Editorial to the Invited Special Section "Advancing methods to assess patient-reported outcomes: Lessons learned from the Patient- Reported Outcomes Measurement Information System® (PROMIS®) initiative"

Edward Haksing Ip, Editor, ARCS section

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is one of the largest applications of item response theory (IRT) outside of educational assessment. The PROMIS project developed a plethora of patient-reported outcome (PRO) measures for use in research and clinical assessment. PROMIS is now the gold standard for patient-reported outcome (PRO) measurement.

Numerous publications have been published during the 17 years since the beginning of PROMIS. However, a systematic examination of the various psychometric challenges faced in evaluating and using PROMIS measures has not yet been presented. What are the lessons learned from over a decade of applying IRT to the relatively new field of PROs? What are the new psychometric challenges that arise for such applications? Are there new psychometric approaches to solve emerging problems? Can what has been learned from PROMIS be used by researchers and practitioners from more traditional fields of IRT application? Answers to these questions are of interest to *Psychometrika* readers.

To address these questions, I organized a special section within the journal's Application Reviews and Case Studies (ARCS) section and invited two PROMIS experts—Drs. Bryce Reeve and Ron D. Hays—to be guest editors. Dr. Reeve is Professor of Population Health Sciences, Professor of Pediatrics, and Director of the Center for Health Measurement at Duke University School of Medicine. From 2000 to 2010, he served as Program Director for the U.S. National Cancer Institute (NCI). He was instrumental in the creation of the PROMIS initiative and helped to design the initial psychometric analysis plans for the PROMIS measures. Dr. Hays is Professor in the UCLA Department of Medicine, Professor in the Department of Health Policy and Management, and Affiliated Adjunct Research at the RAND Corporation. He is Co-Editor-in-Chief of the *Journal of Patient-Reported Outcomes*, serves on the editorial boards of *Quality of Life Research* and *Applied Research in Quality of Life*, and is a member of the special methodology panel for the *Journal of General Internal Medicine.* Drs. Reeve and Hays are both highly cited researchers.

The guest editors invited investigators from PROMIS to contribute to the special section. They also invited commentaries for the invited articles. Our aim is to provide readers with a range of articles that focus on the psychometric advances and challenges in the applications of IRT to PROs.

It is my hope that this special section will stimulate psychometric research in two directions: 1) raising broader interest in applying psychometric methods, including IRT, to fields beyond psychology and education, and 2) bringing methodologic innovations in measurement from other fields, including PRO, back to "mainstream" psychometrics.

Guest Editors' Introduction to the Invited Special Section

Bryce B. Reeve
Duke University School of Medicine

Ron D. Hays
UCLA

The National Institutes of Health (NIH) initiated the Patient-Reported Outcomes Measurement Information System® (PROMIS®) collaborative in 2004 to develop and provide access to standardized state-of-the-science health-related quality of life (HRQOL) measures for use in health research and clinical practice (Cella et al., 2007; Cella et al., 2010). The success of the NIH's PROMIS project is due to the involvement of a broad range of experts in measurement methods and clinical research from academia, government, and industry working together to build, refine, and implement the PROMIS measurement tools in research and healthcare delivery settings. PROMIS includes over 300 measures of physical, mental, and social aspects of HRQOL that may be used in the general population and for individuals with chronic and acute health conditions. This includes self-report measures for adults (18 years or older), self-report measures for children and adolescents (between 8 and 17 years) and proxy-report measures by caregivers for children between 5 and 17 years of age. The adoption of PROMIS measures by the international community is evidenced by the existence of over 50 translations of at least one of the PROMIS measures (HealthMeasures, 2021; Alonso et al., 2013). As of December 2020, there were well over 2000 publications in the scientific literature about PROMIS.

The high quality of the PROMIS measures is due to the multi-method approaches used by the multi-disciplinary experts to design the measures. Initially, PROMIS investigators examined previous research and vetted existing HRQOL measures to identify the salient concepts that should be measured. Next, they derived an initial set of questions (or "items") to capture each concept following best practices for patient-reported health surveys. Importantly, they conducted multiple rounds of cognitive testing to make sure the PROMIS items are clear, relevant to the patient experience, and content valid (DeWalt et al., 2007; Irwin et al., 2009). Then, they used a wide range of psychometric methods to evaluate the item and scale properties and to calibrate the items to enable the application of computerized adaptive testing (CAT) and static short form development (Reeve et al., 2007; Cella et al., 2007; Liu et al., 2010).

For each PROMIS HRQOL domain (e.g., fatigue, depression, physical functioning), there is an item bank that includes a large number of items that capture the salient concepts it intends to measure (i.e., to be content valid) and to estimate the respondent's level on the HRQOL domain across a broad range of the continuum. Each item underwent extensive evaluation using qualitative and psychometric methods to make sure it is appropriate for measuring the HRQOL domain of interest (DeWalt et al., 2007; Reeve et al., 2007). Items were subsequently calibrated with unidimensional item response theory (IRT) models. For each method, different approaches were applied recognizing each had their strengths and limitations. For example, tests for differential item functioning (DIF) included IRT-based and structural equation modeling (SEM)-

based methods. The PROMIS research team held multiple scientific meetings to discuss their approach and seek feedback from the community.

It's been 17 years since the initiation of PROMIS and over 10 years since its primary measures were released to the public through the HealthMeasures.net website. It has received unprecedented attention for its quality and standards. It has also allowed enough time for the broader scientific community to consider alternate psychometric methods that may give more insight on the item and scale performance and to consider how PROMIS measures may be used in clinical research to assess treatment efficacy. The set of papers in this special section of *Psychometrika* discuss some of the lessons learned and identify future psychometric directions for HRQOL researchers.

Teresi et al. (2021) included authors who were original architects for the approaches used to test for differential item functioning (DIF) for the items included in the PROMIS HRQOL item banks. In their recent article, they summarize the strengths and limitations of some of these approaches including IRT-based and SEM-based methods (Teresi et al., 2021). They highlight future work to examine DIF through the lenses of models that account for the multidimensional nature of the HRQOL data.

Schalet et al. (2021) discuss approaches to link PROMIS measures with established ("legacy") PRO measures to allow the comparison or combination of data from multiple studies that use different PRO measures of the same HRQOL construct. Schalet et al. (2021) highlight the strengths and limitations of equipercentile, unidimensional IRT-based calibration, and calibrated projection methods.

Cai and Houts (2021) highlight the value of modeling HRQOL longitudinally. They summarize psychometric methods of growth models, multilevel models and latent variable models and provide examples with HRQOL data collected by PROMIS measures used in clinical trials.

Hays et al. (2021) contrast the IRT-based and the classical test theory approaches to evaluate individual change in HRQOL data. Using PROMIS data from a longitudinal study of chronic low back pain and chronic neck pain patients, Hays et al. (2021) find the CTT-based approach to over-estimate change relative to the IRT-based approach.

Finally, Reise et al. (2021) address the critical issue for all modeling methods to make sure the selected approach should be based on a deep understanding of the concept and its distribution in the target population. Reise et al. (2021) contrast with PROMIS data Samejima's graded response model with the log-logistic model to illustrate how two methods (with the log-logistic model a non-linear transformation of the graded response model with equivalent fit) provide different interpretations of the performance of the items (or set of items) for the HRQOL construct being modeled. These are important considerations to make when thinking about constructs that may be continuous in nature versus constructs (e.g., pain) that may be unipolar and skewed.

We hope that this series of papers provides food for thought and stimulates future efforts to apply the most psychometrically appropriate methods in research and clinical practice with PROMIS and other HRQOL measures.

References

Alonso, J., Bartlett, S. J., Rose, M., Aaronson, N. K., Chaplin, J. E., Efficace, F., ... & Forrest, C. B. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. *Health and quality of life outcomes*, *11*(1), 1-5.

Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika*. (in press)

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(1), 133-141.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., ... & Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical care*, *45*(5 Suppl 1), S3.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... & PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of clinical epidemiology*, *63*(11), 1179-1194.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: the PROMIS qualitative item review. *Medical care*, *45*(5 Suppl 1), S12.

Hays, R. D., Spritzer, K. L., & Reise, S. P. (2021). Using Item Response Theory to Identify Responders to Treatment: Examples with the Patient-Reported Outcomes Measurement Information System (PROMIS®) Physical Function Scale and Emotional Distress Composite. *Psychometrika.* (in press)

HealthMeasures (2021, July 04). *PROMIS: available translations*. https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/available-translations

Irwin, D. E., Varni, J. W., Yeatts, K., & DeWalt, D. A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study. *Health and Quality of Life Outcomes*, *7*(1), 1-10.

Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., & Hays, R. D. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of clinical epidemiology*, *63*(11), 1169-1178.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical care*, S22-S31.

Reise, S. P., Du, H., Wong, E. F., Hubbard, A. S., & Haviland, M. G. (2021). Matching IRT models to patient-reported outcomes constructs: the graded response and log-logistic models for scaling depression. *Psychometrika*. (in press)

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika*. (in press)

Teresi, J. A., Wang, C., Kleinman, M., Jones, B. N., & Weiss, D. J. (2021). Differential item functioning analyses of the patient reported outcomes measurement information system (PROMIS) measures: Methods, challenges, advances, and future directions. *Psychometrika*. (in press)