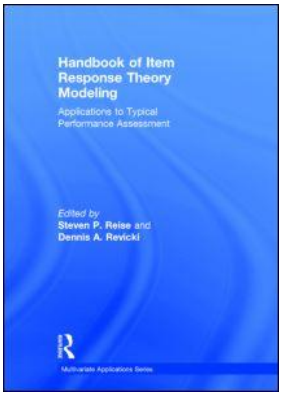

Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

Developing Item Banks for Patient-Reported Health Outcomes

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch16
Dennis A. Revicki, Wen-Hung Chen, Carole Tucker
Published online on: 16 Dec 2014

16 Developing Item Banks for Patient-Reported Health Outcomes

Dennis A. Revicki, Wen-Hung Chen, and Carole Tucker

Introduction

Over the past 30 years, health outcome assessments have been increasingly used in clinical trials and clinical practice settings. Health outcomes assessment is now commonly incorporated into clinical practice and patient health care. In clinical trials and comparative effectiveness research studies, health outcomes assessments are used to demonstrate treatment effectiveness (Ahmed et al., 2012). Health outcomes are also playing a larger role in health system performance assessments by providing critical information on health care quality from the consumer perspective (National Quality Forum, 2013). However, for years, the development of health outcomes assessments was not systematic or consistent, and multiple assessments for different purposes and of varying quality have been developed. Many instruments were developed to meet the immediate needs of a specific clinical study or clinical development program, and often multiple instruments covering the same health domains were simultaneously developed by competing health outcomes researchers. The development of item banks using item response theory (IRT) methods for different health domains provides a way to standardize the health outcomes assessment by drawing from a pool of items that are systematically and consistently developed and that have scores on a standardized metric (Cella et al., 2007; Cella et al., 2010). In addition, a tailored health outcomes assessment can be developed using a subset of items from an item bank so that the assessment is more efficient and less burdensome to patients yet retains adequate precision.

This chapter discusses the necessary components to develop and evaluate the measurement qualities of an item bank. The chapter includes examples from completed research and development activities from the Patient Reported Outcomes Measurement Information System (PROMIS®) network, sponsored by the National Institutes of Health. First, a clear and well-defined health concept needs to be developed because this conceptual framework (content map) provides the fundamental structure for the item bank. Second, factor analyses and item response theory analyses are applied to determine an optimal set of items for the item bank. This chapter summarizes the systematic process used to develop the conceptual framework through qualitative methods, and to describe the item bank construction and calibration phases through quantitative evaluation.

Conceptual Framework

As with the development of any measure that includes a patient-reported outcome (PRO), the initial step in development is identifying the conceptual framework for the item bank. This conceptual framework, or item map, provides a representation of the content of the item bank. The conceptual framework needs to capture the range of experience intended

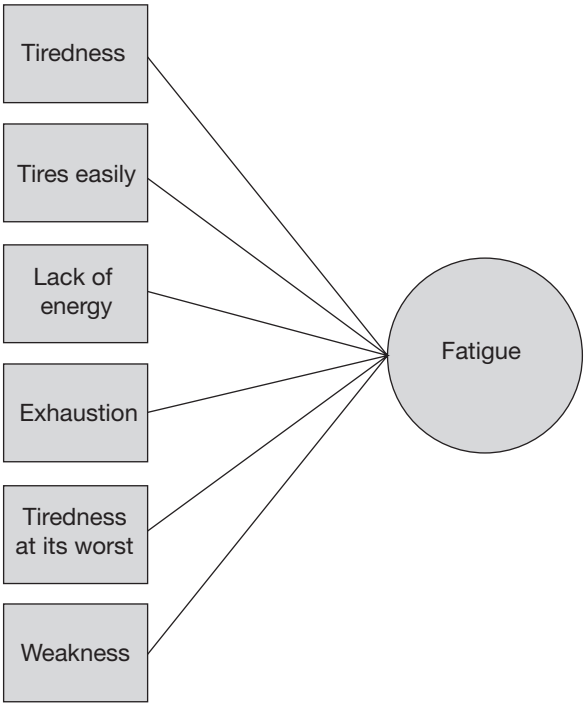

Figure 16.1 Conceptual framework for Anemia Impact Measure Fatigue Scale.

to be reflected in the item bank. The conceptual framework for an item bank is generated based on information from the health-related literature and clinician and patient input (see later in this chapter). An example of the conceptual framework underlying the Anemia Impact Measure (AIM) fatigue experience scale is shown in Figure 16.1 (Kleinman et al., 2012). As can be seen, the tiredness, weakness, lack of energy, and other items fit into the concept of fatigue from the patients' perspective.

The complete PROMIS® domain framework includes physical, mental, and social health (Cella et al., 2010; Riley et al., 2010). This tripartite framework is consistent with the health framework defined by the World Health Organization (World Health Organization, 1958). The PROMIS® domain framework is depicted in Figure 16.2. For example, in the PROMIS® project, emotional distress was conceptualized as including depression, anxiety, and anger domains (Pilkonis, 2011), and the physical function item bank included lower and upper extremity and central neck and back subdomains (Fries, Krishnan, Rose, Lingala, & Bruce, 2011). The conceptual framework for the PROMIS® health domains was originally established based on review of the health outcomes and medical literature, patient qualitative research, and interviews with clinicians and health outcomes researchers. The conceptual framework and item bank content was subsequently rigorously reviewed, and revisions were made to the conceptual framework (Cella et al., 2010; Riley et al., 2010), with more recent revisions made in 2012 (www.nihpromis.org).

Research Methods

Qualitative Item Bank Development

Qualitative development needs to be carefully and systematically conducted to assure and provide evidence of the content validity of a new item bank (Brod, Tesler, & Christensen,

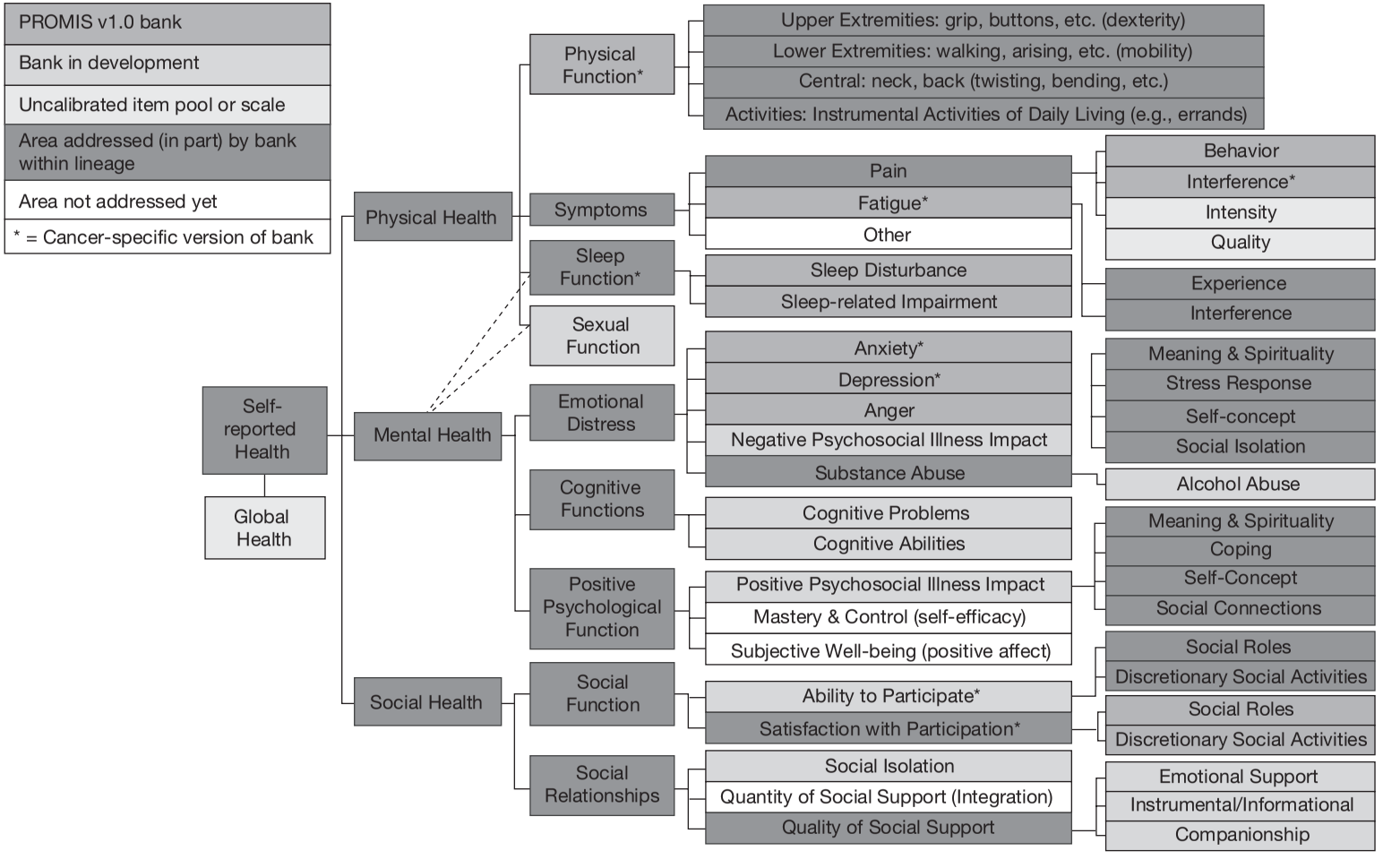

Figure 16.2 PROMIS® conceptual framework.

2009; Lasch et al., 2010; Magasi et al., 2012). Focus groups and cognitive interview study methods are described and examples are provided based on the development research conducted as part of PROMIS® and other projects. This section summarizes the steps in using the qualitative data, combined with information from existing health outcomes instruments, to develop and populate the pool of potential questions for the item bank. Attention to response scales and recall period, depending on the health domain, are also discussed.

The initial stage of item bank development begins with populating the conceptual framework with possible item content. The main sources of potential items are previously developed and published items in the same domain or content area and qualitative interviews. A systematic literature search should be conducted to identify and select possible items for the new item bank. In the PROMIS® project, investigators reviewed the health outcomes, psychological, and medical literature to identify instruments that assessed the target domain (e.g., fatigue, physical function, pain interference, etc.). Given that thousands of items may be initially identified, a system for organizing and evaluating the items is needed. Therefore, the identified items were pooled and a systematic process of "binning and winnowing" was completed followed by item revision, focus group exploration of domain coverage, cognitive interviews on individual items, and final revision before field testing (DeWalt, Rothrock, Yount, Stone, & P.C. Group, 2007).

Binning

Binning refers to a systematic process for grouping items according to content and the specific latent construct (DeWalt et al., 2007). By grouping items systematically, the item bank developer can identify redundancy among different content-relevant items and identify the best potential items based on qualitative characteristics. In the PROMIS® project, the binning of items was accomplished by the domain work groups and the bins were based on previous factor analyses of domain items and theory-based studies of the domain structure. Items that did not fit an existing bin were set aside and reviewed, and new bins were subsequently defined based on emergent groups of these items. Single items that did not fit within a bin were reviewed by content experts, placed within an existing bin, or recorded for possible inclusion later in the conceptual framework. The final set of items in each of the bins was then reviewed to ensure content coverage.

Winnowing

Winnowing refers to the process used to reduce the large item pool to a smaller representative set of items that are consistent with each domain definition and to reduce item redundancy in each domain. Each item was systematically reviewed by two to three team members based on a set of criteria. The specific criteria were:

1. Item content was inconsistent with domain definition.
2. Item was semantically redundant with another item.
3. Item content was too narrow to be universally applicable.
4. Stem of the item was disease specific, reducing general applicability.
5. Item content was confusing.

In the PROMIS® project, across all health domains, about 30 percent of items were removed because of redundancy and about 45 percent of the items were removed because

they did not fit within the domain definitions and structure (DeWalt et al., 2007). During the winnowing process, multiple researchers within the domain workgroup reviewed the items and the decisions about eliminating or including specific items within each domain. For example, in the fatigue item bank the question "How much have you gotten fatigued easily?" was determined to be redundant with other items and was removed, because these other items covered the same concept using simpler language. The item "Do you feel much too much tiredness with normal or soft efforts?" was reviewed and deleted because the content of the item was considered confusing. This process of binning and winnowing resulted in a smaller set of items that was more representative of the structure and definition of each health domain.

Focus Groups

As for any PRO, participant qualitative research is needed to ensure that the individual's experiences with the targeted health domain are covered adequately in the item bank (Brod et al., 2009; Lasch et al., 2010; Magasi et al., 2012; Strauss & Corbin, 1998). Focus group sessions assist the item bank developer to understand and determine the vocabulary and experiences of the targeted group to help inform the development of the domain item bank. The information from focus groups and individual concept elicitation interviews can confirm the domain definitions, identify common language for describing concepts, and identify important gaps in coverage of the targeted domains and concepts. Finally, it is important to cover the patient experience with each domain to ensure that the patient perspective is included in the content of the item banks. Although the PROMIS® project used focus groups for concept elicitation, individual interviews can also be used to identify relevant item bank content for the targeted domain.

A key challenge in developing item banks that will be applied across demographic groups and diseases relates to selecting participants for the focus groups. To adequately cover the experience of all individuals across different demographic and disease groups would likely require thousands of participants, which is not practical. Therefore, some sort of targeting needs to be applied to make sure that age, gender, and racial/ethnic diversity is achieved and that a range of chronic illnesses is covered.

For the PROMIS® project, where multiple item banks were developed to assess multiple domains, three to five focus groups were organized for each domain, except for emotional distress, which held 13 focus groups given the number of domains (i.e., depression, anxiety, anger, alcohol abuse) (DeWalt et al., 2007). Participants were recruited from general medical clinics, outpatient psychiatry clinics, arthritis registries, rehabilitation centers, and schools for the pediatric domains. An attempt was made to achieve diversity in gender, age, and racial/ethnic group diversity (see Table 16.1). Semi-structured interviews were also used to develop concepts within some of the pediatric PROMIS® projects as an alternative approach to focus groups. For these interviews, a script was developed and interviewers were trained to maximally elicit the range of the child's experience within the domains (Forrest et al., 2012).

PROMIS® investigators moderated the focus groups and the participants discussed the range of their experience with the targeted domain (e.g., fatigue, pain, physical function, etc.). After the completion of the focus groups or the interviews, content analysis was performed on the transcripts, investigator notes, and recall of the discussions. The content analysis identified key words and phrases, emergent themes for each domain, and areas incompletely addressed in the domains. This information was then used to enrich and expand the item pools and to revise items to better cover the content from the patient's perspective.

Table 16.1 PROMIS® Network Focus Group Participant Characteristics for Five Health Domains

	Domains				
	Emotional Distress	*Fatigue*	*Social Function*	*Physical Function*	*Pain*
Number groups	13	3	5	3	4
Total participants	104	17	31	15	24
Female, %	50%	65%	65%	80%	79%
Race, %					
White	57%	94%	65%	94%	88%
African American	38%	6%	29%	0%	8%
Other	6%	0%	6%	7%	4%
Hispanic, %	2%	0%	0%	7%	0%
Age, Mean (range)	50 (23–88)	48 (26–65)	53 (23–83)	56 (31–86)	61 (26–76)
Education, %					
<11th grade	5%	6%	3%	7%	0%
High school	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%

Source: DeWalt et al. (2007).

Item Revision, Response Options, and Recall Periods

After the binning and winnowing process and the focus groups, the domain workgroups each had a set of items covering the domain content. However, the set of items included a variety of phrases, recall time frames, response options, and literacy demands (DeWalt et al., 2007). This variation in item content and style would make it difficult to administer as a coherent measure or computerized adaptive test (CAT). The domain item set was then revised to provide a uniform format, make the item content understandable to a range of literacy levels, and minimize ambiguity and cognitive difficulty. A sixth-grade reading level was targeted for all item content in the adult banks, and a third-grade reading level in most pediatric banks. Item stems were revised to be single-barreled (one concept being queried) and grammatically simple. Consideration was also given to consistency in both tense and person within a given item bank.

Next, it is necessary to determine the response options for the items in the domain item bank. For health outcome domains, the variety of different response options can be classified into categories of frequency, duration, intensity, and capability. For any domain item bank, the response scale needs to fit the construct being assessed and make cognitive sense to the respondent. For an item bank, it is important that there is a consistent set of response options and number of response levels within the bank. For item banks, four-to six-level response scales allow for a range of responses and work well for IRT analyses (Bode, Lai, Cella, & Heinemann, 2003). However, more research is needed on response scales and number of response scale levels (Preston, Reise, Cai, & Hays, 2011).

In the PROMIS® project, after review of different response options and psychometric considerations, a small set of different response options were specified (DeWalt et al., 2007) (Table 16.2), which has been further expanded as the number of item banks has grown (www.nihpromis.org). The set of different possible recall periods allowed for flexibility and uniformity among the PROMIS® item banks. The uniformity across different PROMIS® item banks has been key to ensure consistent response options for the profile instruments that use items from several different banks. Most of the PROMIS® item banks used one or more of these response options. The pain behavior item bank used frequency response options, and added, based on cognitive interviews, the response option "*had no pain*" (response scale: 1 (*had no pain*) to 6 (*always*)) (Revicki, Chen, Harnam, et al., 2009).

Finally, the recall period needs to be specified. A variety of recall periods can be used from the past 24 hours, past seven days, past two weeks, or past four weeks. Most important, the recall period needs to fit the health domain. The decision about recall period is complex and there is not much guidance on the best way to select a recall period. For example, there is research evidence that identifies problems associated with respondent recall of health and other concepts related to memory and cognitive heuristics for recalling experiences (Bradburn, Rips, & Shevell, 1987; Erskine, Morley, & Pearce, 1990; Gorin & Stone, 2001; Menon & Yorkston, 2000; Redelmeier & Kahneman, 1996; Robinson & Clore, 2002; Schwartz & Sudman, 1994). Recall accuracy may be improved if a single-day interval is used, but this result may lack generalizability if that particular day was different than most days. Longer recall periods provide a broader report of the behavior, symptom, or experience of interest, but often tend to more heavily weight the most recent days and "estimate" the previous days. This can add to the respondent's cognitive burden by requiring more thought to estimate and recall.

For example, during the development of the AIM, a 24-hour recall period was determined for the anemia-related symptoms (i.e., fatigue, feeling lightheaded, etc.) as symptom experience may vary considerably day to day, while for the daily activities, social activities, cognitive function, and emotions subscales, a seven-day recall period was specified (Kleinman et al., 2012).

For most of the PROMIS® domains, a seven-day recall period was selected. This recall period was selected because it is short enough to minimize biases due to memory and long

Table 16.2 Response Option Examples for PROMIS® Item Banks

Category	*Response Options*	
Intensity	None	Not at all
	Mild	A little bit
	Moderate	Somewhat
	Severe	Quite a bit
	Very severe	Very much
Frequency	Never	Never
	Rarely	Once a week or less
	Sometimes	Once every few days
	Often	Once a day
	Always	Every few hours

Source: DeWalt et al. (2007).

enough to have ecological validity. The main focus was to minimize the patient recall biases, but to have a recall period sufficient to capture experiences that were clinically relevant for health outcomes research. However, not all the PROMIS® domains use a seven-day recall period. For example, the sexual dysfunction domain uses a 30-day recall period, as this longer time period allows for the respondent to experience several sexual events and activities in order to be able to complete the questions (Flynn et al., 2013). The adult physical function domain items do not include any recall period, as the developers considered that function was assessed by self-reported capability rather than self-reported performance (Fries et al., 2011).

Cognitive Interviews

After final revisions to the item content, response scales, and recall period, a cognitive interview study is needed to make certain that respondents understand the meaning of the questions and how to make a response. Cognitive interviewing can be conducted using a variety of techniques, although most studies incorporate a retrospective verbal probing method (Willis, 2005). In retrospective verbal probing, participants are asked to read and complete the item, and then an interviewer asks the participant questions about their understanding of the item content and response scale.

Cognitive interviewing is an integral part of the PRO instrument development process and is essential for ensuring the content validity of the resulting instrument or item bank. Cognitive interviewing examines whether the respondent understands the instrument's instructions, item content, and response scales. Cognitive interviewing focuses on the respondent's comprehension of the item stem and content, the process the respondent uses to retrieve relevant information from memory, decision processes (e.g., social desirability, motivation), and response processes.

The usual process for cognitive interviewing starts with the respondent completing the items or instrument. Next, trained interviewers use a script to query the respondent for other specific information relevant for each item, that is, they probe systematically into the basis for each response (Willis, 2005). Cognitive debriefing studies often include small samples (N = 10–30) depending on the number of items and the complexity of the PRO measure. Often, an iterative approach is taken where each item of the draft instrument is reviewed by 5 to 10 participants, and then revisions, if necessary, are made to the instrument and additional cognitive interviews are completed. The process continues until no further changes are required for the item bank or instrument. The qualitative data from the cognitive interviews are content analyzed to identify misunderstandings, absence of comprehension, and other problems with the item stems and response scales.

For example, in the development of the Anemia Impact Measure (Kleinman et al., 2012), based on the cognitive interviews, the instructions were changed. "Please complete the daily diary every evening after you have eaten dinner" was revised to "Please complete the daily diary every evening at approximately the same time." This was done because patients with chemotherapy-induced anemia reported eating dinner at different times or not eating dinner at all on some days. Also, based on the interviews, the stem of all symptom impact items was changed to read "To what extent during the past 7 days" to improve comprehension of the participants. Several items were also modified based on the cognitive interviews; for example, "Did you need to rest during the day?" was changed to "Did you need to rest more during the day?" The results of cognitive interviewing provide guidance to the item bank developers as to necessary revisions that will ensure comprehension and understanding of the item bank instructions, item content, and response scales.

In the PROMIS® project, sets of 30 items were developed and each item was evaluated by five respondents using cognitive interviewing methods (DeWalt et al., 2007). Trained interviewers used the retrospective verbal probing method for this cognitive debriefing. The small sample size was selected for practical reasons, and each item was reviewed by at least two participants with one of the following: a) education less than high school, b) reading level less than ninth grade, or c) cognitive impairment diagnosis. Interviews were conducted in the South, Northeast, Midwest, and Western regions of the United States. Table 16.3 provides a summary of the demographic characteristics of the PROMIS® adult cognitive interviewing participants by domain. Final revisions to the questions were made, as needed, based on the results of the cognitive interviews. The resulting item banks for the various domains were then ready for field testing and psychometric analyses.

Quantitative Evaluation

Once a draft item bank is developed and available for field testing, it is necessary to collect data to enable psychometric evaluation. Ideally, the data collected for psychometric evaluation will be from the general population to allow for normative scoring and interpretation. A sufficiently large sample is necessary at this stage for evaluating dimensionality, item parameters, model fit, differential item functioning, reliability, validity, and responsiveness.

The recommended sample size for psychometric testing depends on the complexity of domain structure and number of items in an item bank. For most item banks, a sample size of 500 is probably reasonable, and a rule of thumb of 10 participants per item makes

Table 16.3 PROMIS® Network Cognitive Interview Study Participant Characteristics for Five Health Domains

	Domains				
	Emotional Distress	*Fatigue*	*Social Function*	*Physical Function*	*Pain*
Total participants	33	22	21	18	44
Female, %	64%	55%	40%	67%	59%
Race, %					
White	76%	50%	71%	67%	82%
African American	38%	6%	29%	0%	8%
Other	2%	0%	0%	7%	0%
Hispanic, %	2%	0%	0%	7%	0%
Age, Mean (range)	50 (23–88)	48 (26–65)	53 (23–83)	56 (31–86)	61 (26–76)
Education, %					
<11th grade	5%	6%	3%	7%	0%
High school	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%

Source: DeWalt et al. (2007).

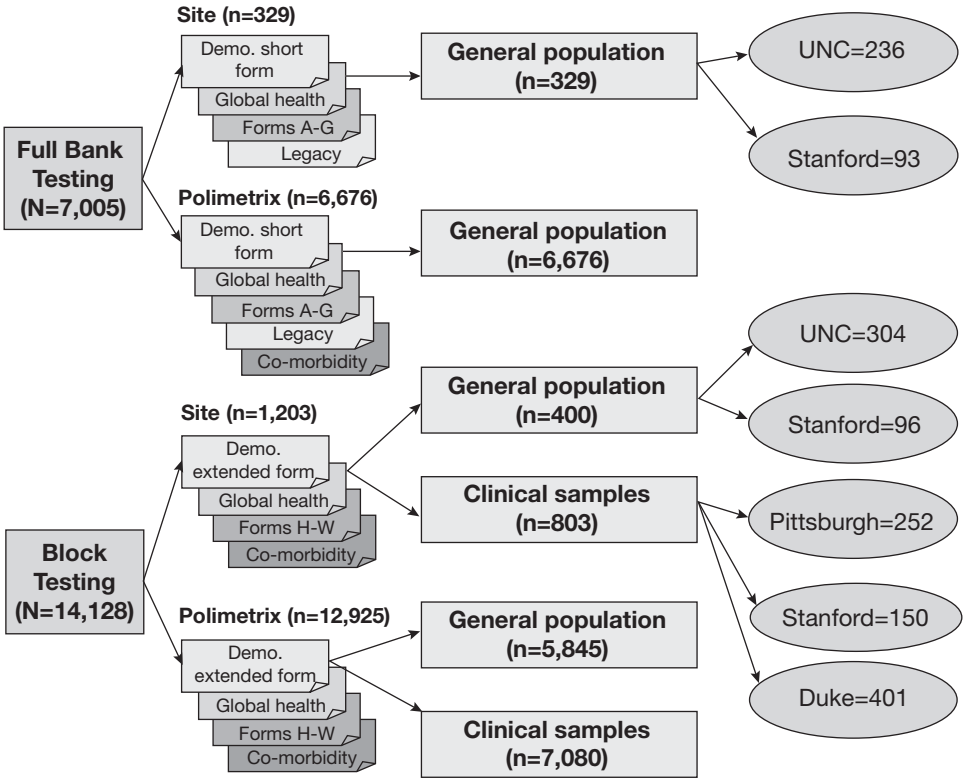

Figure 16.3 PROMIS® Wave 1 samples

psychometric sense. For the multi-domain PROMIS® I project, a sample size of more than 21,000 was recruited because of the large number of domains (14) and items within each domain, to minimize respondent burden, and for the planned psychometric analyses. Because of the number of item banks tested in PROMIS®, a complex data collection strategy was employed (see Figure 16.3). This strategy included two arms and a total sample size of 21,133. A total of 19,601 individuals were recruited by Polimetrix, with the remaining 1,532 recruited from selected PROMIS® primary research sites. Similarly, in the PROMIS® pediatric work, 12,488 children and 5,037 parents were recruited to assess the psychometric properties of the 10 new item banks. An additional 5,000 children and 7,500 parents were recruited for prospective norming of the pediatric item banks.

Unidimensionality

One critical assumption of IRT models relates to the unidimensionality of the set of items, that is, whether the items represent a single underlying construct. No set of items will ever perfectly meet strictly defined unidimensionality assumptions (McDonald, 1981; Reise et al., this volume). All health outcome data contain some multidimensionality. The objective is to assess whether scales are "essentially" or "sufficiently" unidimensional (McDonald, 1999; see also Chapter 2) to allow unbiased scaling of individuals on a common latent trait or construct. One important criterion is the robustness of item parameter

estimates, which can be examined by removing items that may represent a significant secondary dimension. If the item parameters (in particular the item discrimination parameters or factor loadings) significantly change, then this may indicate insufficient unidimensionality (Drasgow & Parsons, 1983; Harrison, 1986). A number of researchers have recommended methods and considerations for evaluating essential unidimensionality (Lai et al., 2006; McDonald, 1981, 1999; Roussos & Stout, 1996; Stout, 1987; see also Chapter 2).

Classical Test Theory Methods to Assess Unidimensionality

Prior to assessing dimensionality, it is recommended that several basic classical test theory statistics are estimated in order to provide descriptive information about the performance of the items in the bank. Item-level descriptive statistics are useful for identifying patterns of skewed and missing responses. Based on the PROMIS® Network experience with the pain behavior item bank, it is essential to identify skewed response patterns and sparse data in the more severe response levels. For example, in the general population, few participants endorsed response options that indicated more frequent pain behaviors. Two distributions of responses were observed, one involving no pain and at least some pain, which was dichotomous, and then a more normal distribution of responses among those reporting some pain (Revicki et al., 2009). Clearly, response distributions of this type have implications for IRT modeling and analyses because some response categories may not be adequately used to provide robust and stable item parameter estimates (see Chapter 13).

Additional analyses may include inter-item correlations, item-scale correlations, and internal consistency reliability. Cronbach's coefficient alpha (Cronbach, 1951) can be used to examine internal consistency with 0.70 to 0.80 as an accepted minimum for group-level measurement and 0.90 to 0.95 as an accepted minimum for individual-level measurement (Hays & Revicki, 2005). Cronbach's alpha provides a measure of the reliability of the composite, which is influenced by all sources of common variance. With large item banks (i.e., > 20 items) of small to moderately correlated items, the coefficients are often very large even when subsequent factor analyses reveals significant multidimensionality. Based on PROMIS® and other experience (Cortina, 1993), coefficient alpha is not a very useful indicator of unidimensionality in health domain item banks.

Factor Analytic Methods to Assess Unidimensionality

Confirmatory factor analysis (CFA) is recommended to evaluate the extent that the item pool measures a dominant dimension that is consistent with the content experts' definition of the domain (Reeve et al., 2007). Exploratory factor analysis (EFA) is recommended and is often used as the first step to explore dimensionality. In very specific situations, CFA can be selected as the first step when the pool of items are carefully developed to represent a dominant construct based on an exhaustive literature review and qualitative research. However, in most cases it is advisable to start with EFA to understand potential multidimensionality in the item bank data. Because of the ordinal nature of the patient-reported outcome data, appropriate software (e.g., MPLUS (Muthén & Muthén, 1998) or LISREL (Jöreskog, Sörbom, & Du Toit, 2003)) should evaluate polychoric correlations using an appropriate estimator. Polychoric correlations are used because of the ordinal nature of the data. Recommendations from PROMIS® include weighted least squares with adjustments for the mean and variance estimator in MPLUS or diagonally weighted least squares estimator in LISREL for the confirmatory factor analysis.

The CFA model fit needs to be assessed by examining multiple indices. Given that statistical criteria like the chi-square statistic are sensitive to sample size, a range of practical fit indices needs to be examined such as the comparative fit index (CFI > 0.95 for good fit), root mean square error of approximation (RMSEA < 0.06 for good fit), Tucker-Lewis Index (TLI > 0.95 for good fit), standardized root mean residuals (SRMR < 0.08 for good fit), and average absolute residual correlations (< 0.10 for good fit) (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2010; McDonald, 1999; Reeve et al., 2007). However, these CFA fit indices do not always provide useful information on unidimensionality when alternative multidimensional structures may underlay the data (Reise et al., 2012).

EFA needs to be conducted to explore unidimensionality and to determine the multidimensional structure of the data. The magnitude of eigenvalues for the larger factors (at least 20 percent of the variability on the first factor is especially desirable), differences in the magnitude of eigenvalues between factors, scree test, parallel analysis, correlations among factors, and factor loadings may be inspected to determine the underlying structural patterns. The concept of simple structure needs to be considered in interpreting the derived factor structure. Simple structure is present when items load primarily on only one factor, with minimum loadings on other factors in an EFA.

An alternate method to determine whether the items are "sufficiently" unidimensional is the bifactor item factor analysis model (Gibbons, Hedeker, & Bock, 1992; McDonald, 1999). The bifactor approach to assessing unidimensionality is to assign each item to a specific subdomain based on theoretical considerations. A model is then fit with each item loading on a common factor and on a specific subdomain ("group" factor). The common factor is defined by all the items, while each subdomain is defined by subsets of items in the bank. The factors are constrained to be mutually uncorrelated so that all covariance is partitioned either into loadings on the common factor or onto the subdomain factors. If the standardized loadings on the common factor are all salient (i.e., greater than 0.30) and substantially larger than loadings on the group factors, the item bank is considered to be "sufficiently homogeneous" (McDonald, 1999). In addition, the researcher can compare individual scores under a bifactor and unidimensional model. If scores are highly correlated (e.g., $r > 0.90$), this provides further evidence that the effects of multidimensionality are ignorable (Reise & Haviland, 2005).

For example, a single factor CFA was fit to the PROMIS® pain interference and pain behavior item banks (Amtmann et al., 2010; Revicki et al., 2009). Table 16.4 summarizes the fit statistics for the two item banks. Clearly, there is evidence supporting essential unidimensionality for both item banks. In addition, the first extracted factor in an exploratory

Table 16.4 Confirmatory Factor Analysis Goodness of Fit Statistics for PROMIS® Pain Interference and Pain Behavior Item Banks

	Pain Interference	*Pain Behavior*
Number of items	41	52
CFI	0.97	0.90
TLI	0.99	0.99
RMSEA	0.175	0.156
SRMRR	0.033	0.035

Source: Revicki et al. (2009); Amtmann et al. (2010).

factor analysis accounted for 86 percent and 90 percent of the variance for pain interference and pain behavior, respectively. The RMSEA values exceed recommended criteria; however, this finding is not surprising given that Cook, Kallen and Amtmann (2009) found that RMSEA values tend to be elevated when there are larger numbers of items, and that this statistic may not be appropriate for evaluating dimensionality of large item banks.

Caution should be exercised when interpreting fit based on the usual CFA fit indices (i.e., CFI, RMSEA, etc.), as they may indicate relatively good fit, when more multidimensional models may more likely fit the data. For example, Reise and colleagues (2012) demonstrated that the usual structural equation modeling fit indices such as CFI or SRMR routinely reject unidimensional measurement models even in contexts in which the structural coefficient bias is low. If there is any indication of multidimensionality, Reise and colleagues (2012) recommend examining factor strength indices (e.g., explained common variance, OmegaH) and applying bifactor models to determine if the data is sufficiently unidimensional for IRT analyses.

Local Independence

Local independence assumes that once the dominant factor influencing an individual's response to an item is controlled, there should be no significant association among item responses (Steinberg & Thissen, 1996; Wainer & Thissen, 1996; Yen, 1993). The presence of local dependence influences IRT parameter estimates and represents a problem for scale construction and computerized adaptive test (CAT) applications. Uncontrolled local dependence among items in a CAT assessment may result in a score unrelated to the PRO construct being assessed. Scoring respondents based on mis-specified models will result in inaccurate estimates of their level on the underlying trait (specifically, the standard errors will be too small). This problem occurs frequently in health outcomes measurement because of narrowly defined constructs and similar repeated items. For example, two items on current pain (i.e., How intense is your pain right now? What is your level of pain right now?) showed evidence of local dependency.

Identification of local dependence among polytomous response items includes examining the residual correlation matrix produced by the single factor CFA. High residual correlations (greater than 0.2) are flagged and considered for possible local dependence. In addition, IRT-based tests of local dependence can be utilized including Yen's (1984) Q3 statistic and Chen and Thissen's (1997) local dependence indices. These statistics are based on a process that involves fitting a unidimensional IRT model to the data, and then examining the residual covariation between pairs of items, which should be zero or near zero if the unidimensional model fits.

The modification indices (MIs) of structural equation modeling (SEM) software may also be used as statistics to detect local independence. When inter-item polychoric correlations are fitted with a one-factor model, the result is a limited information parameter estimation scheme for the graded normal ogive model. The MIs for such a model are one degree of freedom chi-square scaled statistics that suggest unmodeled excess covariation between items, which, in the context of item factor analysis, is indicative of local independence.

Items flagged as locally dependent are then examined to evaluate their potential effect on IRT parameter estimates. One test is to remove one of the locally dependent items, and to examine changes in IRT model parameter estimates and in factor loadings for all the other items.

To handle local dependence on item and person parameter estimates, one of the items can be removed from the item bank. If this is not feasible because both items provide a substantial amount of contextual information, then locally dependent items can be

marked as "enemies," preventing them from being administered in a single assessment to any individual. Local dependence needs to be controlled in the calibration step to remove the influence of these highly correlated items. In all cases, the locally dependent items should be evaluated to understand the source and resultant impact of this dependency.

Monotonicity

In the context of health outcomes research, the assumption of monotonicity means that the probability of endorsing or selecting an item response indicative of better health status should increase as the underlying level of health increases. This is a basic requirement for IRT models of items with ordered response categories. Approaches for evaluating monotonicity include examining graphs of item mean scores conditional on "rest-scores" (i.e., total raw scale score minus the item score) using the MOKKEN package in R, or fitting a nonparametric IRT model (Ramsay, 1997) to the data that yields initial IRT probability curve estimates. A nonparametric IRT model fits trace lines for each response to an item without any a priori specification of the order of the responses. The data analyst then examines the fitted trace lines to determine which response alternatives are associated with lower levels of the domain and which are associated with higher levels. The shapes of the trace lines may also indicate other departures from monotonicity, such as bimodality. While nonparametric IRT may not be the most efficient way to produce the final item analysis and scores for a scale, this type of analysis can be very informative about the tenability of the assumptions of parametric IRT.

Item Calibration and Item Response Theory Model

Once the assumptions have been confirmed, IRT models are fit to the data for item and scale analysis as well as for item calibration. IRT refers to a family of models that describe, in probabilistic terms, the relationship between a person's response to a question and his or her level on the PRO latent construct (e.g., physical function, fatigue, etc.) that the scale measures (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). For every item in the item bank, item parameters are estimated. The item slope or discrimination parameter describes how well the item performs in the scale in terms of the strength of the relationship between the item and the underlying scale. The item difficulty or threshold parameters identify the location along the construct's latent continuum where the item best discriminates among individuals.

The PROMIS® network evaluated both a general IRT model—the graded response model (GRM; Samejima, 1969, 1997)—and two models based on the Rasch model framework—the partial credit model (Masters, 1982) and the rating scale model (Andrich, 1978; Wright & Masters, 1982). Based on these analyses, PROMIS® network recommended using the GRM in item bank development work. The GRM is a very flexible model of the parametric, unidimensional, polytomous-response IRT family of models. Because it allows discrimination to vary item by item, it typically fits response data better than a one-parameter model (Embretson & Reise, 2000; Thissen & Orlando, 2001). Compared to alternative models such as the generalized partial credit model, the GRM is relatively easy to understand and illustrate and retains its functional form when response categories are merged. The GRM offers a flexible framework for modeling the participant responses to examine item and scale properties, to calibrate the items of the item bank, and to score individual response patterns in the health outcomes assessment. In PROMIS®, other IRT models were fit as needed. For example, in the analysis of the pain

behavior item bank, nominal and hybrid nominal and partial credit IRT models were examined (Revicki et al., 2009). The GRM was the primary IRT method used in PROMIS® (Cella et al., 2010).

The unidimensional GRM is a generalization of the IRT two-parameter logistic model for dichotomous response data. The GRM is based on the logistic function that describes, given the level of the trait being measured, the probability that an item response will be observed in *category k or higher*. For ordered responses $X = k$, $k = 1,2,3, \ldots, m_i$, where response m reflects the highest theta (latent ability) value, this probability is defined (Samejima, 1969, 1997; Thissen, 2001) as:

$$P(X_i = k \mid \theta, b_i, a_i) = \frac{1}{1+\exp\left[-a_i(\theta - b_{i,k-1})\right]} - \frac{1}{1+\exp\left[-a_i(\theta - b_{i,k})\right]}.$$

This function models the probability of observing each category as a function of the underlying construct. The subscript on m indicates that the number of response categories does not need to be equal across items. The discrimination (slope) parameter a_i varies by item i in a scale. The threshold parameters b_{ik} varies within an item with the constraint $b_{k-1} < b_k < b_{k+1}$, and represents the point on the theta axis where the probability passes 50 percent that the response is in category k or higher.

IRT model fit is assessed using a number of indices. Residuals between observed and expected response frequencies by item response category are compared, and the fit for different models is based on analyses of the relative magnitudes of the differences (residuals). IRTFIT (Bjorner et al., 2006) is used to assess model fit for each item and computes the extension of S-X^2 and S-G^2 for items with more than two responses (Orlando & Thissen, 2000, 2003). These statistics estimate the fit of the item responses to the IRT model, that is, whether the responses follow the pattern predicted by the model. Statistically significant differences indicate poor fit. The S-X^2 (a Pearson X^2 statistic) and S-G^2 (a likelihood ratio G^2 statistic) are fit statistics that use the sum score of all items and compare the predicted and observed response frequencies for each level of the scale sum score. The ultimate goal is to determine the extent to which misfit affects model performance in terms of the valid scaling of individual differences (Hambleton & Han, 2005).

For example, the IRT parameters and fit statistics for selected pain behavior items are summarized in Table 16.5 for participants who reported more than mild pain (Revicki et al., 2009). The item on "isolating myself from others" had a slope (a) of 2.71 and assessed a broad range of the construct of pain behavior. Although the item on "moving slowly" had a similar slope (2.67), a narrower range of the pain behavior construct was assessed.

Once the fit of the IRT model to the response data is considered satisfactory, attention is shifted to analyzing the item and scale properties of the PROMIS® domains. The psychometric properties of the items will be examined by review of their item parameter estimates, CRCs, and item information curves (Reeve, 2003; Reeve & Fayers, 2005). Information curves indicate the range of theta where an item best discriminates among individuals. Higher information indicates greater precision for measuring a person's domain level. The height of the curves (denoting more information) is a function of the discrimination power (*a* parameter) of the item. The location of the information curves is determined by the threshold (*b*) parameter(s) of the item. Information curves indicate which items are most useful for measuring different levels of the measured construct. The information curves for selected items from the PROMIS® pain behavior item bank are depicted in Figure 16.4.

Table 16.5 IRT Item Parameters and Fit Statistics for PROMIS® Pain Behavior Items, Excluding Subjects With 0 or 1 Global Pain Responses: Pooled WAVE-1 and ACPA Data (N = 9,589)

ID	*Item*	*Slope*[1]	*Category Threshold*[1]					*Item Fit Statistics*[2]				*df*
		A	*b1*	*b2*	*b3*	*b4*	*b5*	*S-G2*	*Prob_G2*	*S-X2*	*Prob_X2*	
PAINBE2	When I was in pain I became irritable	2.90	–2.16	–0.82	–0.07	0.81	1.64	55.97	0.2936	52.29	0.4236	51
PAINBE3	When I was in pain I grimaced	2.55	–2.24	–0.86	–0.03	1.02	2.03	46.83	0.6770	47.95	0.6338	52
PAINBE8	When I was in pain I moved extremely slowly	2.67	–2.22	–0.93	–0.18	0.68	1.49	59.31	0.2266	57.11	0.2911	52
PAINBE24	When I was in pain I moved stiffly	2.15	–2.37	–1.11	–0.42	0.57	1.48	73.67	0.0317	67.45	0.0875	53
PAINBE25	When I was in pain I called out for someone to help me	2.51	–2.01	0.58	1.25	2.07	2.83	35.09	0.8792	35.73	0.8625	46
PAINBE31	I limped because of pain	1.81	–2.41	–0.47	0.06	0.93	1.71	55.35	0.3859	55.70	0.3735	53
PAINBE37	When I was in pain I isolated myself from others	2.71	–2.26	–0.24	0.26	0.92	1.85	76.82	0.0039	64.82	0.0433	47

[1] Using IRT software MULTILOG (Thissen et al., 2002).
[2] Using SAS macros IRTFIT (Bjorner et al., 2006).
*Two highest response categories are collapsed for Items 40 and 41.

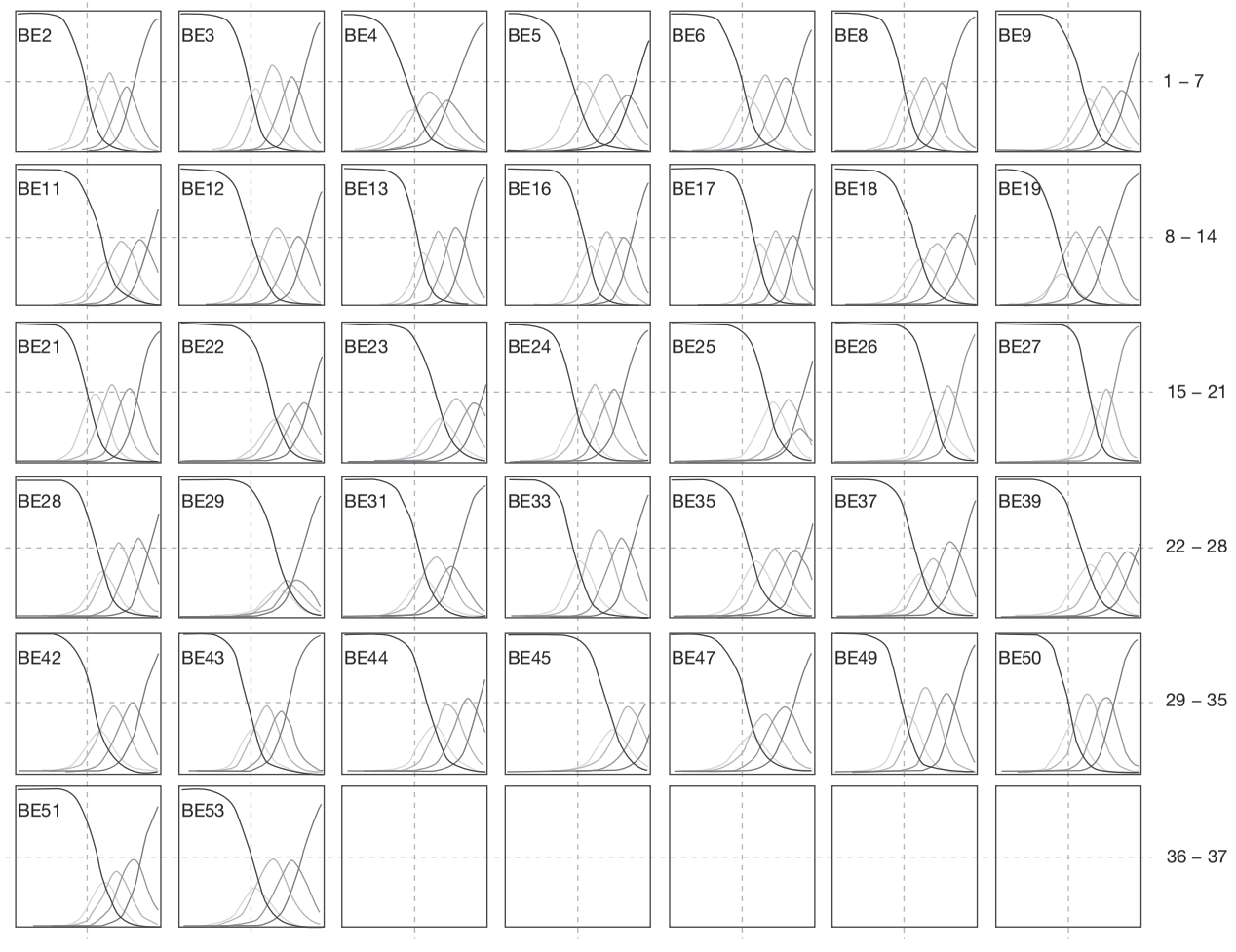

Figure 16.4 Category response curves for pain behavior items.

Poorly performing items are reviewed by content experts before the item bank is established. Misfitting items may be retained or revised when they are identified as clinically relevant and when alternative, better-fitting items are not available. Low discriminating items in the tails of the theta distribution (at low or at high levels of the construct being measured) may be retained or revised to add information for extreme scores though they would not have been retained in regions of the continuum better populated by items.

Differential Item Functioning

Differential item functioning (DIF) refers to the situation where members from different groups (i.e., age, gender, race, education, culture) on the same level of the latent trait have a different probability of giving a certain response to a particular item. DIF is a threat to the validity of any health outcome instrument. DIF occurs when subjects on the same level of the latent trait, such as disease severity, answer the same item differently depending on their group membership (Chang, 2005; Holland & Thayer, 1988). The concept of uniform and nonuniform DIF was first introduced by Mellenbergh (Mellenbergh, 1982). Uniform DIF occurs when the discrepancy between groups is constant across the range of the latent construct. Whereas, with nonuniform DIF, the discrepancy between groups differs depending on the level of the latent construct. The validity of the PRO measure may be compromised because the response to the item with DIF may be due to some external factor other than the intended construct. For example, crying spells is one of the symptoms of patients with depression, but this concept is endorsed more often by women than men with the same level of depression severity (Teresi et al., 2009). Therefore, the crying item is said to exhibit DIF due to gender for assessing depression.

DIF is a discrepancy between two groups of subjects in the conditional probability of a response to an item conditioned on the level of the latent construct. Because IRT consists of mathematical models already expressed as the conditional probability of a certain item response given the latent trait, its application to DIF detection is straightforward. Two methods commonly used for DIF detection for polytomous response items are the IRT log-likelihood ratio (IRTLR) (Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Thissen et al., 1986; Thissen et al., 1993) and ordinal logistic regression (Zumbo, 1999). The IRTLR is based on a multidimensional model of DIF that formally defines a second latent trait that contributes to the DIF. The ordinal logistic regression method of detecting DIF is a straightforward application of logistic regression modeling. Because an interaction term of latent trait by groups can be included in the model, the logistic regression method can be used to detect the nonuniform DIF directly. A method based on complex multiple indicators, multiple causes (MIMIC) confirmatory factor analysis model can also be used for DIF detection (Finch, 2005).

In the PROMIS® project, most of the domain item bank investigators used either ordinal logistic regression or IRTLR modeling; however, the investigators primarily used logistic regression and ordinal logistic regression using an observed conditioning score. A modification, IRTOLR (Crane, van Belle, & Larson, 2004; Crane, Gibbons, Jolley, & van Belle, 2006), was used in some analyses. Estimates from a latent variable IRT model, rather than the traditional observed score, are used as the conditioning variable; this method incorporates effect sizes into the uniform DIF detection procedure. This allows for DIF criteria specification, that is, statistical tests of uniform and nonuniform (Swaminathan & Rogers, 1992), or an effect size modification based on changes in the pseudo-R^2 in nested model (Zumbo, 1999).

An alternative method involving IRTLR likelihood ratio tests (Cohen et al., 1996; Kim & Cohen, 1998; Thissen et al., 1986; Thissen et al., 1993) in IRTLRDIF, MULTILOG (Thissen & Orlando, 2001; Thissen et al., 2002), and IRTPRO (Cai et al., 2011), can be used for DIF detection, accompanied by magnitude measures (Teresi, Kleinman, & Ocepek-Welikson, 2000), such as the non-compensatory DIF index (Flowers, Oshima, & Raju, 1999; Raju, Van Der Linden, & Fleer, 1995). Scale-level impact can be assessed using expected scale scores, expressed as group differences in the total test (scale) response functions, that show the extent to which DIF cancels at the scale level.

Given that numerous factors can impact DIF assessment, it is difficult to recommend a single, best method for evaluating DIF. However, generally, IRT-based methods are recommended. The best recommendation is to have a primary method and secondary method used for sensitivity analyses. Moreover, the magnitude of DIF should be assessed together with both scale and individual impact.

Multiple factors may impact DIF and the assessment of PROs. In many cases, differences in item responses are expected for members of different groups. An item may cover very important content for assessing a health outcome despite different groups of participants responding to it differently. In this case, the item should be retained in the item bank, but each individual's response should be compared within his or her reference group. This requires the item to be treated differently according to group membership, for example, different item scores between male and female. An ideal assessment of health outcomes involves tailoring the items to the unique characteristics of the individual for maximum information. This is possible by using CAT with a large item bank. In this case, items with DIF are no longer a threat to the validity of the test and become assets of the instrument.

Evaluating Psychometric Characteristics

The calibrated item bank can be used to develop dynamic (CAT) or fixed length static forms for application in clinical trials and other studies. As with any PRO measure, the measurement properties of the CAT or static forms need to be evaluated. The most relevant psychometric characteristics that need to be evaluated include test-retest reliability, construct validity, and responsiveness. While a comprehensive review of these measurement properties is beyond the scope of this chapter, we will briefly summarize each approach. More comprehensive reviews are available (De Vet, Terwee, Mokkink, & Knol, 2011; Fayers & Machin, 2007; Hays & Revicki, 2005; Nunnally & Bernstein, 1994).

Reliability

In classical test theory, the concept of reliability is at the scale score level. For IRT, the concept of reliability is conceptualized as "information" and examines measurement precision that can differ across the levels of a construct. The relationship between information and standard error (SE) is defined by the formula $SE(\theta) = 1/\sqrt{I(\theta)}$, where θ is the estimated trait level, SE is the standard error of θ, and I is information. As the formula indicates, increased scale information is associated with smaller SEs and, therefore, greater precision.

Figure 16.5 summarizes the standard errors over the range of pain behavior scores for the total item bank, a seven-item short form, and a seven-item computerized adaptive test (Revicki et al., 2009). For the full item bank, reliability is 0.90 or greater across most of the score distribution, and the short form and CAT have reliabilities exceeding 0.80 across the majority of the score distribution. In comparison, Cronbach's alpha for the full pain behavior item bank was 0.98.

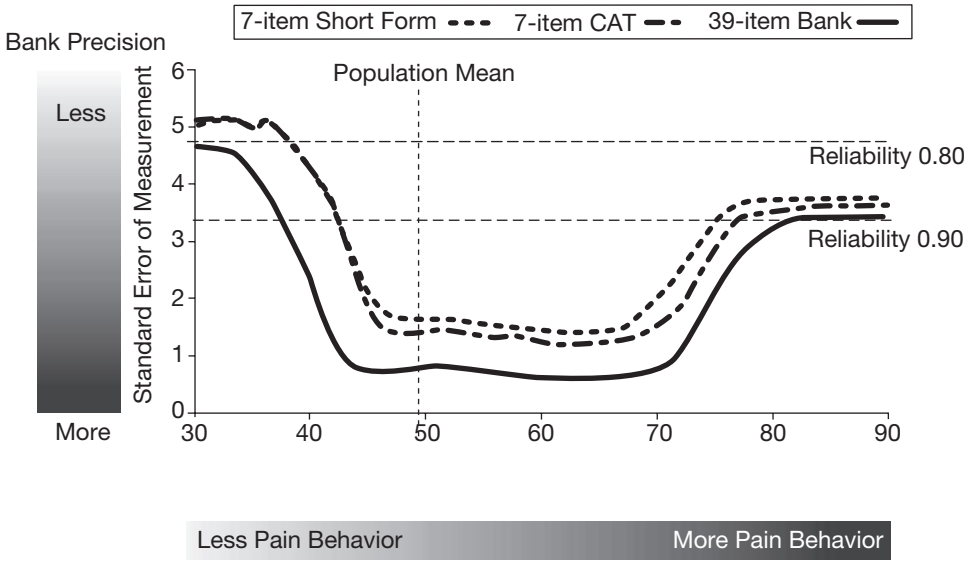

Figure 16.5 SEM for pain behavior item bank, short-form and CAT.

Test-Retest Reliability

Test-retest reliability, or reproducibility, is the relationship between scores obtained by the same individual on two or more separate occasions (Hays & Revicki, 2005). Most often test-retest reliability is tested over short periods of time (i.e., one to two weeks) in individuals who are not expected to change. Intraclass correlation coefficients (ICCs), based on one-way, two-way fixed, or two-way random analysis of variance models depending on data structure, are normally used to assess test-retest reliability (Hays & Revicki, 2005). For PROMIS® CAT and static scores, test-retest reliability is evaluated in a number of clinical samples. For example, the pain behavior domain scores are evaluated in patients with low back pain, osteoarthritis, and multiple sclerosis.

Currently, there is some evidence supporting the stability of the short form and CAT scales based on the PROMIS® item banks. Pilkonis and colleagues (2011) examined the test-retest reliability of the emotional distress and sleep function scales. Based on a seven- to 14-day period, test-retest reliability (ICCs) was 0.79 to 0.89 for anger, 0.85 to 0.86 for anxiety, 0.76 to 0.90 for depression, 0.89 to 0.93 for sleep disturbance, and 0.83 to 0.95 for sleep impairment. In a recent study, Bajaj and colleagues (2011) evaluated the psychometric characteristics of multiple PROMIS® CATs (i.e., anger, anxiety, depression, fatigue, physical function, pain interference, pain behavior, social activities, social roles, sleep disturbance, sleep impairment) in 200 patients diagnosed with cirrhosis. These investigators found that test-retest reliability (ICCs) ranged from 0.76 to 0.99 for the CAT scores over a 14-day interval.

Construct Validity

Validity represents the extent to which a PRO measure reflects what it is intended to measure rather than something else (Hays & Revicki, 2005; Nunnally & Bernstein, 1994). The

process of evaluating the validity of PRO measures involves accumulating evidence that reflects the degree to which the measures denote what they were intended to represent. To evaluate construct validity, a series of hypotheses are generated as to how measures should "behave" and then observed data are used to test these hypotheses (Cronbach & Meehl, 1955). Hypotheses are stated regarding the direction (and sometimes the strength) of relationships that might be expected, and validity is supported when the associations are consistent with hypotheses. The evaluation of construct validity is iterative by its nature with empirical results accumulating to provide further support and increasing confidence in the validity of the PRO measure. The main types of validity assessed for PRO measures include concurrent, predictive, and known-groups validity.

Both concurrent and known-groups validity are being used to evaluate the measurement properties of the CAT and static form scores from the PROMIS® item banks (Cella et al., 2010; Rothrock et al., 2010). Concurrent validity, for most of the item banks, is evaluated by correlations with legacy instruments that assess comparable health outcome domains (Cella et al., 2010). Table 16.6 provides a summary of the concurrent validity findings for several of the PROMIS® domains. The PROMIS® pain interference scores were correlated with Brief Pain Inventory interference scores (Amtmann et al., 2010). The PROMIS® pain interference score was correlated 0.90 with the BPI pain interference scale, 0.48 with pain intensity, –0.84 with the SF-36 bodily pain scale, and 0.55 with PROMIS® physical function scores.

Known-groups validity, that is, whether the PRO score varies significantly by levels on a criterion variable (i.e., clinical severity) or some other characteristic, has been evaluated for many of the PROMIS® item bank scores (Cella et al., 2010). For example, pain interference and behavior scores have been compared by levels of pain intensity (Amtmann

Table 16.6 Concurrent Validity of Selected PROMIS® Domains

PROMIS® Domain		*Legacy Measure*
Physical Function		
–0.88		Health Assessment Questionnaire
0.88		SF-36 Physical Function Scale
Fatigue		
0.95		FACIT-Fatigue Scale
0.88		SF-36 Vitality Scale
Anxiety	**Depression**	
0.75	0.83	Center for Epidemiologic Studies-Depression Scale
0.80	0.72	Mood and Anxiety Symptom Questionnaire
Pain Interference		
0.81		Brief Pain Inventory-Severity Scale
0.85		Brief Pain Inventory—Interference Scale
–0.83		SF-36 Bodily Pain Scale
Sleep Disturbance		
0.85		Pittsburgh Sleep Quality index
0.70		Epworth Sleepiness Scale

et al., 2010; Revicki et al., 2009). PROMIS® pain behavior scores varied significantly by ratings of pain intensity with higher pain behavior scores in those with greater pain intensity numerical rating scale (NRS) (see Figure 16.6). Mean pain interference scores varied significantly by levels of general health status (see Figure 16.7). Rothrock and colleagues (2010) compared various PROMIS® domain scores by chronic disease groups and those without reported chronic diseases to evaluate the validity of the PROMIS® scores. The results indicated that across the different PROMIS® domains, the presence of chronic conditions was associated with more impaired scores compared with those without chronic medical conditions.

Responsiveness

Responsiveness is critical for clinical trial applications, and refers to the ability of a PRO measure to detect changes in health status in those individuals who are changing in clinical status (Hays & Revicki, 2005; Revicki et al., 2008). Responsiveness of health outcome assessments is evaluated by using anchor-based or distribution-based methods. However, the anchor-based methods provide the most insight into responsiveness with the distribution-based methods providing additional supportive information on the magnitude of effects. In the anchor-based approach, multiple relevant anchors are identified and changes in health outcome scores are evaluated by these anchors.

The selected anchors can be clinical measures (i.e., hemoglobin levels, clinician severity ratings, signs and symptoms, etc.) or other patient-reported measures (i.e., patient ratings of disease severity, other measures with known response scales and interpretation). Often, clinicians and patients are asked to rate the change in the patient's health

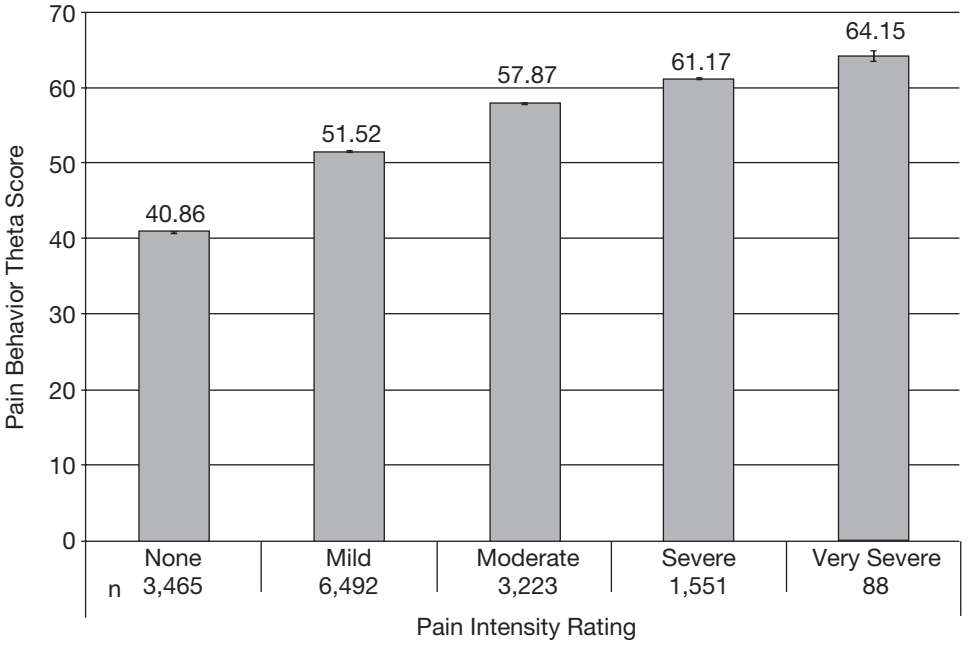

Figure 16.6 Mean pain behavior T-scores by pain intensity.

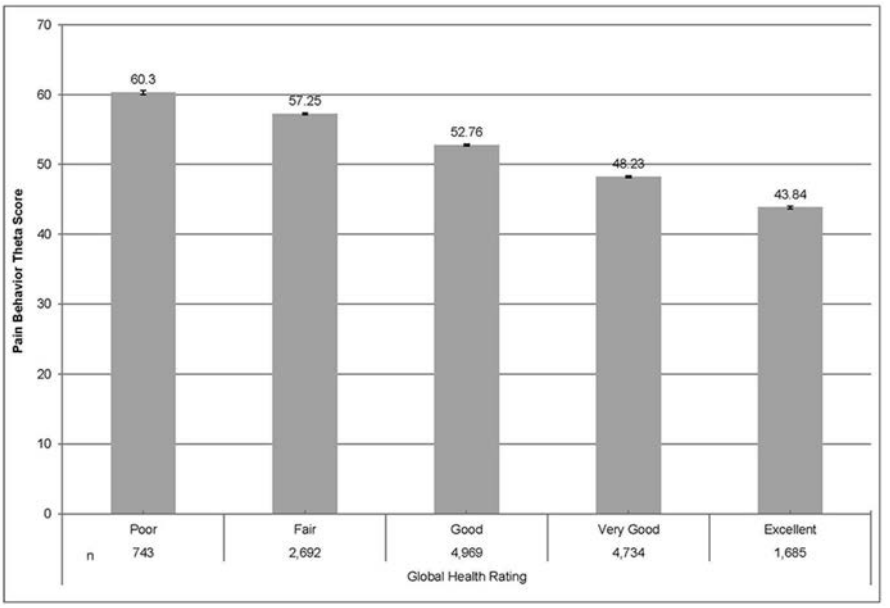

Figure 16.7 Mean pain interference T-scores by general health status.

status from baseline in a longitudinal study. However, limitations in recall for baseline health status may limit the usefulness of patient global ratings of change, and variations in clinician understanding of facets of the patient's health status may also limit the clinician ratings. Anchors should be correlated at least moderately (> 0.35) with the targeted health domain. Ideally, multiple anchors are available and responsiveness can be estimated across the anchors to get a sense of the responsiveness of the item bank derived CAT or short-form scale.

A number of studies have been completed or are under way to document the responsiveness of the PROMIS® measures in patients with low back pain, osteoarthritis, depression, COPD, heart transplantation, and various cancer diagnoses. The PROMIS® depression scores have been demonstrated to be responsive to antidepressant treatment in patients with major depressive disorder (Pilkonis et al., 2011), across multiple domains after heart transplantation (Weinfurt, 2011), in rheumatoid arthritis patients (Fries et al., 2011), and for pain interference after treatment for low back pain (Revicki & Cook, 2011). In fact, the PROMIS® 20-item physical function scale outperformed the Health Assessment Questionnaire in sensitivity to changes in clinical status among rheumatoid arthritis patients (Fries et al., 2011). For low back pain in those patients who achieved a 50 percent improvement in pain intensity NRS, the PROMIS® pain interference scale demonstrated a 1.72 effect size compared with a 0.97 effect size for the Brief Pain Inventory interference scale (Revicki & Cook, 2011).

Yost, Eton, Garcia and Cella (2011) recruited 101 cancer patients and administered the PROMIS® fatigue, pain interference, physical function, anxiety, and depression short form scales at baseline and after 6 to 12 weeks to evaluate responsiveness and to determine minimum important differences. All the PROMIS® scales were responsive to changes in

clinical status, and minimal important differences ranged from 2.5 to 6.0 points on the T-score metric for different domain scale scores.

Further Development and Evolution of Existing Item Banks

Item banks may need to be updated and revised over time when there are changes in the context of measurement and in the population. Some of the questions in the item bank may reference everyday activities or items that change over time. For example, some older items may reference technologies that become outdated, such as rotary landline telephones changing to cellular telephones and smartphones. Health domain item banks should be reviewed on a regular basis to ensure that the items include culturally relevant content. Item shift may occur over time; thus, developers need to review their item banks to make sure they are relevant for respondents. There need to be iterations between qualitative development and quantitative evaluation, and continued focus on the evolution of health domain item banks.

Challenges and Future Development of Item Banks

Challenges for Item Bank Development

There are several obstacles and challenges associated with developing health domain item banks, especially for item banks that will be used to assess outcomes in both the general population and chronic disease populations. First, the content of the item bank needs to be relevant across the continuum of the targeted health domain. Therefore, the initial development of the item bank needs to focus on the health experience of the diverse representatives of the general population and selected chronic disease samples that represent the range of impairments in the targeted domain. The main challenge is in identifying these representative patient groups and engaging them in qualitative research.

Second, some health domains are fairly narrow in definition and in some cases only a limited number of unique items can be generated. Health domains with narrow bands are challenging such as depression, anxiety, and anger, and may result in small sets of items that cover the range of the trait continuum. These nearly redundant, narrow item sets may result in problems of local dependence. These narrower constructs also challenge the need for other IRT applications, such as computerized adaptive testing.

Third, there are limitations in the number of DIF analyses that may be feasible given the size of typical psychometric evaluation studies. At a minimum, DIF should be assessed for gender, age groups, race/ethnicity, education level, and language groups. However, a number of other evaluations of DIF might also be relevant to understand the performance and bias in item banks, including comparing different chronic disease groups and other demographic characteristics. Additional DIF analyses can be planned once there are sufficient sample sizes in demographically varied populations. DIF analyses should always be conducted to compare different language translations as they are developed. However, for DIF to be meaningful, it is important to fully understand the dimensionality of an item bank.

Fourth, existing item banks may need to be extended to cover a broader range of the health domain continuum depending on the planned applications for short form scales and CATs based on the item banks. Many of the existing item banks cover unidirectional concepts (i.e., pain behavior, depression, anxiety, fatigue, etc.), and it is difficult to extend these banks further into the "healthy" end of the continuum of the latent construct. For example, the existing PROMIS® item banks are beginning to be administered to military personnel

who may have physical functioning and other domain levels at the highest end of the health domain. Additional items may need to be developed and calibrated to enable discrimination of functioning at these higher levels of functioning. There may be cases where it is necessary to extend the item banks to better cover the most impaired part of the health domain. As with better functioning, additional items may need to be developed and calibrated to cover the more impaired part of the health domain continuum. As research is completed with different chronic disease groups, additional items may need to be added to existing item banks. For example, research with individuals with spinal cord injury and HIV disease may result in the identification of additional concepts in a depression domain that may require new items. As further research is conducted with different chronic disease groups, more items may need to be included in existing item banks to extend relevant domain coverage. These items will then need to be calibrated and included in the item banks.

Finally, as additional items are included in existing item banks, item calibration will need to be conducted and updated. Methods will need to be applied to ensure that the revised item calibrations continue to maintain the original domain item bank metrics. This is important for maintaining continuity with the initial item banks, short form scales, and CAT forms. The score metric is important for linking the domain scores over the years and for continued interpretation of these scores across studies and across time.

Summary

Item banks and computerized adaptive testing based on these item banks represent the future of health outcomes assessment. The flexibility and added measurement precision of item banks and short form scales and CAT forms allows for more efficient assessment of health outcomes. Central to PROMIS® are multiple item banks that include a comprehensive set of questions assessing each health domain. Items can be selected from the PROMIS® item banks to create targeted static short forms for specific patient populations. Such static short forms are not adaptive, and consist of the same items, such as an eight-item depression-specific or pain interference measure. Profile measures, covering multiple health domains, can also be developed from selected static short forms. As an alternative, researchers can use the PROMIS® measures to assess a patient's health status using computerized adaptive testing. CAT tailors the items administered to any specific patient to the actual health status of the patient and potentially represents a more efficient assessment method, often with improved precision.

Given the increase in access to computers and other electronic technologies, health assessments based on item banks may be more feasible in the future. The increasing development of the infrastructure for electronic health records may enable the accommodation of health outcome data from multiple domain item banks. Such clinical applications for measures based on item banks may be useful for monitoring the health status of patients in clinical practice and for improving clinicians' understanding of the effectiveness of health care interventions. These health outcome measures may also be used to examine the quality of the delivery of health care services. The increased emphasis on comparative effectiveness research requires access to large data sets that include information on treatments, clinical outcomes, and patient-reported outcomes (Ahmed et al., 2012). The availability of efficient and standardized multi-domain health assessments as part of the medical record may assist in developing a relevant database for conducting comparative effectiveness research.

The value of health assessments based on item banks will be demonstrated with the increased application of short form scales and CAT forms based on multi-domain item banks. Clearly, these health outcomes assessments can be used to increase the precision

and efficiency of patient-reported outcomes in clinical trials, epidemiologic studies, and clinical practice. The availability of comprehensive item banks covering multiple health domains, as with the PROMIS® project, provides for flexible and efficient assessment of health-related outcomes.

References

Ahmed, S., Berzon, R.A., Revicki, D.A., Lenderking, W.R., Moinpour, C.M., Basch, E., . . . International Society for Quality of Life (2012). The use of patient-reported outcomes (PRO) within comparative effectiveness research: Implications for clinical practice and health care policy. *Medical Care, 50*, 1060–1070.

Amtmann, D., Cook, K.F., Jensen, M.P., Chen, W-H, Choi, S., Revicki, D., . . . Lai, J.S. (2010). Development of a PROMIS item bank to measure pain interference. *Pain, 150*, 173–182.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Bajaj, J.S., Thacker, L.R., Wade, J.B., Sanyal, A.J., Heuman, D.M., Sterling, R.K., . . . Revicki, D.A. (2011). PROMIS computerized adaptive tests are dynamic instruments to measure health-related quality of life in patients with cirrhosis. *Alimentary Pharmacology & Therapeutics, 34*, 1123–1132.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychololological Bulletin, 107*, 238–246.

Bjorner, J.B., Smith, K.J., Orlando, M., Stone, C., Thissen, D., & Sun, X. (2006). *IRTFIT: A macro for item fit and local dependence tests under IRT models*. Lincoln, RI, QualityMetric Incorporated.

Bode, R.K., Lai, J.S., Cella, D., & Heinemann, A.W. (2003). Issues in the development of an item bank. *Archives of Physical Medicine & Rehabilatation, 84*(4 Suppl 2), S52–60.

Bradburn, N.M., Rips, L.J., & Shevell, S.K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236*, 157–161.

Brod, M., Tesler, L.E., & Christensen, T.L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Quality of Life Research, 18*, 1263–1278.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cai, L., Thissen, D., & du Toit, S.H.C. (2011). *IRTPRO for windows*. Lincolnwood, IL: Scientific Software International.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . P.C. Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . P.C. Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*, 1179–1194.

Chang, C.H. (2005). Item response theory and beyond: Advanced in patient-reported outcomes measurement. In W.R. Lenderking & D.A. Revicki (Eds.), *Advancing health outcomes research methods and clinical applications* (pp. 37–55). McLean, VA: International Society for Quality of Life Research.

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Educational and Behavioral Statistics, 22*, 264–289.

Cohen, A.S., Kim, S.H., & Wollack, J.A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychololological Measurement, 20*, 15–26.

Cook, K.F., Kallen, M.A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*, 447–460.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Crane, P.K., Gibbons, L.E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care*, *44*(11 Suppl 3), S115–123.

Crane, P.K., van Belle, G., & Larson, E.B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

De Vet, H.C.W., Terwee, C.B., Mokkink, L.B., & Knol, D.L. (2011). *Measurement in medicine: A practical guide*. Cambridge, UK: Cambridge University Press.

DeWalt, D.A., Rothrock, N., Yount, S., Stone, A.A., & P.C. Group. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, *45*(5 Suppl 1), S12–21.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psycholological Measurement, 7*, 189–199.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. New York: Psychology Press.

Erskine, A., Morley, S., & Pearce, S. (1990). Memory for pain: A review. *Pain, 41*, 255–265.

Fayers, P.M., & Machin, D. (2007). *Quality of life: Assessment, analysis, and interpretation*. Chichester, UK: Wiley.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 28*, 278–295.

Flowers, C.P., Oshima, T.C., & Raju, N.S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309–332.

Flynn, K.E., Lin, L., Cyranowski, J.M., Reeve, B.B., Reese, J.B., Jeffery, D.D., . . . Weinfurt, K.P. (2013). Development of the NIH PROMIS (R) sexual function and satisfaction measures in patients with cancer. *Journal of Sexual Medicine, 10*(Suppl 1), 43–52.

Forrest, C.B., Bevans, K.B., Tucker, C., Riley, A.W., Ravens-Sieberer, U., Gardner, W., & Pajer, K. (2012). Commentary: The patient-reported outcome measurement information system (PROMIS(R)) for children and youth: Application to pediatric psychology. *Journal of Pediatric Psychology, 37*, 614–621.

Fries, J.F., Krishnan, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Research & Therapeutics, 13*, R147.

Gibbons, R.D., Hedeker, D., & Bock, R.D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Gorin, A., & Stone, A.A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. In A. Baum, T.A. Revenson, & J.E. Singer (Eds.), *Handbook of health psychology* (pp. 405–413). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W.R. Lenderking & D.A. Revicki (Eds.), *Advances in health outcomes research methods and clinical applications* (pp. 57–78). McLean, VA: International Society for Quality of Life Research.

Hambleton, R.K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91–115.

Hays, R.D., & Revicki, D.A. (2005). Reliability and validity, including responsiveness. In P. Fayers & R.D. Hays (Eds.), *Assessing quality of life in clinical trials* (pp. 25–39). New York: Oxford University Press.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hu, L.T., & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jöreskog, K.G., Sörbom, D., & Du Toit, S. (2003). *LISREL 8: New statistical features*. Lincolnwood, IL: Scientific Software International.

Kim, S.H., & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345–355.

Kleinman, L., Benjamin, K., Viswanathan, H., Mattera, M.S., Bosserman, L., Blayney, D.W., & Revicki, D.A. (2012). The anemia impact measure (AIM): Development and content validation of a patient-reported outcome measure of anemia symptoms and symptom impacts in cancer patients receiving chemotherapy. *Quality of Life Research, 21*, 1255–1266.

Kline, R.B. (2010). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Lai, J.S., Crane, P.K., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Quality of Life Research, 15*, 1179–1190.

Lasch, K.E., Marquis, P., Vigneux, M., Abetz, L., Arnould, B., Bayliss, M., . . . Rosa, K. (2010). PRO development: Rigorous qualitative research as the crucial foundation. *Quality of Life Research, 19*, 1087–1096.

Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R.D., Brod, M., . . . Cella, D. (2012). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research, 21*, 739–746.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

McDonald, R.P. (1981). The dimensionality of test and items. *British Journal of Mathematical Statistical Psychology, 34*, 100–117.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–118.

Menon, G., & Yorkston, E. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgments. In A. Stone, J. Turkkan, & C. Bachrach (Eds.), *The science of self-report: Implications for research and practice* (pp. 63–79). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, L.K., & Muthén, B.O. (1998). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

National Quality Forum (2013). *Patient-reported outcomes in performance measurement*. Washington, DC: National Quality Forum.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298.

Pilkonis, P. (2011, April). *Pittsburgh PROMIS wave II protocols*. PROMIS Steering Committee Meeting, Bethesda, MD.

Pilkonis, P., Choi, S.W., Reise, S.P., Stover, A.M., Riley, A.W., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcome Measurement Information System (PROMIS): Depression, anxiety, and anger. *Assessment, 18*, 263–283.

Preston, K.S.J., Reise, S.P., Cai, L., & Hays, R. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational & Psychological Measurement, 71*, 523–550.

Raju, N.S., Van Der Linden, W.J., & Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368.

Ramsay, J.O. (1997). A functional approach to modeling test data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 381–394). New York: Springer.

Redelmeier, D.A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3–8.

Reeve, B.B. (2003). Item response theory modeling in health outcomes measurement. *Expert Review Pharmacoeconomics & Outcomes Resesearch, 3*(2), 131–145.

Reeve, B.B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers & R.D. Hays (Eds.), *Assessing quality of life in clinical trials* (pp. 131–145). New York: Oxford University Press.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Med Care, 45*(5 Suppl 1): S22–S31.

Reise, S.P., & Haviland, M.G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*, 228–238.

Reise, S.P., Scheines, R., Widaman, K.F., & Haviland, M.G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational & Psychological Measurement, 73*, 5–26.

Revicki, D.A., Chen, W.H., Harnam, N., Cook, K.F., Amtmann, D., Callahan. L.F., . . . Keefe, F.J. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain, 146*, 158–169.

Revicki, D.A., & Cook, K.F. (2011, May). *Development and psychometric evaluation of the PROMIS pain interference item bank*. Presentation at the 30th Annal Scientific Meeting of the American Pain Society. Austin, TX.

Revicki, D.A., Hays, R.D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemioliology, 61*, 102–109.

Riley, W.T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E.A., & Cella, D. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research, 19*, 1311–1321.

Robinson, M.D., & Clore, G.L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*, 934–960.

Rothrock, N.E., Hays, R.D., Spritzer, K., Yount, S.E., Riley, W., & Cella, D. (2010). Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemioliology, 63*, 1195–1204.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355–371.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* 17.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Schwartz, N., & Sudman, S. (1994). *Autobiographical memory and the validity of retrospective reports*. New York: Springer-Verlag.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81–97.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Strauss, A., & Corbin, J.M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Swaminathan, H., & Rogers, H.J. (1992). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

Teresi, J.A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651–1683.

Teresi, J.A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J.P., Crane, P.K., Jones, R.N., . . . Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the

Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychological Science Quarterly, 51*, 148–180.

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*.

Thissen, D., Chen, W.H., & Bock, D. (2002). *MULTILOG*. Lincolnwood, IL, Scientific Software International.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement, 15*, 22–29.

Weinfurt, K. (2011, October). *Validity of PROMIS measures in patients undergoing heart transplantation for congestive heart failure*. PROMIS Steering Committee Meeting, Bethesda, MD.

Willis, G. (2005). *Cognitive interviewing*. Thousand Oaks, CA: Sage.

World Health Organization (1958). *The first ten years of the World Health Organization*. Geneva, Switzerland: World Health Organization.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yen, W.M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Yost, K.J., Eton, D.T., Garcia, S.F., & Cella, D. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemioliology, 64*, 507–516.

Zumbo, W.M. (1999). *Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.