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 raised by the participant(s). **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.

PMC39

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

Gu NY
Pharmerit North America, LLC, Bethesda, MD, USA

OBJECTIVES: To examine item response patterns over time using the SF-12v2™ 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-12v2™ and, had at least one of the top 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 time point 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 52–53 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 health measurement. Hence, cross-sectional health measures should be interpreted with caution.

PMC40

ITEM CALIBRATION OF A GENERIC ROLE FUNCTIONING ITEM BANK

Anatchkova M[1], Bjorner J[2]
[1]University of Massachusetts Medical School, Worcester, MA, USA; [2]National 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-G^2 index. Comparison of group mean bank scores of participants with different self-reported general health status and chronic conditions was used to test the external validity of the bank. **RESULTS:** After excluding items with DIF and poor fit the final item bank had a total of 64 items covering 4 general content areas of role functioning (family, social, occupational, generic). Slopes in the bank ranged between 0.96 and 4.51; the mean threshold range was –0.66 to –1.80. Item bank based scores were significantly different for participants with and without chronic conditions ($F(4, 2488) = 31.48$, $P < 0.0001$) and self-reported general health ($F(4, 2488) = 233.55$, $P < 0.0001$). **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.

PMC41

PREEMPTING DIFFICULTIES IN LINGUISTIC VALIDATION, THE USE OF FACE VALIDATION TO CREATE MORE SOUND TRANSLATIONS

Gawlicki M[1], Handa M[2]
[1]Corporate Translations, Inc, East Hartford, CT, USA; [2]Corporate Translations, Inc, Chicago, IL, USA

OBJECTIVES: The process of linguistic validation is complex especially when working with a variety of languages in widely divergent cultural settings. The ability to clearly delineate concepts and synchronize wording within an instrument before the linguistic validation process begins not only significantly improves the original instrument, but also aids in optimizing its translatability, ensuring greater uniformity between multiple linguistic adaptations and saving time and resources along the way. This paper seeks to explain the benefits provided by the supplemental pre-translation process of face validation. **METHODS:** As part of a case study, face validated questionnaires were compared to the original homegrown versions of the corresponding instruments—questionnaires that were already psychometrically validated were not eligible. Changes that were made as a result of this analysis will be discussed in-depth to clarify difficulties that each issue would have created for the linguistic validation process had they not been corrected. a cost benefit-analysis was also conducted to confirm the value of this supplemental linguistic validation phase. **RESULTS:** While standard elements of the linguistic validation process, such as concept elaboration, international harmonization, survey research expert review, in-country clinician review and cognitive debriefing all assist greatly in creating a quality translation, none of their benefits are a substitute for face validation. Furthermore, cost-benefit analysis reveals that the preemption of linguistic or methodological issues prior to translation and the greater uniformity obtained amongst multiple translations created through face validation save time and money later on in the linguistic validation process, justifying the added up-front costs. **CONCLUSIONS:** As the case studies confirm, taking steps to maximize the translatability of a questionnaire prior to linguistic validation, through face validation in particular, is highly beneficial to the end-products and can also hasten overall project completion and improve the quality of all language versions of the instrument.

PMC42

TO WHAT EXTENT CAN TECHNOLOGY IMPROVE THE VALIDITY OF CLINROS?

Wild D[1], Langel K[2]
[1]Oxford Outcomes Ltd, Oxford, Oxon, UK; [2]CRF Health, Helsinki, Finland

OBJECTIVES: ClinROs are the most commonly observed endpoint in FDA approved product labels but few have been adequately scrutinized in terms of their suitability as endpoints. This study evaluates two widely used ClinROs (the Expanded Disability Status scale (EDSS), and the Hamilton Rating scale for Depression (HAM-D)) and provides an assessment on how migrating the measures onto an electronic platform might be able to improve their validity and reliability. **METHODS:** A literature review was conducted on both measures to evaluate the availability of information on their content validity and reliability and validity. An assessment was made on how the measures could be improved if they were to be migrated onto an electronic platform: **RESULTS:** The EDSS has shown varying results for validity and inter-rater reliability and it involves a complex scoring procedure. The migration of the EDSS onto an electronic format would enable an automated scoring system which could improve its validity. The HAM-D was found to be lacking in evidence of content validity and to have some complexity in the scoring system. Transferring the HAM-D onto an electronic platform could simplify the scoring system which could improve its validity. **CONCLUSIONS:** This study has highlighted some of the issues with validity and reliability of two widely used ClinROs. The migration of ClinROs to an electronic platform in addition to the ePRO migration cognitive debriefing and usability testing might go some way to improving the clarity of ClinROs which may go some way to improving the validity of the measures. It cannot however resolve all of the issues such as lack of content validity and its impact would vary widely according to the complexity of the ClinRO itself.

PMC43

DATA POOLING OF PATIENT-REPORTED OUTCOMES IN CLINICAL TRIALS: EVALUATION OF STATISTICAL TECHNIQUES FOR ASSESSING MEASUREMENT EQUIVALENCE

Nixon M
Quintiles, Bracknell, Berkshire, UK

OBJECTIVES: This analysis describes the development, application and comparison of three different approaches to evaluate measurement equivalence properties of a patient reported outcome (PRO) questionnaire applied to two treatment groups for gastroesophageal reflux disease (GERD). The data used in this analysis was obtained from an on-line patient community, iGuard.org. Patients using either of the two treatments were randomly invited to complete a measure of treatment satisfaction, the Treatment Satisfaction Questionnaire for Medication (TSQM). **METHODS:** Three statistical approaches were used to evaluate the measurement equivalence of the TSQM across the two patient populations: 1) Classical Test Theory (CTT) to assess the internal consistency of the TSQM items within each of the three factors using Cronbach's alpha; 2) Confirmatory Factor Analysis (CFA) using a special case of structural equation modelling (SEM); and 3) Item Response Theory (IRT)—based technique of Differential Item Functioning (DIF). **RESULTS:** All three statistical methods indicated measurement equivalence had been achieved across the two treatment populations for all the three domains of the TSQM. The effectiveness and global