. New information is highlighted in blue in the text accompanying the slides to distinguish it for the reader as new content, but it might be integrated into a future adaptation of the tutorial.

We recorded whether participants asked questions or made comments during the tutorial. After the tutorial, the interviewer asked specifically about whether participants felt they had a basic understanding of an item bank and its novel products. The interviewer then asked what else

investigators would want to know in order to use an item bank, what barriers to adoption they anticipated, and whether they thought an item bank would be useful to them.

Analysis

Two members of the study team independently coded interview content using a coding dictionary that was created and revised iteratively during review of the audio recordings. The reviewers resolved coding discrepancies through discussion. Responses to 2 of the post-tutorial questions overlapped considerably (namely, what else investigators needed to know to use an item bank and barriers to adoption), so we present the results of these questions analyzed together. Given the small sample size and the qualitative nature of the interviews, we present descriptive results only.

RESULTS

The overall response rate was 35% (42/119), although there was variation in response among the different areas of clinical research: the response rate was 29% in cardiology (10/34), 39% in oncology (11/28), 61% in pediatrics (11/18), and 26% in mixed outcomes (10/39). Table 1 shows the characteristics of the sample. The majority of participants were white, non-Hispanic men, mirroring national representation of physicians [14] and clinical investigators [15]. About 15% of participants had PhDs, representing specialties in psychology and epidemiology. The remaining participants had medical degrees, with specialties in cardiology, internal medicine, pediatrics (including subspecialties in pediatric rheumatology, nephrology, pulmonology, and neonatology), oncology (including subspecialties in gynecologic oncology, gastrointestinal oncology, hematology, and radiation oncology), psychiatry, infectious diseases, rheumatology, and palliative medicine. Most participants (90%) had used PROs in clinical research, but the content of this experience ranged from symptom reports, such as adverse event reporting, to formal quality-of-life measures, such as the Medical Outcomes Study Short Form-36. Likewise, although about half of the participants said they had experience in the development of PRO measures, there was a wide range of experience, from modifying existing measures to developing case report forms to developing and validating formal quality-of-life instruments. One quarter of the participants were familiar with the FDA draft guidance on PROs [13], and 6 others asked to be sent the FDA document after the interviewer's question alerted them to it. Most investigators saw themselves (69%) or a committee of investigators (40%) as responsible for choosing PRO measures for trials; however, sponsors also played a significant role (31%), especially in research involving cardiology outcomes (70%).

During the tutorial, fewer than half of the participants spontaneously asked questions (Table 2). After the tutorial, over 90% of the participants said they had a basic understanding of item banks and their potential products. Over half of the participants thought item banks would be useful to them. An additional third of the participants thought an item bank might be useful, and many qualified that response with specific things they would want to see or know about the item bank before they would consider using it (discussed in detail below). Four participants did not think an item bank would be useful to them; of these, 2 had not previously used PROs in research. Only 1 respondent could not think of additional things he would need to know in order to use an item bank to construct a measure for a trial. Likewise, only 2 respondents could not think of any barriers to the adoption of IRT item banks. The remaining participants voiced concerns about a wide variety of issues (Table 3).

Logistical Barriers

Cost—A frequent concern about IRT item banks (29%) was the perceived high cost or economic constraints involved in gaining access to PRO measures or in implementing new technology such as CAT. Many participants thought that switching to IRT item banks would

be, as one investigator expressed it, “horrendously expensive.” However, many others who cited cost as a barrier to the adoption of item banks appeared to be extrapolating from the already difficult time they have obtaining funding for PRO research. One oncologist described what he saw as the “lack of understanding of the complexity of clinical trials and how underfunded they are. The government says pharma makes billion of dollars and they should pay, but the pharmaceutical companies don’t want to spend their money, especially when the FDA doesn’t require PROs.” Similarly, an investigator in pediatrics thought cost was an issue, “especially in an acute care setting, because at the end of the day when you’re doing a measure, nobody is paying for it. Cost plays a big deal into it, because you’re not charging the patient for assessing it, and it’s considered an extra thing that you’re doing.” Another investigator in pediatrics described how cost was not an issue for high-level research like that funded by R01 grants from the NIH, but for pilot work where funding is more limited, “we may tend to err on the side of getting the lab value over the PRO.”

Feasibility—Participants expressed concern about the feasibility of using measures from IRT item banks. While some participants (10%) mentioned feasibility in general, concerns often focused specifically on computers. Not all types of measures developed from IRT item banks require electronic data capture (for example, administration of fixed-item short forms may be paper-based), but computer use is critical for CAT. After hearing about IRT item banks and CAT, 17% of the participants thought computer availability would be a barrier, and 12% mentioned computer literacy (e.g., “The two issues in our trials are rooms without computers and many patients with low education.”). Consistent computer availability was also an issue. As one internist said, “We’re still in paper mode now. In clinical trials, at least, we collect data in a variety of settings, sometimes at bedside, and sometimes in a room with a computer.”

Burden—Some investigators cited burden on patients, sites, or research personnel as potential barriers to use of PROs generally. For example, one cardiologist said, “The major challenge really is whether the investigator’s site is competent enough and has good people who can do these PROs and actually use the instruments and get the patients to understand them. It does take a certain amount of effort on the part of the coordinator who’s administering the instrument. That, I think, is a problem.” Another investigator described how in her cooperative group there was a lot of concern about patient burden; however, she felt that this was unfounded, given the high response they see in their clinical trials. Instead, she thought blaming patient burden was an “easy out” for people who are reluctant to include additional measures “because it’s more work for the CRAs [clinical research associates] and the statisticians.” One big advantage of PRO measures derived from IRT item banks is that they can be shorter than traditional PRO measures while retaining reliability and precision. Many investigators understood that adoption of IRT item banks in clinical trials was one way to alleviate patient burden, e.g., “the issue of patient burden is not as large with computers, where it can be very quick,” and “I think the key is to get them brief enough that the patient burden is as modest as possible, while the information is maximized.”

Scientific Barriers

Validity—While not mentioned by a majority of participants, there was serious concern among some participants, particularly oncologists, about the validation of PRO instruments generally as well as those developed as part of IRT item banks. As one oncologist put it, “Those of us who take care of patients wonder if we can trust the validation process.” Another oncologist had concerns about widespread adoption of IRT item banks because of the current limited understanding of validity for a tailored test compared to standard instruments: “Most people think you shouldn’t tinker with a validated instrument, and picking and choosing items would be looked at as sleazy.”

One fifth of the interview participants said item banks would have to be validated in their specific disease populations before they would be able to use them. As one pediatrician said, "These types of assays are used for so many different conditions or types of studies, that it seems to me that it's going to be hard to develop an item bank that will be widely applicable. Each investigator has their own focus." Others did not comment on this issue specifically.

Clinical Relevance or Meaning—Participants wanted to know that scores from IRT item banks would be easy to interpret and clinically relevant, especially with regard to precision. One participant said that scores from an IRT item bank would have to be able to discern between respondents who were and were not "doing well" to be useful for her research. Conversely, for other participants the concern was that increased precision was not necessarily clinically relevant. One said, "You are making precision discriminations that are probably irrelevant in the clinical arena." Another voiced the need for demonstration of clinical relevance for IRT-based item banks:

> The one thing we need, which is always the case even with the standard ratings is, what's a clinically meaningful difference, as opposed to what's a statistically significant difference. That's very important to add to the effort. Could run the risk of putting a lot of effort into item banks, you increase sensitivity, but the differences that you're detecting are not of clinical significance.

Details About Item Banks—One third of the participants had questions about how measures from IRT item banks would be better than current PRO measures. A few expressed skepticism about the potential advantages of IRT methods, such as the cardiologist who said, "In this day and age, computers are used all the time. I think there should be no problem. But you have to show that this is better than what we've been using before. Clearly what will happen is that the patient has to sit down, may not know how to use a computer, you'll have to put a coordinator there, and so forth." Another perspective was that IRT item banks provide a potential solution to the problems with current PRO measurement:

> The major value of this item bank system is analogous to how I see computers. They have gotten largely idiotproof. The biggest challenge to collecting PROs is that the tools are too difficult. I'm looking for a simple measure, for example, global quality of life on a 1-to-5 scale. The question is, could you use technology and IRT to create tools that would be easy for investigators to use but that would be rigorous, because obviously global quality of life on a scale of 1 to 5 is not rigorous. Ideally you'd measure quality of life in every study, just like we measure mortality in every study.

After the tutorial, two participants had misunderstandings about how IRT and CAT function. For example, one of the oncologists thought items from a CAT would need to be weighted after their administration to patients in order to generate scores, when in fact CAT is only possible because the psychometric properties of each item in the item bank are determined before administration. These misunderstandings were resolved through continued discussion. Six participants wondered about integrating measures from IRT item banks into the electronic trial data collection systems they have already established (customization). Five investigators suggested including accepted measures (what the PROMIS Network calls "legacy instruments") in tandem with measures from an item bank until the research community is comfortable with the new system. For others, the methodological improvements that IRT item banks offer over traditionally scaled PROs was simply not of interest (e.g., "We already collect more data than we can use.").

Barriers to Change

Habit—A perceived barrier to switching to IRT item banks was how investigators would adapt to using a new method over a known method. Some admitted this would be a problem for

themselves, such as the investigator who explained, “You’ve got old guys like me that have been doing this a certain way for a long time, so that’s obviously going to be a barrier.” Others highlighted this problem among their colleagues:

> People tend to get stuck in a rut about an instrument they think they like, regardless of clinical utility. They tend to use the same thing over and over again. You see that in my field all the time, people publishing using PROs that are very dated, that have been criticized on multiple grounds, and yet investigators keep using it.

Convincing Others—Convincing others of the value of IRT item banks was a concern cited by one fifth of the participants. For example, “Convincing the statisticians who have to work with this from a remote location that this is not only feasible but better than [Scantron systems] … There will be a huge teaching component for people other than the enthusiastic clinical investigator.” Obtaining sponsor buy-in for the new system was also a concern. For example, “[The investigator] would have to convince all of the people that fund these studies and the decision makers that this is the best method of measuring PROs. If the funders require it, and the people doing the trials are in favor of it, it will happen.” Likewise, one pediatrician explained, “Drug companies are not interested, if the FDA won’t buy it. The same thing applies to submitting grants to NIH. The reviewer of the grant may not know about it or be as expert in an area as you are.” Others mentioned this barrier as it relates to publishing specifically or to acceptance of PROs more generally.

DISCUSSION

The adoption of any new technology or methodological approach requires that the eventual end users are educated and that potential barriers to adoption are addressed. The brief tutorial evaluated in this study was sufficient to introduce the idea of item banks to clinical investigators from different backgrounds, and we have included some suggested changes for future versions of the tutorial. The investigators we talked to expressed varying degrees of enthusiasm for the possibilities associated with such item banks. In addition to making comments supportive of item banks, investigators cited logistic and scientific barriers to the adoption of IRT item banks, as well as barriers related to habits or attitudes. Many of these barriers can be addressed through education, such as providing investigators with training on how to use and customize IRT item banks and analyze PRO data collected through CAT. One source for such training is the PROMIS website and the PROMIS Statistical Coordinating Center’s workshops on IRT, CAT, and Assessment Center, a dynamic web-based software application that allows researchers to develop, administer, and analyze customized PRO instruments (http://www.nihpromis.org/WebPages/AssessmentCenter.aspx). Addressing other barriers will require additional work on the part of item bank developers, such as validating item banks in specific populations of patients. A logical first application of item banks will be to promote the inclusion of fixed-item, customized short forms in clinical trials, which can be administered either electronically or on paper.

It is less obvious how to address barriers such as the need to convince multiple stakeholders besides clinical investigators of the acceptability and feasibility of PRO measurement using IRT item banks. These stakeholders include the FDA, federal and industry sponsors, trial operations staff, and fellow clinical researchers. Some investigators noted that it is already difficult to conduct research using PROs and raised the concern that radically modifying PRO instrument development might create difficulties for obtaining research funding and labeling claims from the FDA. According to the draft guidance on PROs, the FDA considers changes to the order of items or deleting portions of a questionnaire to be a modification that requires additional validation [13]. This would suggest that the FDA may not be open to CAT. However, a recent paper coauthored by FDA staff suggests otherwise [6]. This paper says the FDA

encourages demonstrations of IRT-based instruments and CAT in the clinical trial setting and that evaluation of IRT strategies as compared to known PRO measurement paradigms will assist in their evaluations of these new methods. The burden of such comparisons lies with developers of IRT item banks and underscores the critical importance of educating stakeholders about IRT-based measurement of PROs and discussing the benefits and risks of using item banks in clinical research.

The specific percentages of investigators in this study who had various concerns is likely affected by selection bias, as only 35% of investigators who were contacted completed the interview (2 of whom were identified by a PROMIS project investigator). One might expect those who did participate to be more enthusiastic about PRO assessment and more willing to learn new methods. A potentially compounding limitation was that, in an effort to keep the IRT tutorial brief, there was no discussion of some of the potential difficulties, such as the complexity of statistical analyses in repeated measurements [16]. Both selection bias and the brief nature of the tutorial might have led to more positive responses to the concept of item banks in clinical trials. It is noteworthy, however, that participants still listed many concerns about item banks and fairly strong criticism about the quality of PRO measurements to date. An alternative possibility is that more negative responses may have been elicited due to the brevity of the tutorial, since it may not have sufficiently described all of the advantages of IRT and CAT.

Future educational efforts should address technical details that illustrate how, for example, the increased precision of an IRT-based short form can lead to a reduction in the sample size required to detect clinically meaningful differences between groups for trials in which the PRO is the primary outcome [17,18]. Additional tools are also necessary to help clinical investigators decide when increased costs associated with using CAT or related technologies will be offset by cost savings associated with smaller sample sizes, less missing data due to reduced patient burden, and other factors. Such decision support tools are important to ensure that scarce operational resources in a clinical trial are allocated effectively.

Our findings demonstrate a willingness on the part of clinical investigators to learn more about IRT-based measures of PROs. The questions and concerns raised by these investigators merit the attention of those attempting to promote the acceptability and use of item banks in clinical research.

Acknowledgements

Funding/Support: Supported by grant 5U01AR052186 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.

References

1. Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Control Clin Trials 2004 Dec;25(6):535–52. [PubMed: 15588741]
2. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007 May;45(5 Suppl 1):S3–S11. [PubMed: 17443116]
3. Becker J, Schwartz C, Saris-Baglama RN, Kosinski M, Bjorner JB. Using item response theory (IRT) for developing and evaluating the Pain Impact Questionnaire (PIQ-6). Pain Med 2007;8(s3):S129–S44.
4. Petersen MA, Groenvold M, Aaronson N, Blazeby J, Brandberg Y, de Graeff A, et al. Item response theory was used to shorten EORTC QLQ-C30 scales for use in palliative care. J Clin Epidemiol 2006 Jan;59(1):36–44. [PubMed: 16360559]

5. Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, et al. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. Qual Life Res 2004 Dec;13(10):1683–97. [PubMed: 15651539]
6. Reeve BB, Burke LB, Chiang YP, Clauser SB, Colpe LJ, Elias JW, et al. Enhancing measurement in health outcomes research supported by Agencies within the US Department of Health and Human Services. Qual Life Res 2007;16(Suppl 1):175–86. [PubMed: 17530449]
7. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000 Sep;38(9 Suppl):II28–42. [PubMed: 10982088]
8. Nunnally, JC.; Bernstein, IH. Psychometric Theory. Vol. 3. New York: McGraw-Hill, Inc; 1994.
9. Reeve, BB.; Masse, LC. Item response theory modeling for questionnaire evaluation. In: Presser, S.; Rothgeb, JM.; Couper, MP.; Lessler, JT.; Martin, E.; Martin, J., editors. Methods for testing and evaluating survey questionnaires. Hoboken, NJ: John Wiley & Sons, Inc; 2004.
10. Schaeffer GA, Bridgeman B, Golub-Smith ML, Lewis C, Potenza MT, Steffen M. Comparability of paper-and-pencil and computer adaptive test xcores on the GRE General Test: ETS. 1998Report No.: RR-98-38
11. Newton DW, Boyle M, Catizone CA. The NAPLEX: evolution, purpose, scope, and educational implications. Am J Pharm Educ 2008 Apr 15;72(2):33. [PubMed: 18483600]
12. Federation of State Medical Boards & National Board of Medical Examiners. United States Medical Licensing Examination bulletin of information for computer-based testing. Philadelphia, PA: FSMB and NBME Joint Publication; 1999.
13. Food and Drug Administration. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2006. cited; Available from: http://www.fda.gov/CDER/GUIDANCE/5460dft.pdf
14. Community Tracking Study: Physician Survey. 2000. cited; Available from: http://ctsonline.s-3.com/psurvey.asp
15. Getz K, Faden L. Racial disparities among clinical research investigators. Am J Ther 2008 Jan-Feb; 15(1):3–11. [PubMed: 18223347]
16. te Marvelde JM, Glas CA, Van Landeghem G, Van Damme J. Application of Multidimensional Item Response Theory Models to Longitudinal Data. Educational & Psychological Measurement 2006;66 (1):5–34.
17. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol 2007 Jun;34(6):1426–31. [PubMed: 17552069]
18. Fries JF, Bruce B, Bjorner J, Rose M. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. Ann Rheum Dis 2006 Nov; 65(Suppl 3):iii16–21. [PubMed: 17038464]

Table 1

Characteristics of Sample (N=42)

Characteristic	N	%
Gender		
Female	12	29
Male	30	71
Race/ethnicity		
White, non-Hispanic	36	86
White, Hispanic	2	5
Asian	4	9
Degree		
MD	36	86
PhD	6	14
Year graduated (median)	1981	
% research time (median)	63	
Ever used PRO in trial[a]	38	90
Ever developed PRO measure[b]	22	52
Familiar with FDA guidance	11	25
PRO Instrument selection		
Consult expert	7	17
Clinical investigator	29	69
Sponsor	13	31
Committee, team	17	40
Standard	5	12

[a] Wide range from adverse event reporting to formal measures like SF-36

[b] Wide range from modifying existing measure to developing and validating formal instrument

Table 2

Investigator Reactions to Tutorial (N=42)

Reaction	N	%
Asked questions during tutorial	16	38
Understood tutorial[a]		
Yes	39	93
No	1	2
Sort of	2	5
Would an item bank be useful to you?		
Yes	25	60
No	4	10
Maybe	13	31

[a]When asked specifically if they had a basic understanding of item banks and their products

Table 3

Investigator Concerns after Tutorial (N=42)

Concern	N	%
Logistic barriers		
Cost	12	29
Feasibility	4	10
Computer availability	5	12
Computer literacy	7	17
Burden	6	14
Patient burden	5	12
Site/research personnel burden	3	7
Scientific barriers		
Validity	3	7
Face validity	4	10
Selection bias	1	2
Translated/culturally validated	2	5
Surrogate reporters	1	2
Reliability	2	5
Reproducibility	2	5
Sensitivity to change	1	2
Validity in specific population	9	21
Clinical relevance/meaning	6	14
Details about item banks	7	17
How is item bank better?	14	33
Item response theory	2	5
Computer adaptive testing	2	5
Customization	6	14
Legacy items	5	12
Barriers to change		
Habit	5	12
How to convince others	9	21
Publishing	3	7
Expectations of scientific community	3	7
Expectations of FDA	2	5
Acceptance of PROs generally	1	2