Psychological Test and Assessment Modeling, Volume 58, 2016 (1), 31-35

Overview to the two-part series: Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short forms

Bryce B. Reeve[1,2] *& Jeanne A. Teresi*[3,4,5]

Abstract

Measurement equivalence across differing socio-demographic groups is essential for valid assessment. This is one of two issues of Psychological Test and Assessment Modeling that contains articles describing methods and substantive findings related to establishing measurement equivalence in self-reported health, mental health and social functioning measures.

The articles in this two part series describe analyses of items assessing eight domains: fatigue, depression, anxiety, sleep, pain, physical function, cognitive concerns and social function. Additionally, two overview articles describe the methods and sample characteristics of the data set used in these analyses. An additional article describes the important topic of assessing magnitude and impact of differential item functioning. These articles provide the first strong evidence supporting the measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short form measures in ethnically, socio-demographically diverse groups, and is a beginning step in meeting the international call for further study of their performance in such groups.

Key words: PROMIS, short form measures, patient-reported outcomes, measurement equivalence, differential item functioning

[1] *Correspondence concerning this article should be addressed to:* Bryce Reeve, PhD, Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 1101-D McGavran-Greenberg Hall, 135 Dauer Drive, CB 7411, Chapel Hill, NC 27599-7411, USA; email: bbreeve@email.unc.edu

[2] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill

[3] Columbia University Stroud Center at New York State Psychiatric Institute

[4] Research Division, Hebrew Home at Riverdale; RiverSpring Health

[5] Division of Geriatrics and Palliative Medicine, Weill Cornell Medical College, New York

Measurement equivalence across differing socio-demographic groups is essential for valid assessment. This is one of two issues of Psychological Test and Assessment Modeling that contains articles describing methods and substantive findings related to establishing measurement equivalence in self-reported health measures. Differential item functioning (DIF) is a method for examining equivalence of items across groups, and is a technique used in the articles contained in this and the second issue of Psychological Test and Assessment Modeling.

In 2004, the United States National Institutes of Health (NIH) launched an initiative to enhance and standardize the way that patient-reported outcomes (PROs) such as physical function, fatigue, depression and anxiety are measured in populations with different diseases and conditions. The goal of the NIH Patient Reported Outcomes Measurement Information System® (PROMIS®) initiative was to provide researchers and clinicians access to a set of reliable and valid PRO measures that can be used in clinical research or healthcare delivery settings (Cella et al., 2007).

The PROMIS measures were developed following recommended questionnaire design principles and have undergone extensive qualitative and quantitative evaluation in populations with respect to disease, race, ethnicity, age, gender, and education (Reeve et al., 2007). Further, PROMIS measures have been translated into multiple languages to enhance their use in research studies globally (Alonso et al., 2013). Originally developed in English, the PROMIS item banks have been translated into numerous languages including for example: Spanish, German, Mandarin, and Dutch. According to the NIH PROMIS webpage (http://www.nihpromis.org/measures/translations), translation of some PROMIS item banks into several other languages (e.g., Portuguese, Hebrew) is currently in progress. Examples of published translations are given in Paz, Spritzer, Morales, and Hays (2013) and Terwee et al. (2014). The PROMIS translation methodology is described in Eremenco, Cella, and Arnold (2005). However, these measures have received little formal evaluation of DIF across ethnically diverse groups. An international consortium of researchers from the European Union, the United States, Canada, China, and other countries involved in evaluating PROMIS measures internationally concluded that “differential item functioning analyses will be the most important analytical strategy” (Alonso et al., 2013).

PROMIS is a unique measurement system in that its item banks serve as the warehouse or library from which all PROMIS measures are designed. Each PROMIS item bank (there is a bank for each PRO measured) includes multiple items that vary in terms of content and severity. For example, the physical functioning item bank includes items that assess basic physical functioning like standing or picking up an object, to higher levels of physical functioning such as walking for more than a mile. Each item has been reviewed in cognitive interviews to ensure that patients of diverse backgrounds can comprehend the question and provide a response that accurately reflects their experiences or perspectives as they relate to the measured PRO. In addition, the items in the banks have undergone rigorous evaluation to ensure strong psychometric properties. All items are calibrated with item response theory (IRT) models and normed to the US general population that places the items on a common T-score metric with mean 50 and standard deviation of 10. Subsets of questions can be selected from the item bank to create a PROMIS short form instrument (e.g., a 6-item fatigue measure), or based on computerized-adaptive testing (CAT) technology, to tailor the PRO assessment to the level of the respondent. Scores from different PRO measures that come from the same PROMIS item bank can be compared or combined together because they all have been IRT-calibrated to be on the same metric.

While PROMIS measures have been evaluated in large datasets that have included individuals with different types of conditions and diseases, it is also important to examine the performance of the measures in groups with specific diseases/conditions. This approach is consistent with recommendations from the Food & Drug Administration (FDA) who request supporting evidence for the performance of the PRO measure within the population under study. Thus, the following articles in this two-part series focus on evaluating the psychometric properties of the PROMIS measures with a large sample from a diverse population of cancer patients. This evidence along with other supporting psychometric studies will enhance the adoption of PROMIS measures for research and in clinical settings.

These articles describe the first systematic examination of DIF in the PROMIS short-form measures among ethnically diverse groups and among patients with cancer. A focus of these articles is also to describe state-of-the art approaches to examination of measurement equivalence including those based on IRT (e.g., Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Mair & Hatzinger, 2007; Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Rasch, 1960; Teresi, Kleinman, & Ocepek-Welikson, 2000; Thissen, Steinberg, & Wainer, 1993), multiple group confirmatory factor analyses (MGCFA; Jöreskog, 1971; Meredith, 1964), multiple indicators, multiple causes (MIMIC; Jöreskog & Goldberger, 1975; Jones, 2006; Muthén, 1984) and ordinal logistic regression (OLR; Zumbo, 1999) using latent variable models (Crane, Van Belle, & Larson, 2004). Challenges to applications of these methods are also discussed.

This two part series contains articles describing analyses of eight domains: fatigue, depression, anxiety, sleep, pain, physical function, cognitive concerns and social function. Additionally, two overview articles describe the methods and sample characteristics of the data set used in these analyses. An additional article describes the important topic of assessing magnitude and impact of DIF. These articles provide the first strong evidence supporting the measurement equivalence of the PROMIS short form measures in ethnically, socio-demographically diverse groups, and is a beginning step in meeting the international call for further study of their performance in such groups.

Acknowledgements

The editors would like to thank the international team of over 30 reviewers for this series of articles who provided blinded reviews. They are recognized leaders and experts in the field of measurement statistics and/or assessment of patient-reported outcomes.

Partial funding for these analyses was provided by the United States National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971 (PI: A. Potosky, C. Moinpour).

References

Alonso, J., Bartlett, S. J., Rose, M., Aaronson, N. K., Chaplin, J. E., Efficace, F., … Forrest, C. B. for the PROMIS International group. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. *Health and Quality of Life Outcomes, 11,* 210. http://www.hqlo.com/content/11/1/210 doi: 10.1186/1477-7525-11-210

Cella, D. F., 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-S11. doi: 10.1097/01.mlr.0000258615.42478.55

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256. doi: 10.1002/sim.1713

Eremenco, S. L., Cella, D., & Arnold, B. J. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation & the Health Professions, 28*(2), 212-232. doi: 10.1177/0163278705275342

Hambleton, R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications, Inc.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. *Medical Care, 44*(11 Suppl 3), S124-S133. doi: 10.1097/01.mlr.0000245250.50114.0f

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 408-426. doi: 10.1002/j.2333-8504.1970.tb00790.x

Jöreskog, K. & Goldberger, A. (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *10*, 631-639. doi:10.1080/01621459.1975.10482485

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: the R package Rm for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20. doi: 10.18637/jss.v020.i09

Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika, 29,* 187–206. doi: 10.1007/BF02289700

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132. doi: 10.1007/BF02294210

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: applications to the Mini-Mental State Examination. *Medical Care, 44,* S134-S142. doi: 10.1097/01.mlr.0000245251.83359.8c

Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Evaluation of the patient-reported outcomes information system (PROMIS®) Spanish-language physical functioning items. *Quality of Life Research*, *22*(7), 1819-1830. doi: 10.1007/s11136-012-0292-6

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmarks Paedagogiske Institut (Danish Institute of Educational Research).

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, 45*(5 Suppl 1), S22-S31. doi: 10.1097/01.mlr.0000250483.85507.04

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651-1683. doi: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H

Terwee, C. B., Roorda, L. D., de Vet, H. C. W., Dekker, J., Westhovens, R., van Leeuwen, J., ... & Boers, M. (2014). Dutch–Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS). *Quality of Life Research*, *23*(6), 1733-1741. doi: 10.1007/s11136-013-0611-6

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In: P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum, Inc.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.