

Building, Testing, and Improving an Item Bank for Chronic Disease PROs: *Keeping a PROMIS*

**Joseph Lipscomb, PhD
Chief, Outcomes Research Branch
Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute**

A. A national item bank that paves the way for computer adaptive assessment (“testing”) (CAT) of patient-reported chronic disease outcomes can potentially strengthen outcomes assessment within clinical research studies and in other applications.* A landmark effort to put this idea into practice is the Patient-Reported Outcomes Measurement Information System (PROMIS), the cornerstone of the NIH Roadmap initiative on “Dynamic Assessment of Patient-Reported Disease Outcomes.” As envisioned, PROMIS will facilitate administration of tailored questionnaires for measuring a patient’s health status on a number of domains, including symptoms; collect PRO data for research and, in the process, continue to update and refine the measurement system; and has the potential to provide health status reports to patients and health care providers to enhance decision making. Moreover, one can further envision how PROMIS could eventually facilitate collection of population-based data on chronic disease outcomes, permitting large-scale monitoring of health status for whole populations or defined subgroups.

Now, there is much more to PROMIS than item banking and CAT. The initiative will pave the way for a range of analyses exploring the potential advantages and opportunities offered by modern measurement approaches like item response theory modeling, e.g., investigating differential item functioning, cross-walking instrument scores, and so on. The outcomes measurement research agenda of any NIH institute (like the National Cancer Institute) should be

informed by, and seek to capitalize on new and promising developments in IRT-based approaches for assessing patient-reported impacts of disease and interventions. Still, the centerpiece of PROMIS is a national item bank for PRO assessment, and the discussion below is oriented accordingly.

Such a national-level item bank – indeed, dynamic measurement system – will involve no small amount of resources over time in terms of dollars, administrative support, and the need (in principle) to coordinate and obtain cooperation from multiple public and private organizations. What do we hope to get for it? In the end, we seek PRO measures that are superior to what is currently available from fixed-instrument approaches according to some “weighted average” of the following criteria, which derive in part from principles promulgated by the Medical Outcomes Trust¹: validity (particularly construct and criterion in the present case), reliability, responsiveness, interpretability, comparability, simplicity, and feasibility.² (Not all experts would break out the relevant criteria into precisely these categories,³ but most would concur that the concepts embodied herein span the relevant concerns.) The relative “weights” attached to these sometimes-conflicting criteria should be applied, as explicitly as possible, by the decision maker at hand in arriving at a final judgment.

This leads to quite specific evaluative questions. On balance (that is, weighing the various criteria), do CAT-based measures of PRO beat out fixed-instrument approaches, either IRT-based or Classical Test Theory (CTT)-based? The fact that CAT approaches may add apparent complexity to the measurement process is not *ipso facto* a reason to reject the approach, or even to lean excessively against the wind *a priori*. Rather, such concerns should lead us naturally to other, constructive questions. After a system like PROMIS is in place and

functioning, and analysts and users gain experience with the methodology and its terminology, will the analyses and the interpretation of findings still be regarded as more “complex” than CTT true-score results? Even if that is the case, is the additional complexity worth it or not, as adjudicated *via* application of the criteria noted above? Specifically, what is the value-added, in terms of producing better patient-level or group PRO scores, of banks and CATs?

B. For this potential to be achieved, what must be accomplished?

1. *Before they can come, you first must build it.* That’s what the PROMIS does. In due course, this dynamic assessment system will or should handle a range of types of PROs, from symptom-oriented to subjective patient evaluation of HRQOL. It should also be applicable to a number of high-prevalence, high-burden chronic diseases.
2. *Then you must test it.*
 - a. Validation should be undertaken at two levels: first, through the scientifically rigorous construction of the item bank itself, and second, through ongoing construct validation and (if applicable) criterion validation.
 - b. Head-to-head comparison studies with leading fixed-item instruments, whether scored through IRT or CTT, are critically important. By the same token, these fixed-item instruments should be subject to the same level of construct and criterion validity testing as the item bank. How an item bank for latent variables such as HRQOL stacks up against an item bank for an educational testing construct like math ability is interesting to ponder, but not directly relevant. The important question here is: for measuring patient-

reported health outcomes, how does the item bank perform compared with contending alternative approaches, e.g., CTT-based fixed-item instruments, according to well-defined evaluation criteria?

3. *Item bank must be maintained and improved over time.* We must encourage participation not only by the initial PROMIS investigators, but top-flight measurement, social science, and clinical researchers worldwide. This will require careful attention to incentive structures and mechanisms. Over time, we will want to: (1) augment the existing bank with new items; (2) consider development of “branch banks” for specific diseases; and (3) encourage application of innovative methods and exploration of new issues, e.g., multidimensional IRT and how to “bank” it, the relationship of preference-based measures to latent variable measures estimated from banks, extended notions of construct validation, e.g., cast in terms of how well scale scores predict patient decision making (either revealed-preference or contingent valuation).
4. *All the while, we don't want to stifle or preclude start-up efforts to build “competing” PRO item banks.* True, multiple national item banks could put a damper on achieving maximal comparability across studies. If study A used item bank X and study B used item bank Y, findings from A and B would not be directly comparable (unless some form of item equivalence linking was done across banks). However, competition between banks could spur improvements in each one individually, with the overall state of the science advancing more rapidly than if there was but one bank (with quasi-monopoly status). At the same time, some researchers and end-users might cite additional, practical concerns. Would

a national item bank effectively “crowd out” smaller, privately developed banks in the competition for public and private research support and clients (end users)? Would a national bank stifle the development of additional disease-specific or condition-specific items or questionnaires that happen not to be well developed through that national bank?

This is not at all to imply that a national bank is akin to the “evil monopolist” that must be vigilantly monitored. Indeed, the more apt economic metaphor is that of the public good; it is unlikely that anything of the scope and scale of the PROMIS would have emerged in the absence of NIH leadership. Rather, the important issue at play is how to capitalize on the strong scale economies of a national item bank – both in terms of resource deployment and concentration of intellectual firepower – while keeping the door ajar, and perhaps encouraging entry of innovative researchers and end users. The aim here is not the proliferation of “small” banks that would add little, substantively, beyond what can be found in a well-constructed national bank. Instead, one might look to encourage small banks that offer a genuinely differentiated product compared with the national bank, e.g., the availability of items especially tailored to specific diseases, health conditions, or other circumstances.

5. *Ongoing public support of national item bank and CAT may be required, post-development, until there has been adequate field testing of validity and feasibility.* Practically, “adequate” may be defined in terms of when major regulatory bodies, like the FDA, judge they have enough evidence to make a dispassionate decision about whether CAT-based PROs will be accepted for decisions about product

approval and marketing. When that occurs, industry is much more likely to embrace CAT, at least for some studies; and the foundations are strengthened for a public-private (NIH-industry) partnership to support and nurture a national bank.

6. All the while, we should *encourage a range of CAT applications of PROs*, not only in clinical research studies, but to foster better provider-patient communications and decision making (micro-level applications) and also to monitor population trends in health status (macro-level, or surveillance, applications).

C. Keeping our PROMIS, today and tomorrow: what will need to be done?

1. *There must be adequate public funding* (hopefully augmented with private dollars) to maintain a national item bank not only through development, but validation and testing.
2. *Regulatory agencies, and other important public and private organizations (e.g., purchasers) that may face a choice about whether to accept CAT-generated PRO data, should work towards developing clearly articulated scientific criteria to guide their decisions.* Only then can those developing, validating, and testing the national item bank(s) know exactly where the bar is being set, and why. It would be unfortunate if the bar, rather than being clearly articulated in psychometric terms that are transparent to all, is left rather vaguely “out there, somewhere,” and forever just beyond the reach of this (or any) new, complex approach to measurement. Rather,

IRT-based CAT, IRT-based pencil-and-paper instruments, and CTT-based pencil-and-paper instruments should be judged together according to how well the PROs they yield measure up to the relevant criteria: validity, reliability, responsiveness, interpretability, feasibility, and indeed the other elements put forth by the Medical Outcomes Trust. For that matter, the same holds for computer-based assessments that are not *adaptive* as such, but rather present the respondent with a fixed set of items, which may be IRT based or CTT based. In any event, the relative weighting accorded to these evaluation criteria is, of course, the agency's decision. And it is a decision that should be clearly articulated when the moment of truth arises.

3. Incentives will be required to induce the best measurement and clinical scientists to contribute to the national item bank's empirical and methods work over time.

Creative, flexible strategies must be considered that give due weight to intellectual property rights, professional career development (including publication and tenure), financial factors, and the opportunity to play an important role in and contribute to a major national measurement project. One strategy for promoting inclusiveness and broad participation early on might be the creation of "user groups" or "working groups" to test, apply, and provide feedback on the products being created by the national bank.

4. NIH institutes and study sections must carefully balance the desire to encourage broad use of the national item bank and CAT in supported research studies with the flexibility to allow other, traditional measurement approaches. If CAT advocates get too far ahead of the curve on implementation, there is the risk of backlash from investigators, especially if the scientific motivation for the new approaches is not yet

well understood. Similarly, one must guard against reflexive rejection or undue criticism of CTT-based measures, which may perform quite admirably in many circumstances, simply because they are not obviously on the “cutting edge.” On the other hand, for Institutes and study sections to resist IRT-based CAT because, and only because, they do not in fact adequately understand its strengths (as well as its limitations) is not in the spirit of publicly supported scientific research. Complexity for complexity’s sake is of course not good. Simplicity for its own sake, however, is not obviously optimal either. Rather, the economist’s decision-theoretic principle should hold forth in this sphere of measurement models and approaches: add modeling complexity up to the point where benefits and costs balance on the margin.

In sum, a major effort to build a national item bank for patient-reported chronic disease outcomes has been launched by the NIH – the CAT is out of the bag. But critically important tasks, issues, and questions remain. The bank must be solidly constructed, validated, and tested. These analyses must be sufficiently rigorous and dispassionate to allow key agencies and other organizations – which are presently waiting and watching with elements of hope, skepticism, optimism, and concern – to draw defensible conclusions about whether item banking and CAT represent the new state of the science for PRO assessment, or rather work still in progress.

* The author has benefited from very useful comments by Ron Hays, Eleanor Perfetto, and Bryce Reeve.

REFERENCES

1. Lohr KN. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research* 11:193-205
2. Lipscomb J, Gotay CC, Snyder C. (2004). Introduction to *Outcomes Assessment in Cancer*. In *Outcomes Assessment in Cancer*, J. Lipscomb, C.C. Gotay, C. Snyder (Eds), Cambridge University Press. In press.
3. Hays RD, Hadorn D. (1992). Responsiveness to change: an aspect of validity, not a separate dimension. *Quality of Life Research* 1:73-75.