The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008

David Cella,*, William Riley, Arthur Stone, Nan Rothrock, Bryce Reeve, Susan Yount, Dagmar Amtmann, Rita Bode, Daniel Buysse, Seung Choi, Karon Cook, Robert DeVellis, Darren DeWalt, James F. Fries, Richard Gershon, Elizabeth A. Hahn, Jin-Shei Lai, Paul Pilkonis, Dennis Revicki, Matthias Rose, Kevin Weinfurt, Ron Hays on behalf of the PROMIS Cooperative Group

Medical Social Sciences, Northwestern University, Chicago, IL, USA
National Heart, Lung, and Blood Institute, Bethesda, MD, USA
Psychiatry, Stony Brook University, Stony Brook, NY, USA
National Cancer Institute, Bethesda, MD, USA
Rehabilitation Medicine, University of Washington, Seattle, WA, USA
Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, PA, USA
Health Behavior and Health Education Department, University of North Carolina, Chapel Hill, NC, USA
General Internal Medicine and Cecil G. Sheps Center for Health Service Research, University of North Carolina, Chapel Hill, NC, USA
Department of Medicine, Stanford University, Palo Alto, CA, USA
United BioSource Corporation, Bethesda, MD, USA
Department of Medicine, Hamburg University, Hamburg, Germany
School of Medicine, Duke University, Durham, NC, USA
Department of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Accepted 8 April 2010

Abstract

**Objectives:** Patient-reported outcomes (PROs) are essential when evaluating many new treatments in health care; yet, current measures have been limited by a lack of precision, standardization, and comparability of scores across studies and diseases. The Patient-Reported Outcomes Measurement Information System (PROMIS) provides item banks that offer the potential for efficient (minimizes item number without compromising reliability), flexible (enables optional use of interchangeable items), and precise (has minimal error in estimate) measurement of commonly studied PROs. We report results from the first large-scale testing of PROMIS items.

**Study Design and Setting:** Fourteen item pools were tested in the U.S. general population and clinical groups using an online panel and clinic recruitment. A scale-setting subsample was created reflecting demographics proportional to the 2000 U.S. census.

**Results:** Using item-response theory (graded response model), 11 item banks were calibrated on a sample of 21,133, measuring components of self-reported physical, mental, and social health, along with a 10-item Global Health Scale. Short forms from each bank were developed and compared with the overall bank and with other well-validated and widely accepted ("legacy") measures. All item banks demonstrated good reliability across most of the score distributions. Construct validity was supported by moderate to strong correlations with legacy measures.

**Conclusion:** PROMIS item banks and their short forms provide evidence that they are reliable and precise measures of generic symptoms and functional reports comparable to legacy instruments. Further testing will continue to validate and test PROMIS items and banks in diverse clinical populations.

**Keywords:** Outcome measures; Quality of life; Chronic disease; Item response theory; Patient-reported outcomes; PROMIS

1. **Introduction**

Clinical outcome measures, such as radiographic imaging and laboratory tests, have minimal immediate relevance to the day-to-day functioning of patients with chronic
What is new?

Key finding
PROMIS item banks have demonstrated reliability, precision, and construct validity.

What this adds to what was known?
PROMIS measures provide a common metric for assessment of physical function, pain, fatigue, emotional distress, social function, and sleep/wake disturbance.

What is the implication, and what should change now?
PROMIS measures are available for use in clinical research.

This initiative—the Patient-Reported Outcomes Measurement Information System (PROMIS)—establishes a national resource for precise and efficient measurement of patient-reported symptoms, functioning, and health-related quality of life, appropriate for patients with a wide variety of chronic diseases and conditions. The main goal of the PROMIS initiative is to develop and evaluate, for use in clinical research and health care delivery settings. This initiative—the Patient-Reported Outcomes Measurement Information System (PROMIS)—establishes a national resource for precise and efficient measurement of patient-reported symptoms, functioning, and health-related quality of life, appropriate for patients with a wide variety of chronic diseases and conditions. The main goal of the PROMIS initiative is to develop and evaluate, for use in clinical research and health care delivery settings. This initiative—the Patient-Reported Outcomes Measurement Information System (PROMIS)—establishes a national resource for precise and efficient measurement of patient-reported symptoms, functioning, and health-related quality of life, appropriate for patients with a wide variety of chronic diseases and conditions. The main goal of the PROMIS initiative is to develop and evaluate, for use in clinical research and health care delivery settings.

During the first 2 years of support, the PROMIS network developed a domain framework (Fig. 1) that focused efforts to organize item pools for wave 1 testing. This framework begins on the left side of the figure with three broad aspects of self-reported health: physical, mental, and social. Each of these aspects, in turn, is comprised of components, or “domains,” of health-related quality of life (HRQL). In the first year of PROMIS, investigators working within the consensus-based framework decided to initiate work in at least one domain from each broad aspect of health (physical, mental, and social). Specific domains selected for development were physical function, fatigue, pain, emotional distress, social function, and global health. The framework in Fig. 1 represents the March 2010 version, which has been modified over time based on empirical results (including some reported herein). Content elaborations on the right half of the figure represent functioning banks (green background), components of functioning banks (gray background), item banks in development (yellow background), and uncalibrated items pools and scales (blue background). Conceptual definitions that guided the development of the proposed wave 1 domains are as follows.

1.1. Wave 1 domains and definitions

During the first 2 years of support, the PROMIS network developed a domain framework (Fig. 1) that focused efforts to organize item pools for wave 1 testing. This framework begins on the left side of the figure with three broad aspects of self-reported health: physical, mental, and social. Each of these aspects, in turn, is comprised of components, or “domains,” of HRQL. In the first year of PROMIS, investigators working within the consensus-based framework decided to initiate work in at least one domain from each broad aspect of health (physical, mental, and social). Specific domains selected for development were physical function, fatigue, pain, emotional distress, social function, and global health. The framework in Fig. 1 represents the March 2010 version, which has been modified over time based on empirical results (including some reported herein). Content elaborations on the right half of the figure represent functioning banks (green background), components of functioning banks (gray background), item banks in development (yellow background), and uncalibrated items pools and scales (blue background). Conceptual definitions that guided the development of the proposed wave 1 domains are as follows.

1.1.1. Physical health (including physical function, physical symptoms, sleep function, and sexual function)

1.1.1.1. Physical function. Physical function is defined as one’s ability to carry out various activities that require physical capability, ranging from self-care (activities of daily living) to more vigorous activities that require increasing degrees of mobility, strength, or endurance [6–10]. Physical function is conceptually multidimensional, with four related subdomains: mobility (lower extremity function), dexterity (upper extremity function), axial (neck and back) function, and ability to carry out instrumental activities of daily living [11].

1.1.1.2. Fatigue. In the health-outcomes measurement perspective, fatigue is defined as an overwhelming, debilitating, and sustained sense of exhaustion that decreases one’s ability to carry out daily activities, including the ability to work effectively and to function at one’s usual level in family or social roles [12–14]. Similar subjective feelings, yet fewer behavioral impacts, are associated with lower levels of fatigue. Fatigue is divided conceptually into the experience of fatigue (such as its intensity, frequency, and duration), and the impact of fatigue on physical, mental, and social activities.

1.1.1.3. Pain. Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue...
damage or is described in terms of such damage [15–18]. Pain is what the respondent says it is, that is, the “gold standard” of pain assessment is self-report [19]. Pain is divided conceptually into components of quality (referring to the nature, characteristics, intensity, frequency, and duration of pain); impact on physical, mental, and social activities; and behaviors one engages in to avoid, minimize, or reduce pain.

1.1.1.4. Sleep disturbance and sleep-related impairment.

Sleep and wakefulness are the two fundamental behavioral states of human beings. Sleep is a rapidly reversible, recurrent state of reduced (but not absent) awareness of and interaction with the environment. Wakefulness is a behavioral state of active engagement and interaction with the environment, including the perception and processing of stimuli and the production of cognitive, emotional, and behavioral responses.

The PROMIS sleep disturbance item bank focuses on perceptions of sleep quality, sleep depth, and restoration associated with sleep; perceived difficulties with getting to sleep or staying asleep; and perceptions of the adequacy of and satisfaction with sleep. The sleep disturbance item bank does not include symptoms of specific sleep disorders nor does it provide subjective estimates of sleep quantities (e.g., the total amount of sleep, time to fall asleep, or amount of wakefulness during sleep).

The PROMIS sleep-related impairment item bank focuses on perceptions of alertness, sleepiness, and tiredness during usual waking hours, and on functional impairments during wakefulness that are associated with sleep problems or impaired alertness. The sleep-related impairment item bank does not directly assess cognitive, affective, or performance impairments. The sleep-related impairment item bank measures the level of waking alertness, sleepiness, and function within the context of overall sleep–wake function.

1.1.2. Mental health (includes emotional distress, cognitive function, and positive psychological function)

1.1.2.1. Emotional distress. Emotional distress is an important component of emotional health and is comprised typically of aspects of anxiety, depression, and anger. Given the overlap among these symptoms, a number of conceptual models have been proposed to account for the shared vs. unique variance captured in measures of negative affect. PROMIS adopted a hierarchical structure to explain the relationships between self-reported symptoms of anxiety, depression, and anger [20,21]. This structure includes a second-order, nonspecific factor reflecting high levels of negative affect—or “general distress”—common to all these emotions. Anger tends to have smaller loadings on the general factor than anxiety and depression, but it still is a strong marker of emotional distress. The PROMIS item
banks emphasize the cognitive and affective components of these concepts. Both psychometric considerations (e.g., skewed distributions for high-threshold behavioral items, the need to fit item-response theory [IRT] models to coherent unidimensional concepts) and considerations regarding validity (e.g., potential confounding between somatic symptoms of emotional distress and markers of physical disease) led us to this emphasis.

1.1.2.1. Depression. The PROMIS item bank for depression focuses on negative mood (e.g., sadness, guilt), decrease in positive affect (e.g., loss of interest), information-processing deficits (e.g., problems in decision-making), negative views of the self (e.g., self-criticism, worthlessness), and negative social cognition (e.g., loneliness, interpersonal alienation).

1.1.2.1.2. Anxiety. The PROMIS item bank for anxiety focuses on fear (e.g., fearfulness, feelings of panic), anxious misery (e.g., worry, dread), hyperarousal (e.g., tension, nervousness, restlessness), and somatic symptoms related to arousal (e.g., cardiovascular symptoms, dizziness).

1.1.2.1.3. Anger. The PROMIS item bank for anger focuses on angry mood (e.g., irritability, reactivity), negative social cognition (e.g., interpersonal sensitivity, envy, venegfulness), verbal aggression, and efforts necessary to control angry mood.

1.1.3. Social health

Social health is defined as perceived well-being regarding social activities and relationships, including the ability to relate to individuals, groups, communities, and society as a whole. Components of social functioning include understanding and communication, getting along with people, participation in society, and performance of social roles. Additional conceptualizations of social functioning focus on the quality, reciprocity, and size of an individual’s social network [22,23]. Although social function was the initial focus of PROMIS investigation, several other aspects of social health are noteworthy. These include social support and interpersonal attributes independent of particular roles, such as intimacy, assertiveness, sociability, submissiveness, and interpersonal control [24].

1.1.3.1. Social function. Social function is defined by PROMIS as involvement in, and satisfaction with, one’s usual social roles in life’s situations and activities. These roles may exist in dyadic or family relationships, parental responsibilities, work responsibilities, and social activities [25,26]. Social function has also been referred to with terms, such as role participation and social adjustment [26]. Qualitative and quantitative analysis of PROMIS and archival data collected before the current study (Cella et al., 2007) [27,28] led us to hypothesize a conceptual division of social function into “ability to participate” and “satisfaction with participation.” Each of these two components has subcomponents that divide social roles, such as work and family responsibilities, and more discretionary social activities, such as leisure activity and relationships with friends.

1.1.4. Global health

Global health refers to a person’s general evaluations of health rather than any of its specific components. The global health items include global ratings of the five primary PROMIS domains (physical function, fatigue, pain, emotional distress, and social health) and general health perceptions that cut across domains. Global items allow respondents to weigh together different aspects of health to arrive at a “bottom-line” indicator of their health status. Global health items have been found to be consistently predictive of important future events, such as health care utilization and mortality [8,14,18]. Results from wave 1 testing of the global items are reported elsewhere [29,30].

1.2. Item development

Each domain listed earlier was assigned a team of PROMIS investigators, consisting of experts in the measurement and assessment of the domain area. These teams identified, evaluated, and revised an exhaustive set of extant questionnaire items and wrote new items when necessary to form a core item pool for each domain. Six phases of item development were documented: identification of existing items, item classification and selection, item review and revision, focus group input on domain coverage, cognitive interviews with individual items, and final revision before field testing [2].

To inform PROMIS item selection and development, we analyzed 11 large data sets with self-report data on the five broad PROMIS core domains: pain, fatigue, emotional distress, physical function, and social function [1,29,31,32]. Sleep disturbance and sleep-related impairment were not included, as their development was focused at a single PROMIS site rather than as a full network effort. Psychometric results from these analyses were reviewed collectively by the analysis team, and summaries were presented to the appropriate domain working group. The primary goal was to use these archival data to better understand the dimensional structure of items that tap one of the five selected PROMIS domains. Secondarily, we aimed to inform the revision of items in the item pools, identify the best performing sets of response options, and guide new item construction in preparation for the first wave of PROMIS testing [1].

1.3. Recall period

Although some data suggest that recall periods beyond 1 day may introduce bias into the reporting of symptoms [33], a recent study of pain and fatigue [34] suggests
reasonably high correspondence between real-time symptom reports and 7-day recall of the same symptoms. In addition, Revicki et al. [35] found that gastrointestinal symptom scores based on a daily diary had a correlation greater than 0.90 with a 2-week recall instrument, suggesting minimal recall bias. Thus, from a practical viewpoint, a 7-day recall period provides a sufficiently long interval to capture a clinically relevant window of time and experience with minimal bias. Based on these studies, we opted for the 7-day option as optimal in most cases. “In the past 7 days” is the reference period for all items in anxiety, anger, depression, fatigue, pain quality, pain interference, pain behavior, satisfaction with participation in discretionary social activities, satisfaction with participation in social roles, sleep disturbance, and sleep-related impairment. An exception is physical function, which emphasizes current capabilities and, therefore, does not use a recall period. Item stems begin with phrases, such as “Does your health now limit you” or “Are you able to.” Some global health items use a 7-day recall period, whereas others do not use a recall period and emphasize current status in general.

1.4. Response options

Most of the PROMIS items use response scales with five options (e.g., 1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much). This number of response options was selected after extensive discussion based on prior work [36] and analyses of available large data sets, in which five response options produced data sets with ample responses in each option for IRT analysis; provided good discrimination in item characteristic curves without producing failures of monotonicity, scalability, or item misfit; and performed well in cognitive testing. Pain behavior uses six response options to allow for respondents to endorse “had no pain.” In this way, we could differentiate those with no pain from those who report no such behavior in response to pain. Each of the 10 PROMIS global health items has five response choices, except the 11-point pain intensity item (“How would you rate your pain on average” with 0 = no pain and 10 = worst imaginable pain). All modifications to existing items regarding the number and wording of response options were made with permission of the source item developer. To ease respondent burden, the wording of response categories was kept consistent within banks, and a limited degree of variation in response options was used across banks. Some flexibility in response choices within banks was considered important, however, to capture the range of patient experience in a domain (e.g., intensity, frequency, duration). Therefore, for example, most banks used a common set of response options for intensity (i.e., “not at all” to “very much”) and frequency (i.e., “never” to “always”). The selected response categories were pretested with cognitive interviews to confirm patient comprehension before field testing for item calibration.

1.5. Wave 1 testing

After the extensive literature review to identify items for each bank, review of the items by experts and patients, and standardization of the questions and response format, the next phase of PROMIS included the large wave 1 testing of the items to collect patient-reported data to allow quantitative evaluation and calibration of the PROMIS items.

1.5.1. Sampling and analysis framework

From July 2006 to March 2007, data were collected from the U.S. general population and multiple disease populations. A sampling plan was developed for collecting responses to the candidate items from the targeted PROMIS domains. This plan was designed to accommodate multiple objectives: (1) obtain item calibrations for each domain; (2) estimate profile scores for various disease populations; (3) create linking metrics to legacy questionnaires (e.g., Short Form [SF]-36); (4) confirm the factor structure of the domains; and (5) conduct item and bank analyses. Because of the large total number of items (>1,000), it was unreasonable to ask participants to respond to the entire pool of items. We estimated that participants would respond to approximately four questions per minute and limited the maximum number of items administered to about 150, for an estimated average response time of 37 minutes.

Figure 2 outlines the two arms of the sampling design: “full-bank” and “block” administration. There were 14 candidate item banks (three physical functioning banks, anxiety, depression, anger, alcohol abuse, fatigue interference, fatigue experience, social role performance, social role satisfaction, pain interference, pain quality, and pain behavior). All 56 items for each of the two PROMIS candidate item banks (112 PROMIS items) were administered to a subset of individuals in the full-bank arm. They also completed appropriate “legacy” questionnaires (well-validated and widely used measures of the same concept). Another subset of the PROMIS wave 1 sample was administered blocks of seven items selected from each of the 14 candidate item banks (98 PROMIS items). All participants completed a clinical form consisting of approximately 25 auxiliary items measuring global health perceptions and sociodemographic variables, including age, income, number of hospitalizations, disability days, use of prescription medication, height, weight, gender, race/ethnicity, relationship status, educational attainment, and employment status. This clinical form also included a series of health questions about the presence and degree of limitations related to 25 chronic medical conditions: hypertension, angina, coronary artery disease, heart failure, heart attack, stroke or transient ischemic attack, liver disease, kidney disease, arthritis or rheumatism, osteoarthritis, migraines, asthma, chronic obstructive pulmonary disease, diabetes, cancer, depression, anxiety, alcohol or drug problems, sleep disorder, human immunodeficiency virus/acquired immunodeficiency...
syndrome, spinal cord injury, multiple sclerosis, Parkinson’s disease, epilepsy, and amyotrophic lateral sclerosis.

We organized the sampling frame and item administration according to two types: full-bank and block administration. These are described in detail later. The full-bank administration provided data for evaluating dimensionality and calibrating within item banks (domains). The block administration provided data for evaluating associations among domains. Blocks of PROMIS items were administered both to general population and clinical samples. The sampling design ensured that each item was administered to at least 900 respondents from the general population (some of whom reported having chronic medical conditions) and 500 respondents with known chronic medical conditions.

Most of the response data were collected by YouGovPolimetrix (www.polimetrix.com; also see www.pollingpoint.com), a polling firm based in Palo Alto, CA. YouGovPolimetrix operates PollingPoint.com, a centralized portal that allows interested individuals to provide their views about public policy and other current issues. The respondents for a typical YouGovPolimetrix Internet survey are selected from the PollingPoint panel, a panel of more than 1 million respondents who have provided YouGovPolimetrix with their names, street addresses, e-mail addresses, and other information, and who regularly participate in online surveys. Panelists were recruited by a variety of methods, including e-random digit dialing, invitations by means of web newsletters, and Internet poll–based recruitment, where panelists have opted to participate in a survey advertised on the World Wide Web. Panel members receive modest compensation (less than $10 value) when they participate.

YouGovPolimetrix uses a sample-matching procedure to select representative samples. The sample-matching algorithm starts with a listing of all respondents in the desired or target population. Next, a random sample of the desired size is selected from the population listing (the “target sample”). Third, for each element of the target sample, the closest match is selected from the PollingPoint panel. This method has been shown to give accurate results in a wide variety of contexts, even for groups significantly underrepresented on the Internet [37]. The validity of the approach depends on the panel being sufficiently large and diverse, not on Internet usage or other types of behaviors. For PROMIS, we specified targets in terms of gender (50% female), age (20% in each of five age groups: 18–29, 30–44, 45–59, 60–74, older than 75 years), race/ethnicity (12.3% African American and 12.5% Latino/Hispanic to match the U.S. census), and education (10% less than high school graduate). To supplement these specifications, we developed a subset representative of the U.S. general population [38].

1.6. Wave 1 sample

The PROMIS wave 1 sample included 21,133 respondents. Of these, 1,532 were recruited from primary research sites associated with PROMIS network sites and the remainder (19,601) from YouGovPolimetrix’s panel sample. Figure 2 describes the samples. These are broken down by source and type of respondent (clinical vs. general population). The PROMIS steering committee chose to anchor
the calibration of the first wave of PROMIS items on the U.S. population (unselected for any specific health problem). Therefore, all full-bank respondents were drawn from nonclinical samples, which we refer to as “general population.” The clinical population supplied by YouGovPolimetrix for block testing was identified through a presurvey of 250,000 YouGovPolimetrix panel members. These respondents completed the PROMIS clinical form described earlier. Persons were included in the clinical sample associated with a particular condition if they reported having received the diagnosis from a physician. The general population sample included people with reported conditions. They were administered the clinical form, but their responses did not exclude them from participation in the general population sample.

YouGovPolimetrix sample data were collected using their Web site on a secure server. PROMIS network site data were collected using a Web-based platform created by PROMIS. On completion of data collection, the PROMIS Statistical Coordinating Center received de-identified data sets from YouGovPolimetrix. Full banks were administered to 7,005 individuals (6,676 from YouGovPolimetrix, 236 from University of North Carolina, and 93 from Stanford University). Block administration included 14,128 individuals (6,245 from general population and 7,883 from clinical samples). The clinical samples 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 disease (n = 1,214), spinal cord injury (n = 531), and other conditions (n = 560). Participants with comorbidities were included. Figure 2 details which of these clinical samples came from each of the PROMIS sites.

1.6.1. Sample demographics

The overall sample (n = 21,133) was 52% female. The median age was approximately 50 years. The breakdown by age range was as follows: 18–29, 12%; 30–39, 12%; 40–49, 16%; 50–64, 32%; and 65 and older, 28%. Eighty-two percent were whites, 9% blacks, 8% multiracial, and 1% others (Asian/Pacific Islanders and Native Americans). The sample was 9% Latino/Hispanic. Highest educational attainment of the participants included 3% less than high school, 16% with terminal high school diploma, 39% with some college but no degree, 24% with a college degree, and 19% with a post-baccalaureate degree. The combined sample was used primarily for calibrating item parameters and setting the optimum location for establishing the midpoints of the score range for each calibrated item bank when it was time to derive scores. This would enable the comparison of item bank scores with general-population benchmark values.

1.6.2. Scale-setting subsample

Calibrations of scores based on IRT models yield scores in logits and typically range from around −4 to +4. Most researchers apply a linear transformation to scores (e.g., to create an approximate range of 0–100). PROMIS investigators decided that all PROMIS measures would use the T-score metric [39], in which scores have a mean of 50 and a standard deviation (SD) of 10 compared with the general population. For example, a person who has a PROMIS-pain interference score of 70 is reporting adverse pain interference 2 SDs worse than the general-population average.

The scale-setting PROMIS wave 1 general-population sample was obtained to represent the marginal distributions of race/ethnicity (white vs. black, Latino/Hispanic, others) and education (high school or less vs. more than high school) as reflected in the 2000 U.S. census [38]. The percentages by gender, age, race, and education in the 2000 census were as follows: 52% female; 22% aged 18–29 years, 32% aged 30–44 years, 24% aged 45–59 years, 14% aged 60–74 years, and 8% aged 75 years and older; 74% white, 11% black, 11% Latino/Hispanic, and 4% other; and 51% more than high school. The distribution of characteristics for the PROMIS scale-setting subsample (n = 5,239) were as follows: 57% female; 15% aged 18–29 years, 22% aged 30–44 years, 28% aged 45–59 years, 22% aged 60–74 years, and 13% aged 75 years and older; 74% white, 10% black, 11% Latino/Hispanic, and 4% other; 51% had more than a high school education.

1.6.3. Additional chronic pain sample

The distribution of pain in the PROMIS wave 1 data proved highly skewed, because few people reported moderate to severe pain. We were concerned that item calibrations from the available data would be unreliable, and the full continuum of pain severity would not be precisely measured, particularly in the moderate-to-severe pain range. Therefore, we collected additional pain item responses from individuals with chronic pain. These respondents were recruited by Web site invitation in collaboration with the American Chronic Pain Association (ACPA). To be eligible, participants had to be 21 years of age or older and have at least one chronic pain condition for at least 3 months before participating in the survey. Those who met eligibility criteria provided Institutional Review Board–approved, online informed consent. The survey was posted on the Web site of the ACPA from September 2007 to March 2008.

The 967 participants responded to 47 pain interference, 42 pain behavior, and 41 pain quality items, and one global average pain intensity item through online administration (some of the 56 items in the original candidate bank were dropped based on preliminary psychometric analyses). The average age of the chronic pain sample was 48.2 years (SD = 11.1). Eighty-one percent were females, 91% were whites, 1.5% were blacks, and 5% were Latino/Hispanics. Eighty-one percent of the participants had a high school education or greater. The data were combined with wave 1 full-bank data to calculate pain item calibrations for the pain item banks.
1.6.4. Sleep sample

Respondents for the sleep disturbance and sleep-related impairment items were collected by the University of Pittsburgh research site as an independent research project. A total of 128 sleep disturbance/sleep-related impairment items were administered to 1,993 individuals from YouGovPolimetrix (1,259 from general population and 734 with self-identified sleep problem). Clinical sites at Pittsburgh collected responses from 259 individuals with sleep disorders. The overall sample \( n = 2,252 \) was 44% female. The median age was 52 years; 21% of these were 65 years and older. Eighty-two percent were whites, 13% blacks, 3% Native Americans or Alaskans, 0.4% Native Hawaiians or Pacific Islanders, and 6% others. Ten percent of the sample was Latino/Hispanic. Distribution of educational attainment was 14% with high school or less, 39% with some college, 28% with a college degree, and 20% with an advanced degree. Item-response data from the overall sample (2,252 individuals) were used for item calibration.

1.7. Analysis plan and item calibrations

Data analyses were driven by a statistical analysis plan [5] for evaluating IRT modeling assumptions (unidimensionality and local dependence), IRT model fit, monotonicity, scalability, item fit, and differential item functioning (DIF). To aid decisions regarding item bank composition, statistical and psychometric results were provided to the domain teams responsible for the development of each bank. These results were discussed, and decisions were made regarding each item. Typically, a first wave of item “cuts” was made; that is, the most problematic items were eliminated from further consideration (e.g., outside of concept; problematic item wording). The result of the analyses described earlier was a set of item-by-item analysis and discussion with content (domain) experts, regarding each item. Through this process of iterative decision making, the most problematic items were excluded from further analyses to help arrive at decisions regarding item bank composition. (1) Not calibrated but retained for possible future calibration (e.g., items consistent with the domain being measured but having local dependence; responses concentrated in few of the available response options); or (3) excluded from further consideration (e.g., outside of concept; problematic item wording).

2. Results

The result of the analyses described earlier was a set of 11 calibrated item banks that would support computerized adaptive testing (CAT) and development of multiple short forms of varying length [27]. A version 1.0 short form ranging from 6 to 10 items was created from each item bank. Items that represented the range of item bank content and difficulty, had high information, and no evidence of DIF were selected. PROMIS item banks and short forms available since December 2008 are listed in Table 1, along with the correlation between the short form and the entire bank. All instruments can be accessed within Assessment Center (https://www.assessmentcenter.net).

2.1. Validity and interpretation tables

Initial evidence in support of the reliability and validity of IRT-derived summary scores for PROMIS item banks and scales is provided in Tables 1–9. Table 1 reports correlations between scores on the PROMIS full-item banks and scores on the short forms from each domain. With the exception of fatigue, which was developed to sample across content without regard for degree of information provided by each item, all correlations \( r \) were greater than 0.95. This suggests that the short form is reliably measuring the same thing as the item bank from which it was drawn. Table 2 provides each item bank’s standard error and reliability coefficients, by \( T \)-scores, from the wave 1 data collection. Reliability (defined here as measurement precision along the continuum) remained high for all banks from scores at the mean to 2 or more SD units worse than the mean. Table 3 displays the calibration sample \( T \)-score means, SDs, and distributions by percentile. The consistently low standard errors across most the measurement continuum provides confidence in the precision of score estimates, even at the individual level.

Tables 4–9 provide construct validity information based on correlations between scores on item banks, item bank short forms, and legacy measures included in wave 1 testing. The original physical functioning item pool was too long (56 × 3 = 168 items) to administer in total to any one person; hence, it was split into two separate “full-bank” administrations (112 in one set and 56 in another). Therefore, none of the participants answered the complete set of all physical function items in the pool. To estimate correlations between full-bank information and the developed short form, participants were required to respond to at least 93 of 124 items to calculate a full-bank score and 7 out of 10 items to estimate a short-form score. The item parameter estimations were done like all other item banks using the complete information from the block and the full-bank data as described in the analysis plan [5].

2.1.1. Physical function

The physical function item bank is the largest PROMIS bank at 124 calibrated items, including a 10-item short form (Table 4). The full bank is correlated at \( r = 0.96 \) with the short form and –0.80 to –0.88 with legacy measures (Health Assessment Questionnaire and SF-36, respectively). The full bank’s reliability is greater than 0.96 for scores 4 SDs less than the mean (poor functioning) to 1 SD greater than the mean.
2.1.2. Fatigue

The fatigue item bank consists of 95 items assessing the intensity, frequency, and impact of fatigue (Table 5). A seven-item short form [40], created to sample from both fatigue experience and interference, correlated with the full bank at $r = 0.76$. The reliability of measurement was greater than 0.91 for scores ranging from 2 SDs less than the mean to 4 SDs greater than the mean. The fatigue item pool was tested with the Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue Scale, and they were correlated at 0.95. Some of the FACIT-Fatigue items were included in the final calibrated item bank. This calibrated bank had a correlation of 0.89 with the SF-36 Vitality Scale.

### Table 1

<table>
<thead>
<tr>
<th>Item bank</th>
<th>$N$</th>
<th>No. items in full bank</th>
<th>No. items in short form</th>
<th>Correlation with short form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional distress—anger</td>
<td>858</td>
<td>29</td>
<td>8</td>
<td>0.96</td>
</tr>
<tr>
<td>Emotional distress—anxiety</td>
<td>788</td>
<td>29</td>
<td>7</td>
<td>0.96</td>
</tr>
<tr>
<td>Emotional distress—depression</td>
<td>782</td>
<td>28</td>
<td>8</td>
<td>0.96</td>
</tr>
<tr>
<td>Fatigue</td>
<td>9,047</td>
<td>95</td>
<td>7</td>
<td>0.76</td>
</tr>
<tr>
<td>Pain behavior</td>
<td>794</td>
<td>39</td>
<td>7</td>
<td>0.98</td>
</tr>
<tr>
<td>Pain interference</td>
<td>796</td>
<td>41</td>
<td>6</td>
<td>0.95</td>
</tr>
<tr>
<td>Physical function*</td>
<td>737</td>
<td>124</td>
<td>10</td>
<td>0.96</td>
</tr>
<tr>
<td>Satisfaction with participation in discretionary social activities</td>
<td>729</td>
<td>12</td>
<td>7</td>
<td>0.99</td>
</tr>
<tr>
<td>Satisfaction with participation in social roles</td>
<td>726</td>
<td>14</td>
<td>7</td>
<td>0.99</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>2,252</td>
<td>27</td>
<td>8</td>
<td>0.96</td>
</tr>
<tr>
<td>Sleep-related impairment</td>
<td>2,252</td>
<td>16</td>
<td>8</td>
<td>0.98</td>
</tr>
<tr>
<td>Global health</td>
<td>N/A</td>
<td>10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Abbreviation: N/A, not available.

* Because of the study design, none of the participants answered all 124 physical function items or all short-form items. To estimate a correlation between the full bank and the short form, a participant needed to provide complete responses to at least 93 out of 124 items for the full bank and 7 out of 10 items for the short form.

### Table 2

<table>
<thead>
<tr>
<th>Item bank</th>
<th>$N$</th>
<th>SE</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional distress—anger</td>
<td>856</td>
<td>SE</td>
<td>0.93</td>
<td>0.76</td>
<td>0.45</td>
<td>0.23</td>
<td>0.17</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.13</td>
<td>0.43</td>
<td>0.80</td>
<td>0.95</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.93</td>
</tr>
<tr>
<td>Emotional distress—anxiety</td>
<td>788</td>
<td>SE</td>
<td>0.98</td>
<td>0.90</td>
<td>0.62</td>
<td>0.26</td>
<td>0.14</td>
<td>0.13</td>
<td>0.13</td>
<td>0.15</td>
<td>0.33</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.04</td>
<td>0.19</td>
<td>0.62</td>
<td>0.93</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.89</td>
</tr>
<tr>
<td>Emotional distress—depression</td>
<td>782</td>
<td>SE</td>
<td>1.00</td>
<td>0.97</td>
<td>0.73</td>
<td>0.27</td>
<td>0.13</td>
<td>0.11</td>
<td>0.12</td>
<td>0.16</td>
<td>0.45</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.06</td>
<td>0.47</td>
<td>0.92</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
<td>0.80</td>
</tr>
<tr>
<td>Fatigue</td>
<td>781</td>
<td>SE</td>
<td>0.86</td>
<td>0.58</td>
<td>0.23</td>
<td>0.09</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.11</td>
<td>0.31</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.66</td>
<td>0.95</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>Pain behavior</td>
<td>820</td>
<td>SE</td>
<td>1.00</td>
<td>0.96</td>
<td>0.51</td>
<td>0.11</td>
<td>0.10</td>
<td>0.08</td>
<td>0.08</td>
<td>0.21</td>
<td>0.78</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.08</td>
<td>0.74</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.96</td>
<td>0.40</td>
</tr>
<tr>
<td>Pain interference</td>
<td>844</td>
<td>SE</td>
<td>1.00</td>
<td>1.00</td>
<td>0.94</td>
<td>0.48</td>
<td>0.10</td>
<td>0.07</td>
<td>0.07</td>
<td>0.19</td>
<td>0.58</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.00</td>
<td>0.01</td>
<td>0.11</td>
<td>0.77</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.97</td>
<td>0.66</td>
</tr>
<tr>
<td>Physical function</td>
<td>1,700</td>
<td>SE</td>
<td>0.13</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.09</td>
<td>0.21</td>
<td>0.41</td>
<td>0.68</td>
<td>0.93</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.96</td>
<td>0.83</td>
<td>0.54</td>
<td>0.14</td>
</tr>
<tr>
<td>Satisfaction with participation in discretionary social activities</td>
<td>814</td>
<td>SE</td>
<td>0.99</td>
<td>0.83</td>
<td>0.30</td>
<td>0.14</td>
<td>0.14</td>
<td>0.16</td>
<td>0.58</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.02</td>
<td>0.32</td>
<td>0.91</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.67</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Satisfaction with participation in social roles</td>
<td>816</td>
<td>SE</td>
<td>0.98</td>
<td>0.73</td>
<td>0.21</td>
<td>0.13</td>
<td>0.14</td>
<td>0.18</td>
<td>0.66</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.03</td>
<td>0.47</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.57</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>2,252</td>
<td>SE</td>
<td>0.91</td>
<td>0.68</td>
<td>0.34</td>
<td>0.20</td>
<td>0.16</td>
<td>0.16</td>
<td>0.17</td>
<td>0.28</td>
<td>0.49</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.16</td>
<td>0.54</td>
<td>0.88</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.92</td>
<td>0.76</td>
</tr>
<tr>
<td>Sleep-related impairment</td>
<td>2,252</td>
<td>SE</td>
<td>0.91</td>
<td>0.68</td>
<td>0.41</td>
<td>0.31</td>
<td>0.18</td>
<td>0.17</td>
<td>0.17</td>
<td>0.29</td>
<td>0.63</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td>0.17</td>
<td>0.53</td>
<td>0.84</td>
<td>0.91</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.92</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Higher scores indicate more of that domain. Therefore, high scores for anger, anxiety, depression, fatigue, pain behavior, pain interference, sleep disturbance, and sleep-related impairment indicate worse functioning or more symptoms. High scores for physical function, satisfaction with participation in discretionary social activities, and satisfaction with participation in social roles indicate better functioning or more satisfaction. A $T$-score distribution has a mean of 50 and standard deviation of 10. A 50 score here is the mean of the U.S. general population based on 2000 census data (see Liu et al. [38], this issue).

Sample size ($N$) refers to the number of people who completed full-bank testing from wave 1 (Fig. 2).

Abbreviation: SE, standard error.
2.1.3. Pain

Two pain banks, pain behavior and pain interference, were created (Table 6). Pain intensity and quality were not calibrated as banks, but one item from the Global Health Scale reflecting pain intensity was used in analyses (Table 6). The final pain behavior item bank contains 39 items covering different pain-related behaviors [35]. A seven-item short-form pain behavior scale is available for research studies. The short-form scale has a correlation of 0.98 with the full-item bank. For the full-item bank, reliability is 0.90 or greater across most of the score distribution, and the short-form and CAT scales have reliabilities exceeding 0.80 across most of the score distribution. Pain behavior scores have a correlation of 0.77 with pain interference scores and 0.69 with a pain intensity score (Table 6). Mean pain behavior scores vary significantly by levels of pain intensity (P!0.0001) and global health status (P<0.0001) [35].

The pain interference bank consists of 41 items assessing the extent to which pain interferes with functioning. A six-item short form is also available and has a correlation of 0.95 with the full-item bank. Responses to the items of the final bank were strongly unidimensional (e.g., ratio of first and second eigenvalue of 35), and all items had good fit to the graded response model. Nine items exhibited statistically significant DIF, one with respect to gender and the others with respect to age. However, adjusting for DIF had little practical impact on score estimates. Scores provided substantial information across levels of pain interference observed in the wave 1 and supplementary pain data. Full-bank reliability is 0.97 or greater for scores greater than or equal to the mean. Reliability is 0.77 for scores 1 SD below the mean (less pain interference). Pain interference scores discriminated among persons with different numbers of chronic conditions, disabling conditions, and levels of self-reported health (P<0.0001). Patterns of correlations with other health outcomes supported the construct validity of the item bank (r=0.81 with Brief Pain Inventory severity; r=0.85 with Brief Pain Inventory interference; r=-0.86 SF-36 Bodily Pain Scale). Pain

### Table 3

<table>
<thead>
<tr>
<th>Item bank</th>
<th>No. Items</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>P5</th>
<th>P10</th>
<th>P25</th>
<th>P50</th>
<th>P75</th>
<th>P90</th>
<th>P95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional distress—anger</td>
<td>29</td>
<td>858</td>
<td>47.2</td>
<td>9.4</td>
<td>27.9</td>
<td>34.9</td>
<td>41.3</td>
<td>47.3</td>
<td>53.6</td>
<td>58.5</td>
<td>62.3</td>
</tr>
<tr>
<td>Emotional distress—anxiety</td>
<td>29</td>
<td>788</td>
<td>48.5</td>
<td>9.8</td>
<td>31.6</td>
<td>35.2</td>
<td>41.8</td>
<td>48.2</td>
<td>54.6</td>
<td>61.6</td>
<td>65.5</td>
</tr>
<tr>
<td>Emotional distress—depression</td>
<td>28</td>
<td>782</td>
<td>49.3</td>
<td>9.6</td>
<td>33.5</td>
<td>37.7</td>
<td>42.4</td>
<td>48.6</td>
<td>55.3</td>
<td>62.0</td>
<td>66.0</td>
</tr>
<tr>
<td>Fatigue</td>
<td>95</td>
<td>14,931</td>
<td>51.2</td>
<td>9.4</td>
<td>36.5</td>
<td>30.0</td>
<td>44.3</td>
<td>50.8</td>
<td>57.8</td>
<td>64.0</td>
<td>67.3</td>
</tr>
<tr>
<td>Pain behavior Silence</td>
<td>39</td>
<td>15,834</td>
<td>55.6</td>
<td>8.2</td>
<td>36.2</td>
<td>39.3</td>
<td>52.5</td>
<td>57.3</td>
<td>61.4</td>
<td>63.9</td>
<td>65.2</td>
</tr>
<tr>
<td>Pain interference</td>
<td>41</td>
<td>15,903</td>
<td>55.9</td>
<td>10.8</td>
<td>40.0</td>
<td>40.0</td>
<td>47.5</td>
<td>55.2</td>
<td>65.3</td>
<td>70.3</td>
<td>72.8</td>
</tr>
<tr>
<td>Physical function</td>
<td>124</td>
<td>124</td>
<td>50.0</td>
<td>10.0</td>
<td>27.6</td>
<td>37.0</td>
<td>47.4</td>
<td>53.4</td>
<td>56.9</td>
<td>58.3</td>
<td>58.8</td>
</tr>
<tr>
<td>Satisfaction with participation in discretionary social activities</td>
<td>12</td>
<td>712</td>
<td>49.3</td>
<td>10.7</td>
<td>32.1</td>
<td>36.6</td>
<td>42.0</td>
<td>48.5</td>
<td>55.1</td>
<td>68.9</td>
<td>68.9</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>27</td>
<td>2,252</td>
<td>49.8</td>
<td>10.3</td>
<td>33.4</td>
<td>36.5</td>
<td>42.1</td>
<td>49.4</td>
<td>57.2</td>
<td>63.4</td>
<td>66.5</td>
</tr>
<tr>
<td>Sleep-related impairment</td>
<td>16</td>
<td>2,252</td>
<td>50.0</td>
<td>9.8</td>
<td>34.6</td>
<td>37.4</td>
<td>43.0</td>
<td>49.5</td>
<td>56.9</td>
<td>63.2</td>
<td>66.7</td>
</tr>
</tbody>
</table>

T-Score means, SDs, and T-scores by percentile are computed for the full-calibration sample to describe this sample. The normative sample from which the T-scores were derived was based on a subsample of the full-calibration sample matched to U.S. census demographics. Therefore, the means and SDs for the full sample are not 50 and 10, respectively, and the values reported for the full-calibration sample should not be used as norms (see Liu et al. [38], this issue).

**Abbreviation:** SD, standard deviation.

### Table 4

<table>
<thead>
<tr>
<th>Measure</th>
<th>HAQ 1</th>
<th>SF36-PF 2</th>
<th>Physical Function 3</th>
<th>N 4</th>
<th>Correlation with Short Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAQ</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td>-0.88</td>
</tr>
<tr>
<td>SF36-PF</td>
<td></td>
<td>-0.80</td>
<td>1.00</td>
<td>711</td>
<td>-0.88</td>
</tr>
<tr>
<td>Physical Function Bank</td>
<td></td>
<td>-0.88</td>
<td>0.88</td>
<td>1.00</td>
<td>708</td>
</tr>
</tbody>
</table>

1 Health Assessment Questionnaire (HAQ)-Disability Scale (scored without devices & aids)
2 SF-36 Physical Function Scale
3 Physical Function Bank (at least 93 items non missing)
4 N=sample size

Correlations are all based on full-bank data as legacy scales were only administered on FORM-D based on full-bank data
2.1.4. Sleep disturbance and sleep-related impairments

The sleep disturbance item bank includes 27 items reflecting difficulties with sleep, whereas the 16-item sleep-related impairment bank consists of items capturing the negative daytime consequences of poor sleep (e.g., cognitive and emotional problems, feeling sleepy) (Table 7). The banks are correlated at $r = 0.75$. Each bank has an eight-item short form. The sleep disturbance short form has a correlation of 0.96 with the full bank. The bank’s reliability is greater than 0.88 across most of the score distribution. It is correlated at $r = 0.85$ with the Pittsburgh Sleep Quality Index and $r = 0.25$ with the Epworth Sleepiness Scale. The sleep-related impairment short form is correlated with the full-wake bank at 0.98. The reliability is greater than 0.84 across most of the distribution. It is correlated with the Pittsburgh Sleep Quality Index at $r = 0.70$ and the Epworth Sleep Quality Index at $r = 0.45$.

2.1.5. Anger, anxiety, and depression

The emotional distress domain includes final item banks for anger, anxiety, and depression (Table 8). The anger

---

Table 5

**Fatigue: correlations with legacy measures and short forms**

<table>
<thead>
<tr>
<th>Measure</th>
<th>FACIT¹</th>
<th>SF36-VT²</th>
<th>Fatigue³</th>
<th>Fatigue⁴</th>
<th>N³</th>
<th>Correlation with Short Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACIT</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF36-VT</td>
<td>0.88</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fatigue Bank³</td>
<td>0.96</td>
<td>0.89</td>
<td>1.00</td>
<td></td>
<td>9047.00</td>
<td>0.76</td>
</tr>
<tr>
<td>Fatigue Bank⁴</td>
<td>0.95</td>
<td>0.88</td>
<td>1.00</td>
<td>1.00</td>
<td>9047.00</td>
<td>0.76</td>
</tr>
</tbody>
</table>

¹FACIT-Fatigue Scale  
²SF36 Vitality Scale  
³Including legacy items (i.e., FACIT-F)  
⁴Excluding legacy items (i.e., FACIT-F)  
⁵N=sample size  
Correlations are all based on full-bank data only

---

Table 6

**Pain: correlations between Patient-Reported Outcomes Measurement Information System banks, legacy measures, and short forms**

<table>
<thead>
<tr>
<th>Measure</th>
<th>BPI¹ (Severity)</th>
<th>BPI¹ (Interference)</th>
<th>SF36-BP²</th>
<th>Pain Behavior³</th>
<th>Pain Interference³</th>
<th>N³</th>
<th>Correlation with Pain Behavior Short Form⁴</th>
<th>N³</th>
<th>Correlation with Pain Interference Short Form⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPI (Severity)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPI (Interference)</td>
<td>0.84</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF36-BP</td>
<td>-0.85</td>
<td>-0.84</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain Behavior Bank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain Interf. Bank</td>
<td>0.81</td>
<td>0.85</td>
<td>-0.86</td>
<td>0.77</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain Intensity Item</td>
<td>0.83</td>
<td>0.74</td>
<td>-0.78</td>
<td>0.69</td>
<td>0.76</td>
<td>794</td>
<td>0.60</td>
<td>795</td>
<td>0.69</td>
</tr>
</tbody>
</table>

¹Brief Pain Inventory  
²SF36 Bodily Pain Scale  
³N=sample size  
⁴Based on calibrations drawn from general population and a chronic pain community sample

---

Other notes:
- Correlations are all based on full-bank data only.
- Data not available is indicated.
- Based on block-testing data and full-bank data.
bank’s 29 items and the eight-item short form have a correlation of 0.96. The full bank’s reliability is greater than 0.93 across most of the score distribution. Anger bank scores have a correlation of 0.59 with the anxiety bank and 0.60 with the depression bank. The correlation with the Aggression Questionnaire was 0.51. The final anxiety bank included 29 calibrated items with a seven-item short form that together correlated at 0.96. Reliability was greater than 0.89 for most of the score distribution. The anxiety bank had a correlation of 0.81 with the depression bank and 0.59 with the anger bank. Correlations with legacy measures were strong (r = 0.80 with Mood and Anxiety Symptom Questionnaire; r = 0.75 with the Center for Epidemiological Studies Depression Scale). The depression bank (28 items) also had a high correlation (0.96) with its eight-item short form. The reliability was greater than 0.92 for most of the score distribution. In addition to the correlations with other emotional distress banks described earlier, it correlated strongly with legacy measures (r = 0.83 with the Center for Epidemiological Studies Depression Scale; r = 0.72 with the Mood and Anxiety Symptom Questionnaire).

2.1.6. Social health

After analyses, two item banks were constructed from the social health satisfaction item pool (Table 9). These were satisfaction with participation in discretionary social activities (e.g., leisure, recreation) and satisfaction with participation in social roles (e.g., family, household responsibilities, work). The banks are the smallest at 12 and 14 items, respectively, and are correlated with each other at 0.83. Short forms of seven items each were constructed and they correlated at 0.99 with their respective item bank. Reliability between 2 SDs less than the mean (poorer satisfaction) and 1 SD greater than the mean was greater than 0.91 for satisfaction with participation in discretionary

Table 7
Sleep and sleep-related impairment: correlations between Patient-Reported Outcomes Measurement Information System banks, legacy measures, and short forms

<table>
<thead>
<tr>
<th>Measure</th>
<th>PSQI^1</th>
<th>ESS^2</th>
<th>Sleep Dist</th>
<th>Wake Dist</th>
<th>N^3</th>
<th>Correlation with Short Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSQI</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESS</td>
<td>0.28</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleep Dist. Bank</td>
<td>0.85</td>
<td>0.25</td>
<td>1.00</td>
<td></td>
<td>2252</td>
<td>0.96</td>
</tr>
<tr>
<td>Wake Dist. Bank</td>
<td>0.7</td>
<td>0.45</td>
<td>0.75</td>
<td>1.00</td>
<td>2252</td>
<td>0.98</td>
</tr>
</tbody>
</table>

^1 Pittsburgh Sleep Quality Index  
^2 Epworth Sleepiness Scale  
^3 N = sample size  

based on full-bank data

Table 8
Anger, anxiety, and depression: correlations between Patient-Reported Outcomes Measurement Information System banks, legacy measures, and short forms

<table>
<thead>
<tr>
<th>Measure</th>
<th>AQ^1</th>
<th>CES-D^2</th>
<th>MASQ^3</th>
<th>Anger</th>
<th>Anxiety</th>
<th>Depression</th>
<th>N^4</th>
<th>Correlation with Short Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQ</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CES-D</td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MASQ</td>
<td></td>
<td></td>
<td>0.78</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger Bank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.51</td>
<td>1.00</td>
<td></td>
<td>858</td>
</tr>
<tr>
<td>Anxiety Bank</td>
<td></td>
<td></td>
<td></td>
<td>0.75</td>
<td>0.80</td>
<td>0.59</td>
<td>1.00</td>
<td>788</td>
</tr>
<tr>
<td>Depression Bank</td>
<td></td>
<td></td>
<td>0.83</td>
<td>0.72</td>
<td>0.60</td>
<td>0.81</td>
<td>1.00</td>
<td>782</td>
</tr>
</tbody>
</table>

^1 Aggression Questionnaire  
^2 Center for Epidemiological Studies Depression Scale  
^3 Mood and Anxiety Symptom Questionnaire  
^4 N = sample size  

Based on full-bank data

Based on block-testing data  
Based on full-bank data
social activities. Outside of that range, reliability was lower. A similar pattern exists for satisfaction with social roles with reliability greater than 0.96 for the same range. Several legacy items were administered, including the FACIT-Functional Well-Being Scale and the SF-36 Role Physical, Role Emotional, and Social Functioning Scales. For satisfaction with participation in social roles bank, correlations with the SF-36 scales ($r = 0.57$–$0.59$) were less than the FACIT-Functional Well-Being Scale ($r = 0.76$). For satisfaction with discretionary social activities, correlations with the SF-36 ranged from 0.44 (Role Physical) to 0.53 (Social Functioning). The correlation with the FACIT-Functional Well-Being Scale was 0.76.

3. Discussion

PROMIS provides item banks that offer the potential for efficient (minimizes item number without compromising reliability), flexible (enables optional use of interchangeable items), and precise (has minimal error in estimate) measurement of commonly studied PROs. We summarized the domain framework, definitions, and sampling plan that guided the development, testing, and calibration of the first (version 1.0) PROMIS item banks. Item calibrations and statistics are available on the PROMIS Web site through Assessment Center (www.assessmentcenter.net/ac1). Item bank and short-form reliabilities, and their correlations with one another, are presented in Tables 1 and 2. Item bank score distributions for the entire calibration sample, presented in Table 3, form a useful basis for comparison for future research efforts. Further details are available on the PROMIS Web site (www.nihpromis.org).

Initial evidence in support of construct validity (comparison with legacy measures) is presented in Tables 4–9. As of 2008, there are 11 item banks available for public use. Data based on these items proved to have sufficient unidimensionality to be treated as a measure of a single defined concept. We describe the calibration sample for these item banks. In all, 912 items were tested, and based on analysis of responses from 21,133 people, 454 items became part of these calibrated banks. From the 11 banks, we derived version 1.0 static short-form measures for each domain and have preliminary evidence supporting the reliability and construct validity of these item banks. Numerous additional study-tailored short forms can be derived from a single bank to accommodate the special needs or preferences of individual investigators. For each PROMIS short form, a scoring table has been developed to associate short-form scores onto a $T$-score metric, which is referenced to (and centered on) the U.S. general population (see Liu et al. [38]). In addition, each of the PROMIS item banks can be administered using a CAT, in which the assessment is tailored to each individual based on responses to previously administered items. CAT administration reduces test length dramatically without compromising measurement precision [27]. Based on wave 1 testing experience, respondents completed an average of five items per minute, suggesting, for example, that a CAT administration of all 11 banks, with an average of five items administered per bank, would take about 11 minutes to complete. CAT simulations in support of this degree of measurement efficiency have been published on PROMIS banks [41]. By design, CAT administration assumes that the order of administration of items does not have a substantial effect on the score derived at the end of the administration. Although not the usual way
that HRQL instruments have been applied in clinical research or practice, this assumption is testable, and initial simulation studies have been very encouraging regarding the absence of effect of item selection or order administration on the derived score [41].

The PROMIS Cooperative Group developed and tested several hundred items measuring the 11 health domains described in this article. These core PROMIS domains reflect common, generic symptoms and experiences that likely apply to people in a variety of contexts or with a variety of diseases. In each case, items were worded so that a given respondent, with or without a given health condition, could respond. None of these questions carry attributions to a specific condition, although some do refer to specific symptoms, such as pain or fatigue. These item banks, therefore, permit a wide range of respondents to report their symptoms, function, or health perceptions, without needing to make attributions to a specific condition. This approach has the advantage of working in populations with multiple chronic illnesses and allows comparability of experiences across diseases. These banks are not intended to differentiate different subtypes of a symptom, to the extent they might exist (e.g., fatigue from fibromyalgia vs. fatigue from multiple sclerosis). Instead, they aim to differentiate severity levels of the symptom or functional ability. In all cases, the assumption of universality of these banks is testable by evaluating DIF [42,43] across diseases or other contexts.

This is testable in future research. Current analyses of available data, and future research in this area, will help determine the extent to which the generic symptoms and functional reports made possible by these item banks are generalizable. In cases where generalizability is compromised, those items that demonstrate DIF can be removed or recalibrated to apply to the specific disease or context [44]. When this is done, the same metric for each of these 11 domains can be applied after recalibrating affected items as needed and modifying how they contribute to the standard score. Identifying the extent of generalizability of these banks across diseases, and DIF-based item recalibrations to retain comparability across diseases where possible, will be a major research emphasis of the PROMIS network and, we hope, others, in the future.

These initial PROMIS item banks have demonstrated reliability, precision, and construct validity based on their correlation with legacy instruments [45–48]. Evidence for validity in longitudinal clinical research (e.g., responsiveness to change) is yet to be demonstrated with PROMIS instruments, but clinical validation studies are underway in PROMIS “wave 2” studies in rheumatoid arthritis, depression, back pain, cancer, heart failure, and chronic obstructive pulmonary disease. However, there is no reason to believe that the PROMIS item banks and derived short-form scales will be any less responsive than the existing legacy measures.

In addition, a 10-item PROMIS short-form global item scale has been developed and tested, and it provides single-item measures of mental and physical health summary scores [29]. This instrument efficiently assesses functioning and well-being and may be most useful for large epidemiologic and observational studies for monitoring or assessing the health of populations. Based on these global items and summary scores [30], the estimated health preference—based scores allow for the calculation of preference scores for application in health economic and comparative effectiveness research.

PROMIS version 1.0 instruments were developed based on data collected on an Internet survey platform. As such, they can be considered appropriate for Internet or personal computer-based applications with screen presentations of individual items. Comparability of results obtained using paper or telephone administration cannot be assumed without further testing. An ongoing PROMIS study is evaluating paper-and-pencil administration and palm device administration of PROMIS measures compared with the currently validated Internet computer interface version. Results from these efforts are forthcoming. Similarly, most PROMIS items use a 7-day recall period, which has been the most common recall period in health status and HRQL questionnaires. Further research is needed to evaluate the validity of this recall period and potential for meaningful bias introduced by the demand/expectation that people can reliably recall experiences over this time frame.

The PROMIS item banks have been released for public use (www.assessmentcenter.net/ac1) to encourage researchers in various settings with a range of patient populations to further validate these banks in specific patient populations. With additional validation work, these banks can provide a common metric of represented constructs across a range of patient groups, reducing the cacophony of disparate measures currently being used in clinical research and allowing researchers to compare these constructs within and across patient groups in different studies.

Acknowledgments

The Patient-Reported Outcomes Measurement Information System (PROMIS) is an NIH Roadmap initiative to develop a computerized system measuring PROs in respondents with a wide range of chronic diseases and demographic characteristics. This work was funded by cooperative agreements to a Statistical Coordinating Center (Northwestern University, PI: David Cella, PhD, U02AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project have included Deborah Ader, PhD, Susan Czajkowski, PhD,
Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Susana Serrate-Sztein, MD, and James Witter, MD, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee before external peer review. See the Web site at www.nihpromis.org for additional information on the PROMIS initiative.

References


