Published in final edited form as:
J Clin Epidemiol. 2010 November ; 63(11): 1179–1194. doi:10.1016/j.jclinepi.2010.04.011.

Initial Adult Health Item Banks and First Wave Testing of the Patient-Reported Outcomes Measurement Information System (PROMIS™) Network: 2005–2008

David Cella, Ph.D.[1], **William Riley, Ph.D.**[2], **Arthur Stone, Ph.D.**[3], **Nan Rothrock, Ph.D.**[1], **Bryce Reeve, Ph.D.**[4], **Susan Yount, Ph.D.**[1], **Dagmar Amtmann, Ph.D.**[5], **Rita Bode, Ph.D.**[1], **Daniel Buysse, M.D.**[6], **Seung Choi, Ph.D.**[1], **Karon Cook, Ph.D.**[5], **Robert DeVellis, Ph.D.**[7], **Darren DeWalt, M.D.**[7], **James F. Fries, M.D.**[8], **Richard Gershon, Ph.D.**[1], **Elizabeth A. Hahn, M.A.**[1], **Jin-Shei Lai, Ph.D.**[1], **Paul Pilkonis, Ph.D.**[6], **Dennis Revicki, Ph.D.**[9], **Matthias Rose, M.D.**[10], **Kevin Weinfurt, Ph.D.**[11], and **Ron Hays, Ph.D.**[12] **on behalf of the PROMIS Cooperative Group**

[1] Northwestern University, Chicago, IL

[2] National Heart, Lung, and Blood Institute, Bethesda, MD

[3] Stony Brook University, Stony Brook, NY

[4] National Cancer Institute, Bethesda, MD

[5] University of Washington, Seattle, WA

[6] University of Pittsburgh, Pittsburgh, PA

[7] University of North Carolina, Chapel Hill, NC

[8] Stanford University, Palo Alto, CA

[9] United BioSource Corporation, Bethesda, MD

[10] Hamburg University, Germany

[11] Duke University, Durham, NC

[12] University of California, Los Angeles, Los Angeles, CA

Abstract

Objective—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 PRO measurement that is *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 sub-sample 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 to the overall bank as well as with other well-validated and widely accepted ("legacy") measures. All item banks demonstrated good reliability across the majority 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 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

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 diseases such as arthritis, multiple sclerosis, cancer and asthma, or conditions characterized by chronic pain and fatigue. Often, the best way patients can judge the effectiveness of treatments is by perceived changes in symptoms, distress or function. In late 2004, a group of scientists from several US-based academic institutions and the National Institutes of Health (NIH) formed a cooperative group funded under the NIH Roadmap for Medical Research Initiative (http://www.nihroadmap.nih.gov) to revolutionize the assessment of patient-reported outcomes for use in clinical research and healthcare 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 the clinical research community, a set of publicly available, efficient and flexible measurements of PROs, including health-related quality of life (HRQL).

This article summarizes PROMIS network research during the period from 2005–2008, which includes six primary research sites and a statistical coordinating center. This summary builds upon a previously-published summary of the processes that defined the activity of PROMIS from 2004–2006.[1] The previous report[1] reviewed the PROMIS conceptual framework and defined the prioritization of patient-reported outcome (PRO) domains to be initially developed by PROMIS. This paper also builds on previous articles that have described the qualitative review process of PROMIS' item pools [2–4] and the proposed quantitative methods [5] to be used to evaluate the large scale data collected by PROMIS for item evaluation and calibration for PROMIS item banks. This paper describes and defines the domains first developed and tested by the PROMIS network, summarizes the sampling strategy used for our first wave of testing, and provides summary data based upon initial item calibrations and U.S. general population PROMIS scores.

Wave One Domains and Definitions

During the first two years of support, the PROMIS Network developed a domain framework (see Figure 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, social). Specific domains selected for development were physical function, fatigue, pain, emotional distress, social function, and global health. The framework in Figure 1 represents the March 2010 version, which has been modified over time based upon empirical results (including some reported herein). Content elaborations on the right half of the figure represent functioning banks (green background), components of functioning banks (grey background), item banks in development (yellow background), and uncalibrated item pools and scales (blue background).

Conceptual definitions that guided the development of the proposed Wave 1 domains were as follows:

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

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]

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 upon physical, mental, and social activities.

Pain: Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or 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 upon physical, mental and social activities, and behaviors one engages in to avoid, minimize, or reduce pain.

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 bank measures the level of waking alertness, sleepiness, and function within the context of overall sleep-wake function.

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

Emotional distress: Emotional distress is an important component of emotional health, 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 versus 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.

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).

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).

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

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]

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 prior to the current study (see 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 sub-components that divide *social roles* such as work and family responsibilities, and more *discretionary social activities* such as leisure activity and relationships with friends.

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, 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]

Item Development

Each domain listed above 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]

Recall Period

Although some data suggest that recall periods beyond one 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 correlated 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 employ 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 while others do not employ a recall period and emphasize current status in general.

Response Options

The majority of the PROMIS items employ 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 upon 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. The 10 PROMIS Global Health items each have 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 employed 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 pre-tested with cognitive interviews to confirm patient comprehension, prior to field testing for item calibration.

Wave 1 Testing

Following 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.

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., 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 (> 1000), it was unreasonable to ask participants to respond to the entire pool of items. We estimated that participants would respond to approximately 4 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 (3 physical functioning banks, anxiety, depression, anger, alcohol abuse, fatigue interference, fatigue experience, social-role performance, social-role satisfaction, pain interference, pain quality, pain behavior). All 56 items for each of 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 7 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, socio-demographic 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, HIV/AIDS, 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 below. 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 over one million respondents who have provided YouGovPolimetrix with their names, street addresses, email 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 via 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 upon the panel being sufficiently large and diverse, not upon Internet usage or other types of behavior. For PROMIS, we specified targets in terms of gender (50% female), age (20% in each of 5 age groups: 18–29, 30–44, 45–59, 60–74, over 75), race/ethnicity (12.3% African American; 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 US general population [38].

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 versus general population). The PROMIS steering committee chose to anchor the calibration of the first wave of PROMIS items on the United States population (unselected for any specific health problem). Therefore, all full bank respondents were drawn from non-clinical samples, which we refer to as "general population." The clinical population supplied by YouGovPolimetrix for the block testing was identified through a pre-survey of 250,000 YouGovPolimetrix panel members. These respondents completed the PROMIS clinical form described above. 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 website on a secure server. PROMIS network site data were collected using a web-based platform created by PROMIS. Upon completion of data collection, the PROMIS Statistical Coordinating Center received de-identified datasets 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 general population, 7,883 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.

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 white, 9% Black, 8% multi-racial, and 1% other (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 came time to derive scores. This would enable comparison of item bank scores to general population benchmark values.

Scale setting sub-sample—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 to 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 relative to the general population. For example, a person who has a PROMIS-Pain Interference score of 70 is reporting adverse pain interference two standard deviations 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 versus Black, Latino/Hispanic, Other) and education (High School or less versus more than high school) as reflected in the 2000 United States census [38]. The percentages by gender, age, race, and education in the 2000 census were: 52% female; 22% 18–29, 32% 30–44, 24% 45–59, 14% 60–74 and 8% 75 and greater years old; 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 sub-sample (n = 5,239) was: 57% female; 15% 18–29, 22% 30–44, 28% 45–59, 22% 60–74 and 13% 75 and greater years old; 74% white, 10% Black, 11% Latino/Hispanic, and 4% other; 51% had more than a high school education.

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 website 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 prior to participating in the survey. Those who met eligibility criteria provided IRB-approved, online informed consent. The survey was posted on the website 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 female, 91% were white, 1.5% were Black, and 5% were Latino/Hispanic. 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.

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 old; 21% of these were 65 and older. Eighty-two percent were white, 13% Black, 3% Native American or Alaskan, 0.4% Native Hawaiian or Pacific Islander, and 6% other. 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.

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 and the reduced-length item pools were subjected to follow up analyses to help arrive at decisions regarding each item. Through this process of iterative analysis and discussion with content (domain) experts, item-by-item level decisions were made as to whether an individual item should be: (1) calibrated and included in the bank, (2) 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).

Results

The result of the analyses described above was a set of 11 calibrated item banks that would support computerized adaptive testing 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[SM] (https://www.assessmentcenter.net).

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 were above r=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 two or more standard deviation units worse than the mean. Table 3 displays the calibration sample T score means, standard deviations and distributions by percentile. The consistently low standard errors across the majority of 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, so 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]

Physical function (Table 4)—The physical function item bank is the largest PROMIS bank at 124 calibrated items including a 10-item short form. 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 above 0.96 for scores four standard deviations below the mean (poor functioning) to one standard deviation above the mean.

Fatigue (Table 5)—The fatigue item bank consists of 95 items assessing the intensity, frequency, and impact of fatigue. A 7-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 above 0.91 for scores ranging from two standard deviations below the mean to four standard deviations above the mean. The fatigue item pool was tested with the 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 was correlated 0.89 with the SF-36 Vitality Scale.

Pain (Table 6)—Two pain banks, pain behavior and pain interference, were created. Pain intensity and quality were not calibrated as banks, but one item from the Global Health scale reflecting pain intensity was utilized in analyses (See Table 6). The final pain behavior item bank contains 39 items covering different pain-related behaviors [35]. A 7-item short form pain behavior scale is available for research studies. The short-form scale is correlated 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 the majority of the score distribution. Pain behavior scores are correlated 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 6-item short form is also available and is correlated 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 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 at or higher than the mean. Reliability is 0.77 for scores one standard deviation 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 interference scores are correlated 0.76 with pain intensity scores.

Sleep disturbance and sleep-related impairment (Table 7)—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). The banks are correlated at $r=0.75$. Each bank has an 8-item short form. The sleep disturbance short form is correlated at 0.96 with the full bank. The bank's reliability is above 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 above 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 $r=0.45$.

Anger, anxiety, and depression (Table 8)—The emotional distress domain includes final item banks for anger, anxiety, and depression. The anger bank's 29 items and the 8-item short form are correlated 0.96. The full bank's reliability is above 0.93 across most of the score distribution. Anger bank scores correlated 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 7-item short form that together correlated 0.96. Reliability was above 0.89 for the majority of the score distribution. The anxiety bank correlated 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 8-item short form. The reliability was above 0.92 for most of the score distribution. In addition to the correlations with other emotional distress banks described above, 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).

Social health (Table 9)—Following analyses, two item banks were constructed from the social health satisfaction item pool. 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 7-items each were constructed and correlate at 0.99 with their respective item bank. Reliability between two

standard deviations below the mean (poorer satisfaction) and one standard deviation above the mean was above 0.91 for satisfaction with participation in discretionary social activities. Outside of that range, reliability was lower. A similar pattern exists for satisfaction with social roles with reliability above 0.96 for the same range. Several legacy items were administered including the FACIT-Functional Well-Being scale, the SF-36 Role Physical, Role Emotional, and Social Functioning scales. For the satisfaction with participation in social roles bank, correlations with the SF-36 scales (r=0.57 to 0.59) were lower 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.

Discussion

The Patient-Reported Outcomes Measurement Information System (PROMIS™) provides item banks that offer the potential for PRO measurement that is *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™ website 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, comprise a useful basis for comparison for future research efforts. Further detail is available on the PROMIS™ website (www.nihpromis.org).

Initial evidence in support of construct validity (comparison with legacy measures) are 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 forms 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 upon) the US General population (See Liu et al paper). [38] In addition, each of the PROMIS item banks can be administered using a computerized adaptive test (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 Wave1 testing experience, respondents completed an average of 5 items per minute, suggesting, for example, that a CAT administration of all 11 banks, with an average of 5 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 order of administration of items does not have a substantial effect upon the score derived at the end of the administration. While 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 absence of effect of item selection or order administration upon the derived score. [41]

The PROMIS Cooperative Group developed and tested several hundred items measuring the 11 health domains described in this paper. 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 or treatment, 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 versus 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 differential item functioning [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 upon their correlation with legacy instruments (Tables 4–8). Evidence for validity in longitudinal clinical research (e.g., responsiveness to change) has 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 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] estimated health-preference based scores thus allowing 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 pen administration as well as palm device administration of PROMIS measures compared to 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 health-related quality of life 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 a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes 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 prior to external peer review. See the web site at www.nihpromis.org for additional information on the PROMIS initiative.

Reference List

1. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. Med Care 2007;45(5 Suppl 1):S3–S11. [PubMed: 17443116]
2. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of Item Candidates: The PROMIS Qualitative Item Review. Med Care 2007;45(5 Suppl 1):S12–S21. [PubMed: 17443114]
3. Castel LD, Williams KA, Bosworth HB, Eisen SV, Hahn EA, Irwin DE, et al. Content validity in the PROMIS social-health domain: A qualitative analysis of focus-group data. Qual Life Res 2008;17(5): 737–49. [PubMed: 18478368]
4. Christodoulou C, Junghaenel DU, DeWalt DA, Rothrock N, Stone AA. Cognitive interviewing in the evaluation of fatigue items: Results from the patient-reported outcomes measurement information system (PROMIS). Qual Life Res 2008;17(10):1239–46. [PubMed: 18850327]
5. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):S22–S31. [PubMed: 17443115]
6. Haley SM, Coster WJ, Binda-Sundberg K. Measuring physical disablement: The contextual challenge. Phys Ther 1994;74(5):443–51. [PubMed: 8171106]
7. Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47 (6):671–84. [PubMed: 7722580]
8. Stewart, AL.; Kamberg, C. Physical functioning. In: Stewart, AL.; Ware, JE., editors. Measuring functioning and well-being: the medical outcomes study approach. Durham, NC: Duke University Press; 1992. p. 86-142.
9. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995;273(1):59–65. [PubMed: 7996652]
10. Fries JF, Bruce B, Bjorner J, Rose M. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. Ann Rheum Dis 2006;65:16–21.
11. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61(1):17–33. [PubMed: 18083459]
12. Glaus, A. Fatigue in patients with cancer: Analysis and assessment. Heidelberg, Germany: Springer-Verlag Berlin; 1998.

13. North American Nursing Diagnosis Association. Nursing diagnoses: Definition and Classification, 1997–1998. Philadelphia, PA: McGraw-Hill; 1996.
14. Stewart, AL.; Hays, RD.; Ware, JE. Measuring functioning and well-being: the medical outcomes study approach. Durham, NC: Duke University Press; 1992. Health perceptions, energy/fatigue, and health distress measures; p. 143-72.
15. Chang H. Cancer pain management. Med Clin North Am 1999;83(3):711–36. [PubMed: 10386122]
16. Merskey, H.; Bogduk, N. Classification of chronic pain: Descriptions of chronic pain syndromes and definitions of pain terms. Seattle, WA: IASP Press; 1994.
17. Meuser T, Pietruck C, Radbruch L, Stute P, Lehmann KA, Grond S. Symptoms during cancer pain treatment following WHO-guidelines: a longitudinal follow-up study of symptom prevalence, severity and etiology. Pain 2001;93(3):247–57. [PubMed: 11514084]
18. Sherbourne, CD. Pain measures. In: Stewart, AL.; Ware, JE., editors. Measuring functional status and well-being: The Medical Outcomes Study Approach. Durham, NC: Duke University Press; 1992. p. 230-4.
19. Turk DC, Dworkin RH, Revicki D, Harding G, Burke LB, Cella D, et al. Identifying important outcome domains for chronic pain clinical trials: an IMMPACT survey of people with pain. Pain 2008;137(2):276–85. [PubMed: 17937976]
20. Clark LA, Watson D. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. J Abnorm Psychol 1991;100(3):316–36. [PubMed: 1918611]
21. Watson D, Clark LA. Affects separable and inseparable: On the hierarchical arrangement of the negative affects. J Pers Soc Psychol 1992;62(3):489–505.
22. Beels CC, Gutwirth L, Berkeley J, Struening E. Measurements of social support in schizophrenia. Schizophr Bull 1984;10(3):399–411. [PubMed: 6382589]
23. Brekke JS, Long JD, Kay DD. The structure and invariance of a model of social functioning in schizophrenia. J Nerv Ment Dis 2002;190(2):63–72. [PubMed: 11889358]
24. Horowitz LM, Rosenberg SE, Baer BA, Ureno G, Villasenor VS. Inventory of interpersonal problems: Psychometric properties and clinical applications. J Consult Clin Psychol 1988;56(6):885–92. [PubMed: 3204198]
25. Dijkers MP, Whiteneck G, El Jaroudi R. Measures of social outcomes in disability research. Arch Phys Med Rehabil 2000;81(12 Suppl 2):S63–S80. [PubMed: 11128906]
26. McDowell, I. Measuring health: a guide to rating scales and questionnaires. Oxford; New York: Oxford University Press; 2006.
27. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16(Suppl 1):133–41. [PubMed: 17401637]
28. Hahn, EA.; Cella, D.; Bode, RK.; Hanrahan, RT. Soc Indic Res. 2009. Measuring Social Well-being in People with Chronic Illness. Epub 2009 May 24
29. 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. Qual Life Res 2009;18(7):873–80. [PubMed: 19543809]
30. Revicki DA, Kawata AK, Harnam N, Chen WH, Hays RD, Cella D. Predicting EuroQol (EQ-5D) scores from the Patient-Reported Outcome Measurement Information System (PROMIS) global items and domain item banks in a United States sample. Qual Life Res 2009;18(6):783–91. [PubMed: 19472072]
31. Chen WH, Revicki DA, Lai JS, Cook KF, Amtmann D. Linking pain items from two studies onto a common scale using item response theory. J Pain Symptom Manage. Epub 2009 July 3.
32. Hays RD, Liu H, Spritzer K, Cella D. Item Response Theory Analyses of Physical Functioning Items in the Medical Outcomes Study. Med Care 2007;45(5 Suppl 1):S32–S8. [PubMed: 17443117]
33. Stone AA, Broderick JE, Schwartz JE, Shiffman S, Litcher-Kelly L, Calvanese P. Intensive momentary reporting of pain with an electronic diary: Reactivity, compliance, and patient satisfaction. Pain 2003;104(1–2):343–51. [PubMed: 12855344]
34. Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. Pain 2008;139(1):146. [PubMed: 18455312]

35. Revicki DA, Chen WH, Harnam N, Cook K, Amtmann D, Callahan L, et al. Development and Psychometric Analysis of the PROMIS Pain Behavior Item Bank. Pain. 2009 Forthcoming.

36. Alwin DF, Krosknick JA. The Reliability of survey attitude measurement: the influence of question and respondent attributes. Sociological Methods Research 1991;20(1):139–81.

37. Rivers, D. Sample matching: representative sampling from Internet panels. Palo Alto, CA: Polimetrix, Inc; 2006.

38. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the PROMIS Internet Panel. J Clin Epidemiol. 2009

39. McCall, WA. How to measure in education. New York: The Macmillan Company; 1922.

40. Garcia SF, Cella D, Clauser SB, Flynn KE, Lai JS, Reeve BB, et al. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. J Clin Oncol 2007;25(32):5106–12. [PubMed: 17991929]

41. 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. Qual Life Res 2010;19:125–136. [PubMed: 19941077]

42. Holland, PW.; Wainer, H. Differential item functioning. Hillsdale, NJ: Lawrence Earlbaum Associates; 1993.

43. Camilli, G.; Shepard, LA. Methods for identifying biased test items. London, England: Sage Publications; 1994.

44. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. Med Care 2006;44(11 Suppl 3):S115–S23. [PubMed: 17060818]

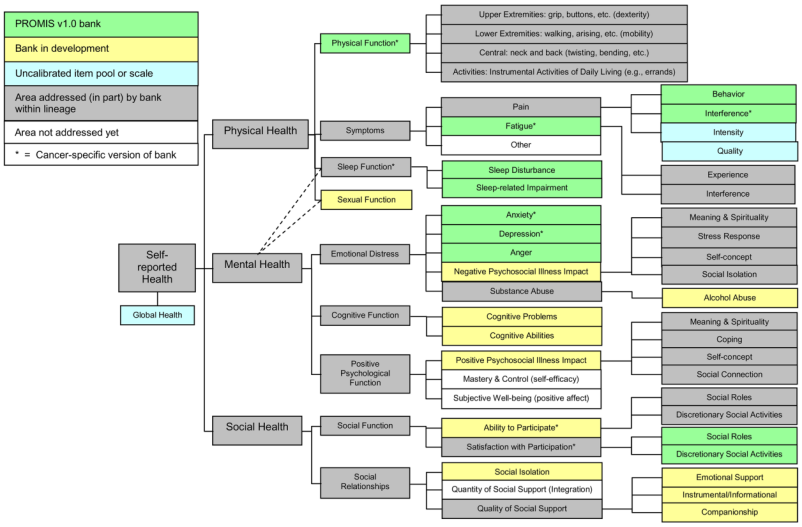

Figure 1.
PROMIS Adult Health Domain Framework, March 2010

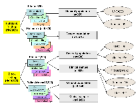

Figure 2.
PROMIS Wave 1 Sample

Table 1

Correlations of PROMIS Full Item Banks with Short Forms

Item Bank	N	# Items Full Bank	# Items Short Form	Correlation with Short Form
Emotional Distress – Anger	858	29	8	0.96
Emotional Distress – Anxiety	788	29	7	0.96
Emotional Distress – Depression	782	28	8	0.96
Fatigue	9047	95	7	0.76
Pain Behavior	794	39	7	0.98
Pain Interference	796	41	6	0.95
Physical Function*	737	124	10	0.96
Satisfaction with Participation in Discretionary Social Activities	729	12	7	0.99
Satisfaction with Participation in Social Roles	726	14	7	0.99
Sleep Disturbance	2252	27	8	0.96
Sleep-related Impairment	2252	16	8	0.98
Global Health		n/a	10	n/a

*Due to 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

PROMIS Item Bank Standard Error and Alpha Reliability by T-scores

Item Bank	N		T-Scores								
			10	20	30	40	50	60	70	80	90
Emotional Distress - Anger	856	SE	0.93	0.76	0.45	0.23	0.17	0.16	0.16	0.16	0.26
		Reliability	0.13	0.43	0.80	0.95	0.97	0.98	0.98	0.97	0.93
Emotional Distress – Anxiety	788	SE	0.98	0.90	0.62	0.26	0.14	0.13	0.13	0.15	0.33
		Reliability	0.04	0.19	0.62	0.93	0.98	0.98	0.98	0.98	0.89
Emotional Distress – Depression	782	SE	1.00	0.97	0.73	0.27	0.13	0.11	0.12	0.16	0.45
		Reliability	0.01	0.06	0.47	0.92	0.98	0.99	0.99	0.97	0.80
Fatigue	781	SE	0.86	0.58	0.23	0.09	0.06	0.06	0.06	0.11	0.31
		Reliability	0.27	0.66	0.95	0.99	1.00	1.00	1.00	0.99	0.91
Pain Behavior	820	SE	1.00	0.96	0.51	0.11	0.10	0.08	0.08	0.21	0.78
		Reliability	0.00	0.08	0.74	0.99	0.99	0.99	0.99	0.96	0.40
Pain Interference	844	SE	1.00	1.00	0.94	0.48	0.10	0.07	0.07	0.19	0.58
		Reliability	0.00	0.01	0.11	0.77	0.99	1.00	0.99	0.97	0.66
Physical Function	1700	SE	0.13	0.07	0.06	0.06	0.09	0.21	0.41	0.68	0.93
		Reliability	0.98	0.99	1.00	1.00	0.99	0.96	0.83	0.54	0.14
Satisfaction with Participation in Discretionary Social Activities	814	SE	0.99	0.83	0.30	0.14	0.14	0.16	0.58	0.97	1.00
		Reliability	0.02	0.32	0.91	0.98	0.98	0.97	0.67	0.06	0.00
Satisfaction with Participation in Social Roles	816	SE	0.98	0.73	0.21	0.13	0.14	0.18	0.66	0.98	1.00
		Reliability	0.03	0.47	0.96	0.98	0.98	0.97	0.57	0.04	0.00
Sleep Disturbance	2252	SE	0.91	0.68	0.34	0.20	0.16	0.16	0.17	0.28	0.49
		Reliability	0.16	0.54	0.88	0.96	0.97	0.98	0.97	0.92	0.76
Sleep-related Impairment	2252	SE	0.91	0.68	0.41	0.31	0.18	0.17	0.17	0.29	0.63
		Reliability	0.17	0.53	0.84	0.91	0.97	0.97	0.97	0.92	0.61

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 US general population based upon 2000 census data (see Liu et al, this issue).

Sample size (N) refers to the number of people who completed full bank testing from wave one (see Figure 2)

Table 3

PROMIS Item Bank Calibration Sample T-Score Means and Standard Deviations, and Distributions by Percentile

Item Bank	# Items	N	Mean	SD	P5	P10	P25	P50	P75	P90	P95
Emotional Distress – Anger	29	858	47.2	9.4	27.9	34.9	41.3	47.3	53.6	58.5	62.3
Emotional Distress – Anxiety	29	788	48.5	9.8	31.6	35.2	41.8	48.2	54.6	61.6	65.5
Emotional Distress – Depression	28	782	49.3	9.6	33.5	37.7	42.4	48.6	55.3	62.0	66.0
Fatigue	95	14,931	51.2	9.4	36.5	30.0	44.3	50.8	57.8	64.0	67.3
Pain Behavior	39	15,834	55.6	8.2	36.2	39.3	52.5	57.3	61.4	63.9	65.2
Pain Interference	41	15,903	55.9	10.8	40.0	40.0	47.5	55.2	65.3	70.3	72.8
Physical Function	124	1700	50.0	10.0	27.6	37.0	47.4	53.4	56.9	58.3	58.8
Satisfaction with Participation in Discretionary Social Activities	12	712	49.3	10.7	32.1	36.6	42.0	48.5	55.1	68.9	68.9
Satisfaction with Participation in Social Roles	14	712	49.4	10.7	31.9	36.2	41.8	49.1	55.9	67.9	67.9
Sleep Disturbance	27	2252	49.8	10.3	33.4	36.5	42.1	49.4	57.2	63.4	66.5
Sleep-related Impairment	16	2252	50.0	9.8	34.6	37.4	43.0	49.5	56.9	63.2	66.7

T-score means, standard deviations 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, this issue)

Table 4

Physical Function: Correlations with Legacy Measures and Short Forms

	HAQ[1]	SF36-PF[2]	Physical B Function[3]	N[5]	Correlation with Short Form
HAQ	1.00			711	−0.88
SF36-PF	−0.80	1.00		708	0.90
Physical Function Bank	−0.88	0.88	1.00	737	0.96

[1] HAQ-Disability Scale (scored without devices & aids)

[2] SF-36 Physical Function Scale

[3] Physical Function Bank (at least 93 items non missing)

[5] 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

Table 5

Fatigue: Correlations with Legacy Measures and Short Forms

	FACIT[1]	SF36-VT[2]	Fatigue[3]	Fatigue[4]	N[5]	Correlation with Short Form
FACIT	1.00				740.00	0.91
SF36-VT	0.88	1.00			737.00	0.85
Fatigue Bank[3]	0.96	0.89	1.00		9047.00	0.76
Fatigue Bank[4]	0.95	0.88	1.00	1.00	9047.00	0.76

[1] FACIT-Fatigue Scale

[2] SF36 Vitality Scale

[3] Including legacy items (i.e., FACIT-F)

[4] Excluding legacy items (i.e., FACIT-F)

[5] N=sample size

Correlations are all based on full-bank data only

based on block-testing + full-bank data

based on full-bank data

Table 6

Pain: Correlations between PROMIS Banks, Legacy Measures and Short Forms

	BPI[1] (Severity)	BPI[1] (Interference)	SF36-BP[2]	Pain Behavior[4]	Pain Interference	N[3]	Correlation with Pain Behavior Short Form[4]	N[3]	Correlation with Pain Interference Short Form
BPI (Severity)	1.00					N/A	N/A	776	0.76
BPI (Interference)	0.84	1.00				N/A	N/A	774	0.84
SF36-BP	−0.85	−0.84	1.00			N/A	N/A	724	−0.82
Pain Behavior Bank				1.00		794	0.98	N/A	N/A
Pain Interf. Bank	0.81	0.85	−0.86	0.77	1.00	N/A	N/A	796	0.95
Pain Intensity item	0.83	0.74	−0.78	0.69	0.76	794	0.60	795	0.69

[1] Brief Pain Inventory

[2] SF36 Bodily Pain Scale

[3] N=sample size

[4] Based on calibrations drawn from general population and a chronic pain community sample

data not available

based on block-testing data

based on full-bank data

Table 7

Sleep and Sleep-related Impairment: Correlations between PROMIS Banks, Legacy Measures and Short Forms

	PSQI[1]	ESS[2]	Sleep Dist.	Wake Dist.	N[3]	Correlation with Short Form
PSQI	1.00					
ESS	0.28	1.00				
Sleep Dist. Bank	0.85	0.25	1.00		2252	0.96
Wake Dist. Bank	0.7	0.45	0.75	1.00	2252	0.98

[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 PROMIS Banks, Legacy Measures and Short Forms

	AQ[1]	CES-D[2]	MASQ[3]	Anger	Anxiety	Depression	N[4]	Correlation with Short Form
AQ	1.00							
CES-D		1.00						
MASQ		0.78	1.00					
Anger Bank	0.51			1.00			858	0.96
Anxiety Bank		0.75	0.80	0.59	1.00		788	0.96
Depression Bank		0.83	0.72	0.60	0.81	1.00	782	0.96

[1] Aggression Questionnaire

[2] Center for Epidemiological Studies-Depression Scale

[3] Mood and Anxiety Symptom Questionnaire

[4] N=sample size

data not available

based on block-testing data

based on full-bank data

J Clin Epidemiol. Author manuscript; available in PMC 2011 November 1.

Table 9

Social Function: Correlations between PROMIS Banks, Legacy Measures and Short Forms

	FACIT-FWB[1]	SF36-RP[2]	SF36-RE[3]	SF36-SF[4]	SR[5]	DSA[6]	N[7]	Correlation with Short Form	
								SSR[5]	SDSA[6]
FACIT-FWB	1.00						708	0.74	0.75
SF36-RP	0.47	1.00					712	0.56	0.43
SF36-RE	0.62	0.46	1.00				709	0.57	0.51
SF36-SF	0.62	0.52	0.61	1.00			706	0.57	0.52
SR Bank	0.76	0.57	0.59	0.58	1.00		726	0.99	0.82
DSA Bank	0.76	0.44	0.52	0.53	0.83	1.00	729	0.82	0.99

[1] FACIT - Functional Well-Being

[2] SF36 -Role Physical

[3] SF36 -Role Emotional

[4] SF36 -Social Functioning

[5] PROMIS Satisfaction with Social Roles (SSR)

[6] PROMIS Satisfaction with Discretionary Social Activities (SDSA)

[7] N=sample size

Based on full-bank data