Editorial

Items, Instruments, Crosswalks, and PROMIS

In the English language, there are 165 published questionnaire instruments intended to assess "disability" or "physical function" health outcomes, containing 1860 items[1]. If "health related quality of life" (HRQOL) questionnaires are included, the numbers are yet larger. With integration of the new sciences of Item Response Theory (IRT)[2] and Computerized Adaptive Testing (CAT)[3] into instrument development, the numbers of items and instruments could again grow larger. This unbounded proliferation of health status instruments is problematic and raises both serious issues and intriguing opportunities.

For understanding these issues, some knowledge of IRT and CAT is required. IRT[2,4] works at the level of the specific item, which has measurable characteristics such as information content, degree of difficulty, reliability, clarity, ease of translation, performance in different populations, importance to the subject, and others. IRT is sometimes termed "latent trait theory" as a major application of IRT is to estimate the value of a trait (or domain) such as "disability" or "quality of life," where the trait itself cannot be directly observed. Two IRT requirements are that the items aggregated to estimate a trait are unidimensional in that they measure a single concept, and that they are not redundant ("locally dependent") with other items in the group[2]. Given sufficient information about each item, one can predict the performance of one outcome assessment instrument compared with another. For example, one can quite readily, by selecting the better items from an item bank, create instruments that make more precise outcome assessments than the Health Assessment Questionnaire Disability Index (HAQ-DI)[5] or the 6-item Short Form Health Survey (SF-6D) HRQOL[6] instruments. In turn, this permits major increases in study statistical power or allows use of many fewer items for the same level of precision[7].

CAT replaces the concept of an instrument with a fixed set of items, offering a dynamic selection of the best items for each subject, based on the subject's responses to prior items[4]. This allows efficient estimation of the latent trait for an individual using only a few items and over a wider range of disease severity. Using IRT and CAT, a new generation of better items, better algorithms, and more definitive results can appear.

ISSUES

Unnecessary proliferation of outcome assessment instruments serves us poorly, yielding a Tower of Babel, where all speak a somewhat different language and in which small differences are exploited, validations are fewer and weaker, translations are less available, benchmarks generally absent, and the latent traits not consistently or clearly defined. This is the province of the "splitter," where nuance crowds out substance, and at some point the liberty to innovate becomes license. On the other hand, the alternative view of the "lumper" brings another set of problems, of the "one size fits none" type. Excessive standardization prevents the tailoring of instruments to different diseases or populations and at some level stifles innovation. It implicitly suggests that improvement of present instruments is not possible. We need to harmonize these approaches.

Another issue involves the definitions of the domains themselves. Physical function and disability scales usually contain items about the ability to "walk a block" or "dress yourself"; this is an "ability" domain, and it is the most commonly studied. Alternatively, you could ask whether the individual had actually performed the activity in the past week; this is a "performance" physical function domain. Similarly, there could be a "social participation" physical function domain or a "satisfaction with physical performance" domain. Or, there could be a "HRQOL" domain estimated from the SF-6D or other QOL instrument. There are advocates for each of these domain constructs, and a case can be made for each of them as desirable trial endpoints under particular circumstances, such as study of antidepressants, physical therapy, or rehabilitation. Yet, there are large areas of overlap between these 5 domains and many items that are very similar in instruments intended to estimate dif-

See Sensitivity of HAQ-DI SF-6D in early aggressive RA, *page 1150*

ferent latent traits. The domain concepts are sufficiently nuanced that many lay subjects may have trouble understanding the differences and may give the same answer to conceptually different items. Again, there is the issue of splitting versus lumping; the scholar wants the nuanced distinctions and the untutored subject may be unable to make these distinctions.

OPPORTUNITIES

There are many and varied research questions, including those above, that accompany the transformation of outcome assessment into an IRT-based environment with greatly enhanced capabilities. These capabilities are ultimately based more on the item than on the instrument. Given a universe of a few thousand items, the number of possible instruments that represent combinations of these items is very large indeed, and with the maturation of CAT applications almost every patient will have a unique instrument crafted just for them in real time.

In approaching solutions to these issues, a first liberating step would be to have all agree that items and their IRT item characteristics should be in a common accessible item bank. This item bank should have defined methods for adding or subtracting items over time. This requires, among other things, that items be considered to reside in the public domain. Instruments, on the other hand, can in some instances represent intellectual property and (a very few instruments) have been licensed or marketed by their developers, who retain some rights to derivative instruments.

A second liberating step is the "crosswalk," an unfamiliar term that will become more familiar over time, as it addresses most of the issues raised here. The crosswalk estimates scores on one item or instrument from scores on another, and back again. It could be the HAQ and the SF-36, the SF-36 and the Arthritis Impact Measurement Scale, or the HAQ and the SF-6D, representing HRQOL. Or, one can crosswalk domains, as with translating an "ability" physical function bank into one representing literal performance of the activity. Crosswalks can be performed in a number of ways, many largely unexplored, including mapping, ranking, correlation, and multiple regression. A crosswalk will sometimes reveal large areas of redundancy and small differences in scores across instruments or domains. Thus, the academic distinctions may be found to be quite unimportant in practice. Or, the crosswalk may document that only a few items from one instrument predict scores on the other and use only these items for the crosswalk.

A third liberating step is PROMIS (Patient-Reported Outcomes Measurement Information System), a large multicenter US National Institutes of Health (NIH) Roadmap initiative designed to improve the infrastructure of clinical science through improved outcome assessment, including item improvement, IRT, and CAT (www.nihPROMIS.org). Some 200 participating investigators and collaborators have worked over the past 5 years to develop domain hierarchies and definitions, IRT calibrated item banks, a variety of improved short-form instruments, and CAT applications. PROMIS item banks are in the public domain, as are PROMIS short-forms of 4, 6, 8, 10, and 20-item lengths, and soon a PROMIS CAT. The PROMIS Assessment Center (www.nihPROMIS.org) is an open resource that can collect data online for studies and can create new short-form instruments for specific purposes. The PROMIS item banks include the IRT calibrations on the items, drawn from validation studies of over 21,000 individuals. The next 4 years of PROMIS are focused upon building collaborative efforts to support the field and resources to make scientific inquiries more efficient, including investigation of issues as discussed here. We ask for your help in extending these efforts and their applications. You are welcome.

In this issue of *The Journal*, Amjadi, *et al* report a "crosswalk" study (although they do not use this term) supporting the use of the HAQ-DI derived SF-6D in RA cohorts and clinical trials that lack preference-based measures[8]. The study is from a well respected and experienced clinical and HRQOL group and is carefully performed and conservatively interpreted. They report that you can use selected HAQ items to estimate quality-adjusted life-years with acceptable accuracy. Thus, a disability measure can be translated into a preference/utility measure, a crosswalk between 2 seemingly quite different domains. This is one of the first of such studies and requires confirmation; it is consistent with some prior, more primitive work, for example, correlation of HAQ patient global scores with the Torrance feeling thermometer, a HRQOL instrument[9]. This kind of effort, we believe, will help blur domain boundaries and serve as a uniting focus for the future. Crosswalks, common item banks, and the PROMIS resource will help us better understand our outcome domains and their relationships to each other, reduce redundancy in domains and instruments, and act to moderate excessive instrument proliferation.

JAMES F. FRIES, MD,
Professor of Medicine;
ESWAR KRISHNAN, MD;
BONNIE BRUCE, DrPH,
Senior Research Scientist,
Division of Immunology and Rheumatology,
Stanford University School of Medicine,
Palo Alto, California, USA

REFERENCES

1. Bruce B, Fries JF, Ambrosini D, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Rheum 2009; [submitted]
2. Embretson SE, Reise SP. Item response theory for psychologists. London: Lawrence Erlbaum Associates; 2000.
3. Wainer H, Dorans N, Flaugher R. Computerized adaptive testing: A primer. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 2000.
4. 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:S22-31.
5. 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.
6. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1:4.
7. Cella D, Yount S, Rothrock N, et al. The Patient Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45:S3-S11.
8. Amjadi SS, Maranian PM, Paulus HE, et al. Validating and assessing the sensitivity of the Health Assessment Questionnaire-Disability Index-derived Short Form-6D in patients with early aggressive rheumatoid arthritis. J Rheumatol 2009;36:1150-7.
9. Fries JF, Ramey DR. "Arthritis specific" global health analog scales assess "generic" health related quality-of-life in patients with rheumatoid arthritis. J Rheumatol 1997;24:1697-702.

J Rheumatol 2009;36:1093–5; doi:10.3899/jrheum.090320