Published in final edited form as:
Expert Rev Pharmacoecon Outcomes Res. 2013 April ; 13(2): 183–186. doi:10.1586/erp.13.10.

INTRODUCTION TO PATIENT-REPORTED OUTCOME ITEM BANKS: ISSUES IN MINORITY AGING RESEARCH

Thomas N Templin[*],
Thomas N Templin, Office of Health Research, College of Nursing, Wayne State University, 5557 Cass Avenue, Room 321, Detroit, MI, 48083, USA

Ron D Hays,
Ron D Hays, Department of Medicine, 911 Broxton Avenue, Room 110, University of California Los Angeles, Los Angeles, CA, 90024-2801, USA, RAND, Santa Monica, CA., drhays@ucla.edu; hays@rand.org

Richard C Gershon,
Richard C Gershon, Medical Social Sciences, Northwestern University, 625 North Michigan Avenue, Suite 2700, Chicago, IL, 60611, USA, Gershon@northwestern.edu

Nan Rothrock,
Nan Rothrock, Medical Social Sciences, Northwestern University, 625 North Michigan Avenue, Suite 2700, Chicago, IL, 60611, USA, n-rothrock@northwestern.edu

Richard N Jones,
Richard N Jones, Institute for Aging Research, Hebrew Senior Life, 1200 Centre Street, Boston, MA 02131, USA, Phone: 617-971-5323, jones@hsl.harvard.edu

Jeanne A Teresi,
Jeanne A Teresi, Columbia University Stroud Center and Research Division, Hebrew Home at Riverdale, 5901 Palisade Avenue, Riverdale, New York, 10471, USA, Teresimeas@aol.com; jat61@Columbia.edu

Anita Stewart,
Anita Stewart, University of California, San Francisco, School of Nursing, Institute for Health & Aging, 3333 California Street, LHts-340, San Francisco, CA, 94143, USA, anita.stewart@ucsf.edu

Robert Weech-Maldonado, and
Robert Weech-Maldonado, Department of Health Services Administration, School of Health Professions, University of Alabama at Birmingham, 520 WEBB Building, 1530 Third Avenue South, Birmingham, Alabama, 35294, USA, rweech@uab.edu

Steve Wallace

[*]Author for correspondence: Wayne State University, College of Nursing, Office of, Health Research, Detroit, MI., 48202, USA, Tel.: +1 313 577 7992, Fax.:+1 313 577 5777, t.templin@wayne.edu.

Financial & competing interest disclosure

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.

Steve Wallace, University of California, Los Angeles, Center for Health Policy Research, 10960 Wilshire Blvd., Suite 1550, Los Angeles, CA 90024, USA, swallace@ucla.edu

Abstract

In 2004 NIH awarded contracts to initiate the development of high quality psychological and neuropsychological outcome measures for improved assessment of health-related outcomes. The workshop introduced these measurement development initiatives, the measures created, and the NIH supported resource (Assessment Center) for internet or tablet-based test administration and scoring. Presentation covered: (a) item response theory (IRT) and assessment of test bias, (b) construction of item banks and computerized adaptive testing, and (c) the different ways in which qualitative analyses contribute to the definition of construct domains and the refinement of outcome constructs. The panel discussion included questions about representativeness of samples, and assessment of cultural bias.

Report

Presenters were key personnel in one or more of the item bank projects or core faculty associated with the NIH funded Resource Center for Minority Aging Research (RCMAR). Ron Hays was the program Chair. Nan Rothrock, Scientific Director for Assessment Center, described the methodology used in creating the Patient-Reported Outcomes Measurement Information System (PROMIS) and the Quality of Life in Neurological Disorders (Neuro-QOL) item banks; Richard Gershon, PI for the NIH Toolbox and for the PROMIS Technology Center, described the methodology used in creating the NIH Toolbox for the assessment of neurological behavior and function (Gershon et al., 2010 [1]). Rothrock and Gershon also demonstrated computer adaptive testing (CAT) and the *Assessment Center* [101] - software created at Northwestern University designed to administer outcome measures created from recent federally funded initiatives including PROMIS [102], Neuro-QOL [103], and the NIH Toolbox [104] [2]. Assessment Center allows creation of study-specific URLs for administering self- or proxy-report short form and CAT instruments from its library (PROMIS, Neuro-QOL, NIH Toolbox) as well as custom instruments. Multiple time points, study arms, scoring, real time data export, and security precautions including storing protected health information in a separate database are supported. All data is stored on a server at Northwestern University and the security precautions are described in detail elsewhere [105]. This was the first workshop to cover all three initiatives.

Item banks

Rothrock outlined three characteristics of high quality item banks: individual items are easy to understand, have a shared meaning across individuals, and measure the target domain. A multi-step, rigorous approach utilizing both qualitative and quantitative techniques was used in PROMIS [3,4] and Neuro-QOL [5]. Rothrock summarized the process in 16 steps: definition of construct, identification of existing measures, archival data analysis, patient focus groups, expert review/consensus/revision, sorting/selecting best questions, literacy level analysis, translatability review, expert item revision, cognitive interviews, large scale testing (500 participants per item), statistical analysis, intellectual property permission, final

decisions about inclusion/exclusion and scoring, validation studies, and revision of measure as needed throughout its lifespan.

Both PROMIS and Neuro-QOL developed item banks in physical, mental and social health domains that can be administered as CATs or fixed length short forms. Adult and pediatric measures are available in English and Spanish. PROMIS instruments are intended for use in general populations and for those with chronic conditions. Neuro-QOL instruments were developed for use within neurologic conditions.

Gershon illustrated the creation of NIH Toolbox item banks using the vocabulary and reading comprehension tests from the Cognition domain normed in the US population for ages 3 to 85. Gershon explained the huge gains in efficiency from item response theory (IRT) modeling. To enhance motivation among study participants, the target correct proportion used for CAT was adjusted to 75% for children and 68% for adults. To reduce response time, automatic advance to next item with the option to “go back” was introduced [6]. In designing the vocabulary test, 625 words with four photographs for distracters were administered using a design based on 50% content overlap on adjacent lists and 80 to 100 observations per items (N = 1,100 overall). Many other details were covered including the voice tone used by the actor for adults vs. children. Convergent validity with the Peabody Picture Vocabulary Test-4 was supported by a product-moment correlation of 0.78. The technical manuals summarize psychometric evaluation, norming, and scoring of the measures.

While PROMIS and Neuro-QOL are based exclusively on self- or proxy-reports, the NIH Toolbox also includes proctor-administered instruments. Measures address function in cognition, emotion, motor, and sensation. Most NIH Toolbox measures take less than 5 minutes to administer. The battery of tests within a domain can be administered in about 30 minutes rather than several hours needed for conventional testing.

IRT methodology

Presentations from Ron Hays and Richard Jones covered technical aspects of IRT modeling and methodological issues specific to the development of item banks for computerized adaptive testing (CAT). Hays gave a brief review of evaluating IRT assumptions (dimensionality, local independence, montonicity, person fit) and introduced some of the features of the methodology that are especially helpful in evaluating survey items. For example, he discussed how category response curves can inform about the functioning of different response options. Hays also noted how IRT provides the most efficient administration approach possible that reduces response burden in achieving a target level of reliability or information. He noted how the response curves for different subgroups are indicative of differential item functioning—the curves for two groups should overlap completely if items are functioning equivalently in two subgroups because this means that the probability of responding in each category are the same in the two subgroups, conditionally on the estimated level on the construct being measured (“theta”).

Richard Jones reviewed IRT describing a heuristic approach to understanding item discrimination parameters and item difficulty parameters. He discussed the history of parameter estimation techniques, which has evolved to address the main challenge of simultaneously estimating unknown person variables (underlying, latent ability) and unknown item parameters. He presented an overview of a common modern approach to item parameter estimation, marginal maximum likelihood estimation with an expectation maximization algorithm (MMLE/EM), and some emerging alternatives. He concluded with some advice on judging the adequacy of item parameter estimates.

Qualitative analysis

Presentations by Anita Stewart and Robert Weech-Maldonado reviewed the growing literature on use of qualitative methods in item bank development which provides investigators with details on how the methods are applied. Stewart's presentation summarized the role of qualitative methods in developing item banks. Qualitative methods are applied during concept development (domain mapping and definitions), creating an item pool, standardizing and pretesting items including item revisions throughout the process, and assuring that the domain name/definition accurately reflects the final item pool. The most common qualitative methods used include judgment and consensus by item bank investigators, review of items by content experts, focus groups, and cognitive interview pretesting. There are some differences from how qualitative methods are applied in developing classical measures [7]. During concept development, focus groups are conducted with patients after domains are defined by item pool investigators in order to refine domain definitions. Judgment and consensus by investigators and expert review are both used to classify items in the item pool and delete items not meeting established criteria (e.g., do not measure concept, redundant, poorly worded). Because items are pooled from a variety of measures, investigators specify a standard for the item bank instructions, item stems, and response choices, and all items are revised to be consistent with that standard. Cognitive interview pretesting is essential to improve the clarity of items by iterative revisions to item wording [8].

Robert Weech-Maldonado outlined a framework for the cross-cultural adaptation of survey measures that consists of instrument translation, qualitative review and modification of translated version, and field test of modified translation [106]. This process consists of two or more forward translations, independent review of translation by bilingual experts, and committee review and decision on final translated instrument. He concluded that there has been limited research examining the cultural adaptation of item banks, and further research is needed in this area.

Cross-cultural validity

Conceptual and psychometric measurement equivalence of scales are fundamental requirements for valid cross-cultural and demographic subgroup comparisons. Jeanne Teresi reviewed briefly the different methodological approaches to evaluate measurement equivalence. She focused on methods that use latent conditioning variables. Latent variable models used to examine measurement invariance include IRT [9] and structural equation

modeling [10], such as multiple group confirmatory factor analyses [11], similarities and differences are summarized in several articles [12-18].

Differential item functioning (DIF) analysis is commonly used to study the performance of items in scales. Different methodologies for detecting DIF have been summarized and compared [19]. DIF is observed when the probability of item response differs across comparison groups such as gender, country or language or race/ethnicity, after conditioning on (controlling for) level of the state or trait measured, such as depression or physical function.

Teresi reviewed steps required for proper assessment of measurement invariance, broadly categorized as (1) qualitative methods, including selection of groups to be studied and generation of DIF hypotheses relevant to these groups; (2) tests of model assumptions and fit; (3) tests of DIF; (4) examination of magnitude (effect sizes associated with DIF); (5) evaluation of aggregate and individual impact of DIF; (6) expert review and disposition regarding items with DIF.

In the initial phase of PROMIS, DIF studies were performed but the samples were not ethnically diverse, and were characterized by individuals with higher educational levels. Teresi reviewed briefly the studies of PROMIS item banks and short-forms, including pain, fatigue, depression, anxiety, physical and social functioning. There are few studies extant that include examination of different ethnic groups; however, one study examined language of assessment [20]. She discussed opportunities for examination of measurement equivalence in later PROMIS efforts which include a large study of 4000 ethnically diverse individuals. She also noted that item banks and short forms derived from these banks, including PROMIS, Neuro-Qol and NIH Toolbox will not be accepted widely if evidence regarding measurement equivalence across ethnically diverse groups is not provided.

The workshop concluded with a panel discussion. Questions were raised about representativeness of panel company samples, and some of the complexity involved in the assessment of cultural bias. Conference slides are posted at online [107].

Acknowledgments

The workshop organizers were current or former members of RCMAR measurement cores including: Jack Goldberg, Ron Hays (Chair), Judy Shea, Anita Stewart, Thomas Templin, Jeanne Teresi, Steven Wallace, and Robert Weech-Maldonado.

The workshop was supported with funding from the National Institute of Aging Grant R13-AG023033. In addition, investigators were supported by the following grants: NIA-2P30AG015281-16 (T Templin); Toolbox HHSN260200600007C, PROMIS Technical Center 5U54AR057943-04, PROMIS Technical Center Supplement to NIH Toolbox 3U54AR057943-04S1, PROMIS Technical Center Supplement 3U54AR057943-04S2 (Gershon); NCI-U01AR057971, NIMHD-P60MD00206, NIA-P30 AG028741 (J Teresi); NIA-P30AG021684, NIMHD-P20MD000182 (R Hays), NIA-P30AG15272 (A Stewart), P30AG031054 (R Weech-Maldonado), NIH-U54 AR057943, NIH-U05 AR057951 (N Rothrock).

References

1. Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural function: the NIH Toolbox. Lancet Neurology. 2010; 9(2):138–139. [PubMed: 20129161]

2. Gershon RC, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment Center. Quality of Life Research. 2010; 19:677–685. [PubMed: 20306332]

3. Dewalt DA, Rothrock N, Yount S, Stone A, On Behalf of the Promis Cooperative Group. Evaluation of item candidates – the PROMIS qualitative item review. Med. Care. 2007; 45:S12–S21. [PubMed: 17443114]

4. Cella D, Riley W, Stone A, et al. Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005 - 2008. J. Clin. Epidemiol. 2010; 63:1179–1194. [PubMed: 20685078]

5. Cella D, Nowinski C, Peterman A, et al. The neurology quality of life measurement initiative. Arch Phys Med Rehabil. 2011; 92(Suppl 1):S28–S36. [PubMed: 21958920]

6. Hays RD, Bode R, Rothrock N, Riley W, Cella D, Gershon R. The impact of next and back buttons on time to complete and measurement reliability in computer-based surveys. Quality of Life Research. 2010; 19:1181–1184. [PubMed: 20552282]

7. Magasi S, Ryan G, Revicki D, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Quality of Life Research. 2012; 21(5):739–746. [PubMed: 21866374]

8. Christodoulou C, Junghaenel DU, Dewalt DA, Rothrock N, Stone AA. Cognitive interviewing in the evaluation of fatigue items: results from the patient-reported outcomes measurement information system (PROMIS). Quality of Life Research. 2008; 17(10):1239–1246. [PubMed: 18850327]

9. Lord, FM. Applications of item response theory to practical testing problems. Lawrence Erlbaum Assoc.; Hillsdale, New Jersey: 1980.

10. Muthén B. Beyond SEM: General latent variable modeling. Behaviormetrika. 2002; 29(1):81–117.

11. Jöreskog, K.; Sörbom, D. LISREL 8: Analysis of linear structural relationships: Users Reference Guide. Scientific Software International, Inc.; Chicago, Ill: 1996.

12. Mcdonald RP. A basis for multidimensional item response theory. Appl. Psychol. Meas. 2000; 24(2):99–114.

13. Meade AW, Lautenschlager GJ. A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/lnvariance. Organizational Research Methods. 2004; 7(4):361–388.

14. Mellenbergh GJ. Generalized linear item response theory. Psychol. Bull. 1994; 115(2):300–307.

15. Millsap RE, Everson HT. Methodology review: Statistical approaches for assessing measurement bias. Appl. Psychol. Meas. 1993; 17(4):297–334.

16. Raju NS, Laffitte LJ, Byrne BM. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. Journal of Applied Psychology Vol 88(4). 2002; 88(4):517–528.

17. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychol. Bull. 1993; 114(3):552–566. [PubMed: 8272470]

18. Takane Y, De Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987; 52:393–408.

19. Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Med. Care. 2006; 44:S152–170. [PubMed: 17060822]

20. Paz S, Spritzer K, Morales L, Hays R. Evaluation of the Equivalence of English- and Spanish-Language Patient-Reported Outcomes Information System (PROMIS®) Physical Functioning Items. Quality of Life Research. in press.

Websites

101. Assessment Center. www.assessmentcenter.net

102. PROMIS. www.nihpromis.org

103. Neuro-QOL. www.neuroqol.org

104. NIH Toolbox. www.nihtoolbox.org

105. Assessment Center documentation. http://www.assessmentcenter.net/ac1/AssessmentCenter_Manual.pdf

106. Weech-Maldonado, R.; Weidmer, BO.; Morales, LS.; Hays, RD. Cross-cultural adaptation of survey instruments: The CAHPS Experience. In: Cynamon, M.; Kulka, R., editors. Seventh Conference on Health Survey Research Methods. DHHS; Hyattsville, MD: 2001. p. 75-82.http://www.cdc.gov/nchs/data/misc/conf07.pdf

107. Workshop slides. http://www.rcmar.ucla.edu/GSA_Precon_12/materials