population. We typically recruit between 10–20 patients in which half the participants receive the ePRO first and the other half the paper version. Between administrations participants complete a distraction task. Interviews are recorded and a content analysis conducted to identify key issues. RESULTS: The mix of think-aloud and retrospective probing has worked well in a number of studies across disease areas to ensure equivalence, high usability, and no unforeseen issues unique to ePRO such as screen glare or difficulty holding a PDA device. Some patients have difficulty with the “think-aloud” approach and so the retrospective probing is a useful check against issues not spontaneously brought up by the participants. CONCLUSIONS: Increased use of ePRO questionnaires necessitates a robust methodology for demonstrating equivalence during migration from paper versions. A mix of concurrent “think-aloud” and retrospective probing following completion of both PRO formats has shown to be a useful method for establishing validity of electronic outcome measures.

EXAMINING ITEM RESPONSE PATTERNS OVER TIME IN A HEALTH PROFILE MEASURE USING US NATIONAL REPRESENTATIVE SAMPLES: A MULTI-FACET MODEL APPROACH

Stevens D1, Dewey C1,2

1University of California, Los Angeles, CA, USA; 2National Center for Health Statistics, U.S. Department of Health and Human Services, Hyattsville, MD, USA

OBJECTIVES: To examine item patterns over time using the SF-12v2TM from a measurement perspective using US national representative samples. METHODS: Four panel data with two-year repeated measures on each respondent were extracted from the Medical Expenditure Panel Survey (MEPS). Respondents were included if they were ≥18 years, had completed SF-12v2TM and had at least one of the ten most prevalent health conditions identified using ICD-9-CM. Three-facet measurement model was used to parameterize time as a distinct facet in the model, in addition to person and item facets. Interactions between time and the twelve items were examined at each point in time in all panels. Goodness-of-fit of the items to the model was examined in repeated measures as well as in point-in-time measures. INFIT mean-square (MnSq) > 1.40 was used as an item fit indicator. Cross-validations were conducted in each disease groups. RESULTS: Four panels were comparable in their distributions in health conditions, socio-demographics (mean ages were 32±3 years, and about 76–77% were white) and, sample sizes (2003–04, n = 2,124; 2004–05, n = 2,070; 2005–06, n = 2,148 and 2006–07, n = 2,329). Consistently in all panels, significant time and item interaction biases were found at time 1, especially on mental health items (p < 0.01). On the other hand, interaction biases between time and items at time 2 were not significant (p > 0.05). All items fit the model in repeated measures where time was parameterized as a facet (INFIT MnSq ≤ 1.40). The mental health item “Have you felt calm and peaceful?” consistently showed misfit in all point-in-time measures (INFIT MnSq > 1.40). Similar findings were noted in sub-samples. CONCLUSIONS: Findings from this study suggest consistent learned response patterns over time, especially the responses to mental health item, which give rise to the importance of inter-temporal health context in measurement health. Hence, cross-sectional health measures should be interpreted with caution.

ITEM CALIBRATION OF A GENERIC ROLE FUNCTIONING ITEM BANK

Anatchkova M1, Bjorner J2

1University of Massachusetts Medical School, Worcester, MA, USA; 2National Research Centre for the Working Environment, Copenhagen, Denmark

OBJECTIVES: Role functioning (RF) is a key component of social well-being and thus an important outcome in health research. The aim of this study was to calibrate on a common metric newly developed items assessing the impact of health on RF. The items were developed based on review of the literature and focus group interviews and were found to be sufficiently unidimensional for item response theory applications. METHODS: Two thousand five hundred participants completed a battery of measures including 77 items in a RF bank, covering the impact of health on family, occupational and social role functioning. Each new item covered only one of the content areas. Items were evaluated for potential DIF by demographic variables (gender, age, and chronic condition) using a logistic regression approach. To estimate the item parameters for each domain on a common metric we used the generalized partial credit model. Item fit was evaluated using the S-G2 index. Comparison of group mean bank scores of participants with different self-reported general health status and chronic condition was used to assess the external validity of the bank. RESULTS: After item calibrating using data of three different approaches to evaluate measurement equivalence properties of a item bank (with and without chronic conditions (F(4, 2488) = 31.48, P < 0.0001) and self-reported general health (F(4, 2488) = 23.55, P < 0.0001). CONCLUSIONS: An item bank assessing health impact on RF across 4 content areas has been successfully calibrated. Using computerized adaptive assessment, respondents will only need to answer items regarding relevant roles, while IRT score estimation still allows for scoring all respondents on the same common metric.