The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes

J.F. Fries [1], B. Bruce [1], D. Cella [2]

[1]*Department of Medicine, Stanford University School of Medicine, Stanford, California; and* [2]*Northwestern University, Evanston, Illinois, USA.*

Bonnie Bruce, DrPH, MPH, RD, Senior Medical Scientist; James F. Fries, MD, Professor of Medicine; David Cella, PhD, Director, Center on Outcomes, Research and Education.

This work was supported by an NIH Roadmap Project ([Fries, James F.] Patient-Reported Outcomes Measurement Information System: PROMIS) and the Stanford PROMIS Primary Research Site (NIAMS AR052158).

Please address correspondence and reprint requests to: Dr. James F. Fries, 1000 Welch Road, Suite 203, Palo Alto, CA 94304, USA.

Clin Exp Rheumatol 2005; 23 (Suppl. 39): S53-S57.

Key words: Patient reported outcomes, Patient Reported Outome Measurement Information System, PROMIS, item banking, Item Response Theory, Computerized Adaptive Testing.

ABSTRACT

PROMIS (Patient-Reported-Outcomes Measurement Information System) is an NIH Roadmap network project intended to improve the reliability, validity, and precision of PROs and to provide definitive new instruments that will exceed the capabilities of classic instruments and enable improved outcome measurement for clinical research across all NIH institutes.

Item response theory (IRT) measurement models now permit us to transition conventional health status assessment into an era of item banking and computerized adaptive testing (CAT). Item banking uses IRT measurement models and methods to develop item banks from large pools of items from many available questionnaires. IRT allows the reduction and improvement of items and assembles domains of items which are unidimensional and not excessively redundant. CAT provides a model-driven algorithm and software to iteratively select the most informative remaining item in a domain until a desired degree of precision is obtained.

Through these approaches the number of patients required for a clinical trial may be reduced while holding statistical power constant. PROMIS tools, expected to improve precision and enable assessment at the individual patient level which should broaden the appeal of PROs, will begin to be available to the general medical community in 2008.

Introduction

A quarter of a century ago, Patient-Reported Outcomes (PROs) were of only marginal interest to rheumatologists (1,2). The term "outcome" itself was little used. We had "dependent variables" for our clinical trials, which were laboratory-measured or physician-observed. The "gold standards" were the tender joint count and swollen joint count, the physician global assessment, grip strength, ring size, the timed 50-foot walk, the sedimentation rate, X-rays, and disease markers such as the antinuclear antibody titer and rheumatoid factor titer. Now, while some of these measures have survived and even prospered, a new "gold standard" for many if not most rheumatologists has become the patient's own self-report. These measures are truly "outcomes". They are about things that affect patients' lives in major ways. They measure the impact of the disease process, and they reflect patient values (3). Perhaps closer to the heart of some trialists, they often have better measurement characteristics than the more traditional clinical variables, and may in some cases be more reliable, more valid, more meaningful, and less expensive to obtain.

The major PRO instruments in rheumatology and many other disciplines include the Health Assessment Questionnaire (HAQ) and the SF-36, derived from the Medical Outcomes Study, and we write from a long and generally successful perspective on these instruments, which we have long used (4, 5). These instruments have been employed in thousands of studies, undergone hundreds of separate validations, and each has been translated into more than 50 languages and cultures. They have become important standards for the Food and Drug Administration (FDA), the American College of Rheumatology (ACR), and OMERACT, among others. However, they are over 25 years old and new measurement sciences have evolved. We have computers, the Internet, and wireless communications. Scientific advances in measurement and the maturation of consumerism in health care require us to re-examine PRO assessment and "raise the bar" to further extend the application of these concepts (6, 7).

PROMIS (Patient-Reported Outcome Measurement Information System)

In May 2002, National Institutes of Health Director Elias Zerhouni convened a series of meetings to chart an "NIH Roadmap Initiative" for medical research in the 21st century. The purpose of the Roadmap was to identify opportunities and gaps in biomedical research that no single NIH institute could tackle alone, and which could make a significant impact on the progress of medical research. A major intent was to catalyze changes that must be made to transform our new scientific knowledge into tangible benefits for people. A major part of the Roadmap deals with re-engineering the clinical research enterprise, and prominent within this effort is PROMIS, intended to bring the new sciences of Item Response Theory (IRT) (8) and Computerized Adaptive Testing (CAT) (9), long used in educational testing settings, to important new uses in medicine.

PROMIS has been charged with developing large item banks, improving these items, and using IRT and CAT to develop next generation outcome measures which are meaningful, precise, and require fewer patients in a trial to achieve the same statistical power. The PROMIS project began in late 2004, with primary research sites at Stanford University, Duke University, University of North Carolina-Chapel Hill, University of Washington, University of Pittsburgh, and Stonybrook University, and with a statistical coordinating center at Northwestern University. This paper is intended in part to provide some background regarding this project and to introduce IRT and CAT concepts to many readers, but also to indicate some early results in defining a process; to illustrate a number of issues, such as unidimensionality, reliability, and responsiveness that are essential to optimal instrument development; to present a Domain Hierarchy with broad applicability (6, 7); and to identify the effect of these issues on reducing the sample sizes required in clinical trials.

The requisite processes to develop optimal PROs are illustrated in Figure 1, and these steps order the discussions which follow here. The process to develop item banks is complicated and labor-intensive. Qualitative item review and improvement must precede data analysis and the resultant computational and statistical analysis. To develop an item bank that enables short, efficient and precise assessment, lengthy preparation is necessary.

The "Domains" exist on a continuum from more the general to the more specific (8, 9), and the more specific may be collapsed into the more general. Figure 2 shows the evolving PROMIS Domain Framework, representing input from hundreds of people and major national and international organizations. Three levels of the hierarchy are shown. The first, to the left, defines "Health" as consisting of Physical, Mental, and Social dimensions, following the model of health adopted by the World Health Organization and many other groups. At the second level, Physical Health (for example), is conceived as consisting of Physical Function (or Disability) and Symptoms. At the third level of the hierarchy, Function/ Disability conceptually consists of the postulated domains of Mobility (lower extremity), Dexterity (upper extremi-

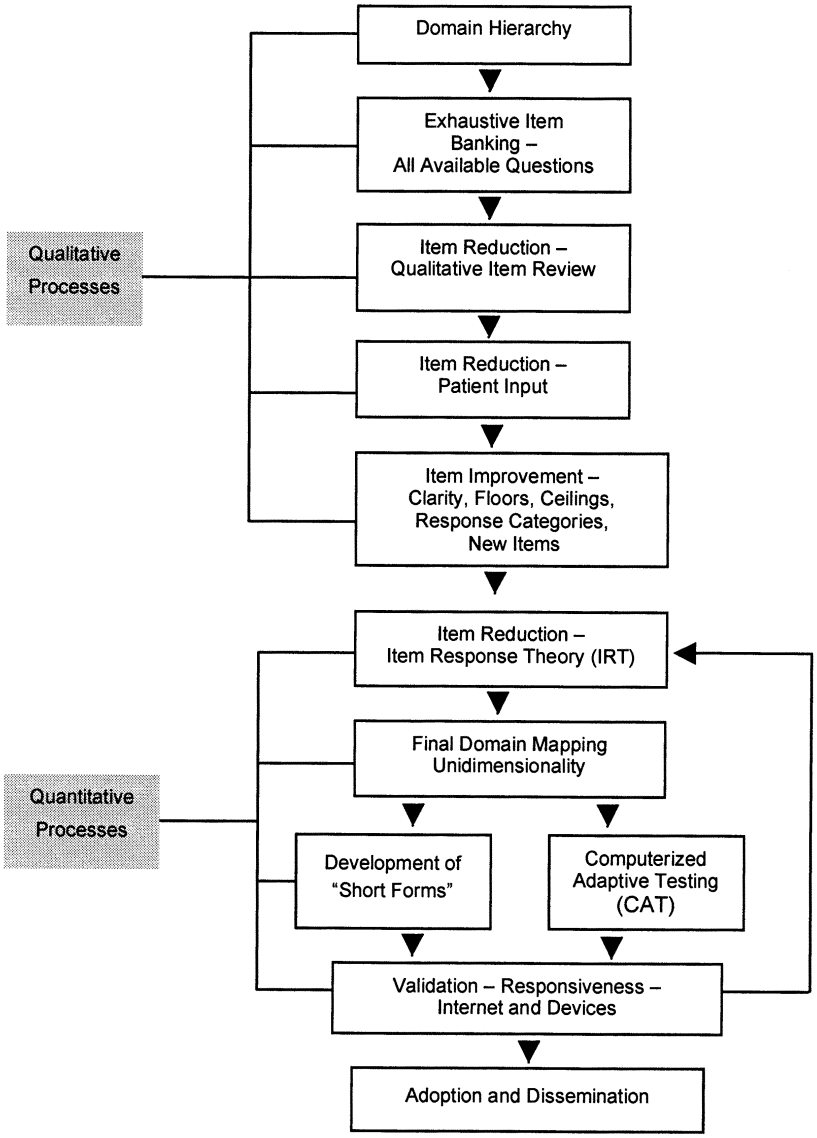

Fig. 1. The Process of PROMIS. The development of improved items and improved ways of using these items proceeds through a number of qualitative steps followed by a number of quantitative steps, each described in text.

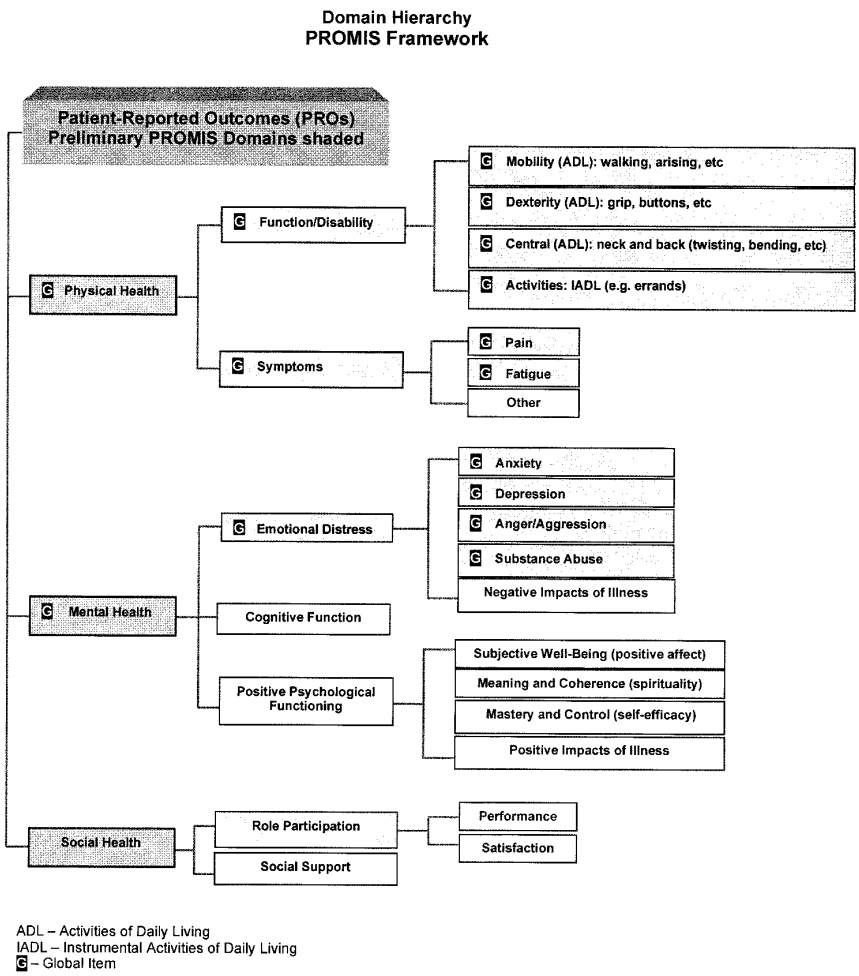

Fig. 2. Domain Framework. This preliminary "Domain Map" proceeds from the more general and conceptual (on the left) to the more data-based and empiric (on the right). The branching process stops when the right-hand boxes are unidimensional by IRT analyses.

ty), Central (neck and back), and Activities, often described as instrumental activities of daily living, or IADL. Symptoms include Pain, Fatigue, and other symptoms. The shaded domains are currently being studied by PROMIS investigators.

As this domain map matures, IRT techniques will be used to assess the domains to the right in order to determine whether the items in that domain all measure the same construct, or whether two or more constructs are included. This amounts to a test for unidimensionality, which is a requirement for most IRT models. If these domains contain only a single dimension, then the mapping task is completed. If there are more dimensions, then the map will continue to expand to the right until unidimensional domains are found. The objective is to have domains, which to the greatest degree possible are mutually exclusive and collectively exhaustive.

It is desirable for CAT applications to have as few domains as possible. Thus, the mapping process begins with conceptual and qualitative decisions and ends with quantitative, evidence-based decisions. For example, the Health Assessment Questionnaire Disability Index (HAQ-DI) has eight domains under Physical Function. This map postulates four, reasoning that you might estimate a stair-climbing score from knowledge of walking ability or a hygiene score from a dressing and grooming score, but you cannot reliably predict hand function from a walking score. Analyses will confirm or deny such postulates with empiric data.

The implications of a Domain Hierarchy Map are of profound importance. They explicitly define the ultimate goals of health care and health policy and the steps by which these goals may be achieved. It is important to "get it right" as completely as possible, and PROMIS investigators will be soliciting continued input over the next several years. Several other domain-mapping efforts are in process nationally and internationally, at slightly earlier stages, and appear to be converging toward eventual consensus with the one outlined here.

Exhaustive item banking

The long journey to practical IRT-based CAT applications begins with the first step of exhaustive "item" identification. A PRO "item" is a question for a patient. As displayed in Figure 3, an item has a context (sometimes termed a "super-stem") which may apply to a group of items in a questionnaire instrument, and a stem, a time frame, and a set of possible responses. It has a domain and a category or sub-domain to which it is temporarily (qualitatively) or permanently (IRT-based) assigned. An item is different from another if it has a different context, stem, time frame, or response options.

A definitive approach to IRT and CAT applications requires that all previously proposed items from all known instruments be considered, in order to reduce bias and to document the breadth of the search. All questions in a bank are carefully scrutinized. This task is not trivial and corresponds to the initial step of meta-analysis where all previous work is identified to the greatest degree possible. In the assessment of physical function alone, for instance, the current PROMIS item banks include approximately 2000 items from over 350 English language questionnaires.

Qualitative item review (QIR)

Initial item pools, the source for item banks, are larger than necessary for a working bank. At the same time they contain many items that are sloppy, imprecise, redundant, grammatically incorrect or potentially offensive, with inappropriate response options or ab-

sent time frames, or are perhaps directed at too high a reading level. Such items may be deleted after an expert review process, where at least three trained raters independently apply a defined set of rules to each item.

The reduced item pool is then tested in the field with patients to ensure that patient inputs and patient values are substantially represented in the final item pools. Techniques include focus groups, cognitive interviews, and patient surveys. Information sought includes the importance of the item to the patient, the clarity of the item, and the ability of the patient to describe the idiomatic meaning of the item. With many items to evaluate, different patient samples will evaluate different item sets, with some common items being used to anchor the evaluations. Weak items uncovered during review by patients will be further culled.

With these inputs noted, the remaining items will be improved further where possible. Changes may be made to improve and then standardize response categories, time scales, item wording, and so forth. Gaps and omissions (particularly with regard to the difficulty of items) will be filled by the writing of new items. Global items to introduce each domain, indicated on Figure 2 by a "G", will be written.

Context:	"Considering your arthritis, are you able to:"
Stem:	"walk a block on level ground"
Time Frame:	"over the past (24 hours, week, two-week, month, six-months, year)"
Response Options:	"normally, with some difficulty, with much difficulty, unable to do (category)" Yes/No, 5-category, 7-category Likert, 10-category scale, 100 category analog scale

Fig. 3. The Anatomy of an Item. An item has a context, a stem, a time frame, and response options; changing any of these makes a new item.

Item response theory (IRT)

Quantitative item analysis using IRT also proceeds through several steps, all of which depend upon the gathering of large amounts of data on individual items from large and diverse patient data sets. The location of each item on the underlying trait is analyzed. This is similar to the difficulty of an item on a test, and indicates how many people can, for example, perform a queried function. Within a given domain, it is useful to have items with a wide range of difficulty. Then items may be screened by correlating them with each other and with an overall index of the domain. Identification of redundant items characterized by high correlation coefficients with each other will allow the deletion of items and reduce their "local dependence"; IRT requires that items in a domain be locally independent of one another. Items which do not correlate well with the other items in a domain do not belong in that domain; if content inspection also confirms a poor fit, such items may be transferred to another domain (8).

Remaining items proceed through formal IRT analyses to confirm the unidimensionality of a proposed domain with a single principal component. These may employ Rasch modeling or more complex models that, for example, take into account the discrimination power (slope) of an item, as well as its difficulty. The generalized partial credit model is one example (10-12). The strategy is to continue to add domains to the right of Figure 2 until domains that are both conceptually and quantitatively unidimensional are found to have no further branches. The underlying hope is that this process will confirm something not too different from the map of Figure 2.

Computerized adaptive testing (CAT)

The term "short-forms", as in the SF-36, is used to denote questionnaire instruments that are not dynamic, but are sufficiently short so that the questionnaire burden to the patient is reasonable. Typically there may be 5-10 items in a domain-specific short form. The ability to create better short forms capable of a more precise estimation of, for example, functional capability because of the use of stronger individual items with closely defined characteristics, is one goal of PROMIS. The techniques above allow instruments that surpass the current standards to be developed, and they also permit the translation of literature results from old to new metrics. For example, the physical function scale of the SF-36 may be calibrated to the HAQ-DI, allowing data to be re-analyzed with instruments which had not been administered, and allowing comparison of literature results across studies which had not been previously possible (13).

The more important advance, however, is the transition to "dynamic" or "adaptive" testing made possible by CAT techniques. As in educational testing applications, CAT makes it possible for everyone to receive a different test, and yet to more precisely estimate the ability of the individual. This enables the achievement of far greater precision without increasing the questionnaire burden, and advances PRO assessment to the age of the computer, the Internet, and the specialized hand-held device (9, 13-15).

Out of many items come fewer. Consider, for example, a domain termed "walking". The HAQ-DI contains two items on walking, and they generate one of four numbers corresponding to normal, mild, moderate, and severe impairment. On a ten-centimeter ruler they might (but actually don't) correspond to 0, 33, 67, or 100 millimeters. In contrast, CAT will first employ a screening question on walking, perhaps estimating one of eleven (0-10 centimeter) points on the scale and will then sequentially ask narrower questions which span this point and then even narrower questions which span the resulting point, until a pre-defined level of precision is obtained, for example, 53 millimeters with a standard deviation of three, after only 3 or 4 questions. The HAQ-DI, in contrast, might have yielded a score of 67 with a standard deviation of at least 20. Having scored one domain, CAT moves to

the next. If the screening question results in the pre-set "floor" of zero, only one question for that domain will be required in most applications. After scoring all of the domains at one level, CAT computes a score for the domain at the next higher level, physical function/disability. Within the questionnaire burden of the "short-forms", CAT can generate far more precise estimates.

A major contribution of the PROMIS project will be the development of PROs which permit smaller sample sizes in clinical trials while retaining the same statistical power (16). A full discussion of this point is beyond the scope of this paper, but power is strongly related to the standardized effect size, which in turn is strongly influenced by the standard deviation of the estimate (17, 18). CAT will reduce the standard deviation and – in some settings, perhaps most – sample size requirements may be reduced by up to one-half. The resulting improved efficiency of trial designs will ease recruitment difficulties, reduce the number of centers required, and decrease the cost of the trial by a proportion similar to the proportion of reduction in the sample size. Over a decade, such methods could save significant money and time in the pursuit of answers to clinical research questions.

Transcendent instruments

After 25 years, the time seems ripe for patient-reported outcome assessment to move to the next level. The HAQ, the SF-36, the AIMS, the WOMAC, the Sickness Impact Profile, and other instruments have a variety of attributes that do not reach the standard currently achievable. With better items, a more carefully developed domain structure, better knowledge of item characteristics, and more efficient and accurate methods of combining the items, improved instruments can certainly be introduced, study sample sizes reduced, and PROs brought to the level of the individual patient. This seems very likely to happen (19).

Nevertheless, there are at the same time liabilities and limitations, disruptions and dangers. To develop a consensus in support of sweeping change is not easy, especially when the standard instruments have been used in large segments of the literature and have been implemented in many languages and cultures. A transition to the use of computers, Internet, or handheld devices that operate CAT procedures involves a learning curve extending to even several years from now. If proposed changes are adopted by the FDA, industry, the American College of Rheumatology, and other major organizations, the transition will be easier. Well-documented advantages, such as reduced study sample sizes and better applicability for the individual patient course, will help inform change.

Validation of the degree of improvement thus becomes very important. If the advances are small, change is far less likely. In PROMIS, protocols of evaluation and validation are being developed. Randomized controlled trials will compare the reliability and responsiveness of the traditional measures with these newer techniques. Throughout PROMIS there will be quality-enhancing evaluations aimed at reducing redundancy, informing the CAT sequences, and examining differential item functioning (DIF) across diseases and medical fields. A bottom line evaluation will lie in the extent to which the emerging PROMIS tools facilitate and improve the quality of clinical research funded by the NIH and others.

References

1. FRIES JF, SPITZ P, KRAINES RG, HOLMAN HR: Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980; 23: 137-45.
2. BROOK RH, WARE JE JR, DAVIES-AVERY A: Overview of adult health status measures fielded in Rand's health insurance study. *Med Care* 1981; 17: 1-131.
3. FRIES JF: Toward an understanding of patient outcome measurement. *Arthritis Rheum* 1983; 26: 697-704.
4. BRUCE B, FRIES JF: The Stanford Health Assessment Questionnaire: Dimensions and Practical applications. *Health and Quality of Life Outcomes* 2003, I: 20.
5. WARE JE JR, KOSINSKI M: *SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version I*. 2nd ed. Lincoln, RI, QualityMetric Inc., 2001
6. FRIES JF, SPITZ PW, YOUNG DY: The dimensions of health outcomes: The Health Assessment Questionnaire, disability and pain scales. *J Rheumatol* 1982; 9: 789-93.
7. FRIES JF, RAMEY DR: Platonic outcomes. *J Rheumatol* (Editorial), 1993; 20: 415-7.
8. WARE JR JE, KOSINKSI M, BJORNER JB: Item banking and the improvement of health status measures. *Quality of Life* 2004; 2: 2-5.
9. WARE JE, KOSINKSI M, BJORNER JB *et al.*: Application of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research* 2003; 12: 935-52.
10. CELLA D, LAI J and ITEM BANKING INVESTIGATORS: CORE Item Banking Program: Past, present and future. *Quality of Life* 2004; 2: 5-8.
11. CELLA D, CHANG CH: A discussion of Item Response Theory (IRT) and its applications in health status assessment. *Medical Care* 2000; 38 (Suppl. 2): 1166-72.
12. MCHORNEY CA, COHEN AS: Equating health status measures with Item Response Theory: illustrations with functional status items. *Medical Care* 2000; 38 (Suppl. 2): 1143-9.
13. FISHER WP, EUBANKS RL, MARIER RL: Equating the MOS SF36 and the LSU HIS Physical Functioning Scales. *J Outcomes Measurement* 1997; 1: 329-62.
14. WARE JE, BJORNER JB, KOSINKSI M: Practical implications of Item Response Theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care* 2000; 38 (Suppl. 2): 473-83.
15. WARE JE, BJORNER JB, KOSINKSI M: Practical implications of Item Response Theory and Computerized Adaptive Testing: A brief summary of ongoing studies of widely used Headache Impact scales. *Medical Care* 2000; 38 (Suppl. 2): 1173-82.
16. KRAEMER HC: To increase power in randomized clinical trials without increasing sample size. *Psychopharm Bull* 1991; 27: 217-24.
17. HOLMAN R: How does item selection procedure affect power and sample size when using an item bank to measure health status ? *Quality of Life* 2004; 2: 9-11.
18. KRAEMER HC, THIEMANN S: *How Many Subjects ? Statistical Power Analysis in Research.* Newbury Park, CA: Sage Publications, 1987.
19. RACZEK A, WARE JE JR, BJORNER JB: Comparison of Rasch and summated rating scales constructed from SF-36 Physical Functioning items in seven countries: results from the IQOLA Project. *J Clin Epidemiol* 1998: 51: 1203-14.