UCLA

Title

Introduction to patient-reported outcome item banks: issues in minority aging research

Permalink

https://escholarship.org/uc/item/4kw227k0

Journal

Expert Review of Pharmacoeconomics & Outcomes Research, 13(2)

ISSN

1473-7167

Authors

Templin, Thomas N
Hays, Ron D
Gershon, Richard C
et al.

Publication Date

2013-04-01

DOI

10.1586/erp.13.10

Peer reviewed

EXPERT REVIEWS

Introduction to patient-reported outcome item banks: issues in minority aging research

Thomas N Templin*[1], Ron D Hays[2], Richard C Gershon[3], Nan Rothrock[3], Richard N Jones[4], Jeanne A Teresi[5], Anita Stewart[6], Robert Weech-Maldonado[7] and Steve Wallace[8]

[1]Office of Health Research, College of Nursing, Wayne State University, 5557 Cass Avenue, Room 321, MI 48083, USA
[2]Department of Medicine, 911 Broxton Avenue, Room 110, University of California Los Angeles, CA 90024-2801, USA
[3]Medical Social Sciences, Northwestern University, 625 North Michigan Avenue, Suite 2700, IL 60611, USA
[4]Institute for Aging Research, Hebrew Senior Life, 1200 Centre Street, MA 02131, USA
[5]Columbia University Stroud Center and Research Division, Hebrew Home at Riverdale, 5901 Palisade Avenue, NY 10471, USA
[6]University of California, San Francisco, School of Nursing, Institute for Health & Aging, 3333 California Street, LHts-340, CA 94143, USA
[7]Department of Health Services Administration, School of Health Professions, University of Alabama at Birmingham, 520 WEBB Building, 1530 Third Avenue South, AL 35294, USA
[8]University of California, Los Angeles, Center for Health Policy Research, 10960 Wilshire Blvd., Suite 1550, CA 90024, USA
*Author for correspondence:
Tel.: +1 313 577 7992
Fax: +1 313 577 5777
t.templin@wayne.edu

Pre-Conference Workshop in conjunction with the Annual Meeting of the Geriatrics Society of America
San Diego Convention Center, San Diego, CA, USA, 14 November 2012

In 2004, the NIH awarded contracts to initiate the development of high-quality psychological and neuropsychological outcome measures for the 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. Presentations covered item response theory and assessment of test bias, construction of item banks and computerized adaptive testing, and 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 the assessment of cultural bias.

Presenters were key personnel in one or more of the item bank projects or core faculties associated with the NIH-funded Resource Centers for Minority Aging Research. Ron D 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 C Gershon, principal investigator 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 [1]. N Rothrock and RC Gershon also demonstrated computer adaptive testing (CAT) and the Assessment Center [101] – software created at Northwestern University (IL, USA) designed to administer outcome measures created from recent federally funded initiatives, including PROMIS [102], Neuro-QOL [103] and the NIH Toolbox [2,104]. Assessment Center allows creation of study-specific uniform resource locators (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 are 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
N 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 multistep, rigorous approach utilizing both qualitative and quantitative techniques was used in PROMIS [3,4] and Neuro-QOL [5]. N 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.

RC 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 3–85 years. RC 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 80–100 observations per items (n = 1100 overall). Many other details were covered, including the voice tone used by the actor for adults versus 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, normalization 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 min to administer. The battery of tests within a domain can be administered in about 30 min rather than the several hours needed for conventional testing.

IRT methodology

Presentations from RD Hays and Richard N Jones covered technical aspects of IRT modeling and methodological issues specific to the development of item banks for CAT. RD 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 on the functioning of different response options. RD 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 as 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 ('θ').

RN 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, 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 the use of qualitative methods in item bank development that provide investigators with details on how the methods are applied. A 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). As 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].

R 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 A Teresi briefly reviewed 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) the level of the state or trait measured, such as depression or physical function.

JA Teresi reviewed steps required for proper assessment of measurement invariance, broadly categorized as qualitative methods, including selection of groups to be studied and generation of DIF hypotheses relevant to these groups; tests of model assumptions and fit; tests of DIF; examination of magnitude (effect sizes associated with DIF); evaluation of aggregate and individual impact of DIF; 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. JA Teresi briefly reviewed the studies of PROMIS item banks and short forms, including pain, fatigue, depression, anxiety and physical and social functioning. There are few studies extant that include examination of different ethnic groups; however, one study examined language of assessment [20]. JA Teresi 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 widely accepted if evidence regarding measurement equivalence across ethnically diverse groups is not provided.

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

References

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

2 Gershon R, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment Center. *Qual. Life Res.* 19(5), 677–685 (2010).

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* 45, S12–S21 (2007).

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.* 63, 1179–1194 (2010).

5 Cella D, Nowinski C, Peterman A *et al.* The neurology quality-of-life measurement initiative. *Arch. Phys. Med. Rehabil.* 92(Suppl. 10), S28–S36 (2011).

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. *Qual. Life Res.* 19(8), 1181–1184 (2010).

7 Magasi S, Ryan G, Revicki D *et al.* Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. *Qual. Life Res.* 21(5), 739–746 (2012).

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). *Qual. Life Res.* 17(10), 1239–1246 (2008).

9 Lord FM. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Assoc., NJ, USA (1980).

10 Muthén B. Beyond SEM: general latent variable modeling. *Behaviormetrika* 29(1), 81–117 (2002).

11 Jöreskog K, Sörbom D. *LISREL 8: Analysis of Linear Structural Relationships: Users Reference Guide.* Scientific Software International Inc, IL, USA (1996).

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

13 Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods* 7(4), 361–388 (2004).

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

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

16   Raju NS, Laffitte LJ, Byrne BM. Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *J. Appl. Psychol.* 87(3), 517–529 (2002).

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

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

19   Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Med. Care* 44, S152–S170 (2006).

20   Paz S, Spritzer K, Morales L, Hays R. Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Qual. Life Res.* doi:10.1007/s11136-012-0292-6 (2012) (Epub ahead of print).

Websites

101   Assessment Center. www.assessmentcenter.net

102   PROMIS® – dynamic tools to measure health outcomes from the patient perspective. www.nihpromis.org

103   Neuro-QOL (quality of life in neurological disorders). www.neuroqol.org

104   NIH Toolbox – assessment of neurological and behavioral function. www.nihtoolbox.org

105   Assessment Center documentation. www.assessmentcenter.net/ac1/Assessment-Center_Manual.pdf

106   Weech-Maldonado R, Weidmer BO, Morales LS, Hays RD. Cross-cultural adaptation of survey instruments: the CAHPS Experience. In: *Seventh Conference on Health Survey Research Methods*. Cynamon M, Kulka R (Eds). DHHS, MD, USA (2001). www.cdc.gov/nchs/data/misc/conf07.pdf

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