Next Steps in Use of IRT in the Assessment of Health Outcomes

Ron D. Hays, Ph.D.
UCLA Department of Medicine/RAND Health Program
June 2004 Draft

Health outcome researchers have only recently been exposed to item response theory and its potential. One of the earliest applications in the field was in the creation of the Functional Independence Measure (FIM). The FIM was designed to deal with the problem of lack of a standard measure assessing disability and rehabilitation outcomes. It was developed after a review of 36 functional assessment scales by a national task force (Granger et al., 1986). The FIM measures the extent of independence in self-care, sphincter control, transfers, locomotion, communication, and social cognition. Each item has a raw score ranging from 1 to 7, with 1 representing the need for total assistance and 7 indicating complete independence. There are 18 items in total for a total raw score possible from 18 to 126. Rasch analyses of the FIM lead to the identification of a motor domain having 13 items and a cognitive domain consisting of the other 5 items (Heinemann et al., 1993). Shortly thereafter, Haley, McHorney and Ware (1994) published a Rasch evaluation of the SF-36 physical functioning scale in the Journal of Clinical Epidemiology. The application of IRT to such a widely used health outcome measure began to stir wider interest in the methodology. A few years later, Gonin, Lloyd and Cella (1996) used Rasch modeling to explore the equivalence of the Functional Living Index-Cancer and the Functional Assessment of Cancer Therapy scales. During the last few years the level of interest has accelerated and use of IRT methods among health outcome researchers has increased substantially (e.g., Chan, Orlando et al., 2004; Scholle, Weisman et al., 2004).

This conference represents one of the first systematic opportunities to confront many of the thorny issues in applying item response theory to the patient-reported outcomes field. We began on Thursday morning with Neil Aaronson's review of the current state of the science of health outcomes measurement and a juxtaposition of classical and modern test theory approaches to outcomes measurement by Ronald Hambleton. This was followed with sessions focusing on how IRT can be used to evaluate items and scales and to evaluate the equivalence of measure across populations. Thursday concluded with sessions devoted to use of IRT for linking and equating of measures, and for computer adaptive testing. This morning we turned to concrete demonstrations of the use of IRT to analyze patient-reported outcome questionnaires, followed by discussions of item banking, the possibility of a national item bank, and computer adaptive assessment.

Several potential opportunities for using IRT were noted over the last 2 days in terms of evaluating survey items, improving the efficiency of administration, and increasing the usefulness of health outcomes information in clinical settings. Lingering issues touched upon at the conference include how to best assess local dependence and sufficient unidimensionality, chose between simpler and more complicated IRT models, assess context and sequence effects on item responses produced by computer administration, decide upon the size of differential item and scale functioning that is minimally important, clarify sample size requirements for different IRT models, identify optimal

stopping rules for CAT applications, and investigate the possibilities for public domain versus proprietary item banks. How do we best move ahead to take on the multitude of challenges? I touch upon five issues in the remainder of this paper: 1) collaboration between academia, government and industry; 2) common versus unique item banks; 3) establishing standards for use and reporting of IRT; 4) demonstrating the value of IRT; and 5) continuing efforts to improve the user friendliness of IRT software.

Moving forward will require a collaborative effort between academia, government agencies, and industry to design and conduct research that will help produce the expected benefits. One approach to begin to address the practical issues is to encourage and fund collaborative agreements that bring investigators with complimentary expertise and diverse perspectives together with government project officers and representatives from industry. The collaborative should tackle the most pressing issues and create a framework to continue subsequent progress in integrating IRT into the health outcomes field. The effort would deal with long-standing substantive issues in measuring health outcomes in tandem with facilitating applications of IRT methods to health outcomes data. The overall intent would be to incorporate IRT alongside other methods (e.g., expert and stakeholder input, focus groups, cognitive interviews, classical test theory analyses) to improve the measurement of patient-reported outcomes.

One of the major challenges facing those conducting work in the health outcomes field is selecting the best measure for a particular application. The shift from fixed-length surveys to CAT and item banks provides potentially greater flexibility and a wider pool of item content appropriate to individual respondents, but we still face the possibility of multiple item banks (versus multiple fixed-length generic and targeted survey instrument) that are not linked to one another. Should we work toward a common item bank or in the spirit of competition develop multiple banks? Most likely we will want to do both. A common bank developed with collaboration from investigators from multiple institutions could be very valuable to the field. However, individual investigative teams who have unique and creative ideas for item bank creation and CAT development should be encouraged to pursue their interests and push the envelope from another angle.

The establishment of standards for use and reporting of IRT results in research articles will help users and consumers of the methodology. Existing psychometric guidelines for the health outcomes field such as those published by the Medical Outcomes Trust (Lohr et al., 1996) need to be expanded to incorporate "modern" psychometric methods. These standards should speak to the minimal requirements for evaluating and reporting on tests of model assumptions, model fit, and model parameters.

Clear documentation of how IRT can lead to better patient-reported outcome measures and more accurate understanding of substantive issues will be needed. Because IRT models may better depict actual response patterns and IRT estimates more accurately reflect true status than classical test theory estimates, use of IRT should lead to measurement that is more sensitive to true cross-sectional differences and be more responsive to change in health over time. IRT researchers need to provide examples illustrating when the greater complexity of IRT model-based estimates makes a

difference in understanding the big picture issues the health outcome field is confronting. Grant support of demonstration projects aimed at evaluating the usefulness of IRT (e.g., item banks and CAT, linking, DIF assessment) in improving the assessment of health outcomes for research and clinical applications is needed. Consideration should be given to projects directed at evaluating the cost effectiveness and incremental value of IRT-based CAT estimates of health versus fixed length surveys for clinical trials and assessment of individual patients in clinical practice.

Opportunities for education about IRT are more widespread now as interest in the method has spread and created a demand. The number of available IRT workshops at different venues and applications of the methodology at annual meetings such as the International Society for Quality of Life Research has increased during the last few years. It will be important for IRT experts to communicate with those who do not yet have the background needed to fully appreciate the potential of the methodology. An example of effective communication to build upon is a 1998 newsletter piece written by Bjorner and Ware (1998) on behalf of the Medical Outcomes Trust.

This is the first health outcomes measurement conference devoted solely to IRT. As collective experience is gained and the field matures there will surely be a need for a follow-up conference to revisit the topics touched upon during the last 2 days and discuss new issues. Such a conference might be held about 3 years from now to allow enough time for the emergence of a critical mass of health outcomes researchers who would benefit from sharing of knowledge and experiences from ongoing research and clinical applications.

Government and industry also need to invest in enhancing the user-friendliness of the tools of the trade. The quality of IRT programs has come a long way in recent years. For example, there have been noteworthy improvements in going from Bigsteps to Winsteps, packaging of a cadre of IRT programs by Scientific Software International, and constructing SAS (Christensen & Bjorner, 2003) and STATA (http://freeirt.free.fr/) code to implement Rasch models, but I believe it is still the case that program "documentation is often difficult to read, and finding out the reason for program failures can be time consuming and frustrating" (Hays, Morales, & Reise, 2000, p. II-39). For example, a well-known health outcomes researcher sent me an email on April 14th saying the following:

> I'm mystified by a run error with MULTILOG. I have the following data file (actually it is the 10 PF items from SF-36 for men (group 1) and women (group 2)) … When I try to run this, I get an error ("Error encountered while running the command file: ... see output file for details"). The output file that I get is as follows, and I'm damned if I can figure out what the error is! Help?

The solution to this problem was to substitute (I1, 20a1) for (21a1) that was in the last line of input. Little syntax problems like this are an example of lack of flexibility in existing programs that create big problems when investigators are learning to use IRT.

Hence, SBIR funding to help make IRT programs easier to use could be a good investment of government money. For example, there was once a shortage of good power analysis software available to the general public. The development of the nQuery Advisor® software for conducting power analyses by UCLA statistician Janet Elashoff illustrates one payoff of SBIR funding (http://www.statsol.ie/nquery/nquery.htm). This program and others like it now make conducting power analysis routine and less prone to error than before their development.

Some might argue that this is similar to the problems encountered whenever new software needs to be learned. However, the lack of adequate documentation and quirkiness of the existing software makes the learning process more difficult than it is for more polished software. Nonetheless, there is an onus for those interested in the methodology to invest the time and effort required to master it. After enough exposure and practice, researchers will be able to go beyond questions such as: "What software program is used to run IRT? I'm trying to learn how to do it and wanted to play around with it in an analysis."

References

Bjorner, J. B., & Ware, J. E. (1998, April). Using modern psychometric methods to measure health outcomes. Medical Outcomes Trust Monitor. Boston, Massachusetts. http://www.outcomes-trust.org/monitor/0498mntr.pdf

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiology Studies Depression (CES-D) scale: An item response theory analysis. Medical Care, 42, 281-289.

Christensen, K. B., & Bjorner, J. B. (2003). SAS macros for Rasch based latent variable modeling. Copenhagen: Department of Biostatistics. Technical Report # 03/13.

Gonin, R., Lloyd, S., & Cella, D. (1996). Establishing equivalence between scaled measures of quality of life. Quality of Life Research, 5, 20-26.

Granger, C. V., Hamilton, B. B., Keith, R. A., Zielesny, M., & Sherwin, F. S. (1986). Advances in functional assessment for medical rehabilitation. Top Geriatr Rehabil, 1, 59-74.

Haley, S. M., McHorney, C. A., & Ware, J. E. (1994). Evaluation of the MOS SF-36 physical functional scale (PF-10), I: Unidimensionality and reproducibility of the Rasch item scale. Journal of Clinical Epidemiology, 47, 671-684.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st Century. Medical Care, 38, II-29-II-42.

Heinemann AW, Linacre JM, Wright BD, Granger CV. (1993). Relationships between impairment and physical disability as measured by the Functional Independence Measure. Arch Phys Rehabil, 74, 566-573.

Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., & Roberts, J. S. (1996). Evaluating quality-of-life and health status instruments: Development of scientific review criteria. Clinical Therapeutics, 18, 979-992.

Scholle, S. H., Weisman, C. S., Anderson, R. T., & Camacho, F. (2004). The development and validation of the primary care satisfaction survey for women. Women's Health Issues, 14, 35-50.