The promise of the future, updated: better outcome tools, greater relevance, more efficient study, lower research costs

James F Fries
Stanford University School of Medicine, Department of Medicine, Stanford, CA, USA
Tel.: +1 650 723 6003;
Fax: +1 650 723 9656;
jff@stanford.edu

Better tools are needed to evaluate our treatments for rheumatic diseases. The long-overdue patient-reported outcome measurement information system project is a NIH roadmap initiative to build better instruments for measuring patient-reported outcomes through use of item banking, item response theory and computerized adaptive testing. The resulting tools are intended to supplant current standards, such as the health assessment questionnaire and the SF-36, and will enable greater study power with reduced sample sizes, as well as integrating greater relevance to the patient into the measures.

In some part of early prehistory, it was noticed that clinical studies required 'dependent variables' by which to judge the study results. In complex chronic illnesses, such as rheumatoid arthritis (RA), such dependent variables need to provide summary end points and to reflect the values of patients as well as health professionals. Under the medical model, dependent variables were traditionally typified by the erythrocyte sedimentation rate (ESR) and the number of swollen or tender joints as counted by a physician. Curiously, these dependent variables were accepted as a matter of faith and were very rarely an object of study. Had they been studied, it would have been easily observed that there was an extraordinary amount of noise in these variables, with great variability in repeated measurements, between observers, between the same observer on different days, and across laboratories. The ESR and the joint counts are simply not very reproducible. It would have been observed just as easily that the ESR or joint counts did not correlate well with how the patient felt or functioned; that is, their very relevance could not be well argued.

Compounding these neglected observations was a statistical failure, where the number of patients (sample size) required for a study was calculated under the assumption that the observed variability in the dependent variable was the true variability of the variable, that all of the observed variability was due to differences in treatment effect across patients. There was no place in such sample size calculations for an 'error term'. As a result, the noise in variables was not assessed, and there was little incentive to improve the precision of the variables. If a study lacked statistical power, you could just enroll more patients, regardless of the cost and resource implications.

Ironically, there were already emerging measures that were more responsive than the old. However, a quarter of a century, ago when the SF-36 instrument from the medical outcomes study and the health assessment questionnaire (HAQ) were new instruments, patient-reported outcomes (PROs) were of only marginal interest to rheumatologists [1,2]. The term 'outcome' itself was little used. The 'dependent variables' for our clinical trials were laboratory measured or physician observed, including, in addition to the above, the physician global assessment, grip strength, ring size, timed 50-foot walk, x-ray and immunological tests, such as the antinuclear antibody titer, and the rheumatoid factor titer. Presently, while some of these measures have survived, the new gold standard for many, if not most, rheumatologists has become the patient's own report. PROs can be true outcomes. They are about things that affect patient lives, they measure the impact of the disease process and they reflect patient values [3]. They usually have better measurement characteristics than the more traditional clinical variables, and, in most instances, are more reliable, more valid, more meaningful and less expensive to obtain. However, they are not as good as they now can be.

The major PRO instruments in rheumatology and many other disciplines include the HAQ and SF-36. Our patient-reported outcome measurement information system (PROMIS) research group is led by the developers of these instruments (myself for the HAQ and John Ware for the SF-36), and we proceed from a long and generally successful perspective on these instruments [4,5]. Our instruments have been used in thousands of studies and hundreds of separate validations, and each has been translated into more than 50 languages and cultures. They have become standards for the US FDA, the American College of Rheumatology (ACR) and outcome measure in RA clinical trials (OMERACT), among others. However, these studies are

Keywords: computerized adaptive testing, determination, item response theory, patient-reported outcome measurement information system, sample size

future Medicine

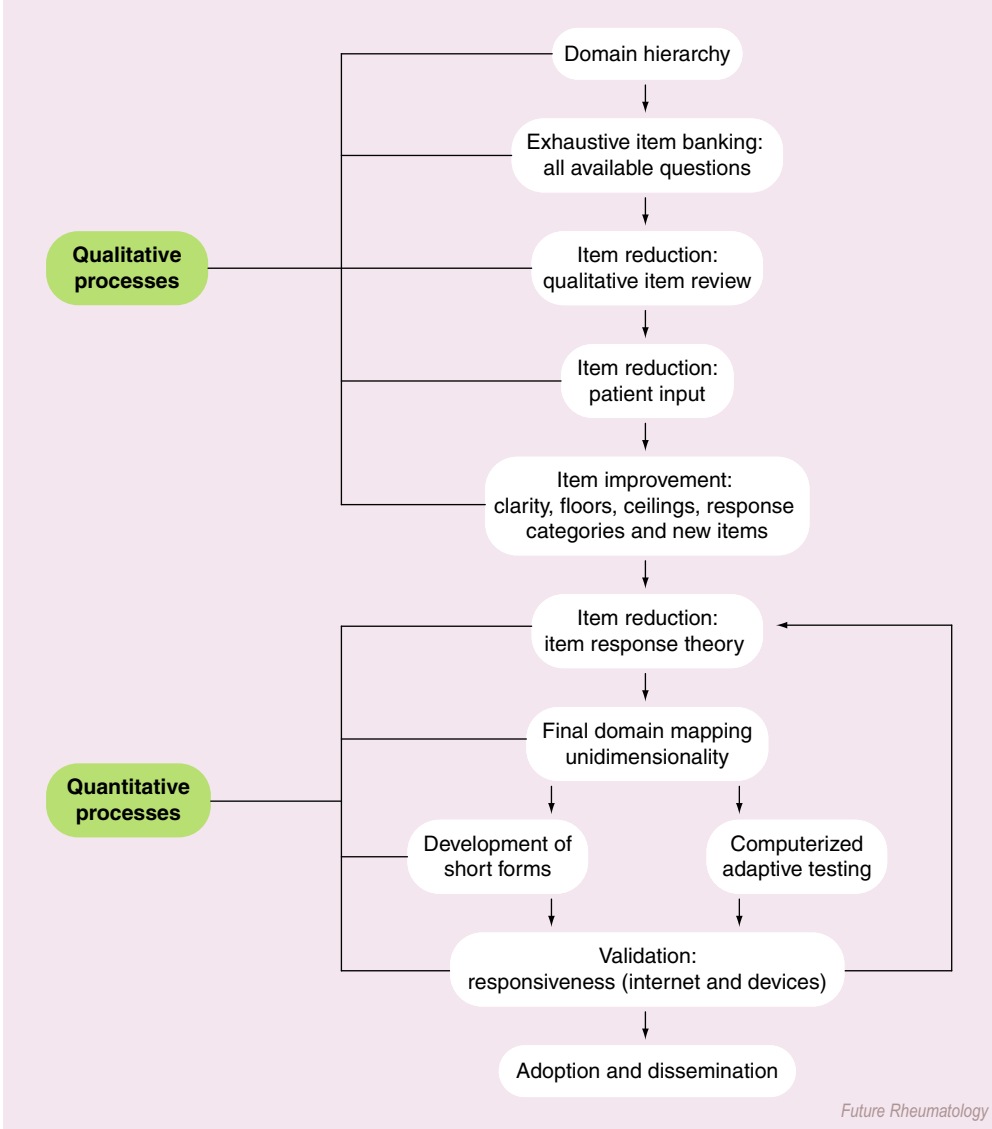
now over 25 years old. New measurement sciences have evolved, which can make good use of computers, the internet and wireless communications. Scientific advances in measurement and the maturation of consumerism in healthcare require us to re-examine PRO assessment and raise the bar in order to extend the application of these concepts [6,7].

PRO measurement information system

The NIH roadmap projects are designed to serve all areas of medicine; to catalyze changes that can transform new scientific knowledge into tangible benefits for people. Part of the roadmap is directed at 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 identifying better treatments.

The process of PROMIS is to develop large item banks of thousands of items, to improve these items and to use IRT and CAT to develop next-generation outcome measures (Figure 1). These new measures will be more meaningful and more precise than previously, and will require fewer patients in a clinical trial to achieve the

Figure 1. 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 the text.



same statistical power. This paper is intended to: explain the process; introduce concepts such as unidimensionality and local independence to many rheumatologists; discuss the concept of a domain hierarchy [6,7]; extend prior discussions of these issues [10]; and estimate the potential effects of these approaches in reducing sample sizes and increasing the precision of clinical trials [11].

The sequence of activities required for the development of more optimal outcome assessment is sometimes tedious, as illustrated in **Figure 1**. The development of large item banks is complicated and labor intensive. Qualitative item improvement must first precede and later accompany quantitative analyses. To develop an item bank that enables short, efficient and precise assessment, lengthy preparation is necessary.

We can think of an item 'domain' as a group of items on a common subject. Such domains exist on a continuum from more general domains to more specific ones [6,7], and the more specific may be collapsed into the more general. The domain hierarchy provides a map to the ordering of domains, and the PROMIS preliminary hierarchy represents analysis of thousands of items and hundreds of instruments, with consultation from hundreds of people and major national and international organizations. Quantitative empirical testing of item clusters has been used to confirm the conceptual framework.

The need for a defined domain hierarchy may not be initially obvious, but there are natural orders for groups of items. 'Walking' may be a part of 'physical function', but physical function cannot be subordinate to walking, nor may walking be subordinate to mental function. Each domain may have a 'score' computed, and the more specific scores reported separately and then 'rolled-up' into more general domain scores.

Three levels of the preliminary PROMIS domain hierarchy illustrate the concept, using physical function/disability as the major example. 'Health', the grandest of domains, is envisioned as consisting of three dimensions of physical, mental and social health, following the definition of health adopted by the WHO and others. At the second domain level, physical health, for example, is considered to consist of the subdimensions of physical function/disability, pain, fatigue and other symptoms. At the third domain level, physical function/disability conceptually divides into domains of mobility (lower extremity items), dexterity (upper extremity items), central (neck and back items) and activities (often described as instrumental activities of daily living [IADL]) (**Figure 2**).

As the domain map continues to divide into ever more narrow domains, IRT techniques are used to determine whether all items in a domain measure the same construct, or whether two or more constructs are required to describe the domain accurately. This involves testing for unidimensionality, a requirement for IRT models, sometimes referring to 'principal component' analysis. If a domain contains only a single construct (such as walking, containing items with a variety of distances, rates and terrains), then the domain-mapping task is completed. All items in walking vary only by their level of difficulty, such as walking a block or walking a mile. If there is more than one principal component present (as with the domain of physical function/disability) then the domain map will continue to be split into narrower subdomains until domains with a single dimension are found.

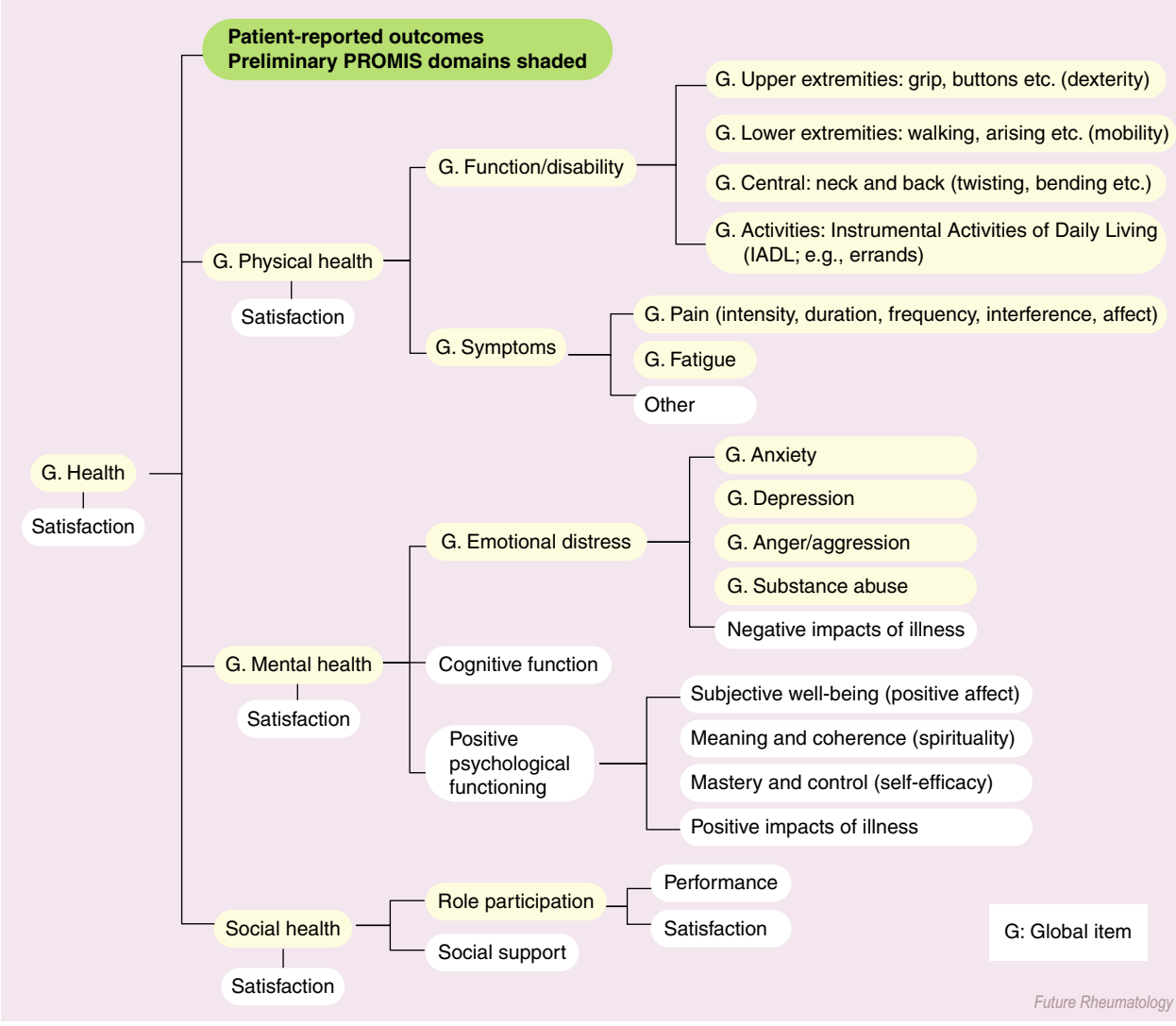
For both CAT and traditional applications, a goal is to have as few and as relevant domains as possible. The domain-mapping process begins with conceptual and qualitative decisions and ends with quantitative, evidence-based decisions. For example, the HAQ disability index (HAQ-DI) has eight domains under physical function; the PROMIS 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.

All items ever used

The item identification process initially establishes the universe of items likely to be useful. An 'item' is a question for a patient. It has four basic components, a 'context' (considering the ways that your arthritis affects you), a 'stem' (are you able to walk a block on level ground), a 'time frame' (considering the past 7 days) and a set of possible 'responses' (without difficulty, with some difficulty, with much difficulty and unable to do). An item has a domain (e.g., mobility) to which it is temporarily (qualitatively) or permanently (quantitatively) assigned. An item is different from another item if it has a different context, stem, time-frame or set of response options.

All previously used items from all known instruments must be initially considered in order to reduce bias and to ensure completeness. Taken together, these items represent the collective wisdom of hundreds of item authors. The item bank development task is not trivial and corresponds to the initial step of a traditional meta-analysis where

Figure 2. The PROMIS domain hierarchy framework (or map).



This model displays the relationships between 'domains', which proceed from more general domains at the left to more specific subdomains to the right. 'G' denotes a domain introduced with a general global item. A 'satisfaction' box is appended to major domains to indicate the importance of this concept and also that it is seldom a primary outcome by itself. The domain map continues to evolve as empiric item response theory data accumulate; this figure reflects status in February 2006.
PROMIS: Patient-reported outcome measurement information system.
With permission from the PROMIS Cooperative Group.

all previous work is identified. In examining the domain of physical function/disability, for instance, the PROMIS item banks identified over 2000 items from over 200 English language questionnaires. To many, this effort may be seen as unexciting, but the strongest buildings must have good foundations.

Losing the bad stuff

These initial item pools are much larger than necessary for a working CAT application, as described later. At the same time, they contain many items that are imprecise, redundant, have inappropriate response options, are grammatically incorrect, are directed at too high a reading level, are absent time-frames or are potentially offensive. Indeed, many items have been found to have been very poorly conceived; for example, 'how much difficulty does your arthritis cause in playing the cello?' or 'how much difficulty do you have with eating or climbing stairs?' 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 with patients, to ensure that patient inputs and patient values are substantially represented in the final item pools. Patient values are assessed by focus groups, cognitive review 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 (back translate) the idiomatic meaning of the item. With many items to evaluate, different patient samples are used to evaluate different item sets, with some common items being used to anchor the evaluations. Weak items uncovered during review by patients are either revised or culled.

After patient views are incorporated, the remaining items are improved further where possible. Changes may be made to standardize response categories, time scales and item presentation. Gaps and omissions, particularly regarding the difficulty of items, are filled by the writing of new items. To date, most identified gaps are regarding the omission of easy items, such as transfer from bed to chair, or hard items, such as jogging two miles. Absence of these items results in floor and ceiling effects, which decrease the reliability of a measure and decrease its generalizability to populations as diverse as nursing-home residents and healthy seniors.

Number-crunching the items

Quantitative item analysis, using IRT, proceeds through several steps, which depend upon the gathering of large amounts of data on individual items from large and diverse patient groups. The location of each item on the underlying trait scale is measured; this is similar to assessing the difficulty of an item on an educational test. Within a given domain, it is useful to have items with a wide range of difficulty. Then, items are screened by correlating them with each other and with an overall index of the domain. Identification of redundant items with high correlation coefficients with each other, enables deletion of items, and reduces 'local dependence' (redundancy); IRT requires that items in a domain be locally independent of each other. Items that do not correlate well with the other items in a domain do not belong in that domain; such items may need to be grouped to form another domain [8].

The items that have survived this gauntlet proceed through formal IRT analyses to confirm the unidimensionality of a proposed domain, seeking a domain with a single principal component. These analyses may employ Rausch modeling or more complex models that, for example, take into account the discriminating power (slope) of an item as well as its difficulty. The generalized partial credit model is one example [12–14]. The strategy is to continue to add levels of ever more specific domains until domains that are both conceptually and quantitatively unidimensional are found, and then to allow no further splitting of domains. The expectation is that this process will confirm a structure not too different from the preliminary PROMIS domain hierarchy [10].

CAT & the mouse

The term short-forms, as in the SF-36, describes questionnaires that are sufficiently short so that the burden to the subject of completing a questionnaire is reasonable. The ability to create better short-forms with more precise estimation of, for example, functional ability, is one goal of PROMIS. IRT techniques combined with item improvements enable development of instruments better than the current standards, and they permit translation of literature results from old to new metrics. For example, the physical function scale of the SF-36 (PF-10) may be calibrated to the HAQ-DI, thereby enabling data to be reanalyzed with instruments that had not been actually administered, and allowing previously impossible comparison of literature results across studies [15].

The more important advance, however, is the transition from static instruments to 'dynamic' or 'adaptive' instruments made possible by CAT. As in educational testing applications, CAT makes it possible for everyone to receive a different and shorter test, as with the graduate record exam (GRE), and also to estimate the ability of the individual more precisely. This enables far greater precision without increasing questionnaire burden, and advances PRO assessment to the age of the computer, internet and specialized hand-held device [9,16,17]. The CAT administers the test, and the subject responds with the mouse.

From 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 no, mild, moderate and severe disability. On a 100-mm ruler, these four response options might (but actually do not) correspond to 0, 33, 67 or 100 mm. By contrast, CAT will first employ a screening question on walking, such as a HAQ-DI item, and will then ask a narrower question

that spans this point, and then an even narrower question that spans the resulting point. This process is continued until a predefined level of precision is obtained, for example, 53 mm with a standard deviation of three, generally achieved after only three or four items. By contrast, the HAQ-DI might have yielded a score of 67 with a standard deviation of 20. Having precisely scored one domain, CAT moves on to the next. If the screening question results in the preset floor of 0, only one confirmatory question for that domain will generally be required. After scoring all of the domains at the third level of mobility, dexterity, central and activities, CAT uses these scores to compute a score for the domain at the second level, physical function/disability. Without additional questionnaire burden, CAT, using the mouse rather than the pencil, can generate far more precise estimates.

Sample size is the cost

The PROMIS project will permit smaller sample sizes in clinical trials while retaining the same statistical power [18]. Full discussion is beyond the scope of this paper, but power is strongly related to the standardized effect size, which, in turn, is a function of the standard deviation of the estimate [11,19,20]. The long neglected fact? As noted above, the standard deviation of an estimate has both a true (latent) term and an error term. Sample size calculations generally assume an error term of zero, even when considering the noisiest of measures. However, the error term is never 0. IRT and CAT will reduce the standard deviation of the estimate, and, in most settings, sample size requirements may be reduced by 25–40% from present levels. The improved efficiency using fewer subjects will facilitate recruitment of subjects, reduce the number of centers required, and decrease the cost of the trial by a proportion only slightly lower than the proportion of reduction of the sample size requirement. Yearly savings in research costs to the NIH alone could be hundreds of millions of dollars.

Transcendent instruments

The new outcome assessment instruments are almost certain to be better than the old. They will use better items in better ways. However, the magnitude of this improvement is important. If the advances are relatively small, profound changes in the way that clinical research is conducted are less likely. Within PROMIS, protocols of rigorous evaluation and validation have been developed. Randomized controlled trials, using different instruments and means of administration as the variables of interest, will compare the reliability and responsiveness of the traditional measures against the newer techniques. PROMIS activities will examine differential item functioning (DIF) across diseases and medical fields, seeking to minimize the number of items that only work well in certain disease populations. A bottom-line evaluation will be the extent to which the PROMIS tools improve the precision and reduce the costs of clinical research. These tools will be in the public domain, and no charges are anticipated for academic uses.

Future perspective

After 25 years, it is time for PRO outcome assessment to move to the next level. The HAQ, the SF-36, Australian Institute of Medical Sciences (AIMS), Western Ontario and Macmaster OA index, sickness impact profile, and other instruments do not reach the standards of precision currently achievable. They are not sufficiently precise to be used to follow the individual patient well. 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 be introduced, study sample sizes reduced and outcome assessment brought to the level of the individual patient [10,21].

However problems will be encountered because of the need for a new consensus in areas less familiar than the old. Acceptance of major change is not often easy to achieve, especially in this setting, where the historical instruments have been used widely and have been implemented successfully in many languages and cultures. The transition to the use of stand-alone computers, internet or hand-held devices that operate CAT involves a considerable learning curve. The clinical trial visit, for example, will include the subject sitting at a keyboard or using a hand-held device to input data. If the proposed changes are adopted by the FDA, industry, ACR, OMERACT 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.

Acknowledgements

This work was supported by a NIH Roadmap Project (PROMIS) and the Stanford PROMIS Primary Research Site (NIAMS AR052158).

Executive summary

- Clinical study evolution has neglected obvious approaches to structural improvements that can enable greater research productivity and more reliable science.
- Patient-reported outcome measurement information system (PROMIS) is a NIH roadmap project to improve the reliability, relevance and efficiency of clinical studies using patient-reported outcomes (PROs).
- PROMIS will provide definitive new PRO instruments that far exceed the capabilities of classic instruments, such as the health assessment questionnaire and the SF-36.
- Item response theory (IRT) now permits us to move forward into an era of item banking and computerized adaptive testing (CAT).
- Item banking uses IRT models to develop higher quality item banks from pools of thousands of items from hundreds of questionnaires.
- IRT enables identification of the best items and assembles domains of items that are unidimensional and not excessively redundant.
- CAT provides software approaches to sequentially select the most informative remaining item in a domain pool until a desired degree of precision is obtained.
- By use of the best items in the best way, the number of patients required for a clinical trial may be reduced by 25–40% while holding statistical power constant, greatly reducing the cost and improving the efficiency of clinical research.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Fries JF, Spitz P, Kraines RG, Holman HR: Measurement of patient outcome in arthritis. *Arthritis Rheum.* 23, 137–145 (1980).
- **Classic original description of development of the health assessment questionnaire.**
2. Brook RH, Ware JE Jr, Davies-Avery A: Overview of adult health status measures fielded in Rand's health insurance study. *Med. Care.* 17, 1–131 (1981).
- **Classic work in the medical outcomes study leading to development of the SF-36.**
3. Fries JF: Toward an understanding of patient outcome measurement. *Arthritis Rheum.* 26, 697–704 (1983).
4. Bruce B, Fries JF: The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health and quality of life outcomes I, 20 (2003).
5. Ware JE Jr, Kosinski M: *SF-36 Physical and Mental Health Summary Scales: a manual for users of version I (2nd Edition).* QualityMetric Inc., Lincoln, RI, USA (2001).
6. Fries JF, Spitz PW, Young DY: The dimensions of health outcomes: The Health Assessment Questionnaire, disability and pain scales. *J. Rheumatol.* 9, 789–793 (1982).
7. Fries JF, Ramey DR: Platonic outcomes. *J. Rheumatol.* (Editorial) 20, 415–417 (1993).
8. Ware JE Jr, Kosinksi M, Bjorner JB: Item Banking and the improvement of health status measures. *Qual. Life* 2, 2–5 (2004).
9. Ware JE, Kosinksi M, Bjorner JB *et al.*: Application of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual. Life Res.* 12, 935–952 (2003).
10. Fries JF, Bruce B, Cella D: The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clin. Exp. Rheum.* 23, S33–S37 (2005).
11. Fries JF, Bjorner J, Bruce B: Reduction in sample size requirements while holding study power constant. *Med. Care* (2006) (In Press).
- **Documentation of abilities to reduce study sample size requirements by 40%.**
12. Cella D, Lai J and Item Banking Investigators: CORE item banking program: past, present and future. *Qual. Life* 2, 5–8 (2004).
13. Cella D, Chang CH: A discussion of item response theory (IRT) and its applications in health status assessment. *Med. Care* 38(Suppl. 9), S1166–S1172 (2000).
- **Excellent introduction to item response theory.**
14. McHorney CA, Cohen AS: Equating health status measures with item response theory: illustrations with functional status items. *Med. Care* 38(Suppl. 9), S1143–S1149 (2000).
15. Fisher WP, Eubanks RL, Marier RL: Equating the MOS SF36 and the LSU his physical functioning scales. *J. Outcomes Measurement* 1, 329–362 (1997).
16. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med. Care* 38(Suppl. 9), SII43–SII59 (2000).
- **Excellent article to introduce cross-calibration of instruments.**
17. Ware JE, Bjorner JB, Kosinski M: Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Med. Care* 38(Suppl. II), 1173–1182 (2000).
- **Excellent introduction to IRT and computerized adaptive testing.**
18. Kraemer HC: To increase power in randomized clinical trials without increasing sample size. *Psychopharmacol. Bul.* 27(3), 217–224 (1991).
19. Holman R: How does item selection procedure affect power and sample size when using an item bank to measure health status? *Qual. Life* 2, 9–11 (2004).
20. Kraemer HC,Thiemann S: *How Many Subjects? Statistical Power Analysis in Research*. Sage Publications, Newbury Park, CA, USA (1987).
- **Definitive approach to sample size issues.**
21. 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.* 51, 1203–1214 (1998).

Affiliation

- *James F Fries*
  Stanford University School of Medicine, Department of Medicine, Stanford, CA, USA
  Tel.: +1 650 723 6003;
  Fax: +1 650 723 9656;
  jff@stanford.edu