AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Psychometric Methods Underlying the Development of ASCQ-Me[1]

This paper details methods the ASCQ-Me research team used in the evaluation of unidimensionality and differential item functioning for ASCQ-Me candidate items, the development of ASCQ-Me calibrated item banks, and the construction of ASCQ-Me short forms.

Unidimensionality

The ASCQ-Me research team used both exploratory and confirmatory factor analyses to examine the unidimensionality assumption for each of the ASCQ-Me item pools. Within each domain—following Cook, Kallen, and Amtmann (2009)— we first identified the optimal number of factors in each item pool by conducting parallel analysis (PA) of the ASCQ-Me data based on principal axis/common factor analysis on permutations of the raw data (Horn, 1965). The PA output does not describe the relationship of items to factors; therefore, we further conducted an exploratory factor analysis (EFA) using the polychoric correlation matrix of the data, restricting the number of factors to those revealed by the PA. We used an oblique solution (Promax rotation method) and examined the standardized coefficients of the regression of items onto factors (i.e., factor loadings). We removed items from the subsequent analyses that were not strongly related to any of the factors (i.e., all loadings smaller than 0.40) and then conducted a second EFA based on the remaining items.

Subsequently, we calculated the Cronbach's alpha coefficient (Cronbach, 1951) for the entire item pool and the item-total correlation for each item corrected for overlap (Howard & Forehand, 1962). We removed any item exhibiting an item-total correlation below 0.40 from the subsequent analyses. This is a commonly used cut-off point because items with lower item-total correlations do not represent the underlying construct well and are likely to have poor performance in factor analyses.

Evaluating unidimensionality of health items is complicated because the number of ways to phrase questions about health is limited. This means that subsets of health questions are likely to share similar word patterns. Covariation among the items that share the same word pattern is to be expected; but this covariation could be independent of the information about the domain contained in each question. An EFA conducted on such data is likely to identify more than one factor and suggest that the data are multidimensional. Thus, it is very unlikely that a set of questions about a particular aspect of health will perfectly meet unidimensionality based on EFA or confirmatory factor analysis (CFA), in which a single dimension is specified. The primary goal must then be to determine whether scales are "essentially" or "sufficiently" unidimensional (McDonald, 1999) to allow unbiased item calibration and scoring of individuals on a common latent trait. In other words, item characteristics and individual scores based on the item's

[1] Authors: San Keller, Manshu Yang, Christian Evensen, and Mary Nishioka, American Institutes for Research, August 24, 2015.

relationship to a single latent trait should not significantly differ from those obtained from a model in which secondary factors are included. The secondary factors should account for the covariation among subsets of items without changing the strength of the relationship of these items to the primary latent trait.

We conducted a bifactor analysis to evaluate the essential unidimensionality of the ASCQ-Me data for each domain by specifying two models (CFAs) of the relationship of items to latent trait(s) and using a polychoric correlation matrix as the input dataset. The first CFA modeled each item response as a function of a single general factor and an error term. The second CFA modeled each item response as a function of a single general factor, a "nuisance" group factor, and an error term. This second CFA is the bifactor model (Reise, Morizot, & Hays, 2007). After both CFA models were fitted to the data, we used fit indices, including the comparative fit index (CFI), the non-normed fit index (NNFI), and root mean square error of approximation (RMSEA) to evaluate the fit of the unidimentional structure to the data. Common current practice with regard to these indications of model fit is (1) to report chi-square values but not to reject models where the *p*-value is < 0.05 in data sets with more than 250 observations, (2) to require CFI and NNFI to be greater than 0.95, and (3) to require RMSEA to be less than 0.06 (Hu & Bentler, 1999; Kenny, 2003; Suhr, 2006). We also compared the standardized regression coefficients (i.e., factor loadings) associated with the general factor for both models. If they did not differ greatly between the two models (i.e., within the range of 0.00 to 0.10), the secondary (group) factors could be interpreted as uninteresting to the underlying construct because they are artifacts of question wording or caused by some other trivial influence. In other words, if the general factor significantly predicted item responses in both models and the relationship of items to this factor did not vary appreciably according to whether the "nuisance" factors were included, we interpreted the results as supporting the essential unidimentionality of the data following Reise and colleagues (2007). We conducted all exploratory and confirmatory factor analyses using SAS software (SAS Institute Inc., 2008).

As a second way to examine the unidimensionality of the item pools, we evaluated IRT item discrimination parameters for a unidimensional IRT model and a bifactor IRT model, respectively. For the unidimensional model, the graded response model (Samejima, 1969) with a single latent trait was fitted to the data. For the bifactor IRT model, each item was allowed to have a discrimination parameter on the general factor and one of the group factors. Using the beta test version of the IRT-PRO software (Thissen, 2009), we fitted both the unidimensional and bifactor IRT models and examined the discrimination parameter estimates for the items with the general factor obtained from both unidimensional and bifactor models for differences. Essential unidimensionality would be supported if the Pearson correlation between the vectors of discrimination parameters under the two models was high (e.g., > 0.90) and the root mean squared difference of discrimination parameters between the two models was comparatively low (Harrison, 1986).

DIF Analysis

Some items might not be equally valid across different types of respondents and lead to bias in measurement. In IRT framework, an item is defined as displaying measurement bias or differential item function (DIF) if the item response curves (i.e., item parameters) are not the same for the reference and focal group (Embretson & Reise, 2000). For health measures,

researchers are usually interested in DIF occurrence across gender and age because these are the sociodemographic variables most consistently and strongly associated with differences in health. We conducted DIF analysis for each of the candidate item banks in ASCQ-Me and removed items showing DIF from the item bank. We used the IRT-based Wald test method (Langer, 2008; Lord, 1977, 1980) to detect DIF for each item.

IRT Calibration

Once the IRT assumptions had been evaluated and confirmed, we fitted the unidimensional Graded Response Model (Samejima, 1969) to the data to estimate item parameters and create individual scores based on item calibrations. For each item in a scale, two item parameters were estimated to describe the item characteristics. To be more specific, the item discrimination parameter describes how well the item can detect differences between persons who are at different levels of health (i.e., regions on the latent trait continuum) (Lord, 1980). In other words, it describes the strength of the relationship between item responses and latent trait scores and is analogous to the item-total correlation. The item difficulty or location parameter represents the location on the latent trait continuum where the item can best discriminate among persons. This is analogous to the mean response to the item across persons.

We estimated item parameters using the marginal maximum likelihood method (Bock & Aitken, 1981) and estimated the psychometric properties of the items to evaluate the efficiency, reliability, and validity of each scale. For each ASCQ-Me item bank, we examined the following properties: (1) item parameter estimates, (2) test and item information curves, (3) the correlation coefficient of estimated individual health scores with the individuals' severity of sickle cell disease according to a medical history checklist, (4) Cronbach's alpha, and (5) the person-item map for each measure. The person-item maps were graphs that showed the location of items and respondents on the same range of scores (from -3.0 to +3.0). This score range represented an underlying health continuum. We produced the maps using a one-parameter model, because the location of the item on the continuum was the main interest and the strength of the relationship of the item to the continuum did not need to be known. Therefore, the Partial Credit model (Masters, 1982) was fitted to data and the estimated difficulty parameters were used to generate person-item maps.

The item information indicates the precision of each item in measuring an individual's latent trait. We used item information curves to identify the most useful items for measuring different levels of the latent health scores. Item information, along with item parameter estimates, was used by the CAT software as criteria to select items to administer to respondents. In addition, we used item information as important criteria to identify the best subset of items from each bank to include in a short form for that bank.

Short Form Construction

After the full ASCQ-Me item banks were defined and all the items had been calibrated, we selected five items from each item bank to create a short form, which would enable users to minimize respondent burden even if they did not have access to the CATs. To guide the item selection, we used four criteria: (1) a reasonable balance of content, (2) examination of item information curves, (3) association between each item and the SCD severity score, and

(4) minimal content overlap. Each of the ASCQ-Me item banks contained content to target specific aspects (subdomains) of that particular domain of health. For example, stiffness questions covered the topics of stiffness upon awaking and stiffness during daily activities. To maintain content balance, we selected at least one item from each subdomain. In addition, based on the item information curves, items were rank-ordered according to the amount of information they provided at different levels of latent health scores. We selected items to maximize information across the entire continuum of latent scores. If two items had similar information curves, then their relationship with the SCD severity score was taken into account and we selected the item that could significantly predict severity scores. If several items had similar content, we selected only the one with the largest information value to avoid content overlap.

Construct Validity

To examine the ability of ASCQ-Me latent health scores to reflect differences among groups of patients which should differ in health, we divided participants into three groups according to their SCD severity scores, representing low, medium, and high level of severity, respectively. The original severity scores had nine possible values ranging from 0 to 8. To define the three larger severity groups, we calculated the percentile corresponding to each of the nine levels of severity in the entire sample and regarded severity scores closest to the 33rd and 66th percentile as the cut-off values to determine low, medium, and high level of severity. We conducted one-way ANOVA for the latent health scores obtained from IRT estimation across the three severity groups. We also examined the correlation of latent scores obtained from the short form and the full scale to determine how well the short form represented the content of the item bank.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18,* 447–460.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists.* Mahwah, NJ: Erlbaum.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91–115.

Horn, J. L. (1965). A rationale and a test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Howard, K. I., & Forehand, G. G. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement, 22*, 731–735.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Kenny, D. A. (2003). *Measuring model fit*. Retrieved from http://davidakenny.net/cm/fit.htm

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral Dissertation).

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, Netherlands: Swets and Zeitlinger.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Earlbaum.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16,* 19–31.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

Suhr, D. D. (2006). Exploratory or confirmatory factor analysis? *SUGI Proceedings, Paper 200–231*.

SAS Institute, Inc. (2008). *SAS 9.2 help and documentation* [Computer software manual]. Cary, NC: SAS Institute Inc.

Thissen, D. (2009). The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system—IRT software. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved 9/4/15 from www.psych.umn.edu/psylabs/CATCentral/