The Patient-Reported Outcomes Measurement Information System (PROMIS)

Progress of an NIH Roadmap Cooperative Group During its First Two Years

David Cella, PhD,† Susan Yount, PhD,*† Nan Rothrock, PhD,*† Richard Gershon, PhD,*† Karon Cook, PhD,* Bryce Reeve, PhD,‡ Deborah Ader, PhD,§ James F. Fries, MD,¶ Bonnie Bruce, DrPH, MPH, RD,¶ and Mattias Rose, MD‖ on behalf of the PROMIS Cooperative Group*

Background: The National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System (PROMIS) Roadmap initiative (www.nihpromis.org) is a 5-year cooperative group program of research designed to develop, validate, and standardize item banks to measure patient-reported outcomes (PROs) relevant across common medical conditions. In this article, we will summarize the organization and scientific activity of the PROMIS network during its first 2 years.

Design: The network consists of 6 primary research sites (PRSs), a statistical coordinating center (SCC), and NIH research scientists. Governed by a steering committee, the network is organized into functional subcommittees and working groups. In the first year, we created an item library and activated 3 interacting protocols: Domain Mapping, Archival Data Analysis, and Qualitative Item Review (QIR). In the second year, we developed and initiated testing of item banks covering 5 broad domains of self-reported health.

Results: The domain mapping process is built on the World Health Organization (WHO) framework of physical, mental, and social health. From this framework, pain, fatigue, emotional distress, physical functioning, social role participation, and global health perceptions were selected for the first wave of testing. Item response theory (IRT)-based analysis of 11 large datasets supplemented and informed item-level qualitative review of nearly 7000 items from available PRO measures in the item library. Items were selected for rewriting or creation with further detailed review before the first round of testing in the general population and target patient populations.

Conclusions: The NIH PROMIS network derived a consensus-based framework for self-reported health, systematically reviewed available instruments and datasets that address the initial PROMIS domains. Qualitative item research led to the first wave of network testing which began in the second year.

Key Words: outcomes, patient reported outcomes, quality of life, health-related quality of life

(*Med Care* 2007;45: S3–S11)

In May 2002, the Director of the National Institutes of Health (NIH) convened a series of meetings to chart a "roadmap" for medical research in the 21st century. The purpose of the roadmap is to identify major opportunities and gaps in biomedical research that no single NIH institute could tackle alone, but that the agency as a whole must address to maximize progress in medical research. The ultimate goal of the roadmap is to catalyze changes that are necessary for transforming new scientific knowledge into tangible benefits for people. It proposes a vision for a more efficient and productive system of medical research and identifies the most compelling opportunities in 3 areas: (1) New Pathways to Discovery, (2) Research Teams of the Future, and (3) Re-engineering the Clinical Research Enterprise.

The theme of New Pathways of Discovery addresses the need to advance our understanding of complex biologic systems, and to build a better "toolbox" for medical research in the 21st century by providing wide access to technologies, databases, and other scientific resources that are more sensitive, robust, and easily adaptable to the needs of researchers. The Research Teams of the Future initiative seeks to encourage scientists to test alternative models for conducting research, including the pursuit of unexplored avenues of research that carry a high potential for failure but also a greater chance for groundbreaking discoveries; stimulating new ways

From the *Evanston Northwestern Healthcare, Evanston, Illinois; †Northwestern University Feinberg School of Medicine, Chicago, Illinois; ‡National Cancer Institute, Bethesda, Maryland; §Samueli Institute, Alexandria, Virginia; ¶Department of Medicine, Stanford University School of Medicine, Stanford, California; and ‖Health Assessment Lab, Boston, Massachusetts.

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant (1U01-AR052177).

Reprints: David Cella, PhD, Center on Outcomes, Research, and Education, Evanston Northwestern Healthcare, 1001 University Place, Suite 100, Evanston, IL 60201. E-mail: d-cella@northwestern.edu.

ISSN: 0025-7079/07/4500-0003

of combining skills and disciplines in the physical and biologic sciences; and encouraging novel partnerships, such as those between the public and private sectors, to accelerate the movement of scientific discoveries from bench to bedside. Re-engineering the Clinical Research Enterprise is designed to accelerate and strengthen the clinical research process by adopting a systematic infrastructure to better and more efficiently serve the field of scientific discovery.

One of the programs within the re-engineering the Clinical Research Enterprise initiative involves dynamic assessment of patient-reported chronic disease outcomes. Supporting this initiative, in late 2004 the NIH initiated a multicenter cooperative group referred to as the Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS will build and validate common, accessible item banks to measure key symptoms and health concepts applicable to a range of chronic conditions, enabling efficient and interpretable clinical trial and clinical practice applications of patient-reported outcomes (PROs).

Applications of item response theory (IRT) and advances in computer technology make it possible to improve health status measurement through the development and maintenance of item banks for measuring specified symptoms and health status domains.[1,2] An item bank is more than a collection of questions about a particular symptom or functional problem; it is comprised of carefully calibrated questions that define and quantify a common concept and thus provide an operational definition of a trait.[3,4] Item banks enable item comparison and selection and computerized adaptive testing (CAT) tools for tailored individual assessment without loss of scale precision or content validity.[5–12] The PROMIS initiative is designed to help realize this potential at a national level, focusing on several important symptoms and health status domains that have relevance across chronic diseases. Valid, generalizable item banks and CAT tools can stimulate and standardize clinical research across NIH-funded research dealing with PROs. They may also assist individual clinical practitioners in assessing patients' responses to interventions and in modifying treatment plans on the basis of these responses.

PROMIS NETWORK STRUCTURE AND ORGANIZATION

The PROMIS network of clinicians, clinical researchers, and measurement experts is organized around 6 primary research sites (PRSs) and a statistical coordinating center (SCC), all of whom work closely with NIH project scientists representing several institutes of the NIH (Table 1). The 6 PRSs include investigators from Duke University, Stanford University, Stony Brook University, University of North Carolina, University of Pittsburgh, and University of Washington, along with several collaborating institutions. The SCC is based at the Center on Outcomes, Research, and Education (CORE) at Evanston Northwestern Healthcare and includes collaborators from UCLA, Rehabilitation Institute of Chicago, United BioSource Corporation and Westat, Inc. Figure 1 displays the organizational structure of the PROMIS.

A steering committee (SC) governs and assumes ultimate responsibility for the priorities and direction of the network. Comprised of the 7 principal grantees and 5 NIH scientists, it is the principal committee through which the NIH interacts and collaborates with the investigators. An independent scientific advisory board (SAB) provides oversight and advice to promote the overall success of the network. It is the responsibility of the SAB to make recommendations that support the exchange of research tools and resources, encourage the adoption of common policies on data sharing, and lead in the creation of item banks. The SAB evaluates the PROMIS activities to ensure that the resources developed will be of maximal utility to the scientific community. The SAB also will solicit input and feedback from stakeholders to ensure the success of the PROMIS initiative. The SAB, appointed by the NIH, consists of 11 experts from academia, government, and industry (http://www.nihpromis.org).

The SCC has responsibility for providing and managing a secure, customizable, coordinated data management system for collection, storage, and analysis of data collected by the PRSs. With guidance from the PRSs and SC, the SCC assembled the PROMIS item banks and other questionnaires to be administered across the network. The SCC also coordinates, facilitates, and maintains information exchange and dissemination across scientific, administrative, and advisory tiers of the PROMIS network; standardizes protocols, study procedures and forms; develops end-user training materials for clinicians who will use the item banks and the CAT system; and works collaboratively with the PROMIS network to develop a public–private partnership to sustain the network beyond the project period. In support of this, the SCC convened a panel of 22 clinical research and health outcomes experts to advise the PROMIS network on relevance and

TABLE 1. PROMIS Network of Investigators

PROMIS Entity	Institution	Principal Investigator
SCC	The Center on Outcomes, Research and Education, Evanston Northwestern Healthcare, Evanston, IL	David Cella, PhD
PRS	Duke University, Raleigh-Durham, NC	Kevin Weinfurt, PhD
PRS	Stanford University, Palo Alto, CA	James F. Fries, MD
PRS	Stony Brook University, Stony Brook, NY	Arthur Stone, PhD
PRS	University of North Carolina, Chapel Hill, NC	Harry Guess, MD, PhD*
PRS	University of Pittsburgh, Pittsburgh, PA	Paul Pilkonis, PhD
PRS	University of Washington, Seattle, WA	Dagmar Amtmann, PhD

*In memoriam. Darren DeWalt, MD, assumed the role of principal investigator in January, 2006.

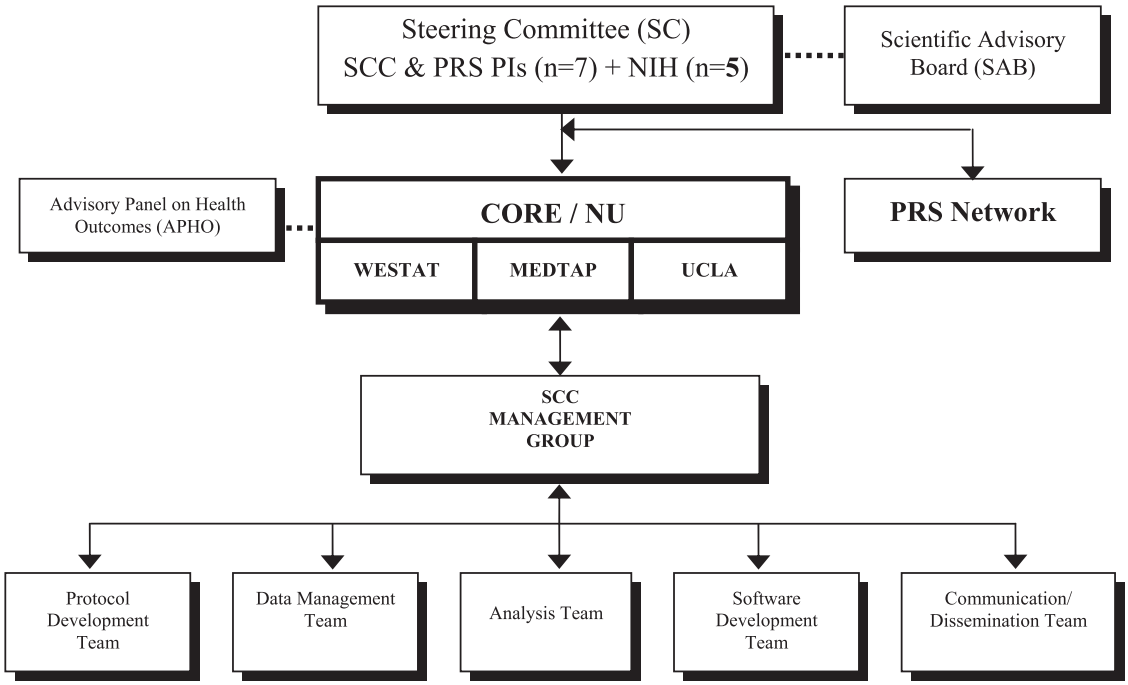


FIGURE 1. PROMIS organizational chart.

feasibility for clinical research. This "Advisory Panel on Health Outcomes" (APHO) includes experienced content experts and clinical trialists in cross-cutting clinical areas such as pain and fatigue; clinical researchers in specialty areas such as oncology, rheumatology, endocrinology, mental health, neurology, cardiovascular disease, respiratory illness, rehabilitation medicine, pediatric, and adolescent medicine; and representatives from the pharmaceutical industry (for APHO members see http://www.nihpromis.org/network_structure/statistical_coordinating_center.asp). During the first year of PROMIS, input from the APHO was formally solicited. Specifically, review and comment were solicited for an early version of the PROMIS domain framework and the 5 health domains selected for study; the definitions of initial PROMIS domains (of critical importance to the item-revision and rewriting process); plans for targeting additional banks/domains in a second wave of item testing; and the process for building the PROMIS item bank. Formal panel input was also solicited from relevant members of APHO regarding issues such as response scale options and the composition of proposed item banks and applications.

PROMIS NETWORK ACTIVITY IN THE FIRST 2 YEARS

PROMIS scientific activity is organized into 2 categories: independent research, conducted by individual PRS groups; and network activity, conducted collaboratively by all PRS group members with the SCC. Table 2 summarizes the independent projects. Articles in this issue[13,14] describe early progress on 2 of these projects. This article focuses on the network activity over the first 2 years, which has been organized around the creation of the PROMIS Item Library and 3 interrelated protocols: (1) Domain Mapping, (2) Archival Data Analysis, and (3) Qualitative Item Review.

The domain groups combined domain-specific content expertise with analytic input from archival data to develop initial item pools. Each domain group is led by a PROMIS investigator with expertise in the area and includes additional members from across the PROMIS network. Each group also has an SCC liaison from the SCC Analysis Team to facilitate communication and coordination of efforts across domains and across network activities. Several protocols were developed to structure and standardize the effort. Domain group activity was guided by a domain framework protocol; the item review and selection/writing was guided by the Qualitative Item Review protocol, and the analysis of archival data to inform the process of item review was guided by the analysis protocol. Interaction and communication across these 3 protocol-driven activities is depicted in Figure 2.

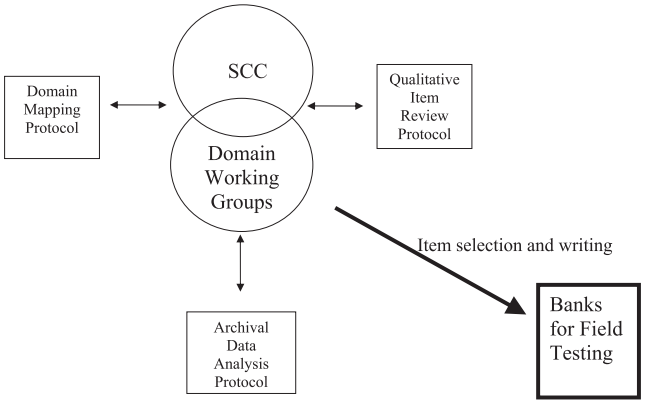
Domain Mapping Protocol

The first task of the PROMIS network was to create a protocol for developing a domain map (framework) that portrayed the structure of each target domain and its conceptual framework or, where applicable, hierarchical structure. Existing outcome assessment questionnaires use an explicit or implicit framework that typically includes the concepts of physical function or limitation, mental health or distress, and social function, with many also including symptoms (eg, fatigue, pain). The SC-approved protocol for the domain mapping activity specified that the preliminary PROMIS framework would be developed through independent literature reviews by the Stanford PRS, the Pittsburgh PRS, and the SCC, followed by a consensus-building Delphi process and

TABLE 2. Independent Research Projects at PROMIS Primary Research Sites

PI, Institution	Independent Project	Patient Population(s)
Harry A. Guess, MD, Darren DeWalt, MD, University of North Carolina at Chapel Hill	Develop PRO item bank measuring important health domains for children; develop and pilot-test CAT tool for children 8–17 yr	Children (8+ yr), range of chronic illnesses and healthy children, large sample of asthma
James F. Fries, MD, Stanford University	Develop improved instruments for arthritis and aging outcome assessment; encourage use of instruments in clinical research	Arthritis, aging
Kevin Weinfurt, PhD, Duke University	Identify and address challenges related to implementation of PROMIS technology in multicenter clinical trials	Chronic pain; psychiatric, cardiovascular disorders; cancer
Arthur A. Stone, PhD, Stony Brook University	Assess and improve ecological validity of PROs by comparing real-time reports of symptoms with retrospective reports as currently assessed with PROs. Understand how testing contexts (including instructional sets and comparison standards) for PRO administration affect responses	Rheumatic disease, community sample
Dagmar Amtmann, PhD, University of Washington	Develop dynamic system for measuring pain and fatigue; increase scientific understanding of pain and fatigue in children and adults with disabilities	MS, SCI, neuro-developmental disorders; neuromuscular disorders; TBI; amputation
Paul A. Pilkonis, PhD, University of Pittsburgh	Development and testing of a core battery for sleep-wake functioning	Psychiatric patients, patients with sleep disorders, community sample

PRO indicates patient-reported outcomes; CAT, computerized adaptive testing; MS, multiple sclerosis; SCI, spinal cord injury; TBI, traumatic brain injury.



NOTE: SCC = Statistical Coordinating Center

FIGURE 2. Protocol interaction to construct banks for field testing. SCC indicates statistical coordinating center.

statistical analysis of available data regarding dimensionality of health status assessment. Early in the first year, the SC endorsed the World Health Organization (WHO) physical, mental, and social framework.[15] Other organizing frameworks were considered, such as the WHO international classification of functioning and a 2-factor model of physical and mental health. However, after discussion and careful consideration, the SC opted to retain the WHO tripartite framework as compelling and sufficiently broad and inclusive to enable important social dimensions of health to be developed further than has been done to date. After achieving consensus on the broad WHO framework, the SC launched the Domain Mapping Protocol. Under this protocol, PROMIS network investigators used a modified Delphi approach to participate in multiple rounds of framework review and revision until consensus was reached on a detailed articulation of subordinate domains beneath the broad physical, mental, and social headings. First published in 2005,[16] this was modified iteratively several times until unanimous agreement was reached from all members of the SC and Domain Committee Chairs. The current version of the PROMIS domain framework and can be viewed at http://www.nihpromis.org/reference_material/domain_framework.asp.

PROMIS experts in these 3 broad domains reviewed and refined the framework by specifying unidimensional subdomains they determined, through the pooling of literature review, data analysis and consensus, to constitute the domain. Physical health was thereby divided into subdimensions of physical function and symptoms. In turn, physical function was divided into lower extremity function (eg, mobility), upper extremity function (eg, dexterity), and central function (eg, bending; twisting), and instrumental activities of daily living. As an example of the next layer of detail, lower extremity function ("mobility") was conceptually subdivided into categories, such as walking and climbing stairs, each of which was assigned specific items. The identified and prioritized symptom subdimension included pain and fatigue. Working definitions for all domains and subdomains were drafted and are available at www.nihpromis.org. Five subdomains were selected as the initial areas for PROMIS item bank construction: (1) Physical Functioning, (2) Fatigue, (3) Pain, (4) Emotional Distress, and (5) Social Role Participation. In addition to building item banks for these 5 domain areas, 1–2 global items were written to capture an overall evaluation of one's location on each domain, and overall health and quality of life.

Archival Data Analysis Protocol

A protocol for Archival Data Analysis was the second network protocol approved by the SC. The protocol specified an IRT-based statistical analysis plan developed with extensive input from a team of 20 coinvestigators and consultants.

The SCC identified and reviewed large datasets that included PROs and, from these, selected 11 for protocol-driven analyses that would inform the building of item bank structure and content (eg, Medical Outcomes Study, Cancer Q-Score Project, Chronic Hepatitis C Study, NHLBI Cardiac Health Study [CHS], World Health Organization's Quality of Life [WHOQOL]-100 database; Table 3). After rigorous data quality review, items were extracted from the datasets and presented verbatim to each domain working group for review. Working groups identified items representing the 5 selected PROMIS domains (physical functioning, fatigue, pain, emotional distress [ED], and social role participation). Items were then subjected to IRT analyses by psychometricians from the SCC and PRSs. All results were reviewed collectively by the SCC analysis team and a summary was presented to the appropriate domain working group. The goal of this process was to better understand dimensionality in the 5 PROMIS domains, inform the revision of items in the item library, inform the identification of the most useful response sets, and guide new item construction. The data analysis plan for PROMIS items, which also guided the archival data analyses done in the first year of activity, is described in detail in the article by Reeve et al[29] in this issue. Finally, the articles by Hays et al[30] and Hill et al[14] in this issue discuss specific applications of IRT for analysis of physical functioning in adults, and multidimensional quality of life in children.

Qualitative Item Review Protocol

The third protocol activated in the first year of PROMIS was the Qualitative Item Review (QIR) protocol. This protocol was supported by a library of items collected through multiple literature reviews and investigator input.

The PROMIS Item Library

A critical first step in the creation of item banks is the building of an "item library." The PROMIS item library is an extensive relational database of items gathered from existing PROs. The purpose of the library is to support the identification, cataloguing, refinement, and writing of items that serve as candidate items for future PROMIS item banks. The relational database includes several variables of interest including item context, item stem, response set, time frame, and instrument of origin with original item number if applicable. All PRS and SCC investigators submitted PRO instruments and items pertaining to any of 5 selected PROMIS domains (as given in the next section). Items were identified through literature review and expert consultation. Where applicable, the item library also includes data on modifications and intellectual property status of items.

The PROMIS item library includes over 10,000 entries, approximately 7000 of which relate to the 5 health domains chosen for initial bank development. Because of the library's size and the amount of content redundancy among items, a selection process was undertaken as part of the QIR protocol. We refer to this process as "binning and winnowing." First, all items in each domain were classified according to content ("binning"). After this "binning" process, a smaller set of items (approximately 1100) were reviewed and revised by domain experts through the QIR process ("winnowing"). The goal of winnowing was to eliminate those items that were either strikingly dissimilar to the identified domain (face invalidity), or highly similar to a better-worded item (redundancy). For detail, see article by DeWalt et al in this issue[31] Information from the binning and winnowing process was catalogued in the Item Library. SCC staff conducted secondary independent item review to ensure consistency of winnowing across domains.

Although the nearly 7000 relevant items were being "binned and winnowed," focus groups were scheduled across multiple diseases at different sites to evaluate the comprehensiveness of the PROMIS domain framework and note any conceptual gaps in the domain definitions. To evaluate comprehension and relevance of items, cognitive assessments including cognitive interviews were scheduled with patient populations. Items were subsequently revised as needed to improve clarity, precision, readability, translatability, and fit to a CAT framework. The final items were then evaluated on

TABLE 3. Archival Analysis Datasets

Name of Dataset	Citation	Domain
Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) Survey	Chen et al[17]	Pain
Northwestern University/Evanston Northwestern Healthcare Pain item bank	Lai et al[18]	Pain
Cardiac Health Study (CHS)	Arnold et al[19]	Social role participation, emotional distress
Northwestern University/Evanston Northwestern Healthcare Social Well-Being item bank	Hahn et al[20]	Social role participation
Chronic Hepatitis C Study (CHC)	Kleinman et al[21]	Fatigue
Northwestern University/Evanston Northwestern Healthcare Fatigue item bank	Cella et al[22]	Fatigue
Medical Outcome Study (MOS)	Stewart and Ware[23] and Tarlov et al[24]. http://www.socio.com/srch/summary/radius/rad3034.htm	Physical functioning
Digitalis Investigation Group (DIG)	The Digitalis Investigation Group[25]	Emotional distress
Cooperative Study of Sickle Cell Disease (CSSCD)	Gaston et al[26]	Emotional distress
World Health Organization Quality of Life-100 (WHOQOL-100)	Szabo[27]	Emotional distress
Q-Score Project (Q-Score)	Chang and Cella[28]	Emotional distress

numerous surface characteristics (eg, item measuring intensity or frequency).

The PROMIS network reached consensus on response options to use in the PROMIS item pool again using a modified Delphi approach. Each domain began by identifying response sets that would be most applicable to their domain. IRT experts provided input regarding the optimal number of response choices within a response set. Also, each domain group identified key instruments to be considered "legacy" instruments. These items will be included in future testing to aid in validation testing of the PROMIS item banks.

Coordinating Early Protocols to Construct Item Banks

As represented in Figure 2, the Domain Mapping, Archival Data Analysis, and QIR processes are integrated and inform one another. The Domain working groups drive the process and decision making as they receive input regarding activity in each of the protocols. The end result of this interactive process is an item bank ready for network testing. The SCC serves as the hub of communication regarding network efforts. As such, the SCC faces the challenge of integrating the analytic expertise of the SCC analysis team and PRS psychometricians, the clinical and conceptual expertise of the Domain Mapping teams, and the methodological expertise of the qualitative item reviewers. The development of the item banks required extensive communication and interaction to inform not only item selection and writing, but also evaluation of the underlying PRO health model itself. We offer an example of this interaction using the case example of defining the subdomains and optimal items for the ED domain.

Interaction Across Protocols to Drive Item Bank Development: A Case Example

The domain team targeting ED measurement is led by Dr. Paul Pilkonis at the University of Pittsburgh. This group's work focuses on 4 subdomains of ED: anxiety, depression, anger, and substance abuse. The team's goal is to develop 4-item banks that are suitable for evaluation of various clinical populations.

Secondary analyses of several instruments informed the work of the ED Domain team, including the WHOQOL instrument, an internationally recognized instrument with content related to ED. The psychometric evaluations specified in the Data Analysis protocol were applied to the WHOQOL. This analysis was completed by Ron Hays and Karen Spritzer from the UCLA group, which was associated with the SCC. The dataset the investigators used is the largest in the United States to include the WHOQOL measure.[32] The WHOQOL instrument includes measures of thinking (cognition, concentration), body image, and affect. An 18-item subset was evaluated using exploratory factor analysis and IRT modeling. Exploratory factor analysis results indicated 1 main factor and 3 additional factors, suggesting both multidimensionality and content heterogeneity. This multidimensionality contributed to poor fit of the data to unidimensional IRT models. Although the results were discouraging with respect to the prospect of applying IRT to this pool of "emotional" items, the evaluation underscores the benefits of conducting secondary data analyses before committing to what might otherwise seem to be a promising item set.

A different experience occurred with the analysis of the ED items from the CHS. This longitudinal, multicenter study monitored the cardiovascular health of more than 5000 individuals with the goal of estimating the incidence and prevalence of coronary heart disease and stroke.[33,34] The ED domain team identified 19 items in the CHS database they believed to be relevant to the ED domain. The percentage of missing item responses ranged from 13% to 19%. For most items, responses were skewed toward lower values; that is, relative to the range of the scale, more of the sample reported lower ED. Three of the 4 items that had more than 5 response options had one or more category that was endorsed by less than 1% of the sample.

Another feature evaluated according to the analysis protocol[29] was item monotonicity. Monotonicity refers to the requirement of IRT models that people who rate an item at the lowest (worst) option for that item indeed have worse scores than those who rate that item at the next (better) option. The probability of endorsing or selecting an item response indicative of better health should increase as one's underlying health increases (monotonically up the set of response options). When monotonicity is met, the proportion of people "passing each step" on the response scale is larger for those with higher scale scores. When the predicted order is reversed, this is a "violation" of monotonicity. In this ED domain analysis of the CHS database, only 1 item failed to increase monotonically ("When you have an important decision to make, do you have someone you can talk to about it?"). This item originally had 6 categories (0 = no, 1 = seldom, 2 = sometimes, 3 = often, 4 = very often, 5 = always). The lower categories were disordered, but by collapsing categories 0–3, monotonicity was restored.

The PROMIS domain framework identifies the subdomains that are hypothesized to comprise each domain. The ED subdomains include "internalizing" symptoms of depression and anxiety, and "externalizing" aspects such as anger and alcohol abuse (see www.nihpromis.org and Ref. 16). Based on this hierarchy, the domain team reviewed CHS items originally judged to represent the ED domain. In the original review, the decision was to err on the side of inclusiveness. In the re-examination, however, 7 items were identified as not consistent with the Emotional Distress domain and were reclassified as "Social Functioning." Not all of the retained items fell neatly within a single subdomain. One item was classified as Anxiety, and 3 others were identified as Anxiety/Depression. Six items were classified as Depression, and 2 were classified as tapping positive psychologic states.

Two fundamental assumptions of unidimensional IRT models are that a single trait determines how people respond to items and that those items are locally independent, that is, there should be little association between responses to 2 items beyond that accounted for by the underlying trait. The PROMIS team evaluated the dimensionality and local independence of the CHS ED items using confirmatory factor analyses (CFA). All CFA's were conducted using MPlus[35]

and tested the fit of a 1-factor model. The estimation procedure used was weighted least squares with mean and variance adjustment. The extent of local dependency was evaluated by examining item residual correlations. When items are perfectly locally independent, correlations between pairs of residuals should be zero. Two separate CFA's were conducted. Item Pool 1 included all items except those reclassified as social support. Item Pool 2 included the subset of items classified either as Anxiety/Depression or Depression Only (excluding social support).

The results for Item Pool 1 indicated poor fit to the unidimensional model and failure of the assumption of local independence. These results supported the decision to exclude the items reclassified as Social Support. The fit for the combined Anxiety/Depression and Depression Only items was below conventional standards for good fit in model testing applications. However, this reduced pool of items more closely approximated a unidimensional model. The items may be "sufficiently unidimensional" for scaling using an IRT model. Analysis of similar data sets may clarify circumstances where it is reasonable to calibrate depression and anxiety items as a single unidimensional pool.

Using Parscale and expected a priori estimation, the 10 Anxiety/Depression and Depression Only items were calibrated using the graded response model. Persons' calibrated Anxiety/Depression scores ranged from −3.93 to 1.33. The item category difficulties, however, ranged only from −1.54 to 0.56. If these items were to serve as the core of a CAT item bank, they would need to be augmented with items that better targeted both the low and high ranges of the trait. Such evidence argues for item bank development, which "builds out" from the core set of questions to provide better coverage of the continuum of measurement.

Another goal of the analysis was to assess the functioning of the response scales. When response categories are functioning well, there will be, for each category, some point along the trait continuum at which that category is the most likely response. Of the 10 Anxiety/Depression and Depression Only items, only half the items meet this criterion. The other items functioned, effectively, as 2 or 3 response category items. All of these findings were shared with the ED Domain working group and helped inform the group's item selection and writing activity for the network testing bank.

Software Development

The SCC software development team started work immediately upon grant award. Early efforts were focused on (1) providing support systems for managing existing datasets; (2) developing the database system used to house the item library; and (3) providing technology systems to support binning, winnowing, and QIR. Existing systems provided by the SCC were modified for use in the scheduled mid-2006 network-wide field testing of all candidate items in the new PROMIS item banks. Network data collection is taking place using internet-based testing. In total, approximately 800 PROMIS bank items are being tested alongside established ("legacy") questionnaires in a cross-country sample in excess of 11,000 individuals. The items are administrated with demographics and general health items.

The software development team is also creating a publicly available system to administer the instruments developed by the PROMIS network for use in clinical research. The system will be designed to enable easy modifications and will allow clinical researchers to access a common repository of items and CATs. Information on the requirements of end-users is currently being gathered from network researchers, nonnetwork researchers, study coordinators, research assistants, psychometricians, statisticians, and technology experts, in addition to members of the PROMIS Scientific Advisory Board, and the NIH. The PROMIS Software system will provide online study-setup and management services for study coordinators in addition to CAT-based assessments and short forms for measuring targeted outcomes. Platforms for the software will include stand-alone computers, websites, personal digital assistants, and integrated voice response. In addition it will be possible to upload data from paper and pencil forms to the PROMIS data repository and aggregate those with data collected electronically. In most cases data will be centralized in real time and made available for individual or group reporting on demand. Under the leadership of an independent project at the University of Washington site, all systems are being developed to insure easy accessibility to people with special needs.

Looking Back/Looking Ahead

The PROMIS roadmap effort is an unprecedented major effort across NIH institutes to standardize and promote a common measurement system for PROs across clinical research. It represents a tremendous opportunity to unify the field of PRO measurement. However, this opportunity is not without its challenges and costs. Early decisions had to be made which by design precluded alternative and arguably equally fruitful directions. For example, the SC decided very early on that network-wide testing of PROMIS item banks should occur within 2 years of PROMIS inception. This decision enabled us to plan for a second wave of validation of PROMIS banks in target clinical populations within the initial 5-year funding period. At the same time, it required us to accelerate the consensus building process surrounding the PROMIS domain framework, and to expedite the QIR protocol activity. As a result, the PROMIS domain framework, which is best viewed as a perpetual work in progress, received only limited input from colleagues outside the PROMIS network and its advisors before widespread testing of constituent item banks. Similarly, although the qualitative research that led to the selection, writing, and rewriting of items was extensive and thorough, we continue to analyze the very rich data even after testing had begun. This implies that there may be some items in the banks being tested that could have benefited from further discussion and refining. We justified this approach with our commitment to revise and retest the item banks in clinical samples to obtain validity data in 2007–2009.

Another reflection with implications for future consequences regards an emphasis placed by the NIH, the PROMIS SAB, and the PROMIS SC on the needs of the clinical researcher. In an ideal world, PROMIS tools will be useful not only in research but also in clinical practice and health

policy applications such as population health measurement or contribution to healthcare reimbursement. Consistent with our roadmap charge, we committed to a clear initial priority by ensuring utility for the clinical researcher. This focus has fueled a productive early emphasis on common chronic medical conditions (eg, cancer, heart disease, depression, arthritis) and clinical trial applications. Collaborations with clinical trials organizations in oncology, diabetes, arthritis, depression, and pelvic floor disorders are underway to test PROMIS measurement tools in clinical trials. In areas outside clinical trials, we have depended upon collaboration with other groups to link PROMIS measures to common measures of population health, or to measure individual patients in clinical practice settings.

Coordinating the PROMIS Cooperative Group in its first 2 years has been challenging and rewarding. The challenge comes from drawing together into a cohesive network several independent research projects led by many of the world's experts in patient-reported outcome measurement and analysis. The reward has been in witnessing how every one of these leaders and affiliated scientists recognize that the PROMIS effort is larger than their individual contributions. Compromise of personal convictions has at times been required. Over these first 2 years this has fostered colleagiality and a conviction that collaboration with respected colleagues outside the network will unify the field.

CONCLUSIONS

Supported by the NIH roadmap effort to re-engineer clinical research, the PROMIS is creating item banks to measure common (generic) health concepts. In our first 2 years, we established an interactive and functional cooperative group governed by a collaborative mix of funded investigators and NIH scientists. Three inter-related research protocols combined expert consensus, rigorous QIR from experts and patients, and empirical analysis of existing data. These efforts informed the building of new item banks that began testing in July 2006. After rigorous psychometric evaluation, the refined item banks will be used to produce clinical research tools such as static short forms and CAT. These tools will be made broadly available for use in clinical research and clinical practice. Collaboration with other groups conducting similar research is strongly encouraged. Information and updates are available at www.nihpromis.org.

ACKNOWLEDGMENTS

Information on the Patient-Reported Outcomes Measurement Information System (PROMIS) can be found at http://nihroadmap.nih.gov/ and http://www.nihpromis.org.

REFERENCES

1. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care.* 2000;38(9 Suppl):II28.
2. Lai JS, Cella D, Chang CH, et al. Item banking to improve, shorten, and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Qual Life Res*. 2003;12:485–501.
3. Choppin B. *Item Banking and the Monitoring of Achievement.* [Research in progress, Series No. 1]. Slough, England: National Foundation for Educational Research; 1978.
4. Choppin B. Educational measurement and the item bank model. In: Lacey C, Lawton D, eds. *Issues in Evaluation and Accountability*. London, England: Methuen; 1981.
5. Bjorner JB, Kosinski M, Ware JE. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Qual Life Res*. 2003;12:913–933.
6. Green B, Brock R, Humphreys L. Technical guidelines for assessing computerized adaptive tests. *J Educ Meas*. 1984;21:347–360.
7. Haley SM, Coster WJ, Andres PL, et al. Score comparability of short forms and computerized adaptive testing: simulation study with the activity measure for post-acute care. *Arch Phys Med Rehabil*. 2004;85: 661–666.
8. Haley SM, Raczek AE, Coster WJ, et al. Assessing mobility in children using a computer adaptive testing version of the pediatric evaluation of disability inventory. *Arch Phys Med Rehabil*. 2005;86:932–939.
9. McKinley R, Reckase MD. Computer applications to ability testing. *Assoc Educ Data Syst J*. 1980;13:193–203.
10. Ware ME, Kosinksi M, Bjorner JB, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual Life Res*. 2003;12:935–952.
11. Weiss DJ. Latent trait test theory and computerized adaptive testing. In: Weiss DJ, ed. *New Horizons in Testing*. New York, NY: Academic Press; 1983.
12. Weiss DJ, Kingsbury G. Application of computerized adaptive testing to educational problems. *J Educ Meas*. 1984;21:361–375.
13. Harniss M, Amtmann D, Cook D, et al. Considerations for developing interfaces for collecting patient-reported outcomes that allow the inclusion of individuals with disabilities. *Med Care.* 2007;45:(Suppl 1):S48–S54.
14. Hill CD, Edwards MC, Thissen D, et al. Practical issues in the application of item response theory: a demonstration using items from the Pediatric Quality of Life Inventory (PedsQL) 4.0 Generic Core Scales. *Med Care*. 2007;45(Suppl 1):S39–S47.
15. World Health Organization. *Constitution of the World Health Organization.* Geneva: WHO; 1946.
16. Fries JF, Bruce B, Cella D. The promise of PROMIS: the new sciences behind patient-reported outcomes. *Clin Exp Rheumatol*. 2005;23:S53–S57.
17. Chen W-H, Lai JS, Cook KF, et al. Evaluation methods for linking pain items from two studies using item response theory analysis. Presented at: The 13th Annual Meeting of the International Society for Quality of Life Research; 2006; Lisbon, Portugal.
18. Lai JS, Dineen K, Reeve BB, et al. An item response theory-based pain item bank can enhance measurement precision. *J Pain Symptom Manage*. 2005;30:278–288.
19. Arnold AM, Psaty BM, Kuller LH, et al. Incidence of cardiovascular disease in older Americans: the cardiovascular health study. *J Am Geriatr Soc*. 2005;53:211–218.
20. Hahn EA, Cella D, Bode RK, et al. Social well-being: the forgotten health status measure. *Qual Life Res*. 2005;14:1991.
21. Kleinman L, Zodet MW, Hakim Z, et al. Psychometric evaluation of the fatigue severity scale for use in chronic hepatitis C. *Qual Life Res* 2000;9:499–508.
22. Cella D, Lai JS, Chang CH, et al. Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer*. 2002;94: 528–538.
23. Stewart AL, Ware JE, eds. *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Durham, NC: Duke University Press; 1992.
24. Tarlov AR, Ware JE Jr, Greenfield S, et al. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *JAMA*. 1989;262:925–930.
25. The Digitalis Investigation Group. The effect of digoxin in mortality and morbidity in patients with heart failure. *N Engl J Med*. 1997;336:525–533.
26. Gaston M, Smith J, Gallagher D, et al. Recruitment in the cooperative study of sickle cell disease (CSSCD). *Control Clin Trials*. 1987;8:131S–140S.

27. Szabo S. The world health organization quality of life (WHOQOL) assessment instrument. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. 2nd ed. Philadelphia, PA: Lippincott-Raven Publishers; 1995:355–362.
28. Chang C-H, Cella D. Equating health-related quality of life instruments in applied oncology settings. *Phys Med Rehabil: State Art Rev.* 1997; 11:397–406.
29. Reeve BB, Hays RD, Bjorner JB, 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(Suppl 1):S22–S31.
30. Hays RD, Liu H, Spritzer K, et al. Item response theory analyses of physical functioning items in the Medical Outcomes Study. *Med Care*. 2007;45(Suppl 1):S32–S38.
31. DeWalt DA, Rothrock N, Yount S, et al. Evaluation of item candidates: the PROMIS Qualitative Item Review. *Med Care*. 2007;45(Suppl 1): S12–S21.
32. WHOQOL Group. The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med*. 1998;46:1569–1585.
33. Ives DG, Fitzpatrick AL, Bild DE, et al. Surveillance and ascertainment of cardiovascular events. The cardiovascular health study. *Ann Epidemiol*. 1995;5:278–285.
34. Psaty BM, Kuller LH, Bild D, et al. Methods of assessing prevalent cardiovascular disease in the cardiovascular health study. *Ann Epidemiol*. 1995;5:270–277.
35. Muthen BO, Muthen LK. *Mplus User's Guide (Version 2)*. Los Angeles, CA: Muthen & Muthen; 2001.

