Published in final edited form as:
Assessment. 2011 September ; 18(3): 263–283. doi:10.1177/1073191111411667.

Item Banks for Measuring Emotional Distress From the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, Anxiety, and Anger

Paul A. Pilkonis[1], **Seung W. Choi**[2], **Steven P. Reise**[3], **Angela M. Stover**[1], **William T. Riley**[4], and **David Cella**[2] **on behalf of the PROMIS Cooperative Group**

[1]University of Pittsburgh Medical Center, Pittsburgh, PA, USA

[2]Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[3]University of California, Los Angeles, Los Angeles, CA, USA

[4]National Heart, Lung, and Blood Institute, Bethesda, MD, USA

Abstract

The authors report on the development and calibration of item banks for depression, anxiety, and anger as part of the Patient-Reported Outcomes Measurement Information System (PROMIS®). Comprehensive literature searches yielded an initial bank of 1,404 items from 305 instruments. After qualitative item analysis (including focus groups and cognitive interviewing), 168 items (56 for each construct) were written in a first person, past tense format with a 7-day time frame and five response options reflecting frequency. The calibration sample included nearly 15,000 respondents. Final banks of 28, 29, and 29 items were calibrated for depression, anxiety, and anger, respectively, using item response theory. Test information curves showed that the PROMIS item banks provided more information than conventional measures in a range of severity from approximately −1 to +3 standard deviations (with higher scores indicating greater distress). Short forms consisting of seven to eight items provided information comparable to legacy measures containing more items.

Keywords

depression; anxiety; anger; item response theory; measurement

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is an NIH Roadmap initiative designed to improve self-reported outcomes using state-of-the-art psychometric methods (for detailed information, see www.nihpromis.org). Adapting the World Health Organization’s (2007) tripartite framework of physical, mental, and social health, PROMIS has developed and calibrated item banks assessing emotional distress, pain, fatigue, sleep disturbance, physical functioning, and social participation (Buysse et al., 2010; Cella et al., 2010; Cella, Yount, et al., 2007; Fries, Cella, Rose, Krishnan, & Bruce, 2009; Revicki et al., 2009). It is the most ambitious attempt to date to apply models from item

Corresponding Author: Paul A. Pilkonis, Western Psychiatric Institute and Clinic, 3811 O’Hara Street, Pittsburgh, PA, USA, pilkonispa@upmc.edu.
Angela Stover, MA, is now affiliated with the Department of Health Behavior and Health Education at the University of North Carolina at Chapel Hill.

Declaration of Conflicting Interests The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

response theory (IRT) to health-related assessment. The PROMIS approach involves iterative steps of comprehensive literature searches; item pooling; development of conceptual frameworks; qualitative assessment of items using expert review, focus groups, and cognitive interviewing; and quantitative evaluation of items using techniques from both classical test theory and IRT (see Cella et al., 2010; Cella, Gershon, Lai, & Choi, 2007; Reeve et al., 2007). We report here on the development and calibration of three item banks capturing the most prominent aspects of emotional distress—depression, anxiety, and anger. We also discuss the conceptual and psychometric challenges that arise when IRT models are applied to constructs assessing psychological symptoms and psychopathology (as contrasted with cognitive skills or aptitudes).

The use of models from IRT to refine measures of psychological symptoms and psychopathology has a 30-year history (Bejar, 1977; D. C. Clark, vonAmmon Cavanaugh, & Gibbons, 1983; de Jong-Gierveld & Kamphuis, 1985; Gibbons, Clark, vonAmmon Cavanaugh, & Davis, 1985; Thissen & Steinberg, 1983), although the conceptual basis for IRT has a longer lineage (see Bock, 1997). The abundance and heterogeneity of measures of emotional distress have made it difficult to determine the comparability of individual scales. Variations in both item content and context (e.g., time frame, response scales, number of response options) further complicate this task. The use of IRT models provides one approach for calibrating different scales on the same metric, but, to date, most applications of IRT to measures of emotional distress have been conducted with single instruments (e.g., Bernstein, Rush, Thomas, Woo, & Trivedi, 2006; Carmody et al., 2006; Gollwitzer, Eid, & Jurgensen, 2005; Pallant & Tennant, 2007). Such analyses are informative about the items from those instruments, but they fail to take full advantage of the potential of IRT-calibrated item banks derived from a comprehensive review of measures assessing the construct of interest (Chang & Reeve, 2005). Item banking and calibration of a large set of items using IRT models can provide a more thorough assessment of a construct, using a common metric that also makes possible the linking of scores between measures used in clinical trials, observational research, and epidemiological studies.

IRT-calibrated item banks underlie the use of computerized adaptive testing (CAT; Kingsbury & Weiss, 1980, 1983; McBride & Martin, 1983; Weiss, 2004) in which the presentation of items is tailored individually to respondents and their levels of the latent construct. The result is an efficient procedure for reducing both the total number of items administered and the measurement error following the administration of each successive item (Gibbons et al., 2008; Lord, 1980; Weiss, 1985, 2004). Simulation studies indicate that CAT using as few as five polytomous items can achieve excellent precision and that scores derived from CAT correlate strongly with the conventional total score from a measure (Bjorner, Chang, Thissen, & Reeve, 2007; Choi, Reise, Pilkonis, Hays, & Cella, 2010; Choi & Swartz, 2009; Gardner et al., 2004; Gardner, Kelleher, & Pajer, 2002). Several studies reporting IRT-calibrated CAT administration of single measures of emotional distress are available (e.g., Fliege et al., 2005; Forbey & Ben-Porath, 2007; Gibbons et al., 2008; Waller & Reise, 1989; Walter et al., 2007). The goal of PROMIS, however, is to move beyond IRT analysis of individual instruments to create item banks that provide a comprehensive profile of health status (including physical, mental, and social health), that are psychometrically sound and that are publicly available on the Internet (Revicki & Sloan, 2007).

Applying IRT models to the measurement of emotional distress involves at least two major challenges: addressing issues of dimensionality and accommodating the asymmetrical nature of the constructs. With regard to the first issue, the conventional wisdom is that traditional tests of ability in the educational literature (e.g., measures of verbal and mathematical proficiency with which the use of IRT has been most common) are more likely to fit unidimensional models than scales of emotional distress (Gibbons et al., 2008). Instruments assessing emotional distress often sample items from multiple domains (e.g., mood, cognition, behavior, somatic symptoms) to capture a comprehensive set of manifest indicators of the latent construct. Therefore, it is common to observe higher correlations within domains than is expected under the conditional independence assumption of a unidimensional IRT model (Bjorner et al., 2007; Steinberg & Thissen, 1996). One of the goals of the current work was to begin with multidimensional conceptual frameworks that were informed by previous empirical (e.g., factor analytic) work and then to shape carefully the most informative unidimensional scales that could be derived for the constructs (Reise, Moore, & Haviland, in press). This process involved both conceptual and psychometric decisions that produced, in an iterative way, the final content of the item banks (a key issue that we discuss below).

The second challenge is the asymmetrical nature of constructs for emotional distress reflected in positively skewed distributions in the general population. In an IRT context, skewed distributions may lessen confidence in parameter estimates (because of the limited representation of high-threshold response choices, even in large samples). Test information functions may become peaked and truncated, with a relatively narrow bandwidth (Reise & Waller, 2009). Because our aim was to create a standard frame of reference for outcomes measurement in diverse settings for clinical research (including clinical trials, observational research, and epidemiological studies), we assembled a sample that was representative of the full range of severity of emotional distress. For this purpose, we relied heavily on an Internet panel, enriched with clinical participants from the PROMIS research sites. The Internet panel provided us with community participants who varied across the full spectrum of health (from no self-reported medical or psychiatric conditions to multiple, comorbid conditions); the clinical samples provided us with bona fide patients who were the most likely to endorse items at the highest levels of severity (see Rothrock et al., 2010).

Most attempts to measure emotional distress begin with moderate to marked indicators of severity, often in the service of identifying people in need of treatment. In the case of depression, these indicators include explicit feelings of sadness, hopelessness, helplessness, and worthlessness recognized as different from one’s usual emotional experience and associated with functional impairment. In this context, the appropriate content (e.g., boredom, wistfulness, nostalgia) and meaning of low-threshold items indicating minimal depression may be problematic. One risk is that the nature of such items may be sufficiently different that they no longer tap the same construct as items of greater severity. For example, items that capture low levels of depression may overlap with transient negative affect, which is universal and may not be as informative specifically about depression. In the current work, we did not alter the conventional content of items assessing emotional distress but rather relied on the use of five-level response options to try to capture low levels of distress —in this case, by endorsements of “never” or “rarely.” Five response choices appear to be a satisfactory number for polytomous items, with the goal of creating items (and scales) that have as large an effective range of measurement as possible (Hawthorne, Mouthaan, Forbes, & Novaco, 2006; Roberson-Nay, Strong, Nay, Beidel, & Turner, 2007).

In summary, this article describes the steps we took to develop comprehensive item pools, to evaluate these items using qualitative methods (focus groups and cognitive interviewing), to administer a reduced number of items to a large calibration sample, and to fit IRT models to the resulting data. The final products were IRT-calibrated banks (suitable for CAT) of 28, 29, and 29 items for depression, anxiety, and anger, respectively, and short forms of seven to eight items that provide information comparable to legacy measures containing more items.

Method

A summary of the PROMIS methodology is provided in Figure 1, and we expand on the major elements from this flowchart in the text.

Development of Item Pool

Comprehensive literature searches—The Pittsburgh research site developed a methodology for performing extensive literature searches in each of the emotional distress domains to ensure content validity. We performed comprehensive literature searches in the MEDLINE, PsycINFO, and Health and Psychosocial Instruments databases. Details of the methodology are reported in Klem et al. (2009), and all search algorithms are available on request. The searches generated 1,204 abstracts for measures of depression (78 scales), 1,738 for measures of anxiety (145 scales), and 1,277 for measures of anger (82 scales). Cited reference searches were run on the primary reference for each measure to determine its acceptance and use by the scientific community. Copies of the measures were gathered from both electronic and print sources, and the measures were then reviewed at the item level.

Conceptual organization of items—The initial emotional distress item pool contained 1,404 items: 518 for depression, 443 for anxiety, and 443 for anger. We organized the items into conceptually meaningful categories using a hierarchical approach informed by previous empirical (e.g., factor analytic) work. Previous work documents that these constructs can be partitioned usefully into subdomains (e.g., mood, cognition, somatic indicators) and factors, and it is also informative about the best manifest indicators (items) for operationalizing such factors. For examples in the area of depression, see Quilty, Zhang, and Bagby (2010), Santor, Gregus, and Welch, (2006), Shafer (2006), and Simms, Gros, Watson, and O'Hara (2008). In the area of anxiety, see Beck, Epstein, Brown, and Steer (1988), Vancleef, Vlaeyen, and Peters (2009), and Zinbarg and Barlow (1996). Studies of the structure of anger include Buss and Perry (1992), Forgays, Spielberger, Ottaway, and Forgays (1998), and Martin, Watson, and Wan (2000).

Our own hierarchical frameworks included domains (depression, anxiety, and anger), subdomains (e.g., mood, cognition, behavior), factors (e.g., within depressed mood, factors included both increased negative affect and decreased positive affect), and facets (e.g., within increased negative affect, facets included sadness, irritability, moodiness, and others). The total number of facets in each hierarchical structure was 46 for depression, 30 for anxiety, and 29 for anger. (Copies of the hierarchies are available on request.)

Qualitative item review—A key step in editing the item bank was qualitative review of the items done by consensus among the members of the Pittsburgh research site (see DeWalt, Rothrock, Yount, & Stone, 2007, for a description of the qualitative procedures used by the PROMIS network). This process involved elimination of redundant items, items that were too narrow (often by virtue of being disease specific), items that were confusing or vague, and items that were poorly written (e.g., double-barreled items). This review reduced the item pool to 457 items: 151 for depression, 170 for anxiety, and 136 for anger.

Focus groups with patients—An important component of content validity is the extent to which items reflect the perspective of the population of interest (Brod, Tesler, & Christensen, 2009; Food and Drug Administration, 2009; Mokkink et al., 2009). Eighty psychiatric and medical outpatients participated in 10 focus groups conducted at two sites (University of Pittsburgh and Duke University) to reveal potential gaps in content in the assessment of depression (*n* = 4 groups), anxiety (*n* = 3 groups), and anger (*n* = 3 groups).

Semistructured scripts were written for the focus groups, and moderators used these scripts to elicit group participation and discussion of specific topics. Open-ended questions focused on the general experience of emotional distress included within each domain, the ways in which this distress was described to others, recent experiences of symptoms, and the effect of symptoms on functioning (e.g., day-to-day activities, interpersonal relationships). (Copies of the scripts used for the focus groups are available on request.) Discussions in the groups were transcribed and coded using Atlas.ti software (Muhr & Freise, 2004).

Kelly et al. (2011) assessed the degree of overlap between our conceptual framework for depression and experiences that patients discussed spontaneously in the four depression focus groups. Participants described 93% of the aspects of depression included in our a priori framework (43 of 46 facets). The large majority of the comments not accounted for by our hierarchical structure for depression fell into the domains of anger, anxiety, or substance abuse, all areas in which we planned to develop separate item banks. Notably, anger was reported more often than any of the symptoms included in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV-TR*; American Psychiatric Association, 2000) diagnosis of major depressive disorder, with the exception of sadness. “Irritability” was a facet of the conceptual structure for depression, but irritability alone failed to capture the intensity of the anger that some participants described. The important summary point is that results from the focus groups did not, in general, require the creation of new items for emotional distress and supported the content validity of the conceptual frameworks and item pools assembled originally. Whenever possible, however, key words used by patients were adopted in the process of editing and rewriting items.

Standardization of items—Final construction of items involved careful consideration of time frames, response sets, verb tense, grammatical structure, and demands on literacy (see DeWalt et al., 2007). We examined precedents for alternative time frames, response sets, and number of response options among the questionnaires in the instrument library. The most common time frame for instruments assessing emotional distress was 7 days (33%), and the most common response sets were severity (52%), frequency (22%), or the presence versus absence of symptoms—for example, “yes/no” or “true/false” (19%). A total of 63% of scales used four or five response options. Based on these data and prior experience with the use of IRT models in assessment of health-related constructs (Bode, Lai, Cella, & Heinemann, 2003), four to six response levels appeared to be optimal. In the area of emotional distress, we adopted a 7-day time frame and a 5-point scale for frequency (*never, rarely, sometimes, often, always*). The 7-day time frame was consistent both with precedents from the research literature and with decisions made for other PROMIS domains, where sensitivity to change in the context of potentially brief clinical trials was a consideration. There was no definitive guidance available from the research literature about the choice between response options reflecting severity or intensity versus those for frequency or duration. There were suggestions, however, that frequency scaling may provide broader coverage (reducing floor and ceiling efforts) for health-related assessment (Chang et al., 2003) and that frequency scaling is more appropriate for short intervals, given the usual “conversational” inferences of respondents, who assume that short reference periods pertain to frequent experiences and that long periods pertain to rare and intense experiences (Winkielman, Knäuper, & Schwarz, 1998).

Based on these considerations, the 457 items in the reduced pool were rewritten in a common format: first person singular, past tense, with a 7-day time frame and a 5-level ordinal scale of frequency for response options (In the past 7 days, I felt depressed: *never, rarely, sometimes, often, always*). The majority of self-report measures of depression and anxiety are written at a reading-grade level that exceeds the mean proficiency in the United States (McHugh & Behar, 2009). We strived to reduce the literacy demand of our items by minimizing the number of words per sentence and choosing simpler rather than more demanding synonyms. More than 50% of items contained five or fewer words, and more than 20% were even simpler, three-word sentences (e.g., I felt sad). The Lexile Framework for Reading (MetaMetrics, 2008), a method for measuring the literacy demands of text, confirmed that the items were easy to read. Lexile text analyses documented that the items required an average first-grade reading level, with a standard deviation of 1.5 grades (Lexile *M* = 180.2, *SD* = 263.7).

Cognitive interviews—Cognitive interviews (based on the recommendations of Willis, 2005) were conducted with 41 participants using 238 items selected for potential inclusion in the item banks. The 238 items were chosen for cognitive interviews on the basis of consensus judgments by the Pittsburgh research team, focused primarily on eliminating redundancy among the items; in their simpler, rewritten form, many items were very similar or identical. A minimum of 5 participants reviewed each of the 238 items selected, and the demographic characteristics of the entire group are described in the appendix. Scripts for the cognitive interviews began with questions about the meaning and clarity of time frames, response options, and item content. Participants were then asked to respond to the actual items that they were reviewing and to reflect on how they understood the item and chose an answer. (Copies of the scripts used for cognitive interviewing are available on request.) When 2 or more participants expressed a concern about an item, it was rewritten and subjected to an additional round of cognitive interviews.

Three rounds of interviews were needed to clarify all items. The first round of interviews revealed 23 items (10%) requiring revision. Three items were split into two new items each to clarify a concept (e.g., distinguishing between mental vs. physical health), 2 items contained examples that led to differences in interpretation, and 18 items were rewritten for clarity. For example, several items were revised when a double negative (a construction that many participants found difficult to understand) occurred between the item stem and response options—for example, “I could not control my temper,” with the response option, “never.” The item was rewritten to read, “I had trouble controlling my temper.” The second round of cognitive interviews included a single block of 23 items arranged into one of six orders randomized for presentation. After the second round of reviews, 3 items still required revision. The third round of interviews was conducted by telephone, and participants judged the items to be satisfactory as rewritten.

Intellectual property issues—A careful review of intellectual property issues was done for all items (see Berzon, Patrick, Guyatt, & Conley, 1994; Revicki & Schwartz, 2009; Ware, 2003). First, we documented the lineage of the items. The large majority of items had been used in multiple scales, most often because they reflected generic aspects of emotional distress and the everyday language in which it is described: sadness, guilt, hopelessness, and helplessness, for example, in the case of depression. Such items appeared in both proprietary and nonproprietary scales (usually without attribution to prior sources). We regarded them as part of the public domain because they reflected common-sense ideas about emotional distress, regardless of the exact format in which they were operationalized in specific scales. Such items were adapted and rewritten using the PROMIS conventions.

Some item stems, however, contained content that was more distinctive. Permission to use such items was sought from the developers or copyright holders, and they were incorporated into the preliminary PROMIS item banks only if such permission was granted. For example, the depression bank includes an item for loneliness (“In the past 7 days, I felt lonely”) but not a more qualified item (“… feeling lonely even when you are with people”) that appears on the Symptom Checklist-90–Revised (Derogatis, 1983) and remains proprietary.

Sampling

A decision was made across the PROMIS network to administer approximately 150 items to each respondent in the calibration sample, anticipating a testing session of about 30 minutes at a response rate of 5 to 6 items per minute. Approximately 50 items were devoted to questions about demographic and social characteristics and medical history, including 10 questions asking about global perceptions of physical, mental, and social health (Hays, Bjorner, Revicki, Spritzer, & Cella, 2009). Because of the multiple domains being tested, final item banks for calibration testing were limited to 56 items in each PROMIS domain. We attempted to choose the final 56 items each for depression, anxiety, and anger on the basis of content balancing (i.e., having a representative group of symptoms and complaints associated with each of these constructs) and balancing with regard to likelihood of endorsement (i.e., having items that represented a lower-to-moderate range of severity as well as items reflecting higher severity).

Inclusive testing format and sample—Initial testing was done in two forms: full-bank testing, in which participants received items soliciting background information, ratings of global health, and two full item banks (112 items) from the PROMIS domains; and block testing, in which participants received the background information items, ratings of global health, and 14 blocks of 7 items each across all the PROMIS domains (98 items). Full-bank testing allowed investigation of the dimensionality within each item bank, and block testing allowed investigation of the convergent and discriminant relationships across the item banks (and PROMIS domains). Participants receiving the full-bank format also completed one or two “legacy” instruments (nonproprietary measures used widely in that domain) to compare results from the PROMIS item banks with traditional benchmarks. For depression, the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) was used as the legacy measure. For anxiety, the general distress (anxiety) scale from the Mood and Anxiety Symptom Questionnaire (MASQ; Watson et al., 1995a, 1995b) was used as the legacy measure. For anger, the combined score from the anger and verbal aggression subscales of the Aggression Questionnaire (AQ; Buss & Perry, 1992) served as the legacy measure.

The large majority of the inclusive testing sample was drawn from an Internet polling company, Polimetrix. Polimetrix (now YouGov Polimetrix, www.polimetrix.com) is a national, web-based polling firm. More than 1 million adult panel members have provided e-mail addresses, contact information, and responses to core profile items in order to receive occasional surveys about a variety of subjects (Polimetrix, 2006). The full-bank testing format was administered to 7,005 participants in the general population (6,676 from Polimetrix and 329 from PROMIS research sites); 1,974 of these participants received the depression, anxiety, or anger banks. The block-testing format was administered to 14,128 persons (6,245 general population and 7,883 clinical participants; 12,925 from Polimetrix and 1,203 from research sites). Because the PROMIS mandate is a broad one—that is, developing item banks that provide a common metric relevant to healthy samples as well as those suffering from medical and psychiatric disorders—the testing sample was selected to include diverse health conditions and to reflect the full range of severity of emotional distress. The Internet panel provided us with community participants who varied across the complete spectrum of health (from no self-reported medical or psychiatric conditions to multiple, comorbid conditions); the clinical participants from the research sites provided us with bona fide patients who were expected to endorse items at the higher levels of severity.

The subsamples of participants with health-related conditions included persons with heart disease (n = 1,156), cancer (n = 1,754), rheumatoid arthritis (n = 557), osteoarthritis (n = 918), psychiatric illness (n = 1,193), chronic obstructive pulmonary disorder (n = 1,214), spinal cord injury (n = 531), and other conditions (n = 560). Participants belonging to these clinical groups were identified on the basis of responses to items assessing chronic health conditions ("Have you ever been told by a doctor or a health professional that you have <condition>?") and their impact on daily functioning ("Are any of your current activities limited by <condition>?"). Further details about the sampling plan are available in Liu et al. (2010) and at www.nihpromis.org/Web%20Pages/PSYCHO%20Metricians.aspx.

The entire testing sample included 21,133 participants (see Table 1 for a summary of their demographic characteristics). Respondents (*n* = 73 in the inclusive sample) flagged by predetermined speed-of-response criteria were excluded from psychometric analyses (average response time per item less than 1 second or 10 consecutive items with a response time less than half a second). Respondents (*n* = 69) with an excessive number of missing responses were also excluded (i.e., 29 or more missing responses for full-bank data and 2 or more missing responses for block-testing data). The IRT parameters reported here were based on a calibration subsample. Whenever possible, the emotional distress items were analyzed using both full-bank and block-testing data. However, analyses of dimensionality, IRT model fit, and differential item functioning (DIF) were conducted using full-bank data only. Table 1 summarizes the size and demographic characteristics of the calibration samples for the item banks.

Scale-setting subsample—A scale-setting subsample (*N* = 5,239) representative of the U.S. population was used to create a *T*-score metric—*M* = 50 and *SD* = 10—for all PROMIS domain (IRT theta) scores (see Table 1). The national distribution of gender, age, race, and education in the 2000 census was 52% female; 22% aged 18 to 29 years, 32% aged 30 to 44 years, 24% aged 45 to 59 years, 14% aged 60 to 74 years, and 8% aged 75 years and older; 74% White, 11% African American, 11% Hispanic, and 4% Other; and 51% more than a high school education. The scale-setting subsample was created with the purpose of representing the marginal distributions of race/ethnicity (White vs. a combined group of African American, Hispanic, and other respondents) and education (high school or less vs. more than high school) from the 2000 census.

Results

Item Selection

A subset of the best performing items was identified for each domain from the pool of 56 items used for calibration testing. These items were chosen using multiple criteria: the results of classical item analysis (inspection of frequency distributions of individual items and adjusted item–total correlations), tests of monotonicity and scalability, and examinations of dimensionality, including exploratory factor analyses (EFAs) and multidimensional scaling (MDS). The primary goal was to achieve sufficient unidimensionality for confirmatory factor analyses (CFAs) and IRT analyses in which the credibility of model parameters relies on the assumption of unidimensionality. Efforts were made to ensure that each bank was suitable for unidimensional scaling without unduly narrowing the construct. Analyses were done iteratively with alternative subsets of items, and, in many cases, items were eliminated on the basis of multiple considerations. We describe here the decisions resulting from the use of these criteria.

Frequency distributions—Summed scores of the 56 items for depression ranged from 56 (the lowest possible score, representing answers of "never" to all items) to 276 (an average score of 4.9 per item or an answer of "always" to almost all items, $\alpha = .98$), 56 to 277 (an average score of 4.9 per item or an answer of "always" to almost all items, $\alpha = .98$) for anxiety, and 59 to 202 (an average score of 3.6 per item or answers of either "sometimes" or "often" to almost all items, $\alpha = .97$) for anger. The summed score distributions were positively skewed for all three banks (depression, skew = 1.23; anxiety, skew = 1.28; and anger, skew = 1.02). Given the positive skew, some sparse cells were found at higher levels of endorsement, most often because of infrequent endorsement of the most severe response option, “always.” Nine depression items were eliminated on this basis, and they were items primarily reflecting suicidality, marked forms of cognitive impairment (difficulties in memory and concentration), and changes in eating. Six anxiety items were eliminated on this basis, and they were items primarily reflecting pronounced physical symptoms (e.g., chest pain, nausea). Sixteen anger items were eliminated on this basis, and they were items primarily reflecting marked forms of verbal and physical aggression.

Adjusted item–total correlations—The mean adjusted (i.e., corrected for overlap) item–total correlation was .72 for depression, .65 for anxiety, and .56 for anger. Items with low item–total correlations (< 0.40) were eliminated—2 items for depression, 5 for anxiety, and 12 for anger, and in almost all cases, these items overlapped with the groups of items with sparse cells.

Monotonicity—The probability of endorsing more severe item responses should increase monotonically as the level of the underlying construct increases if IRT models are to provide a good fit to the data. Monotonicity implies that, apart from sampling error, the proportion of respondents endorsing each successive threshold on the response scale is larger for those with a higher latent trait score. The monotonicity of the items was evaluated using a nonparametric approach (Molenaar & Sijtsma, 2000; Ramsay, 2000) by examining graphs of item mean scores conditional on rest scores (i.e., total raw score minus the item score). For each item, rest scores were calculated for all respondents and transformed into logit scores—that is, the natural logarithm of the rest scores divided by their complements for the maximum possible rest score to create linear rest scores. Ten equal-sized groups were then formed based on the logits. Finally, the median item scores for the ten groups were plotted against their mean logit scores. We examined the rest-score function for each item to ensure that the conditional item means increased monotonically as rest scores increased. Only one item was not monotonic—an item from the anger bank (“It was difficult to let people know I was angry”), and it was eliminated.

Scalability—Loevinger *H* coefficients (Loevinger, 1948) were calculated to examine the scalability of the items (see also Meijer & Baneke, 2004; Sijtsma, Meijer, & van der Ark, 2011). The coefficient is computed as a function of the Guttman errors between pairs of items. A Loevinger *H* coefficient of less than 0.30 is considered unsatisfactory. None of the depression items had an *H* coefficient below 0.3 (*M* = 0.59, *SD* = 0.08). The anxiety bank (*M* = 0.49, *SD* = 0.07) had only one item with an *H* coefficient below 0.3, whereas six items in the anger bank had *H* coefficients less than 0.3 (*M* = 0.41, *SD* = 0.15). The seven items with low *H* coefficients were considered candidates for elimination.

Dimensionality—Exploratory factor analysis was used iteratively to examine the evolving pools of items, and items with modest loadings (< 0.40) for single-factor solutions were eliminated. In addition to EFA, nonparametric MDS solutions were obtained for each of the three domains using polychoric correlations among items as measures of proximity. Two-dimensional plots of MDS solutions provided a useful graphical tool for examining the relationships among items (Roth & Roychoudhury, 1991). A high positive correlation was represented by close proximity in two-dimensional space, and unidimensionality was represented by an elliptical cluster of points (items) arranged along a single line. For depression, somatic items related to weight and eating behavior were outliers. For anxiety, several items reflecting physiological arousal and somatic symptoms and concerns were scattered in the fringes of the MDS space. For anger, behavioral items representing physical aggression and violent acts were outliers. In total, 14 items for depression, 19 for anxiety, and 11 for anger were "suspect" for reasons of dimensionality based on the results of EFA and MDS.

Summary—For each domain, approximately 40% of the items in the original 56-item banks were eliminated based on the iterative application of these decision rules and attention to both statistical and content considerations: 24 items for depression, 20 items for anxiety, and 22 items for anger.

Confirmatory Factor Analysis

A single-factor confirmatory model was fit within each domain using Mplus 4.21 (Muthén & Muthén, 2006) to document unidimensionality. The CFAs were performed on the remaining 32 items for depression, 36 for anxiety, and 34 for anger. The items were treated as categorical variables. Because of planned and systematic missing values present in the block-testing data, only full-bank data were included for CFA. Listwise deletion was used for missing data, with less than 4% removed from each data set. The robust weighted least squares (WLSMV) estimator was used. The fit indices reflected an adequate model fit in all three domains: depression (comparative fit index [CFI] = .929, Tucker–Lewis index [TLI] = .995, and root mean square error of approximation [RMSEA] = .086), anxiety (CFI = .901, TLI = .992, and RMSEA = .082), and anger (CFI = .920, TLI = .989, and RMSEA = .070). Error variances and residual correlations were examined to identify areas of strain (using .20 as a threshold for identifying problematic residual correlations). The largest such correlation (.20) was found between two items in the anxiety set ("I had trouble paying attention" and "I had trouble concentrating"). With the second item removed, the fit indices improved slightly (CFI = .924, TLI = .994, and RMSEA = .072).

IRT Calibration

Items remaining in the pool for each domain were calibrated with the graded response model (GRM) using MULTILOG 7.03 (Thissen, Chen, & Bock, 2003). The convergence criterion for the EM cycles was set to .0001, with the number of cycles set to 100. IRT model fit was examined for each item using the IRTFIT macro program (Bjorner et al., 2006) and the option for the sum-score-based method (Orlando & Thissen, 2003), which uses the sum score instead of theta for computing the predicted and observed frequencies. We examined item misfit using the S-X^2 and S-G^2 statistics (Orlando & Thissen, 2003). Only one item, "I lost my temper easily but got over it quickly" from the anger set, showed misfit ($p < .05$). The test information functions revealed that the measures provided adequate measurement precision (information > 20, or $SE < .22$, corresponding to a classical reliability of .95) between theta scores of −1.1 and +3.1 for depression, −0.8 and +3.4 for anxiety, and −0.8 and +4.0 for anger.

Differential item functioning—DIF occurs when characteristics such as age, gender, or ethnicity, which may seem extraneous to the assessment of cognitive and psychological functioning, actually do have an effect on measurement. An item is identified as functioning differentially if the item is more (or less) difficult to endorse or more (or less) discriminating in some focal group (compared with a reference group) when the different subgroups have been matched on the latent trait under investigation. Demographic and health-related variables, for example, have been found to affect response patterns in depression scales such as the CES-D, the Beck Depression Inventory, and the Geriatric Depression Scale (Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Mui, Burnette, & Chen, 2001; Pedersen, Pallay, & Rudolph, 2002).

We conducted DIF analyses (for both uniform and non-uniform DIF) with three grouping variables: gender, age (younger than 65 years vs. 65 years or older), and education (high school graduate or less vs. education beyond high school). Two different DIF procedures were employed—the IRT likelihood ratio method (Thissen, Steinberg, & Wainer, 1993) and an ordinal logistic regression procedure (Zumbo, 1999)—and items were removed if they showed significant DIF ($p < .01$) by both methods (see also Teresi et al., 2009). Seven items across the three banks exhibited DIF, two for depression, two for anxiety, and three for anger, with age having the greatest influence.

Of the two depression items showing DIF, one item (“I felt like crying”) was endorsed more readily by women than men, even with comparable levels of depression. The second depression item showed age-related DIF (“I had trouble enjoying the things I used to enjoy”), with older respondents more likely to endorse it. Two anxiety items also showed age-related DIF (“I felt I was going crazy” and “I felt shaky”). Younger participants were more likely to endorse the former item, with the opposite pattern occurring for the latter item. Three anger items exhibited age-related DIF (“I was angry when people were unfair,” “I was angry when I did something stupid,” and “I was mad”). Older participants were more likely to endorse the first two items, whereas younger respondents were more likely to endorse the last item. These items displaying DIF were not included in the final calibration pool.

Final Calibrated Items

As described above, one item was removed from the anxiety bank for a large residual correlation following CFA, one item was removed from the anger bank for IRT model misfit, and seven items were removed across the three banks for DIF. Following these decisions, a final review was conducted of the content of the item banks and of the intellectual property status of surviving items. For content reasons, two items were eliminated because of potential overlap with the PROMIS fatigue item bank, one item for depression (“I felt that I had no energy”), and one item for anxiety (“I felt exhausted”). Five items were eliminated to ensure that no intellectual property issues remained. These were items that might still be regarded as too similar to items from proprietary measures (one depression item, three anxiety items, and one anger item).

Thus, final banks of 28, 29, and 29 items were calibrated for depression, anxiety, and anger, respectively, using the GRM. See Tables 2 to 4 for listings of the calibrated items and their corresponding IRT (discrimination and threshold) parameters. Many of the discrimination parameters are large by usual standards, but this did not occur by virtue of conditional dependencies—all residual correlations were less than .20. The magnitude of the discrimination parameters was probably more influenced by the skewness of the underlying distributions, an issue we discuss further below. The calibrated items are available on an Internet testing platform, the PROMIS Assessment Center (www.assessmentcenter.net), with options for administering CAT, the full item banks, or static short forms. Of the 28 depression items, 17 are cognitive, 9 are affective, 1 is behavioral, and 1 reflects passive suicidal ideation. Of the 29 anxiety items, 11 are affective, 8 are cognitive, 8 are somatic, 1 is behavioral, and 1 reflects perceived need for treatment. Of the 29 anger items, 13 are cognitive, 8 are affective, 7 are behavioral, and 1 reflects perceived need for treatment.

Development of Short Forms

Short forms were developed consisting of eight, seven, and eight items for depression, anxiety, and anger, respectively, based on simulations of CAT results, item information, and item content. For this purpose, three primary indicators were computed: the percentage of “respondents” in CAT simulations who would have been presented with a particular item, the expected item information for a standard normal distribution, and the expected item information for a distribution with a mean of 0 and standard deviation of 1.5. Potential items were identified based on their ranks according to these three criteria. Content experts reviewed the potential short form items and made final selections based on these criteria, together with the content of the items. The internal consistency of the short forms was excellent. The mean adjusted item–total correlation for the depression short form was .83; for anxiety, .79; and for anger, .69. Alpha coefficients were .95 for depression, .93 for anxiety, and .90 for anger. We fit one-factor CFA models for each of the short forms, and the fit indices supported their coherence: for depression, CFI = .99, TLI = .99, RMSEA = .11; for anxiety, CFI = .99, TLI = .99, RMSEA = .05; and for anger, CFI = .99, TLI = .98, RMSEA = .08. Correlations between the theta scores derived from the short forms and their corresponding full item banks were very high: .96 for each bank. The same correlations using raw scores were .98 for depression, .96 for anxiety, and .95 for anger.

To examine the full item banks, static short forms, and corresponding legacy measures on the same scale, we calibrated each legacy measure concurrently with the corresponding full bank by fixing the PROMIS item parameters at their bank values. Figures 2 to 4 display the test information curves for the full item banks, the short forms, and the relevant legacy measures. It is clear that the full item banks provide the most information (and greatest precision), but the figures also illustrate that the short forms perform as well as the legacy measures with fewer items. The comparative efficiency of the short forms and the legacy measures should be interpreted with caution, however, because the legacy measure was “projected” onto the PROMIS item bank and, in a strict sense, the two measures may not be assessing the same construct. We have also examined the efficiency of alternative CAT algorithms in comparison with the depression short form and full item bank, demonstrating that CAT provides further improvements over static short forms (Choi et al., 2010).

Validity Evidence

The item response data for the three emotional distress item banks were scored based on the GRM item parameter estimates using the expected a posteriori estimator. The expected a posteriori theta estimates were correlated with a legacy instrument designed to measure a similar construct (convergent measure) and with another measure of a different construct (divergent measure). For depression, the CES-D was used as the convergent measure (r = .83) and the general distress (anxiety) scale from the MASQ was used as the divergent measure (r = .72). Conversely, for anxiety, the general distress (anxiety) scale from the MASQ was used as the convergent measure (r = .80) and the CES-D was used as the divergent measure (r = .75). The depression and anxiety theta scores were themselves highly correlated, r = .81. For anger, the combined score from the anger and verbal aggression subscales of the AQ served as the convergent measure (r = .51) and the global health items from the PROMIS assessment battery (r = .40) served as the divergent measure. The anger theta score correlated .60 with the depression theta score and .59 with the anxiety theta score. Each item was also correlated with its corresponding legacy measure, and the median correlation between the depression items and the CES-D was .72; between the anxiety items and the MASQ general distress (anxiety) subscale, .65; and between the anger items and the AQ subscales for anger and verbal aggression, .37. The large correlations among all measures of depression and anxiety raise the issue of whether these constructs should be modeled as a single, higher order variable. On the other hand, the correlations between the anger items, the legacy measure (AQ), and the measures of depression and anxiety were lower, indicating that the anger items are measuring a somewhat distinct construct.

During the item pooling and standardization process, the content validity of our item banks was established in several ways—through comprehensive literature reviews, patient feedback, and expert consensus. Together with subsequent psychometric analyses, the process resulted in considerable pruning of the original item pool and changes in the relative representation of content from different domains (affective, cognitive, behavioral, and somatic). To provide a further assessment of content validity at the end of the process, we invited nine content experts (not affiliated with the PROMIS network) to review the items in the three emotional distress item banks (blind to the domain names) and to offer a name and brief definition describing what each bank measured. This review provided further confirmation of content validity. In all cases, the outside reviewers provided names and definitions that were matches or close derivatives of the original labels (depression, anxiety, and anger), and their mean ratings of the fit between our domain names and definitions and the final content of the banks ranged from 4.3 to 4.9, using a 5-point scale where 4 = *quite a bit* and 5 = *very much* (see Riley et al., 2010, for details).

Discussion

Our goal was to develop and calibrate item banks for depression, anxiety, and anger as part of the PROMIS. This NIH Roadmap initiative is the most ambitious attempt to date to apply IRT models to the assessment of health status. The PROMIS products are IRT-calibrated item banks (Buysse et al., 2010; Cella et al., 2010; Fries et al., 2009; Revicki et al., 2009), algorithms and an electronic testing platform for computerized adaptive tests (Gershon et al., 2010), and static short forms (for more information, see www.nihpromis.org and www.assessmentcenter.net). The item banks provide a dimensional assessment of emotional distress, applicable across a wide variety of health conditions, and they are not intended currently to be screening or diagnostic tools (although they can now be tested for this purpose). Scores across all the PROMIS domains with current item banks—physical functioning, pain, fatigue, emotional distress, sleep disturbance, and social participation—can be derived with just a few items (in most cases, four to six items using CAT), making them practical to implement in both clinical and epidemiological contexts (Choi et al., 2010). The exhaustive PROMIS process for developing and calibrating item banks has taught us several conceptual and practical lessons, which we discuss here.

Dimensionality of the Constructs

Although recognizing the assumption of unidimensionality required for the use of IRT models, we began with a highly articulated conceptual framework for the three constructs of emotional distress that included 46 facets for depression, 30 for anxiety, and 29 for anger. These facets were organized hierarchically into factors and subdomains (e.g., mood, cognition, behavior, somatic indicators). We found that unidimensional models could be fit satisfactorily with 25 to 30 items for each construct. Certain trade-offs were required, however, to achieve such fit. Thus, more than 90% of the depression items are cognitive and affective indicators. The proportion of affective items remained the same across the progression from the original item pool for depression (34%) to the final item bank (32%), but the proportion of cognitive items increased markedly (34% to 61%). At the same time, behavioral and somatic items were removed, primarily on the basis of psychometric analyses. This result is consistent with other IRT analyses of measures of depression (see Kendel et al., 2010, for a recent example) in which somatic markers often fit poorly, and it is compatible with our emphasis on self-reported outcomes and the internal psychological experiences that they are typically intended to capture. With depression, it can also be argued that the exclusion of many somatic features makes the bank more useful for assessing mood in chronic medical conditions where physical symptoms often confound the measurement of depression.

With anxiety, there is a larger representation of somatic items (28%) in the final item bank, consistent with conceptual discussions of anxiety as distinct from depression by virtue of greater levels of physiological arousal (L. A. Clark & Watson, 1991). At the same time, this percentage is lower than that in the original item pool (44%). Also, similar to the development of the depression item bank, the proportion of affective and cognitive items for anxiety became larger across the item refinement and calibration process. With anger, the most extreme markers of behavioral aggression (physical and verbal) were eliminated from the item bank based on psychometric results (with behavioral items decreasing from 55% of the original pool to 24% of the final item bank). Nonetheless, our final measure of anger includes several indicators of behavioral activation and expression, consistent with a view that anger is associated with stronger action tendencies than depression or anxiety (Cox & Harrison, 2008; Martin, Wan, David, Wegner, Olson & Watson, 2000). The percentage of affective items remained about the same across the development process for anger, but, consistent with both depression and anxiety, the percentage of cognitive items increased (20% to 45%). The important general point is that the proportion of affective and (particularly) cognitive indicators increased systematically as a consequence of our content and psychometric decisions and that our final item banks, as a result, are primarily measures of internalized psychological distress.

Our results regarding unidimensionality are also consistent with a meta-analysis of the factor structures of four popular measures of depression (Shafer, 2006). This analysis documented a hierarchical structure, with one second-order factor of general depression (consistent with an argument for unidimensionality) and several first-order, correlated factors. Some (Gibbons et al., 2007; Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007) have argued that a bifactor model may be warranted in such circumstances (although the bifactor model assumes that its group factors are uncorrelated). The broader issue, however, is that identifying the “true” latent trait underlying heterogeneous indicators is a complex task both conceptually and empirically. The assumption of unidimensionality in IRT models can create a demand in which operationalization of a latent trait may be heavily influenced by the presence of a single predominant factor to develop as homogenous a construct as possible. A bank containing items that are essentially the same single indicator (using different adjectives) may lead to the latent trait being defined by a highly intercorrelated subset of items, with strikingly large discrimination parameters that make other indicators look poorer by contrast. Careful thought must be given to the trade-off between homogeneity and the potential utility of a somewhat diverse pool of indicators. In this context, we believe that our PROMIS item banks for emotional distress do a good job of capturing the internal indicators of depression, anxiety, and anger for which self-reports are primarily intended.

Constructs of Emotional Distress as “Quasi Traits”

The frequency distributions for both individual items and the total scores from the item banks were positively skewed in the calibration sample despite our attempt to enrich the sample with identified psychiatric patients and respondents from the Internet panel who reported that they suffered from one or more chronic health conditions. Thus, the test information functions were displaced to the right, with more information and precision provided in the moderate to marked range of severity (about −1 to +3 standard deviations). Therefore, the PROMIS emotional distress scales may be most valuable for outcome measurement in samples and settings where emotional distress is likely to be present in participants.

Positively skewed distributions motivate a search for low-threshold items intended to add precision closer to the floor, but it is unclear if such items measure the same construct as higher levels of distress. In other terms, there may be a mixture of two distributions: persons with no (state) distress and persons with some to marked distress, who can report changes in characteristic phenomenology or functioning associated with a group of symptoms (or, in some cases, a syndromal diagnosis of clinical depression). Such a mixture model suggests that constructs of emotional distress may function more like “quasi traits” than normally distributed traits (Reise & Waller, 2009). Endorsement of mild depressive symptoms (e.g., boredom, nostalgia) may represent transient mood states or (perhaps) temperamental variables that are distinct from the more notable symptoms of depression typically captured in assessment instruments (including the PROMIS item banks).

From a psychometric perspective, positively skewed distributions can have important implications. They can lead to “artificially” large discrimination parameters, peaked (but narrow) item information functions, and narrow bandwidth for an item bank as a whole. We believe that the PROMIS item banks represent a reasonable compromise in this regard, given the use of polytomous items that provide adequate precision to about −1 standard deviation in severity. The creation of these item banks, however, has increased our awareness of the potential problems of adequate measurement near the floor (low distress) and the need for a careful analysis of what is being measured there. Such analysis also has implications for articulating as clearly as possible the purpose of a measure, its generalizability, and the appropriate calibration samples.

It should also be noted that algorithms and software (RCLOG V2; Woods, 2006) do exist that attempt to estimate simultaneously the shape of the latent distribution and the item parameters. This approach, however, has not become standard practice in the field, and no IRT software estimation package that uses the Bock–Aiken method readily allows a specification of a nonnormal distribution for the latent trait. Also, there is some evidence from simulation studies that the effect of the departure from normality on parameter estimation may be negligible (Stone, 1992). Nonetheless, we did estimate item parameters using RCLOG to explore this issue further. Judging the output is complicated because the program produces multiple solutions, requiring a decision about which model is “best.” This decision is challenging because, on one hand, alternative models often fit similarly, and, on the other, different fit indices may suggest different results. For the depression item bank, the “best” nonnormal model was one with five breaks of order 2 (relative to the normal model with two breaks of order 2). In this model, skewness was estimated to be 0.20, with kurtosis of 3.11. Although the models were significantly different, the item parameters were not substantively different. To document this fact, we plotted the test response curves (i.e., plots of the expected raw scores based on the latent trait score) using the item parameters from both estimation procedures (assuming a normal distribution vs. the “best” nonnormal distribution), and the two curves were virtually identical. In any case, data regarding emotional distress (and health status more generally), which may not be normally distributed in general population samples, pose interesting challenges, and this problem is one that will benefit from more attention moving forward.

Correlations Among the Constructs

Despite successful use of unidimensional models, it is important to note that the IRT theta scores derived from the PROMIS depression and anxiety item banks were highly correlated (r = .81). Given the high rates of comorbidity and the overlap among the symptoms of depression and anxiety, a number of conceptual models have been proposed to account for the shared versus unique variance captured in measures of these constructs. In their tripartite model, Watson and L. A. Clark proposed a hierarchical structure to explain the relationships between symptoms of depression and anxiety (L. A. Clark & Watson, 1991; Watson et al., 1995a, 1995b). They described a second-order, nonspecific factor reflecting high levels of negative affect—or “general distress”—common to both depression and anxiety. This factor is similar to the internalizing spectrum described by Krueger and colleagues (Krueger, 1999; Krueger & Finger, 2001; Krueger, McGue, & Iacono, 2001), and it includes symptoms prevalent in both unipolar depression and anxiety disorders (e.g., negative mood, insomnia, irritability, poor concentration). This factor is also similar to the construct Barlow and colleagues (Brown, Chorpita, & Barlow, 1998; Chorpita, Albano, & Barlow, 1998) labeled *anxious apprehension*, corresponding to symptoms of tension, apprehension, worry, and general distress linked to the behavioral inhibition system (Gray, 1987) and the *DSM* diagnosis of generalized anxiety disorder.

Both Watson and L. A. Clark's tripartite model and Barlow and colleagues' model include first-order factors that are specific to, and may differentiate, depression and anxiety. Symptoms specific to depression are those that reflect low levels of positive affect, including anhedonia, loss of interest or pleasure, lack of energy or motivation, and feelings of futility or hopelessness. Poor mood, anhedonia, and worthlessness are the most frequently assessed symptoms of depression on self-report questionnaires (Santor et al., 2006). In contrast, symptoms that best differentiate anxiety disorders reflect autonomic arousal and other somatic symptoms common to panic attacks or fear reactions. This factor has typically been labeled *anxious arousal* or *physiological reactivity*. Indeed, our final anxiety bank was distinguished by the inclusion of eight somatic items, whereas such items had been edited out of the depression and anger banks because of their poor psychometric characteristics.

The relationship between anger and depression (r = .60) and anger and anxiety (r = .59) was not as strong as that between depression and anxiety, but it was still large. Items in the anger bank also focused primarily on negative moods and thoughts, with items assessing aggression and externalizing behavior largely eliminated for psychometric reasons. Given the large correlations among the three item banks, it may be valuable in the future to model a single construct of internalized distress capturing aspects of depression, anxiety, and anger and using the PROMIS item banks to provide useful indicators of such a dimension. Again, such work would be consistent with the increasing evidence that a single latent trait reflecting propensity toward negative affect may underlie most internalizing disorders (Krueger & Finger, 2001; Lara, Pinto, Akiskal, & Akiskal, 2006).

Future Directions

Validation research on these item banks is ongoing. For example, they are being evaluated for their sensitivity to change, their ability to detect differences between distinct clinical conditions, and their concurrent validity with additional legacy measures in samples of patients with depression and lower back pain. The item banks have also been adopted for the *DSM-5* field trials where they will provide dimensional data on the severity of depression, anxiety, and anger across a broad variety of psychiatric diagnoses. Although the primary purpose of the PROMIS item banks is to create a dimensional metric of the severity of emotional distress, such validation work in various diagnostic (and nonclinical comparison) groups will also allow us to assess the utility of the banks in screening for psychiatric disorders. Given the large correlations among the three constructs (especially depression and anxiety), we have also begun to investigate hierarchical and multidimensional models of depression and anxiety that will allow us to "borrow" information between these constructs and create an even more efficient CAT for their assessment.

Summary

We have reported here on the development and calibration of item banks for three major aspects of emotional distress—depression, anxiety, and anger—using the comprehensive protocol for item banking and qualitative and quantitative analysis developed under the auspices of the PROMIS research network. The final item banks, static short forms, and CAT algorithms, together with the capability to create web-based assessment protocols, are publicly available through the PROMIS Assessment Center (www.assessmentcenter.net). Despite the apparent complexities of IRT-calibrated item banks for those new to this methodology, the resulting CAT functionality and short forms (and their administration and scoring) require no specialized skills. The PROMIS item banks have a number of conceptual and psychometric strengths, and tests of their predictive validity, precision, effect sizes, and sensitivity to change (relative to other conventional self-report measures) are now being conducted. Our overarching goal is to provide a standard frame of reference for outcomes measurement in clinical research, including clinical trials, observational research, and epidemiological studies. At the same time, we recognize potential limitations that may have resulted from the development process, with many of these related to the calibration sample, that is, the pros and cons of using an Internet polling panel to accrue most of the calibration sample, the relatively high educational level in the calibration sample, the limited representation of Asian Americans, and the reliance on self-report to sample most of the “clinical” participants included in initial testing. The PROMIS health status item banks are intended, however, to be a dynamic system, and we are prepared for further evolution of the item banks as new samples and data become available.

Appendix

Demographic Characteristics of the Participants in Cognitive Interviews

Forty-one participants sampled from outpatient mental health clinics reviewed the items (64% women, 24% minority, 24% married, 42% with a college or graduate degree, a mean age of 42 (*SD* = 12), and an age range from 20 to 60 years). Participants also completed the Wide Range Achievement Test (WRAT; Wilkinson & Robertson, 2006) to determine reading proficiency. The mean WRAT score was 46.7 (*SD* = 7.1) or post–high school, with scores ranging from 31 (third grade) to 57 (post–high school). A minimum of 5 participants reviewed each item. At least 2 of the 5 participants were members of underrepresented minorities, at least 2 were men, and at least 1 participant had limited educational attainment —a maximum of a high school education or a WRAT score indicating an eighth-grade reading level or less.

Acknowledgments

We acknowledge the work of our colleagues at or affiliated with the Pittsburgh research site who made a variety of contributions to the development of the item banks for emotional distress: Daniel J. Buysse, MD; Jill M. Cyranowski, PhD; Mary Amanda Dew, PhD; Nathan E. Dodds, BS; Susan V. Eisen, PhD; Anne Germain, PhD; Frank Ghinassi, PhD; Allen Heinemann, PhD; Kelly Johnston, MPH; Morgen Kelly, PhD; Mary Lou Klem, PhD; David J. Kupfer, MD; Douglas Moul, MD; Thomas Rudy, PhD; and Lan Yu, PhD. We also acknowledge the valuable assistance of two colleagues at the Statistical Coordinating Center: Nan Rothrock, PhD; and Susan Yount, PhD.

PROMIS is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring self-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. NIH Science Officers on this project have included Deborah Ader, PhD; Susan Czajkowski, PhD; Lawrence Fine, MD, DrPH; Laura Lee Johnson, PhD; Louis Quatrano, PhD; Bryce Reeve, PhD; William Riley, PhD; Susana Serrate-Sztein, MD; and James Witter, MD, PhD. This article was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See www.nihpromis.org for additional information on the PROMIS Cooperative Group.

Funding The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:

PROMIS was funded initially by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six primary research sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171).

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision. Washington, DC: Author; 2000.

Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: Psychometric properties. Journal of Consulting and Clinical Psychology. 1988; 56:893–897. [PubMed: 3204199]

Bejar II. An application of the continuous response level model to personality measurement. Applied Psychological Measurement. 1977; 1:509–521.

Bernstein IH, Rush AJ, Thomas CJ, Woo A, Trivedi MH. Item response analysis of the Inventory of Depressive Symptomatology. Neuropsychiatric Disease and Treatment. 2006; 2:557–564. [PubMed: 18159226]

Berzon R, Patrick D, Guyatt G, Conley JM. Intellectual property considerations in the development and use of HRQL measures for clinical trial research. Quality of Life Research. 1994; 3:273–277. [PubMed: 7812280]

Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: Item banking and computerized adaptive assessment. Quality of Life Research. 2007; 16:95–108. [PubMed: 17530450]

Bjorner, JB.; Smith, KJ.; Orlando, M.; Stone, CA.; Thissen, D.; Xiaowa, S. IRTFIT: A macro for item fit and local dependence tests under IRT models [Computer software]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory; 2006.

Bock RD. A brief history of item response theory. Educational Measurement: Issues and Practice. 1997; 16:21–33.

Bode RK, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. Archives of Physical Medicine and Rehabilitation. 2003; 84(Suppl 2):S52–S60. [PubMed: 12692772]

Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: Developing best practices based on science and experience. Quality of Life Research. 2009; 18:1263–1278. [PubMed: 19784865]

Brown TA, Chorpita BF, Barlow DH. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. Journal of Abnormal Psychology. 1998; 107:179–192. [PubMed: 9604548]

Buss AH, Perry M. The aggression questionnaire. Journal of Personality and Social Psychology. 1992; 59:73–81.

Buysse DJ, Moul DE, Germain A, Yu L, Stover AM, Dodds NE, Pilkonis PA, et al. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. Sleep. 2010; 33:781–792. [PubMed: 20550019]

Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D, Trivedi MH, et al. The Montgomery Asberg and the Hamilton ratings of depression: A comparison of measures. European Neuropsychopharmacology. 2006; 16:601–611. [PubMed: 16769204]

Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: Item banking, tailored short forms, and computerized adaptive assessment. Quality of Life Research. 2007; 16:133–144. [PubMed: 17401637]

Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Hays RD, et al. 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. 2010; 63:1179–1194. [PubMed: 20685078]

Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Matthias R, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. Medical Care. 2007; 45(Suppl 1):S3–S11. [PubMed: 17443116]

Chang C-H, Cella D, Clarke S, Heinemann A, Von Roenn JH, Harvey R. Should symptoms be scaled for intensity, frequency, or both? Palliative & Supportive Care. 2003; 1:51–60. [PubMed: 16594288]

Chang C-H, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Evaluation & the Health Professions. 2005; 28:264–282.

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. Quality of Life Research. 2010; 19:125–136. [PubMed: 19941077]

Choi SW, Swartz JR. Comparison of CAT item selection criteria for polytomous items. Applied Psychological Measurement. 2009; 33:419–440. [PubMed: 20011456]

Chorpita BF, Albano AM, Barlow DH. The structure of negative emotions in a clinical sample of children and adolescents. Journal of Abnormal Psychology. 1998; 107:74–85. [PubMed: 9505040]

Clark DC, vonAmmon Cavanaugh S, Gibbons RD. The core symptoms of depression in medical and psychiatric patients. Journal of Nervous and Mental Disease. 1983; 171:705–713. [PubMed: 6644280]

Clark LA, Watson D. Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. Journal of Abnormal Psychology. 1991; 100:316–336. [PubMed: 1918611]

Cox DE, Harrison DW. Models of anger: Contributions from psychophysiology, neuropsychology, and the cognitive behavioral perspective. Brain Structure & Function. 2008; 212:371–385. [PubMed: 18197417]

de Jong-Gierveld J, Kamphuis F. The development of a Rasch-type loneliness scale. Applied Psychological Measurement. 1985; 9:289–299.

Derogatis, LR. SCL-90-R: Administration, scoring and procedures manual II. Towson, MD: Clinical Psychometric Research; 1983.

DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: The PROMIS qualitative item review. Medical Care. 2007; 45(Suppl 1):S12–S21. [PubMed: 17443114]

Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Quality of Life Research. 2005; 14:2277–2291. [PubMed: 16328907]

Food and Drug Administration. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. 2009. Retrieved from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf

Forbey JD, Ben-Porath YS. Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. Psychological Assessment. 2007; 19:14–24. [PubMed: 17371120]

Forgays DK, Spielberger CD, Ottaway SA, Forgays DG. Factor structure of the State-Trait Anger Expression Inventory for middle-aged men and women. Assessment. 1998; 5:141–155. [PubMed: 9626390]

Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. Journal of Rheumatology. 2009; 36:2061–2066. [PubMed: 19738214]

Gardner W, Kelleher KJ, Pajer KA. Multidimensional adaptive testing for mental health problems in primary care. Medical Care. 2002; 40:812–823. [PubMed: 12218771]

Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, Frank E. Computerized adaptive measurement of depression: A simulation study. BMC Psychiatry. 2004; 4:13. [PubMed: 15132755]

Gershon RC, Rothrock NE, Hanrahan R, Jansky LJ, Harniss M, Riley W. Development of a clinical outcomes research survey application: Assessment Center. Quality of Life Research. 2010; 19:677–685. [PubMed: 20306332]

Gibbons RD, Bock D, Hedeker D, Weiss DJ, Segawa E, Bhaumik R, Stover A, et al. Full-information item bi-factor analysis of graded response data. Applied Psychological Measurement. 2007; 31:4–19.

Gibbons RD, Clark DC, vonAmmon Cavanaugh S, Davis JM. Application of modern psychometric theory in psychiatric research. Journal of Psychiatric Research. 1985; 19:43–55. [PubMed: 3989737]

Gibbons RD, Hedeker D. Full-information item bi-factor analysis. Psychometrika. 1992; 57:423–436.

Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, Grochocinski VJ, Immekus JC, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. Psychiatric Services. 2008; 59:361–368. [PubMed: 18378832]

Gollwitzer M, Eid M, Jurgensen R. Response styles in the assessment of anger expression. Psychological Assessment. 2005; 17:56–69. [PubMed: 15769228]

Gray, JA. The psychology of fear and stress. 2. Cambridge, England: Cambridge University Press; 1987.

Hawthorne G, Mouthaan J, Forbes D, Novaco RW. Response categories and anger measurement: Do fewer categories result in poorer measurement? Development of the DAR5. Social Psychiatry & Psychiatric Epidemiology. 2006; 41:164–172. [PubMed: 16362166]

Hays RD, Bjorner J, Revicki DA, Spritzer K, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Quality of Life Research. 2009; 18:873–880. [PubMed: 19543809]

Kelly MAR, Morse JQ, Stover A, Hofkens T, Huisman E, Shulman S, Pilkonis PA, et al. Describing depression: Congruence between patient experiences and clinical assessments. British Journal of Clinical Psychology. 2011; 50:46–66. [PubMed: 21332520]

Kendel F, Wirtz M, Dunkel A, Lehmkuhl E, Hetzer R, Regitz-Zagrosek V. Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. Journal of Affective Disorders. 2010; 122:241–246. [PubMed: 19665236]

Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF III. Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. Psychology and Aging. 2002; 17:379–391. [PubMed: 12243380]

Kingsbury, GG.; Weiss, DJ. An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests (Research Report 80-5). Minneapolis, MN: University of Minnesota; 1980.

Kingsbury, GG.; Weiss, DJ. A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In: Weiss, DJ., editor. New horizons in testing: Latent trait theory and computerized adaptive testing. New York, NY: Academic Press; 1983. p. 257-283.

Klem ML, Saghafi E, Abromitis R, Stover A, Dew MA, Pilkonis PA. Building PROMIS item banks: Librarians as co-investigators. Quality of Life Research. 2009; 18:881–888. [PubMed: 19548118]

Krueger RF. The structure of common mental disorders. Archives of General Psychiatry. 1999; 56:921–926. [PubMed: 10530634]

Krueger RF, Finger MS. Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. Psychological Assessment. 2001; 13:140–151. [PubMed: 11281035]

Krueger RF, McGue M, Iacono WG. The higher-order structure of common DSM mental disorders: Internalization, externalization, and their connections to personality. Personality and Individual Differences. 2001; 30:1245–1259.

Lara DG, Pinto O, Akiskal K, Akiskal HS. Toward an integrative model of the spectrum of mood, behavioral, and personality disorders based on fear and anger traits: 1. Clinical implications. Journal of Affective Disorders. 2006; 94:67–87. [PubMed: 16730070]

Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. Journal of Clinical Epidemiology. 2010; 63:1169–1178. [PubMed: 20688473]

Loevinger J. The technique of homogenous tests compared with some aspects of scale analysis and factor analysis. Psychological Bulletin. 1948; 45:507–530. [PubMed: 18893224]

Lord, FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.

Martin R, Wan C, David JP, Wegner E, Olson BD, Watson D. Style of anger expression: Relation to expressivity, personality, and health. Personality and Social Psychology Bulletin. 2000; 25:1196–1207.

Martin R, Watson D, Wan CK. A three-factor model of trait anger: Dimensions of affect, behavior, and cognition. Journal of Personality. 2000; 68:869–897. [PubMed: 11001152]

McBride, JR.; Martin, JR. Reliability and validity of adaptive ability tests in a military setting. In: Weiss, DJ., editor. New horizons in testing: Latent trait theory and computerized adaptive testing. New York, NY: Academic Press; 1983. p. 223-236.

McHugh RK, Behar E. Readability of self-report measures of depression and anxiety. Journal of Consulting and Clinical Psychology. 2009; 77:1100–1112. [PubMed: 19968386]

Meijer RR, Baneke JJ. Analyzing psychopathology items: A case for nonparametric item response theory modeling. Psychological Methods. 2004; 9:354–368. [PubMed: 15355153]

MetaMetrics, Inc. What does the Lexile measure mean?. Durham, NC: Author; 2008.

Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, de Vet HCW, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Quality of Life Research. 2009; 18:313–333. [PubMed: 19238586]

Molenaar, IW.; Sijtsma, K. User's manual for MSP5 for Windows. Groningen, Netherlands: IEC ProGAMMA; 2000.

Muhr, T.; Freise, S. User's manual for ATLAS.ti 5.0. 2. Berlin, Germany: Scientific Software Development; 2004.

Mui AC, Burnette D, Chen LM. Cross-cultural assessment of geriatric depression: A review of the CES-D and GDS. Journal of Mental Health and Aging. 2001; 7:137–164.

Muthén, LK.; Muthén, BO. Mplus user's guide. Los Angeles, CA: Author; 2006.

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

Pallant JF, Tennant A. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology. 2007; 46:1–18. [PubMed: 17472198]

Pedersen RD, Pallay AG, Rudolph RL. Can improvement in well-being and functioning be distinguished from depression improvement in antidepressant clinical trials? Quality of Life Research. 2002; 11:9–17. [PubMed: 12003058]

Polimetrix, Inc. Scientific sampling for online research. Palo Alto, CA: Author; 2006.

Quilty LC, Zhang KA, Bagby MR. The latent symptom structure of the Beck Depression Inventory-II in outpatients with major depression. Psychological Assessment. 2010; 22:603–608. [PubMed: 20822272]

Radloff LS. A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1:385–401.

Ramsay, JO. TestGraph: A program for the graphical analysis of multiple choice test and questionnaire data [Computer software and manual]. Montreal, Quebec, Canada: McGill University; 2000.

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Cella D, et al. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007; 45(Suppl 1):S22–S31. [PubMed: 17443115]

Reise, SP.; Moore, MT.; Haviland, MG. Applying item response theory models to psychological data. In: Geisinger, KF., editor. Handbook of testing and assessment in psychology. Washington, DC: American Psychological Association; IN PRESS

Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research. 2007; 16:19–31. [PubMed: 17479357]

Reise SP, Waller NG. Item response theory and clinical measurement. Annual Review of Clinical Psychology. 2009; 5:25–46.

Revicki DA, Chen W-H, Harnam N, Cook K, Amtmann D, Callahan LF, Keefe FJ, et al. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain. 2009; 146:158–169. [PubMed: 19683873]

Revicki DA, Schwartz CE. Intellectual property rights and good research practice. Quality of Life Research. 2009; 18:1279–1280. [PubMed: 19885743]

Revicki DA, Sloan J. Practical and philosophical issues surrounding a national item bank: If we build it will they come? Quality of Life Research. 2007; 16:167–174. [PubMed: 17468940]

Riley WT, Rothrock N, Bruce B, Christodolou C, Cook K, Hahn EA, Cella D. Patient-Reported Outcomes Measurement Information System (PROMIS) domain names and definitions revisions: Further assessment of content validity in IRT-derived item banks. Quality of Life Research. 2010; 19:1311–1321. [PubMed: 20593306]

Roberson-Nay R, Strong DR, Nay WT, Beidel DC, Turner SM. Development of an abbreviated Social Phobia and Anxiety Inventory (SPAI) using item response theory: The SPAI-23. Psychological Assessment. 2007; 19:133–145. [PubMed: 17371128]

Roth WM, Roychoudhury A. Nonmetric multidimensional item analysis in the construction of an anxiety attitude survey. Educational and Psychological Measurement. 1991; 51:931–942.

Rothrock NE, Hays RD, Spritzer K, Yount SE, Riley W, Cella D. 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 Epidemiology. 2010; 63:1195–1204. [PubMed: 20688471]

Santor DA, Gregus M, Welch A. Eight decades of measurement in depression. Measurement. 2006; 4:135–155.

Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. Journal of Clinical Psychology. 2006; 62:123–146. [PubMed: 16287149]

Sijtsma K, Meijer RR, van der Ark LA. Mokken Scale Analysis as time goes by: An update for scaling practitioners. Personality and Individual Differences. 2011; 50:31–37.

Simms LJ, Gros DF, Watson D, O'Hara MW. Parsing the general and specific components of depression and anxiety with bifactor modeling. Depression and Anxiety. 2008; 25:E34–E46. [PubMed: 18027844]

Steinberg L, Thissen D. Use of item response theory and the testlet concept in the measurement of psychopathology. Psychological Methods. 1996; 1:81–97.

Stone CA. Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. Applied Psychological Measurement. 1992; 16:1–16.

Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, Cella D, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly. 2009; 51:148–180. [PubMed: 20336180]

Thissen, D.; Chen, WH.; Bock, RD. MULTILOG (version 7) [Computer software]. Lincolnwood, IL: Scientific Software International; 2003.

Thissen D, Steinberg L. An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. Applied Psychological Measurement. 1983; 7:211–226.

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum; 1993. p. 67-113.

Vancleef LMG, Vlaeyen JWS, Peters ML. Dimensional and componential structure of a hierarchical organization of pain-related anxiety constructs. Psychological Assessment. 2009; 21:340–351. [PubMed: 19719346]

Waller NG, Reise SP. Computerized adaptive personality assessment: An illustration with the Absorption scale. Journal of Personality and Social Psychology. 1989; 57:1051–8. [PubMed: 2614658]

Walter OB, Becker J, Bjorner JB, Fliege H, Klapp BF, Rose M. Development and evaluation of a computer adaptive test for anxiety (Anxiety-CAT). Quality of Life Research. 2007; 16:143–155. [PubMed: 17342455]

Ware JE. Conceptualization and measurement of health-related quality of life: Comments on an evolving field. Archives of Physical Medicine and Rehabilitation. 2003; 84(Suppl 2):S43–S51. [PubMed: 12692771]

Watson D, Clark LA, Weber K, Assenheimer JA, Strauss ME, McCormick RA. Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptoms. Journal of Abnormal Psychology. 1995a; 104:3–14. [PubMed: 7897050]

Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. Journal of Abnormal Psychology. 1995b; 104:15–25. [PubMed: 7897037]

Weiss DJ. Adaptive testing by computer. Journal of Consulting and Clinical Psychology. 1985; 53:774–789. [PubMed: 3841355]

Weiss DJ. Computerized adaptive testing for effective and efficient measurement in counseling and education. Measurement and Evaluation in Counseling and Development. 2004; 37:70–84.

Willis, GB. Cognitive interviewing: A tool for improving questionnaire design. Thousand Oaks, CA: SAGE; 2005.

Winkielman P, Knäuper B, Schwarz N. Looking back at anger: Reference periods change the interpretation of emotion frequency questions. Journal of Personality and Social Psychology. 1998; 75:719–728. [PubMed: 9781408]

Woods CM. Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. Psychological Methods. 2006; 11:253–270. [PubMed: 16953704]

World Health Organization. Constitution of the World Health Organization: Basic documents. 46. Geneva, Switzerland: Author; 2007.

Zinbarg RE, Barlow DH. Structure of anxiety and the anxiety disorders: A hierarchical model. Journal of Abnormal Psychology. 1996; 105:181–193. [PubMed: 8722999]

Zumbo, BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.

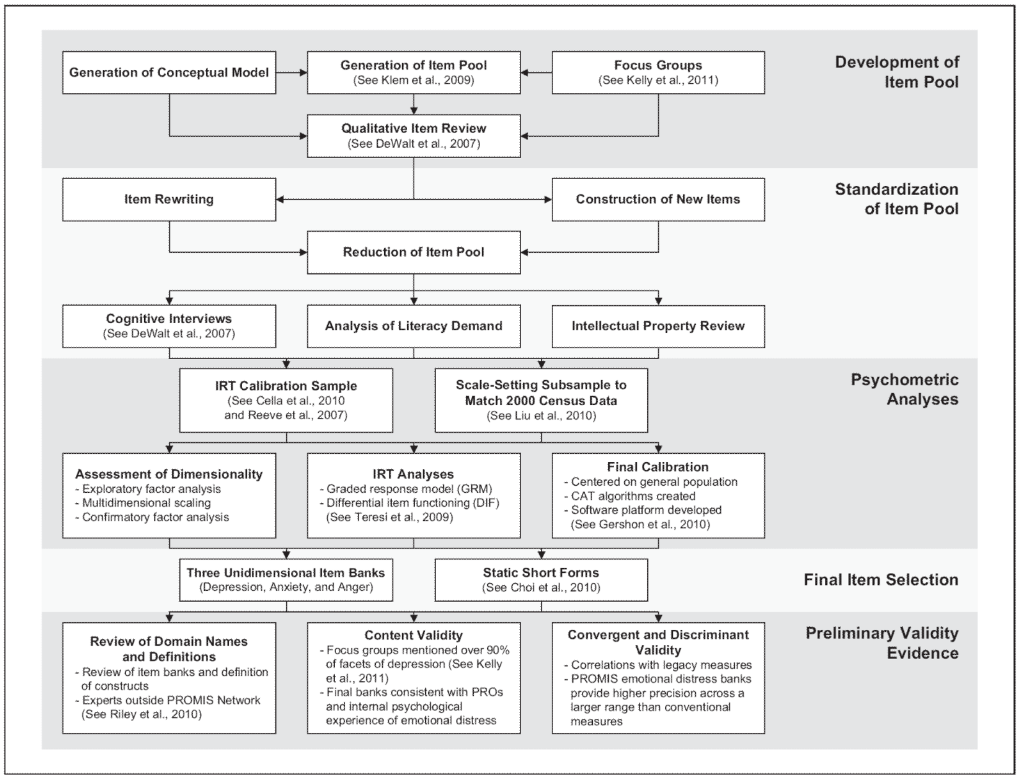
Generation of Conceptual Model
Generation of Item Pool (See Klem et al., 2009)
Focus Groups (See Kelly et al., 2011)
Development of Item Pool
Qualitative Item Review (See DeWalt et al., 2007)
Item Rewriting
Construction of New Items
Standardization of Item Pool
Reduction of Item Pool
Cognitive Interviews (See DeWalt et al., 2007)
Analysis of Literacy Demand
Intellectual Property Review
IRT Calibration Sample (See Cella et al., 2010 and Reeve et al., 2007)
Scale-Setting Subsample to Match 2000 Census Data (See Liu et al., 2010)
Psychometric Analyses
Assessment of Dimensionality
- Exploratory factor analysis
- Multidimensional scaling
- Confirmatory factor analysis
IRT Analyses
- Graded response model (GRM)
- Differential item functioning (DIF) (See Teresi et al., 2009)
Final Calibration
- Centered on general population
- CAT algorithms created
- Software platform developed (See Gershon et al., 2010)
Three Unidimensional Item Banks (Depression, Anxiety, and Anger)
Static Short Forms (See Choi et al., 2010)
Final Item Selection
Review of Domain Names and Definitions
- Review of item banks and definition of constructs
- Experts outside PROMIS Network (See Riley et al., 2010)
Content Validity
- Focus groups mentioned over 90% of facets of depression (See Kelly et al., 2011)
- Final banks consistent with PROs and internal psychological experience of emotional distress
Convergent and Discriminant Validity
- Correlations with legacy measures
- PROMIS emotional distress banks provide higher precision across a larger range than conventional measures
Preliminary Validity Evidence

Figure 1.

Flowchart illustrating the PROMIS methodology

Note. PROMIS = Patient-Reported Outcomes Measurement Information System.

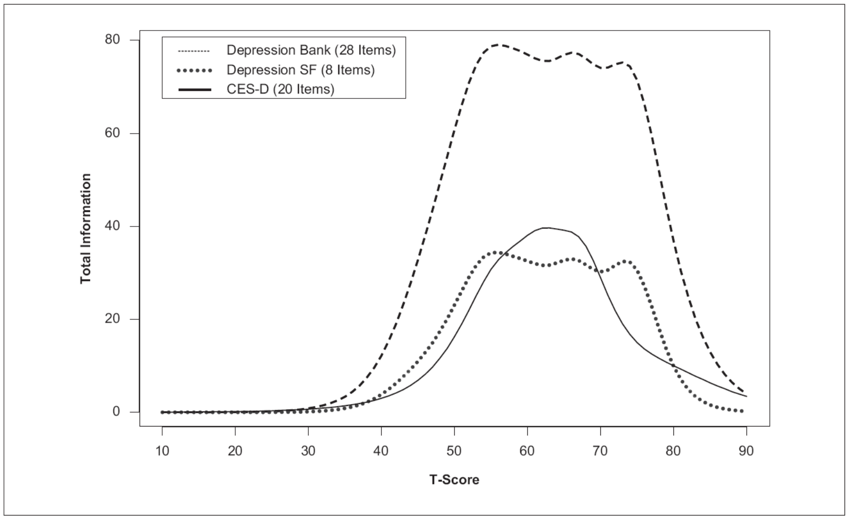

Figure 2.
Test information curves for depression: Full bank, short form, and the CES-D
Note. CES-D = Center for Epidemiologic Studies Depression Scale.

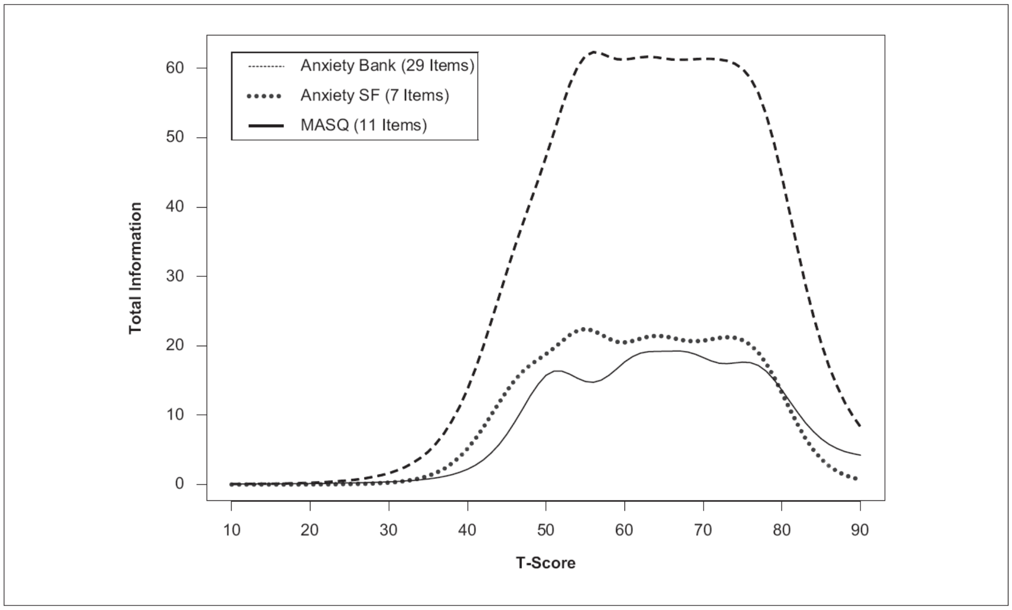

Figure 3.

Test information curves for anxiety: Full bank, short form, and the general distress (anxiety) scale from the MASQ

Note. MASQ = Mood and Anxiety Symptom Questionnaire.

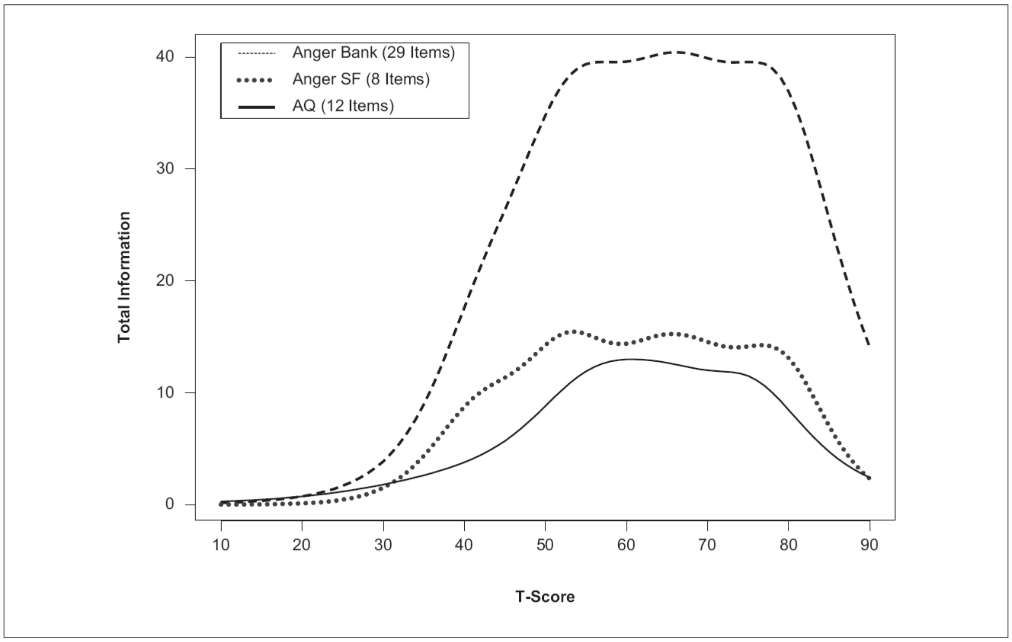

Figure 4.

Test information curves for anger: Full bank, short form, and the combined subscales for anger and verbal aggression from the AQ

Note. AQ = Aggression Questionnaire.

Table 1

Size and Demographic Characteristics of the Testing Samples

	Inclusive sample, *N* = 21,133	**Calibration subsample: Anger, *N* = 14,901**	**Calibration subsample: Anxiety, *N* = 14,836**	**Calibration subsample: Depression, *N* = 14,839**	**Scale-setting subsample, *N* = 5,239**
Sample size					
Testing format, *n*					
Full bank	7,005	858	788	782	2,498
Block	14,128	14,043	14,048	14,057	2,741
Population type, *n*					
General	19,601	13,695	13,631	13,632	4,943
Clinical	1,532	1,206	1,205	1,207	296
Characteristics					
Female, %	52	52	52	52	58
Minority,%	18	17	17	17	17
Hispanic,%	9	8	8	8	9
≤ High school education,%	18	18	18	18	49
Reporting chronic medical condition, %	40	45	45	45	29
Age in years; mean (*SD*)	53 (17)	54 (16)	54 (16)	54 (16)	51 (18)

Note. The inclusive and scale-setting samples represent totals across all Patient-Reported Outcomes Measurement Information System domains and not emotional distress alone.

Table 2

Calibrated Depression Items

Item stem	Slope (discrimination)	Location threshold 1	Location threshold 2	Location threshold 3	Location threshold 4
I felt hopeless[a]	4.46	0.49	1.00	1.71	2.46
I felt depressed[a]	4.35	−0.19	0.53	1.36	2.20
I felt worthless[a]	4.27	0.33	0.90	1.62	2.37
I felt helpless[a]	4.15	0.29	0.84	1.61	2.40
I felt like a failure[a]	3.97	0.13	0.72	1.58	2.22
I felt that I had nothing to look forward to[a]	3.94	0.23	0.84	1.52	2.34
I felt that nothing could cheer me up	3.66	0.24	0.91	1.71	2.50
I felt unhappy[a]	3.49	−0.61	0.28	1.27	2.28
I felt sad[a]	3.28	−0.57	0.33	1.34	2.30
I felt that I wanted to give up on everything	3.24	0.39	0.96	1.76	2.44
I felt that my life was empty	3.19	0.13	0.71	1.45	2.25
I felt discouraged about the future	3.19	−0.33	0.33	1.23	2.06
I felt I had no reason for living	3.13	0.85	1.41	2.09	2.78
I found that things in my life were overwhelming	3.11	−0.03	0.65	1.57	2.40
I felt disappointed in myself	3.10	−0.43	0.34	1.33	2.15
I felt that I was not needed	2.92	0.13	0.82	1.58	2.46
I felt that nothing was interesting	2.84	0.07	0.83	1.77	2.80
I withdrew from other people	2.80	0.08	0.70	1.53	2.46
I felt that I was to blame for things	2.74	0.00	0.74	1.73	2.60
I felt emotionally exhausted	2.69	−0.37	0.35	1.29	2.23
I had trouble making decisions	2.62	−0.09	0.80	1.79	2.75
I felt lonely	2.59	−0.15	0.56	1.41	2.25
I had trouble feeling close to people	2.57	−0.11	0.62	1.58	2.51
I felt upset for no reason	2.55	0.12	0.94	1.94	3.05
I felt pessimistic	2.38	−0.53	0.41	1.47	2.56
I felt ignored by people	2.37	0.14	0.92	1.83	2.86
I felt that I was not as good as other people	2.34	0.12	0.88	1.66	2.56

Item stem	**Slope (discrimination)**	**Location threshold 1**	**Location threshold 2**	**Location threshold 3**	**Location threshold 4**
I felt guilty	2.02	−0.12	0.85	1.93	2.89

Note. Items are rank ordered on the basis of their slope (discrimination) parameters. All items are reprinted with the permission of the Patient-Reported Outcomes Measurement Information System (PROMIS) Health Organization and the PROMIS Cooperative Group.

[a] Items included in the short form.

Table 3

Calibrated Anxiety Items

Item stem	Slope (discrimination)	Location threshold 1	Location threshold 2	Location threshold 3	Location threshold 4
I found it hard to focus on anything other than my anxiety[a]	3.86	0.41	1.20	2.05	2.84
My worries overwhelmed me	3.64	0.29	0.96	1.71	2.56
I felt uneasy[a]	3.64	−0.31	0.52	1.50	2.44
I felt fearful[a]	3.58	0.27	1.02	1.90	2.64
I felt like I needed help for my anxiety	3.53	0.47	0.98	1.80	2.32
I felt frightened	3.43	0.42	1.26	2.09	2.82
I felt nervous[a]	3.38	−0.29	0.56	1.58	2.67
I felt anxious[a]	3.34	−0.27	0.53	1.51	2.38
I felt tense[a]	3.33	−0.59	0.24	1.18	2.23
It scared me when I felt nervous	3.32	0.55	1.17	2.00	2.70
I had difficulty calming down	3.11	0.00	0.87	1.78	2.72
Many situations made me worry	3.03	−0.41	0.48	1.39	2.28
I felt worried[a]	3.02	−0.59	0.24	1.28	2.24
I had sudden feelings of panic	2.97	0.49	1.23	2.09	3.00
I felt something awful would happen	2.85	0.36	1.06	1.97	2.72
I was concerned about my mental health	2.83	0.30	0.99	1.79	2.48
I felt upset	2.72	−0.65	0.40	1.61	2.91
I felt terrified	2.57	1.08	1.75	2.66	3.53
I felt indecisive	2.52	−0.31	0.61	1.73	2.84
I had trouble relaxing	2.40	−0.54	0.33	1.30	2.34
I had trouble paying attention	2.22	−0.38	0.67	1.79	2.87
I felt fidgety	1.94	−0.10	0.81	1.95	3.18
I avoided public places or activities	1.82	0.26	0.91	1.82	2.96
I was anxious if my normal routine was disturbed	1.79	−0.11	0.81	2.05	2.99
I was easily startled	1.71	−0.06	1.10	2.14	3.05
I worried about other people’s reactions to me	1.70	−0.11	0.81	1.94	3.05
I had difficulty sleeping	1.51	−0.91	0.01	1.10	2.32

Item stem	Slope (discrimination)	Location threshold 1	Location threshold 2	Location threshold 3	Location threshold 4
I had a racing or pounding heart	1.42	0.25	1.32	2.68	4.61
I had twitching or trembling muscles	1.26	0.50	1.39	2.62	3.92

Note. Items are rank ordered on the basis of their slope (discrimination) parameters. All items are reprinted with the permission of the Patient-Reported Outcomes Measurement Information System (PROMIS) Health Organization and the PROMIS Cooperative Group.

[a] Items included in the short form.

Table 4

Calibrated Anger Items

Item stem	Slope (discrimination)	Location threshold 1	Location threshold 2	Location threshold 3	Location threshold
I was grouchy[a]	2.93	−0.88	0.22	1.52	2.85
I stayed angry for hours[a]	2.80	0.36	1.24	2.13	3.08
I felt angry[a]	2.77	−0.89	0.29	1.67	2.91
I felt like I was ready to explode[a]	2.77	0.24	1.03	1.96	3.04
I felt like I needed help for my anger	2.73	0.83	1.51	2.37	2.95
I felt angrier than I thought I should[a]	2.56	−0.10	0.71	1.79	2.56
I felt annoyed[a]	2.45	−1.18	−0.06	1.26	2.87
I made myself angry about something just by thinking about it[a]	2.41	−0.47	0.47	1.62	2.64
When I was angry, I sulked	2.34	0.05	0.92	2.02	2.97
I felt like breaking things	2.31	0.81	1.50	2.38	3.34
I was irritated more than people knew[a]	2.30	−0.83	0.01	1.13	2.19
I felt like yelling at someone	2.28	−0.43	0.53	1.72	2.79
Just being around people irritated me	2.26	0.40	1.23	1.94	2.77
I was stubborn with others	2.25	−0.63	0.51	1.90	3.12
Even after I expressed my anger, I had trouble forgetting about it	2.21	−0.27	0.63	1.70	2.63
I felt bitter about things	2.17	−0.24	0.71	1.73	2.79
I felt resentful when I didn’t get my way	2.16	0.08	1.13	2.49	3.49
I held grudges toward others	2.12	0.08	1.14	2.23	3.06
I had a bad temper	2.11	−0.11	1.04	2.26	3.62
I had trouble controlling my temper	2.01	0.60	1.65	2.85	3.88
I felt that people were trying to anger me	1.98	0.35	1.24	2.49	3.48
I felt guilty about my anger	1.93	−0.02	0.76	1.92	2.71
I tried to get even when I was angry with someone	1.80	0.72	1.88	2.98	3.91
I was angry when something blocked my plans	1.79	−0.47	0.62	2.12	3.19
I felt envious of others	1.58	0.06	1.24	2.54	3.68
When I was frustrated, I let it show	1.53	−1.25	0.11	1.80	3.28
I disagreed with people	1.48	−1.91	−0.50	1.88	3.91

Item stem	Slope (discrimination)	Location threshold 1	Location threshold 2	Location threshold 3	Location threshold
I was angry when I was delayed	1.43	−0.35	0.81	2.35	3.97
When I was mad at someone, I gave them the silent treatment	1.38	−0.25	0.78	2.30	3.73

Note. Items are rank ordered on the basis of their slope (discrimination) parameters. All items are reprinted with the permission of the Patient-Reported Outcomes Measurement Information System (PROMIS) Health Organization and the PROMIS Cooperative Group.

[a] Items included in the short form.