Psychological Test and Assessment Modeling, Volume 58, 2016 (1), 37-78

Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, "Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short forms"

Jeanne A. Teresi[1,2,3] *& Richard N. Jones*[4]

Abstract

The purpose of this article is to introduce the methods used and challenges confronted by the authors of this two-part series of articles describing the results of analyses of measurement equivalence of the short form scales from the Patient Reported Outcomes Measurement Information System® (PROMIS®). Qualitative and quantitative approaches used to examine differential item functioning (DIF) are reviewed briefly. Qualitative methods focused on generation of DIF hypotheses. The basic quantitative approaches used all rely on a latent variable model, and examine parameters either derived directly from item response theory (IRT) or from structural equation models (SEM). A key methods focus of these articles is to describe state-of-the art approaches to examination of measurement equivalence in eight domains: physical health, pain, fatigue, sleep, depression, anxiety, cognition, and social function. These articles represent the first time that DIF has been examined systematically in the PROMIS short form measures, particularly among ethnically diverse groups. This is also the first set of analyses to examine the performance of PROMIS short forms in patients with cancer.

[1] *Correspondence concerning this article should be addressed to:* Jeanne A. Teresi, Ed.D, Ph.D., Columbia University Stroud Center at New York State Psychiatric Institute, 1051 Riverside Drive, Box 42, Room 2714, New York, New York, 10032-3702, USA; email: Teresimeas@aol.com; jat61@columbia.edu

[2] Weill Cornell Medical College, Division of Geriatrics and Palliative Medicine

[3] Research Division, Hebrew Home at Riverdale; RiverSpring Health

[4] Department of Psychiatry and Human Behavior, Department of Neurology, Warren Alpert Medical School, Brown University

Latent variable model state-of-the-art methods for examining measurement equivalence are introduced briefly in this paper to orient readers to the approaches adopted in this set of papers. Several methodological challenges underlying (DIF-free) anchor item selection and model assumption violations are presented as a backdrop for the articles in this two-part series on measurement equivalence of PROMIS measures.

Key Words: methods, PROMIS, measurement equivalence, differential item functioning, ethnic diversity

Introduction

Several methods for examining measurement equivalence were used in the papers in this and a second issue of Psychological Test and Assessment Modeling. This set of articles describes the results of analyses of short form scales from the Patient Reported Outcomes Measurement Information System (PROMIS; Cella et al., 2007). The purpose of this overview is to introduce briefly the approaches used to examine differential item functioning (DIF) in these sets of analyses, identifying challenges and new directions. Both qualitative and quantitative methods were used. Qualitative methods focused on generation of DIF hypotheses. The basic quantitative approaches used all relied on a latent variable model, and examined parameters either derived directly from item response theory (IRT; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Lord & Novick, 1968; Rasch, 1960) or from structural equation models (SEM; Jöreskog & Goldberger, 1975; Jöreskog & Sorbom, 1996).

DIF is observed when the probability of item response differs across comparison groups such as country, ethnic group, or language, after conditioning on (controlling for) level of the state or trait measured, such as depression or physical function. Uniform DIF occurs if the probability of response is consistently higher (or lower) for one of the comparison groups across all levels of the state or trait. Non-uniform DIF is observed when the probability of response is in a different direction for groups compared at different levels of the state or trait. For example, the response probability might be higher for Spanish than for English-speakers at higher levels of a measure of depression state, and lower for Spanish than for English speakers at lower levels of depression.

Many reviews of methods to assess DIF exist (Holland & Wainer, 1993; Millsap & Everson, 1993; Potenza & Dorans, 1995; Teresi, 2006; Teresi & Jones, 2013; van de Vivjer & Leung, 1997). PROMIS guidelines and standards provide several evidence-based methods that are recommended for DIF assessment (Reeve et al., 2007; http://www.nihpromis.org/science/publications); the use of these methods were illustrated by Carle et al. (2011), and were those used in the analyses reported in the papers in this series.

These articles represent the first time that DIF has been examined systematically in the PROMIS short form measures. Some studies of DIF have been performed by PROMIS investigators; however, the samples were not ethnically diverse, and were characterized by individuals with higher educational levels. There are practically no studies of PROMIS measures extant that focus on DIF in different racial and ethnic groups. This is also the first set of analyses to examine the performance of PROMIS short forms in patients with cancer. Additionally, a key methods focus of these articles is to describe state-of-the art approaches to the examination of measurement equivalence in health, mental health, and cognition, including those based on IRT (Lord, 1980; Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Teresi, Kleinman, & Ocepek-Welikson, 2000), multiple group confirmatory factor analyses (MGCFA; Jöreskog, 1971; Meredith, 1964), multiple indicators, multiple causes (MIMIC; Jöreskog & Goldberger, 1975; Jones, 2006; Muthén, 1984) and ordinal logistic regression (OLR; Zumbo, 1999) using latent variable models (Crane, Van Belle, & Larson, 2004). Also discussed are challenges to applications of these methods, including selection of DIF-free anchor items that set the metric for comparison groups on a common scale, model assumption violations and missing data.

Qualitative methods

One of the initial steps in DIF analyses is the establishment of an a priori set of hypotheses regarding potential group differences in item response by combining information gathered via two methods: a) qualitatively, from ratings by a panel of content experts, and b) from a review of the literature documenting prior research findings of DIF.

DIF hypotheses were generated by asking a set of clinicians and other content experts to indicate whether or not they expected DIF to be present, and the direction of the DIF with respect to several comparison groups: gender, age, race/ethnicity, language, education, and diagnosis. A definition of DIF was provided, and instructions related to hypotheses generation were given. An illustration of the definition is given using fatigue as the example.

> Differential item functioning means that individuals in groups with the same underlying trait (state) level will have different probabilities of endorsing an item. Put another way, reporting fatigue (e.g., limiting you at work, including work at home) should depend only on the level of the trait (state), e.g., fatigue and not on membership in a group, e.g., male or female or young or old. Very specifically, randomly selected persons from each of two or more groups (e.g., males and females) who are at the same (e.g., mild) level of fatigue should have the same likelihood of reporting limitations working due to fatigue. If it is hypothesized that this is not the case, it would be hypothesized that the item has gender DIF.

The PROMIS items were reviewed qualitatively by nine to twelve content experts regarding potential sources of DIF. The content experts were clinical or counseling psychologists, public health professionals, gerontologists, epidemiologists, and clinicians. Different sets of experts rated each domain. The experts were asked to rate individually each of the items with respect to each of the socio-demographic groups and diagnosis. They provided the hypotheses in terms of presence and direction of DIF. The goal was to identify items that might have a different meaning or not be understood well and/or equivalently by individuals of any of the groups referenced (Malida et al., 2008). A grid containing a row for each of the items and separate columns for each of the referenced groups was distributed to the experts for completion.

Although qualitative prediction of content experts has not always been found to be reliable in educational testing (Frederickx, Tuerlinckx, De Boeck, & Magis, 2010); generation of hypotheses for health and quality-of-life-related constructs may fare better because the item pool is smaller and limited, and the content narrow for a specific construct, e.g., depression. Because the PROMIS measures were translated for this project into languages other than English (Spanish and Mandarin), the hypothesis generation step was considered important in the adaptation and DIF testing process (see also Hambleton & Patsula, 1998, 1999).

Quantitative methods

An IRT-based approach with tests of parameter differences was used in several sets of DIF analyses, namely the papers by Reeve and colleagues examining fatigue (this issue), Fieo and colleagues (in press) examining cognition and Teresi and colleagues, examining depression and anxiety (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim; this issue a, b) and pain (Teresi et al., in press). A combined MGCFA MIMIC approach was used in the paper by Jones, Tommet, Ramirez, Jensen, and Teresi, (in press) examining physical function. Jensen and colleagues (in press) applied MGCFA followed by MIMIC in the analyses of sleep (in press). Hahn and colleagues (in press) used ordinal logistic regression with a latent conditioning variable to examine DIF in social function items. Consistent with the PROMIS psychometric approach, Samejima's Graded Response Model (GRM; Samejima, 1969) was fit to the polytomous response data in the analyses by all investigators except for Jones and colleagues (in press) and Jensen and colleagues (in press) who used a different parameterization.

The primary statistical approaches for examining measurement equivalence used in the analyses were: 1) Wald tests of item parameters from IRT 2) model-based tests of DIF from the MIMIC approach 3) modification indices from MGCFA analyses comparative models and 4) chi-square and pseudo R-square statistics from ordinal logistic regression using latent variable conditioning models. All authors included examination of model assumptions, and considered effect sizes and impact of DIF. Following a best practice recommendation (Hambleton, 2006), most investigators used a purification process to select anchor items that were free of DIF in order to obtain a relatively DIF-free set of items with which to estimate the trait and link the groups studied. Finally, several investigators used an alternative secondary approach in sensitivity analyses.

IRT-based DIF tests

The graded response model (Samejima, 1969) is one of the methods used most often in DIF applications in health, mental health, and many areas of psychology because it permits modeling of polytomous data with multiple ordered response options reflecting symptom severity. This model is the basis for DIF detection using two of the methods applied in these analyses, the Wald test and latent variable ordinal logistic regression. Ordered responses, x=k and k=1,2,...m are assumed, in which a_i is the discrimination (slope) and b_{ik} the difficulty parameters for response category k:

$P(x=k) = P^*(k) - P^*(k+1) = 1 / 1 + \exp[-a_i(\theta-b_{ik})] - 1 / 1 + \exp[-a_i(\theta-b_{ik+1})]$.

$P^*(k)$ is the item characteristic curve (ICC) describing the probability that a response is in category k or higher, for each value of θ (see also Orlando-Edelen et al., 2006; Thissen, 1991).

The discrimination (a_i) parameter, proportional to the slope of the curve, characterizes the strength of the relationship of the item to the underlying attribute measured and how well the item discriminates among individuals at specific levels of that trait. It is equivalent with algebraic manipulation to the factor loading. The severity (location or threshold) parameter (b_i) also known in educational testing as the difficulty parameter represents an inflection or cutting point between two adjacent response categories, and the b parameters collectively indicate the difficulty or severity of the item. Using depression as an example, items may be more or less severe indicators of depression. Admitting to suicidal ideation is a more severe indicator as reflected in higher b parameters. There are k-1 boundary response functions describing the cumulative probability of responding in category k or higher. The degrees of freedom increase with the number of b parameters estimated. There is one less b estimated than there are response categories. If tests of the equivalence of the a parameters (indicative of non-uniform DIF) are not significant, tests of group differences in the b parameters (indicating uniform DIF) are performed, constraining the a parameters to be equal.

Using anxiety as an example, the expectation is that respondents who are anxious would be more likely than those who are not to respond in a symptomatic direction to an item measuring anxiety. A person without anxiety is expected to have a lower probability (than a person with anxiety) of responding in an anxious direction to the item. The item characteristic curve relates the probability of an item response to the underlying state or trait, e.g., anxiety, measured by the item set. According to the IRT model, an item shows DIF if people from different subgroups but at the same level of anxiety have unequal probabilities of endorsement. DIF is demonstrated by ICCs that are different for comparison groups.

Testing item parameters using IRT log-likelihood ratio and Wald tests: The Wald statistic is equivalent to Lord's chi-square (Lord, 1980), which was extended for polytomous data by Cohen, Kim, and Baker (1993). The Wald statistic is also asymptotically equivalent to the likelihood ratio test (Thissen 1991; Thissen, Steinberg, & Wainer, 1993) used in the item response theory likelihood ratio (IRTLR) method. This latter widely used approach tests a series of IRT models established by fixing and freeing parameters. First, a compact (or more parsimonious) model is tested with all parameters constrained to be equal across groups for a studied item, together with the anchor items that are DIF-free (model 1), against an augmented model (2) with one or more parameters of the studied item freed to be estimated distinctly for the two groups. The procedure involves comparison of differences in log-likelihoods (-2LL, distributed as chi-square) associated with nested models; the resulting statistic is evaluated for significance with degrees of freedom equal to the difference in the number of parameter estimates in the two models.

In contrast, the Wald test for DIF follows the model proposed by Lord (1977, 1980) in which vectors of IRT item parameters are compared. The rationale is that if the vectors of item parameters differ significantly across groups, then the item functions differently for the groups. In the context of DIF testing, use of the Wald test based on Lord's chi-square permits testing DIF across multiple groups rather than two groups at a time, as is the case with IRTLR. The final *p* values are adjusted using Bonferroni (Bonferroni, 1936) or other methods such as Benjamini-Hochberg (B-H; Benjamini & Hochberg, 1995; Thissen, Steinberg, & Kuang, 2002).

As summarized in Teresi, Kleinman, and Ocepek-Welikson (2000), Lord (1980, p. 223) proposed a chi-square statistic, $\chi^2 = \underset{\sim i}{\mathbf{v}'}\sum_{i}^{-1} v_i$ testing simultaneously the hypotheses that the *a*'s and *b*'s of group 1 on item *i* are equal to the *a*'s and *b*'s of group 2, where $\underset{\sim}{\mathbf{v}'}$ is the vector $\left\{ \widehat{b}_{i1} - \widehat{b}_{i2}, \widehat{a}_{i1} - \widehat{a}_{i2} \right\}$, and $\sum_{i}^{-1}$ is the inverse of the asymptotic variance-covariance matrix for $\widehat{b}_{i1} - \widehat{b}_{i2}$ and $\widehat{a}_{i1} - \widehat{a}_{i2}$. Because $\widehat{a}_{i1}$ and $\widehat{b}_{i1}$ are independent of $\widehat{a}_{i2}$ and $\widehat{b}_{i2}$, $\sum_{i} = \sum_{i1} + \sum_{i2}$, where $\sum_{i1}$ is the sampling variance-covarance matrix of $\widehat{a}_{i1}$ and $\widehat{b}_{i1}$, and similarly for $\sum_{i2}$.

More advanced estimation procedures (Cai, 2008) were introduced by Langer (2008), and incorporated into Flexible Multilevel Multidimensional Item Analysis and Test Scoring (FlexMIRT; Cai, 2013; Houts & Cai, 2013) and Item Response Theory for Patient Reported Outcomes (IRTPRO; Cai, Thissen, & du Toit, 2012); the latter of which has been compared to IRTLR (Woods, Cai, & Wang, 2013). There are two approaches to the use of the Wald test: the Wald 1 method uses anchor items in DIF detection, while the Wald 2 method does not select for anchor items. As introduced above, and discussed in detail below, anchor items are DIF-free items used to set the metric for group comparisons. As an example, using the Wald test for examination of group differences in IRT item parameters (Lord, 1980; Teresi et al., 2000; Thissen et al., 1993), an overall simultaneous joint test of differences in the *a* or *b* parameters is performed followed by step down tests for group differences in the *a* parameters, followed by conditional tests of the *b* parameters. Uniform DIF is detected when the *b* parameters differ and non-uniform DIF when the *a* parameters differ. Severity (*b)* parameters are interpreted as uniform DIF only if the tests of the *a* parameters are not significant because tests of *b* parameters are performed, constraining the *a* parameters to be equal. As discussed below in the section on anchor items, the procedure is performed iteratively to obtain a purified anchor set. Several sets of investigators (Fieo et al., in press; Reeve et al., 2016; Teresi et al., 2016-a, 2016-b) used IRT Wald tests in IRTPRO (Cai et al., 2012) as the primary DIF detection method.

MIMIC and MGCFA

Also used in this series was MIMIC, a parametric latent variable model related to IRT. The model comes from the tradition of factor analyses and structural equation modeling (see also Jones, 2006). Additionally, multiple group confirmatory factor analyses (CFA) were performed. A unidimensional CFA model estimated for ordinal response data from a matrix of polychoric correlation coefficients with uncorrelated measurement errors is equivalent to a graded response IRT model (Jöreskog & Moustaki, 2001; Mislevy, 1986). The relationship and equivalence between factor analyses based on SEM and IRT has been reviewed and illustrated widely (e.g., McDonald, 2000; Meade & Lautenschlager, 2004; Mellenbergh, 1994; Meredith & Teresi, 2006; Raju, Laffite, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Takane & de Leeuw, 1987).

The measurement model can be represented by $y^* = \Lambda\eta + \in$, where η represents one or more latent variables underlying the item responses. Given a unidimensional model, η has the same meaning as θ in the IRT model. The model is linked to IRT as originally proposed by Birnbaum (Lord & Novick, 1968) because the discrimination parameter can be calculated using the factor loadings (lambda's; see also Thissen et al., 1993; Jones, 2006). As reviewed by Jones (2006) and Teresi and Jones (2013, pg. 152), the outcome variable vector, y^*, contains latent response variables underlying the observed and discrete responses, y. The y^* and y variables have a threshold relationship, where y_j is in category c if y^*_j is greater than threshold τ_c and less than or equal to τ_{c+1}. Λ contains a matrix of linear regression parameters, λ, that are on the scale of factor loadings when the common and latent response variables have unit variance, and describe the per-unit increase in y^*_j per unit increase in η. IRT discrimination parameters (a) can be determined from the factor analysis results in a single factor model using $a_j = \frac{\lambda j}{\sqrt{1-\lambda_j^2}}$ under the standard normal latent trait assumption (see also Lord & Novick, 1968). Boundary (difficulty or severity) parameters are $b_j = -\tau_{jc}\lambda_j^{-1}$ (Muthén & Asparouhov, 2002). Different parameterizations of the measurement model result in more complex linking to IRT parameters, as discussed by Muthén and Asparouhov (2002).

An important point is that for the two methods to be equivalent, parameters must be set in specific ways. In factor analysis, the metric of the latent variable can be set in one of two ways: fixing a factor loading to a constant, usually 1 or fixing the latent trait variance (or residual variance) to a constant, usually 1.0. In MGCFA, the Mplus default is to fix the first factor loading to 1.0 in all groups. Other SEM software packages also set the metric by constraining the first factor loading to 1.0. The variance of the factor is freely estimated. There is another parameter in Mplus categorical factor analysis, a so-called scale parameter (symbolized delta) that does not exist in the IRT framework that must be constrained to be equal across groups for the Mplus SEM model to replicate the IRT model.

An equivalent model is one that estimates all factor loadings and constrains the variance to 1.0. IRT software packages use this approach and assume the underlying latent trait has mean 0 and unit variance for the reference group, while the mean and variance are estimated for the studied group. The equations in this paper that show the relationship between factor analysis parameters and IRT parameters assume that the variance of the latent trait is 1.0. Setting the metric by only constraining the first factor loading to 1.0 in both groups in an MGCFA model will result in the two groups having non-overlapping item response functions because the variances of the latent trait are different.

MIMIC: As presented above, the MIMIC latent trait model is a variant of the factor analytic structural equation model, and except for differences in parameterization is equivalent to an IRT model (Muthén & Muthén, 1998-2013), which assumes that all items load on a single underlying latent attribute such as physical function. MIMIC models allow uniform DIF to be detected across e.g., ethnic/racial groups, after controlling for covariates. A measure of DIF is the direct effect of a studied variable (such as ethnic group membership) on the item response estimated from a model that includes (controls for) the health or mental health status variable. As reviewed in Teresi and Jones (2013, pg. 152) the measurement model, expanded to include direct effects of background variables is: $y^* = \Lambda\eta + Kx + \in$, and the structural equation model: $\eta = \alpha + \Gamma x + \zeta$, where Γ contains regressions of the underlying trait, and describes the effects of covariates (studied group) on the underlying trait. Direct effects (***K***) are estimated from a regression of individual test items' latent response variables on covariates such as background variables (*x*). In the MIMIC model, a significantly non-zero value for ***K*** indicates uniform DIF: an item difficulty shift for members of the group marked by ***x***. Although traditional MIMIC models assess only uniform DIF, with careful construction, the model above can be expanded to include an interaction term for group by trait to capture non-uniform DIF (Woods & Grimm, 2011). MIMIC has evidenced superior performance in DIF detection compared with IRTLR methods, particularly with small studied group sample sizes (Woods, 2009b). Variants of this model were used in the analyses presented in this series of papers by Jones et al. (in press) examining physical function and by Jensen and colleagues (in press) examining sleep.

Multiple Group Confirmatory Factor Analysis: The CFA model can be expanded to test for DIF in multiple groups (MGCFA) and among multiple dimensions using general latent variable modeling approaches (Muthén, 2002). Covariates can also be entered into MGCFA models (and could be called MG-MIMIC models). A measurement model can be estimated separately, but simultaneously, in for example, Black, Hispanic and non-Hispanic White groups. Model identification and measurement model calibration is achieved by imposing equality constraints on the measurement model parameters and variance parameters for the latent, e.g., physical or mental health state across groups. Uniform DIF can be detected by relaxing equality constraints on threshold parameters (τ) and non-uniform DIF by relaxing equality constraints on factor loadings (λ) across groups (Muthén, 1989a). Changes in model modification indices (chi-square scaled derivatives from the model fit function) are examined in DIF analyses (Jones, 2006; Muthén, 1989a). A robust parameter estimation procedure is based on a mean and variance adjusted weighted least squares procedure (WLSMV; Muthén, du Toit, & Spisic, 1997) incorporated into MPlus SEM software (Muthén & Muthén, 2013).

Different levels of equality constraints (subject to model identification) across these models constitute a hierarchy of factorial invariance (e.g., Meredith, 1993). Strong factorial invariance is assumed if groups have equivalent τ (threshold/difficulty) and λ (factor loading) values. Uniform DIF is assessed by relaxing assumptions of group equivalence in the means for the latent response variables or thresholds for observed categorical variables, and non-uniform DIF by relaxing equivalence assumptions for item factor loadings. Further discussion of the levels (hierarchy) of factorial invariance is provided in several reviews (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 2003; Gregorich, 2006; Mellenbergh, 1989; Meredith, 1993; Meredith & Teresi, 2006; Vandenberg & Lance, 2000). The relationship between the MIMIC model and the MGCFA approach is illustrated in Teresi and Jones (2013).

The analyses examining physical function and sleep were both conducted using a structural equation modeling approach. Jensen and colleagues (in press) used both MGCFA (Meredith, 1993; Muthén & Asparouhov, 2002; Muthén & Lehman, 1985) and MIMIC (Jöreskog & Goldberger, 1975; Muthén, 1984) as the DIF detection methods in examining measurement invariance of the PROMIS sleep short forms. The measurement model was tested with respect to 10 sleep short form items using confirmatory factor analysis. Differential item functioning with respect to race/ethnicity (White, Black, Hispanic and Asian) was assessed by fitting separate measurement models for each group. All parameters were free to be estimated in the baseline (first) unrestricted model. The second model specified all factor loadings and thresholds (intercepts) to be constrained simultaneously to equality. Modification indices and expected parameter changes were used to assess whether DIF was present. However, to achieve appropriate fit, four items from the hypothesized model were removed, due to item content overlap. A method effect was also identified due to the use of both positively and negatively worded items; residual covariances were modeled for positively worded items. The factor loadings and intercepts were freed for one item for one comparison group because of large group differences in the loadings, and the establishment of partial strong factorial invariance for the sleep disturbance scale. Standardized residuals, correlations, modification indices, and expected parameter changes were examined to arrive at the six item version that was used for further analyses of race/ethnicity using multiple group CFA. This multigroup CFA model was then extended into a MIMIC model to include age and sex.

The analyses by Jones and colleagues (in press) used a combined MGCFA MIMIC approach which allows thresholds and loadings to differ by group. An iterative multiple group MIMIC analysis, adjusting for the effects of age, sex, race/ethnicity, and education level as appropriate was performed using Mplus/WLSMV. Parameters were tested for DIF. Separate models were constructed for a series of two-group tests contrasting the reference group versus one focal group (e.g., White, Black, Asian, Hispanic). Simultaneous measurement models for two comparison groups were examined for differences in factor loadings (discrimination) and item thresholds (difficulty or severity levels) across a focal and a reference group. All models included adjustment for possible confounders on the level of the latent trait. The MIMIC models include these (mean-centered) covariates as adjustment factors but they were not evaluated for DIF.

Logistic and ordinal logistic regression

The method used as the primary approach to DIF analyses in early studies of PROMIS item banks was logistic regression (LR; Swaminathan & Rogers, 1990) and ordinal logistic regression (OLR; Zumbo, 1999) using an observed conditioning variable. For the OLR formulation proposed by Zumbo (1999) and demonstrated by Gelin & Zumbo (2003), the item response *Y* is specified as a latent continuously distributed random variable. The formula for OLR as given by Zumbo (1999) is: logit $[P(Y \leq k)] = a_k + \beta_1$ trait + β_2 group + β_3 (trait * group), where *Y* denotes item response to category *k*.

Basically three nested models are examined:

Model 1: $\alpha + \beta_1 x_1$;

Model 2: $\alpha + \beta_1 x_1 + \beta_2 x_2$;

Model 3: $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3(x_1 x_2)$;

where x_1 is the trait variable, e.g., anxiety, and x_2 the group or studied covariate, e.g., race/ethnicity. β_1 is the coefficient for trait; β_2 is the coefficient for the group or ordinal studied covariate; and β_3 is the coefficient for the interaction of group * trait. After conditioning on the trait, the main effect of the group variable is tested; this is a test of uniform DIF in the threshold parameters. The significance of the interaction term $\beta_3(x_1 x_2)$ is tested for evidence of non-uniform DIF. The OLR test for DIF uses the cumulative information of the ordinal responses by comparing the odds of endorsing a response less than or equal to *k* versus a response greater than *k*.

Different criteria are used by different investigators in order to identify DIF using OLR. One rule (Gelin & Zumbo, 2003) examines uniform DIF by comparing the R^2 values between the second and first steps in order to measure the unique variation attributable to the group differences over and above that of the conditioning variable. Tests of non-uniform DIF involve the effect of both the group and the interaction, over and above the disability score. Effect size measures are incorporated into the procedure. For example, Jodoin and Gierl (2001) specify two criteria: the two degrees of freedom chi-square test for DIF (testing for the group and interaction effects simultaneously) must have a p value less than or equal to 0.01, accompanied by an R^2 effect size value of at least 0.035.

This method is illustrated in the papers by Reeve et al. (this issue), and by Hahn et al. (in press). However, Reeve and colleagues (this issue) used a standard observed variable OLR approach in sensitivity analyses, while Hahn and colleagues (in press) used OLR with the latent conditioning variable (IRTOLR), described below as the primary DIF approach.

IRTOLR: A modification of the OLR approach, IRTOLR (Crane et al., 2004; Crane, Gibbons, Jolley, & van Belle, 2006) uses estimates from a latent variable IRT model, rather than the traditional observed score conditioning variable, and incorporates effect sizes into the uniform DIF detection procedure. The method allows OLR to be performed with an iteratively purified IRT trait estimated as the matching criterion. A program, lordif was developed (Choi, Gibbons, & Crane, 2011) to perform the analyses; this software uses ltm in R (Rizopoulus, 2006, 2009) to obtain IRT item parameter estimates for the Graded Response Model (Samejima, 1969), and the Design package for OLR (Herrel, 2009).

DIF Flagging Rules: Incorporation of magnitude measures such as R^2 change in LR and OLR can help to reduce flagging of non-salient DIF (Hidalgo, Gómez-Benito, & Zumbo, 2014). For example, Reeve and colleagues (in press) identified salient DIF using the criteria of $R^2 > 0.13$ (Zumbo & Thomas 1997). Simulations by Meade, Lautenschlager, and Johnson (2007) resulted in the recommendation to apply empirically derived DIF cutoff (threshold) values using a Monte Carlo simulation approach. This approach has been incorporated into the DIF detection software (lordif; Choi et al., 2011) used in PROMIS, and in the papers in this series on depression, anxiety, cognition, pain and social function. The lordif software includes a number of DIF effect size measures based on those described above: the change in Beta and pseudo R^2 from models with and without DIF terms. Also estimated are several magnitude and impact indices (Kim, Cohen, Alagoz, & Kim, 2007) described below, which are based on IRT parameters.

IRTOLR and lordif (Choi et al., 2011) were used in sensitivity analyses for the depression, anxiety, cognition and pain domains. Hahn et al. (in press) used IRTOLR and lordif to examine DIF in social function items. These authors applied a magnitude cutoff of McFadden pseudo-R^2 change ≥ 0.010. A series of hierarchical nested models permits DIF evaluation. As presented earlier, the latent social function variable was entered in Model 1; in Model 2 the studied group variable was added to the model (e.g., race/ethnicity, gender, language). Model 3 included social function, group and the interaction term for social function-by-group. Uniform DIF was detected by comparing Model 1 vs. Model 2 and non-uniform DIF by comparing Model 2 vs. Model 3.

Magnitude and impact for IRT-based DIF methods

Because significance tests alone are subject to chance findings, and with large sample sizes, trivial differences in item functioning between groups may be significant; in addition to corrections for multiple comparisons, effect size or DIF magnitude measures are often used in conjunction with statistical tests. Examination of magnitude and impact is important when making decisions about whether to remove an item from a measure or to consider providing separate calibrations for different groups in item banks (Teresi, Ramirez, Jones, Choi, & Crane, 2012). As reviewed above, the OLR approach incorporates effect size measures into the DIF detection method. Most IRT- and SEM-based methods examine magnitude and impact in separate procedures.

DIF Magnitude: The magnitude of DIF relates to the degree of DIF present in an item, and is also referred to as an effect size. IRT-based indices of DIF magnitude are derived from expected item score functions. The item-level expected score is the sum (over categories) of the probability of response in category *k*, weighted by the category score e.g., the ordinal code for the category, such as *not at all* (5), *very little* (4), *somewhat* (3), *quite a lot* (2) and *cannot do* (1).

Several summary measures describe the magnitude of differences between or among the item characteristic curves or expected item score functions for polytomous items. Introduced by Wainer, Sireci, and Thissen (1991) for binary items, these effect size measures are used frequently for DIF magnitude assessment (See also Chang & Mazzeo, 1994; Collins, Raju, & Edwards, 2000; Morales, Flowers, Gutiérrez, Kleinman, & Teresi, 2006; Orlando-Edelen et al., 2006; Steinberg & Thissen, 2006; Teresi et al., 2007). For example, for binary items, the exact area methods compare the areas between the item response functions estimated in two different groups; Cohen et al. (1993) extended these area statistics for the graded response model. Expected item scores are also central in the area statistics and the Differential Functioning of Items and Tests (DFIT) methodology developed by Raju (Raju, 1990; Raju, 1999; Raju, Oshima, & Wolach, 1995a; Raju et al., 2009; Raju, van der Linden, & Fleer, 1995b) and Flowers and colleagues (Flowers, Oshima, & Raju, 1999) to examine the magnitude of the gap between the ICCs for comparison groups. The non-compensatory DIF (NCDIF; Raju et al., 1995b) effect size measure is used widely. The NCDIF index is weighted by the focal (studied) group density such that more weight is given to differences in the region of the trait with the highest frequency in the targeted group. For item *i*, calculated is the average (expected value) of the squared difference between expected item scores for individuals as a members of the focal group and as a members of the reference group.

The average unsigned area difference (AUD) between the expected item response functions weighted by the studied group density, evaluated at various quadrature (theta) points provides another effect size measure (see Raju, 1988; Wainer, 1993; Woods, 2011). This AUD value is similar to NCDIF, but is not squared. Woods (2009a) studied the AUD and found it was not affected by the number of anchor items. An issue is what cutoff values to use for flagging salient DIF. Because of the sensitivity of cutoff thresholds to the distribution of parameter estimates, simulations to derive cutoffs were conducted to establish test-wide critical values to be used across studies and data sets (Flowers et al., 1999). Another approach is to derive sample (data set)-specific cutoff thresholds using Monte Carlo studies. This method has been incorporated into software such as lordif (Choi et al., 2011). Methods such as item parameter replication (IPR; Oshima, Raju, & Nanda, 2006) have also been used to develop empirically derived cutoffs for application to specific data sets with polytomous item responses in the context of DFIT (Raju et al., 2009). The IPR method does not require Monte Carlo studies, but relies on parameter replications based on focal group item parameter variances and covariances (Seybert & Stark, 2012).

The test-wide critical value method proposed by Flowers et al. (1999) produces cutoffs which were used in the analyses presented in this series. This method was found in simulation studies to perform similarly to (Seybert & Stark, 2012) or better than (Raju et al., 2009) the IPR method in terms of power and type 1 error with polytomous data. However, the method must be used under conditions of iterative linking in which equating constants are purified iteratively. The linking process is repeated after removal of items with DIF in previous steps, and new equating constants are derived and used in the DIF analyses, as applied in several of the papers in this series. Details of these measures and formulas are presented in an overview paper (Kleinman & Teresi, this issue) describing magnitude and impact measures. For reviews of effect size measures and DIF detection, see also Steinberg and Thissen (2006) and Monahan, McHorney, Stump, and Perkins (2007).

In the context of OLR, effect sizes are incorporated into the tests of DIF as described above to reduce false DIF detection. In addition, if uniform DIF is observed, the odds ratio can also be used to interpret the direction of the DIF; e.g., what the odds are of the studied group, as contrasted with the reference group responding in a more disordered direction to an item such as fearfulness after conditioning (matching) on overall symptomatology, e.g., anxiety. Use of the delta log odds ratio in the context of binary items was found to reduce false DIF detection (Hidalgo et al., 2014).

Aggregate Scale Impact: Impact refers to the influence of DIF on the scale score. Scale level impact can be expressed as group differences in the total test (scale) response functions. In the context of item response theory, differences in "test" response functions (Lord & Novick, 1968) can be constructed by summing the expected item scores to obtain an expected scale score. The test response function, also called the test characteristic curve (TCC) relates average expected scale scores to theta (the estimate of health or mental health). These latter functions show the extent to which DIF cancels at the scale level (DIF cancellation). (Detailed formulas and calculations can be found in Collins et al., 2000; Kim et al, 2007; Orlando-Edelen et al., 2006; Steinberg & Thissen, 2006; Teresi et al., 2007; Wainer, 1993; Wainer, Sireci, & Thissen, 1991.) Several authors examined the scale level impact of DIF using expected scale score functions for the analyses presented in this series.

Several impact measures, based on examination of group differences in these functions developed for binary items (Wainer, 1993) were expanded for polytomous items (Kim et al., 2007). These impact measures include the expected impact of DIF on scores in terms of absolute group differences between item true-score functions and density-weighted differences between groups. The latter adjusts for the actual distribution of individuals; if few respondents are located at the point where the differences are greatest, the weighted impact will be less.

The DFIT Compensatory DIF index quantifies differences in expected scale scores and is incorporated into the Differential Test Functioning (DTF) method (Fleer, 1993; Flowers et al., 1999; Oshima, Kushubar, Scott, & Raju, 2011; Oshima et al., 2006; Raju et al., 1995 a, b). This method has been evaluated (Meade, Lautenschlager, & Johnson, 2007) and applied to cognitive assessment data (Morales et al., 2006; Teresi et al., 2000; Teresi et al., 1995; Yang et al., 2011). Differential functioning at the test level (aggregated DIF impact) is the sum of differential functioning at the item level, and indicates how much each item's compensatory DIF (CDIF) contributes to differential test functioning of the whole measure. DIF in one item can cancel out DIF in another item; CDIF includes information about bias from other items. The DTF index (Oshima et al., 2011; Raju et al., 1995) is a summary measure of these differences and reflects the aggregated net impact. Plots (for each group) of the expected scale score against the measure of the state or trait (e.g., depression) provides a graphic depiction of the difference in the areas between the curves, and shows the relative impact of DIF.

More recent work on effect sizes is presented in Stark, Chernyshenko, and Drasgow (2004); Steinberg and Thissen (2006); Kim, Cohen, Alagoz, and Kim (2007); and in the article in this issue on magnitude and impact measures. Expected scale score functions were examined in the analyses of PROMIS cognition, depression, pain and anxiety short form data. Hahn and colleagues (in press) examined these plots and used lordif to obtain mean group differences in theta scores, adjusting for DIF or not. ANOVA was used to provide estimates of Cohen's *d* (see Cook et al., 2011).

Impact can also be examined by comparing model-based DIF-adjusted mean scores in the context of MIMIC and MGCFA. MIMIC models assess DIF magnitude through direct effects, and impact by comparing the estimated group effects, i.e., indirect effects in models with and without adjustment for DIF (Jones & Gallo, 2002; Jones, 2006). DIF adjusted and unadjusted effect estimates can be converted to estimated differences in mean scores on the latent variable (Jones, 2006). Jones and colleagues (in press) examined DIF impact by calculating the mean group difference on the underlying latent trait (physical functioning) on a normal metric (i.e., the latent trait is constrained to a unit normal distribution in the reference group). The mean differences were presented for models with and without adjustment for detected DIF. The percent difference in the mean scores was also estimated as a descriptor of DIF impact.

Individual Impact: In addition to aggregate level impact, individual impact can be assessed in the context of latent variable models, by fixing and freeing parameters based on findings of DIF and examining changes in trait score estimates. The unadjusted theta estimates are produced from a model with all item parameters set equal for the two groups. The adjusted thetas are produced from a model with parameters that showed DIF based on the IRT results estimated separately (freed) for the groups. The capacity to fix and free parameters based on DIF, and compare theta estimates is incorporated into software packages such as MULTILOG (Thissen, 1991); IRTPRO (Cai et al., 2012); FlexMIRT (Cai, 2013; Houts & Cai, 2013) and lordif (Choi et al., 2011), and can be coded directly in structural equation modeling software. This method permits comparisons of trait (e.g., mental health status) measure estimates that are DIF free to those with parameters estimated without DIF adjustment. This methodology has been used by several authors to examine the individual impact of DIF (e.g., Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Teresi et al., 2009), and was used in the articles on depression, anxiety, cognition, pain and social function in this series of papers.

This general methodology was extended by Crane and colleagues (2007b); the difference between scores unadjusted for DIF and those that account for DIF (with simultaneous control of covariates) is calculated to examine the cumulative impact of DIF on individual participants. The distribution of these difference scores is then examined; for individual-level DIF impact, a box-and-whiskers plot of the difference scores is constructed. As reviewed in Teresi, Ramirez, Jones, Choi, and Crane (2012), when a minimally important difference (MID) has not been established for the instrument, an approach used in lordif is to plot the differences due to DIF against the median standard error of measurement (SEM). Differences larger than that value are termed salient individual-level DIF impact. Examples of these plots are shown, using the PROMIS Anxiety item bank data (Choi et al., 2011), and with respect to PROMIS physical function items (Paz, Spritzer, Morales, & Hays, 2013) and in the articles in this issue on measuring impact by Kleinman and Teresi (2016). This method was used by Hahn and colleagues (in press). These authors used lordif software to fix and free parameters according to DIF results. Adjusted scores were calculated using common (shared) item parameters for items without DIF across all groups. Unique (group-specific) item parameters were used for items with DIF. DIF impact at the score level was defined as the score difference, which is equivalent to subtracting the adjusted score from the unadjusted score for individuals.

Challenges to accurate DIF detection

DIF detection can be compromised by several factors: lack of purification and improper selection of anchor items; violations of model assumptions, e.g. unidimensionality and local independence; small sample sizes for the number of parameters estimated; unequal sample sizes in subgroups; skewed and sparse item level data; and differences in theta distributions among groups. Differences in estimated mean disability and discrimination distributions between groups can result in inflated type 1 error rates (Li, Brooks, & Johanson, 2012). The shape of the distribution can also affect DIF detection. For example, Woods (2011) compared several non-parametric methods to IRTLRDIF when the latent distributions were not normal for both groups; ordinal response data were examined. She found that latent non-normality affected all procedures, but IRTLRDIF was more robust to latent nonnormality than the nonparametric approaches. Increasing the number of DIF-free anchor items had a mitigating effect on Type 1 error inflation for all methods.

Purification: Purification is the process of iteratively testing items for DIF so that final estimation of the trait can be made after taking this item-level DIF into account. For example, Mazor, Hambleton, and Clauser (1998) proposed a two-stage DIF evaluation: first, all test items are examined for DIF; next, those items showing DIF are removed, and the process is repeated. Such two-stage DIF detection methods were originally proposed by Thissen, Steinberg and Wainer (1988). Stark et al. (2006) and Lopez Rivas, Stark, and Chernyshenko (2009) also proposed DIF-testing using a two-step procedure, first constructing a constrained baseline model in which all item parameters are fixed equal and each of the studied item parameters are freed, followed in stage two by selection of a set of items that are DIF-free. The DFIT method requires purification at the parameter estimation phase (using anchor items), and additional purification of the equating constants through re-equating after items found to exhibit DIF at the first stage are removed.

Simulation studies have shown that many methods of DIF detection are adversely affected by lack of purification. In one simulation comparing parametric methods (Finch, 2005), IRTLR was the most affected (in terms of false DIF detection) by lack of purification and MIMIC least. However, Wang, Shih, and Yang (2009) found that purification outperformed a standard MIMIC approach in terms of power and type 1 error under conditions of low percent DIF for binary items and Wang and Shih (2010) found that a pure anchor method was superior to standard MIMIC approaches for polytomous items.

They recommend iterative purification. Purification is also recommended in the context of logistic regression (French & Maller, 2007).

Anchor Item Selection: Item sets that are used to construct preliminary estimates of the attribute assessed, e.g., depression usually contain items with DIF. Thus, estimation of a person's standing on the attribute may be incorrect, using this contaminated estimate. Anchor items are those items found (through an iterative process or prior analyses) to be free of significant DIF. These items are used to estimate theta, the conditioning variable that links groups compared in terms of the underlying attribute, e.g., depression level. If no prior information about DIF in the item set is available, initial DIF estimates can be obtained by treating each item as a "studied" item, while using the remainder as "anchor" items. The first step is to estimate the mean and variance for the target groups studied (while setting the reference group mean to 0 and variance to 1). The anchor sets the metrics of different groups on a common scale. Typically, multiple anchors are selected in the context of IRT DIF methods, while only one may be used for SEM-based methods. For example, in MGCFA, a single item may be selected to serve as the anchor to fix the scale. However, setting the first item to have the same factor loading of 1.0 in two groups is equivalent to setting an anchor item only if the variance of the latent trait is constrained to be equal across groups. Or, equivalently, the latent trait variance may be set to 1.0 in both groups and constraints relaxed on an item discrimination parameter to be equal across groups.

Use of a reference anchor item or set has been found to improve type 1 error rates in DIF detection for some models, e.g., MIMIC (Wang & Shih, 2010), IRTLR (Woods, 2009a) and hierarchical generalized linear models (Chen, Chen, & Shih, 2013); however, the way in which the item(s) are selected under conditions of differences in the trait distributions for the comparison groups can affect the accuracy of DIF detection (Chen et al., 2013). Selection of the anchor items and the parameters to fix to some value, e.g., loadings and intercepts in order to set the scale can result in different parameter estimates (Little, 2000).

Anchor Item Selection Methods: Best methods for selecting anchor items have been reviewed (e.g., Kopf, Zeileis, & Strobl, 2015a, 2015b; Wang & Shih, 2010; Woods, 2009a), and are summarized briefly below.

1. All-other and all-other with purification

The method used in the analyses presented in several of the articles in this series is the so-called "all-other" anchor method in which initial DIF estimates are obtained by treating each item as a "studied" item, one at a time, while using the remainder as "anchor" items. This is the method used most often in IRTLR (Bolt, 2002; Kim & Cohen, 1998). For each studied item, a model is constructed with all parameters constrained to be equal across groups for the anchor items, with the item parameters of the studied item freed to be estimated distinctly for the comparison groups. The studied item is included in the analyses together with the anchor items. Because of the simultaneous estimation procedures for the two groups, item equating is not required. The more items with DIF in the test, the less accurate this procedure will be. Purification is used to iteratively test for and remove items with DIF from the anchor; however, purification is also affected by the number of items with DIF (Wang, Shih, & Sun, 2012).

An example is embodied in the Wald 1 and Wald 2 test approach. The Wald 2 approach proposed by Langer (2008) performs the all-other selection method without further iterative testing using anchor items. With a single model, the scale is identified by setting the reference group mean to 0 and the standard deviation to 1; parameters for anchor items are fixed as equal and the studied item parameters freed. The Wald 1 test proposed by Cai, Thissen, and du Toit (2012) specifies anchor items using the constant anchor method described below or an all-other ("sweep") approach to arrive at an initial set of potential anchors for use in purification. Then the final iterative purification and DIF analyses can be performed. The two methods have been compared recently (Woods et al., 2013) and as with other approaches that do not use anchor items, Wald 2 resulted in inflated type 1 error and is not recommended. Wald 1 and IRTLR were superior and performed well.

Because there are so many (i * i-1 items) tests of DIF when using the IRTLR approach (as many as 90 tests for a 10 item short form), the false discovery rate is inflated. Thus, to avoid multiple tests, Stark et al. (2006) suggested selecting a single anchor based on a non-iterative rule. Following a test of all items with all others as anchors, one item is selected that has the highest factor loading among the (presumably unbiased) items. The reasoning is that the anchor defines theta (the trait estimate) for the DIF analyses, so the anchor should be highly related to theta. Often this approach is used in factor analyses in selection of the most discriminating item to set the metric for the latent construct. In the context of IRT, other investigators (González-Betanzos & Abad, 2012) have recommended selection as anchor items those with high discrimination parameters. IRTLR tests have greater power when the discrimination parameter is larger (Ankenmann, Witt, & Dunbar, 1999; Lopez Rivas, Stark, & Chernyshenko, 2009).

2. Constant anchor method

An alternative approach is the constant anchor method (Thissen, Steinberg, & Wainer, 1988; Wang & Yeh, 2003) in which selected items are chosen in advance to serve as anchors. This method has been found to perform in a superior fashion in the presence of large amounts of DIF, and longer anchors result in greater power, if they are DIF-free (Thissen et al., 1988; Wang, 2004; Wang & Yeh, 2003). For example, Wang and Yeh (2003) found that in the context of IRTLR a larger number of anchor items (four vs. lower numbers) resulted in greater power for DIF detection. In general, it has been recommended that at least four anchor items be used for adequate power for DIF detection (Wang, 2004; Wang et al., 2012) and for construct measurement integrity (Cohen, Cohen, Teresi, Marchi, & Velez, 1990).

Wang (2004) compared the constant anchor method to the all other. The all-other method did not perform well unless the difference in mean item difficulties between groups approached 0. The constant anchor method yielded unbiased parameter estimates, well-controlled type 1 error and high power of DIF detection regardless of large differences in the mean item difficulties between groups and high percent of items with DIF.

Stark et al. (2006) showed in the factor analytic setting that the free-baseline model (a constant anchor method) was superior to the constrained (parameters set equal) baseline model (all-other) method for both IRT-based and confirmatory factor analytic methods. This finding was confirmed by Shih and Wang (2009) for the MIMIC model and binary items. These authors found that use of a constant anchor following iterative identification of DIF-free items yielded high power and low type I error. Under the constant method, the more anchor items, the higher the power for detecting DIF with MIMIC; however the increase in power diminished with more than four anchors and was trivial after 10 (Shih & Wang, 2009).

Wang et al. (2012) proposed a DIF free then DIF (DFTD) procedure, which is two-step. First, four items are selected as an anchor or 10% to 20% of the items that are less likely to have DIF are selected. Then other items are tested against this anchor. This is similar to the method used by Orlando-Edelen, Thissen, Teresi, Kleinman, and Ocepek-Welikson (2006), in which items are tested iteratively until a minimum number of items without DIF are identified as anchors. Then the remaining items are again tested for DIF.

Based on a simulation study of IRTLR (Lopez Rivas et al., 2009), it was found that in the context of small samples, selection of three to five items without significant DIF and with the largest discrimination parameters was optimal. This result was confirmed by Meade & Wright (2012). However, selection of this subset of items that are DIF-free remains a challenge in many settings. The method of selecting a minimal number of DIF-free items is only practical if there are enough items without DIF. Shorter measures, such as those examined here may not yield enough DIF-free items. Selection of anchor items iteratively may not protect against false positives if there are many items with DIF.

3. Rank order and iterative rank order with purification

Woods (2009a) investigated several anchor methods for use with IRTLR analyses. She recommended selecting anchor items based on a log-likelihood ratio (LR) rank test in order to avoid inflated type 1 error resulting from multiple hypothesis tests. The ratio of likelihood ratio statistics to the number of free parameters (*f*, the degrees of freedom) is calculated from the first all-other model. Items are rank-ordered based on this LR/*f* ratio and because a large LR indicates DIF; items with small values of this ratio are those selected as anchors. Then IRTLRDIF procedures are performed again with the new anchors and the B-H procedure is used in adjustment for final selection. Woods (2009a) found this method to have lower type 1 error rates than the all-other iterative backward method for polytomous items and relatively large sample sizes. However, as expected, the rate of accuracy (type 1 error and power) decreased with greater percentages of items with DIF. With smaller sample sizes (reference group n = 600; studied group n = 200), power was poor with the use of only one anchor. Although research reviewed above has shown that the use of one anchor is sufficient for larger samples, e.g., 1500 and 500; use of only one anchor item is not recommended for smaller samples and there is the danger of construct meaning drift. Group sample sizes for the analyses presented in this series of papers were relatively large, approaching 500 in the studied groups; however such sizes are rare in practice outside of educational testing.

The IRTLR all-other procedure can be used as a first step in selection of anchor items and final DIF tests can be obtained in analyses using the Wald test procedure, as was done with respect to several analyses presented in this series. This approach is similar to the Wald 1 procedure recommended by Cai et al. (2012), in which anchor item selection is incorporated into DIF testing. Final parameter estimates are obtained after the final run and based on the DIF results that inform which parameters can be estimated as equal (DIF free) or separately because of DIF. If there is DIF in any parameter, all parameters are estimated separately.

Because the use of IRTLR with the rank based method of Woods (2009a) is less accurate in the presence of many items with DIF, another method proposed by Wang et al. (2012) is a rank-based scale purification method (RB-S). The all-other method is performed iteratively with purification and the rank order method applied to select a constant anchor set. The procedure is applied to all items until the same set of items are identified with DIF in two consecutive steps. Then anchor items with the smallest LR/*df* ratio are selected: either four or some percentage of items. This method, investigated using IRTLRDIF for binary items, yielded high power and adequate type 1 error control. Applying purification reduced type 1 error and increased power, and was superior to rank order alone in terms of type 1 error under conditions of a large amount of DIF (Wang et al., 2012).

Setodji, Reise, Morales, Fongwam, and Hays (2011) investigated a method for selection of anchor items in the context of the likelihood ratio test. These authors used a semi-parametric permutation test to obtain the empirical distribution of difference statistics. Estimates of differences between parameters obtained from models with parameters constrained as contrasted with those with parameters set free are obtained. Respondents are randomly assigned to two groups, and the randomization repeated a sufficient number of times to obtain the permutated empirical distribution, which is used to obtain a two-sided test statistic. Items with p values > 0.05 are candidate anchor items. However, this method has not been investigated in comparison to other anchor item selection procedures.

4. Iterative forward selection with threshold or significance level criteria

Kopf, Zeileis, and Strobl (2015b) developed and examined several anchor item selection methods in the context of Rasch (Rasch 1960, 1980; see also Glas & Verhelst, 1995; Mair & Hatzinger, 2007) binary item models. These authors investigated various anchor selection strategies to arrive at the first DIF-free anchor. Typically items are selected in an iterative backward procedure, removing items identified with DIF; however, Kopf, Zeileis, and Strobl (2015a) proposed an iterative forward selection method in which items are tested one at a time against a single anchor, using a Wald test of differences in the *b* parameters. Using this single item, all items are tested for DIF against the single anchor and DIF-free items are added iteratively to the anchor. Anchor item selection strategies include ranking the absolute mean of the DIF test statistics or the number of significant results. With the latter approach, the items corresponding to the lowest number of test statistics above the test statistic threshold are chosen as anchors. Their new method, using a threshold criterion outperformed standard methods of selection (Kopf et al., 2015a).

Anchor item methods used in the articles in this series

The anchor item selection method used in many of the papers in this series applies a combination of the first three approaches reviewed above. For example, the analyses examining cognition, depression, anxiety, and pain used a hybrid approach to arrive at a constant anchor. Iterative procedures were used with the goal of identifying a minimum of four items that are DIF-free. The all-other approach was used; however, if very few items were DIF-free the rank-order method was used to select a starting anchor set. The iterative rank order method similar to that described above (Wang et al., 2012) was also used to identify the four items with the least DIF, and this result was examined in sensitivity analyses for consistency. Iterative purification was performed such that the final anchor set contained only items without DIF. The difference between the methods described above and the analyses presented in this series is that although the goal was to have at least four, it was not always possible to have a set number of anchor items. Rather, all items without DIF were selected after the original purification step. However, items that converted from DIF to non-DIF during the iteration process were not added back into the anchor set. Only those items that were consistently DIF-free throughout purification were retained.

As an example, the all-other method in IRTLRDIF (Thissen, 2001), accompanied by tests of LR/*f* was used to select items for iterative purification. IRTPRO option 3, which permits the all-other approach for the multiple group case was used. While IRTLRDIF permits comparison of only two groups at a time, IRTPRO option 3 is an all-other method that can be used with multiple groups, but requires separate runs for each item tested. This procedure is performed iteratively in a purification procedure, such that the analyses are repeated using the final subset of items (at least four if possible) identified as free of DIF as the "purified" anchor set. This (Wald-type) procedure is more robust than just relying on the all-other anchor procedure, and may take several iterations.

In the case of smaller sample sizes and when most items were identified with DIF, as was the case with many of the analyses presented here, the Woods (2009a) rank-order method has been recommended; it has been found that the number of anchor items did not affect adversely the effect size estimates (Egberink, Meijer, & Tendeiro, 2015), which are central to detection of salient DIF. A tradeoff is that too few anchor items compromises power for DIF detection; on the other hand, including items with DIF in the anchor results in inflated type 1 error (false DIF detection). As illustrated in the articles on depression and anxiety, it was not possible to obtain the desired number of anchor items for some subgroup comparisons. Thus, sensitivity DIF analyses were performed to determine the effects of varying sets of anchors on the results.

Although Kopf and colleagues (2015b) found that variants of the all-other approach just described were suboptimal in the context of a Rasch model with binary items, to our knowledge there is no evidence to support this finding in the context of the graded response model. Moreover, DIF detection methods proposed and used in the analyses described in this series incorporate sensitivity analyses and magnitude measures to mitigate false DIF detection.

Testing model assumptions

Dimensionality: Several investigators used exploratory factor analyses (EFA; Asparouhov & Muthén, 2009) and confirmatory factor analysis, respectively to examine the assumption of unidimensionality of the cognition (Fieo et al., in press), social function (Hahn et al., in press), pain (Teresi et al., in press), fatigue (Reeve et al., 2016), depression and anxiety (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016-a, 2016-b) item sets. The weighted least squares with adjustments for the mean and variance (WLSMV) estimation procedure in Mplus (version 7.1, Los Angeles, CA) was applied to the ordered categorical response data. Unidimensional models were examined within each group for several sets of analyses, and the model was identified by fixing the first item to 1 and constraining (fixing) the mean of the trait to 0 for the sample. Random halves of samples were used to examine exploratory and confirmatory models for some analyses (Fieo et al., in press; Reeve et al., 2016; Teresi et al., 2016, a,b; Teresi et al., in press).

Selection of the best methods and criteria for cutoff values for goodness of fit statistics is an area of controversy (e.g., Cook, Kallen, & Amtmann, 2009). Model fit statistics and criteria for goodness of fit (Bentler, 1990; Chou & Bentler, 1990; Chou & Wang, 2010) included the comparative fit index (CFI; Bentler, 1990; CFI > 0.95), Tucker-Lewis Index (TLI; Tucker & Lewis, 1973; TLI > 0.95), standardized root mean residuals (SRMR < 0.08), and the root mean square error of approximation (RMSEA < 0.06). Jensen and colleagues (in press) and Hahn and colleagues (in press) also applied a CFA approach, specifying a single underlying latent construct. Overall fit was evaluated using the chi-square test, RMSEA and the CFI. Following recommendations of Reise (2012), a bifactor model was also examined by several investigators to inform about dimensionality.

Jones and colleagues (in press) examined unidimensionality using permuted parallel analysis (Buja & Eyuboglu, 1992; Horn, 1965). Parallel analysis compares observed eigenvalues from a correlation matrix to eigenvalues that would be expected from a random set of variables. Random eigenvalues were generated by randomly shuffling responses across individuals, and estimating a correlation matrix. The procedure was performed multiple times to generate a permutation distribution for eigenvalues under the assumption of no association. The observed eigenvalues were then compared to the permutation distribution, and a p-value for each eigenvalue obtained.

Dimensionality and Reliability Coefficients from Factor Models: An index of dimensionality, the explained common variance (ECV) was calculated. IRT-based reliability estimates, conditional on the trait were estimated. Additionally, McDonald's (McDonald,

1999) Omega Total (ω_t), a reliability estimate that is based on the proportion of total common variance explained was estimated; ECV and McDonald's omega, generated from exploratory and confirmatory factor analyses, as well as several others recommended by Revelle and Zinbarg (2009) are contained in the "Psych" package that they developed in R (Revelle, 2015; www.R-project.org; R Development Team, 2008).

Cronbach's alpha (Cronbach, 1951) is the most widely used estimate of reliability, although it has significant limitations (e.g., Bentler, 2009; Sijtsma, 2009). This index is based on an unweighted sum score, and is typically calculated using observed response models and Pearson correlations; such values will generally be lower than McDonald's omega under conditions of unidimensionality (Zinbarg, Revelle, Yovel, & Li, 2005). A more appropriate estimate of internal consistency for ordinal data is an ordinal version of alpha (Zumbo, Gradermann, & Zeisser (2007) using polychoric correlations. The polychoric correlation is scaled differently and assumes an underlying unobserved continuous response variable. Polychoric correlations, typically used with polytomous items can be estimated using SEM packages such as MPlus (Muthén & Muthén, 2013). Ordinal alpha estimates will be higher than Cronbach's alpha and more similar in magnitude to McDonald's omega. Because McDonald's omega is typically derived from a latent bifactor model (e.g., Reise, 2012), it is arguably more invariant than values based on observed response models (see also Bentler, 2009). The code for the ordinal version with polychorics given in the R software package was used in the analyses of cognition, depression, anxiety, and pain. (A detailed description of its use can be found in Gadermann, Guhn, & Zumbo, 2012.)

Local Dependency (LD): Local independence requires that all pairs of item responses be independent, conditional on the latent trait. As reviewed in Houts and Edwards (2013), local dependencies can result in biased estimates of theta (Zenisky, Hambleton, & Sireci, 2002) and item parameters (Chen & Thissen, 1997), poor estimates of the parameter standard errors (Junker, 1991) and overestimation of information and coefficient alpha (Sireci, Thissen, & Wainer, 1991). Because local dependencies can result in inflated slopes (Houts & Edwards, 2013) sensitivity analyses were performed by several authors of articles in this series, examining the effects of DIF detection after removal of items with LD.

Numerous methods of LD detection have been proposed. These include for example tests of residuals estimated based on the difference between response and the modeled probability of response conditional on theta (Yen, 1984). A widely used method incorporated into software such as IRTPRO is Chen and Thissen's LD chi-square statistic (Chen & Thissen, 1997). This is based on a comparison of observed and expected frequencies derived from item by item two-way cross-tabulations; the likelihood ratio statistic (G^2) resulting from this comparison is chi-square distributed. These values are approximately z-scores (computed by subtracting the degrees of freedom from the chi-square-distributed statistics and dividing by the square root of 2*df*); higher values are indicative of violations (IRTPRO manual, Cai et al., 2012). The LD statistic is derived from the residual correlation among a pair of items, given a single factor model. The authors of the papers on depression, anxiety, pain, cognition, and fatigue used this approach. In the analyses presented in these papers, it was observed that the LD statistics were affected by sample size, and increased in value with increased sample size.

In the context of IRT binary data models, Liu and Maydeu-Olivares (2012) examined several LD statistics and identified the score tests (Glas & Suárez-Falcón, 2003; Liu & Thissen, 2012, 2014) and standardized bivariate residuals as the most powerful method of evaluating LDs. In the context of the graded response model, Houts and Edwards (2013) found that the best performing methods for detecting LD were the G^2 index (Chen & Thissen, 1997) described above and Jackknife Slope Indices (JSL; Edwards & Cai, 2008). This latter method examines the differences between slope estimates (which are inflated by high LD) before and after removal of the item studied. The modification indices from the structural equation model comparisons of restricted and unrestricted models (after freeing residual correlations for item pairs) evidenced a high type 1 error rate, and were considered as poor measures of LD detection.

Jones and colleagues (in press) although observing a well-fitting unidimensional model without residual covariances, examined possible violations of local independence by estimating residual correlations. Using Mplus, these authors examined two methods of parameterization: delta and theta. Jones and colleagues (in press) discuss that the delta parameterization method specifies a secondary latent variable common to both items in a pair. The theta specification permits estimation of the residual variance/covariance matrix. The delta method is appropriate if one views LD as an indicator of an unmeasured trait (such as response style) rather than content overlap. For example, in the bifactor model formulation, the second factor captures shared residual variance not accounted for by the general factor; the presence of such secondary factors may result in inflated slope (discrimination) parameters (see DeMars, 2014). Delta may also be preferred for ease of translating to the IRT metric, but does not allow access to the off-diagonal elements of the theta parameter matrix capturing the residual variance/covariances for items which can be freed for estimation.

Using a theta parameterization in Mplus, a single common factor model is specified and residual covariance parameters are fixed at zero, and freed iteratively one at a time, based on examination of corresponding modification indices that indicated model misfit. In the formulation of Meredith (1993), loadings and thresholds are first tested for invariance, followed by tests of residuals. This is the logic for the theta parameterization. With the IRT formulation, residuals are assumed to be invariant. However, heterogeneity in residual variances could occur, for example if patients with cancer who are in pain are more prone to variability in response due to distraction than are other groups studied. Woods and Harpole (2015) examined violations of this assumption of homogeneous residuals on DIF detection across several methods, including logistic regression and IRT-log-likelihood ratio tests for binary items. The logistic regression DIF detection method was more affected than was IRT-LR overall; however, both were affected in terms of power.

LD may result from highly similar content as was evidenced in the analyses of sleep, fatigue, and pain. Methods effects such as header/contingency formats in which responses to other items are contingent on the response to a header item may result in LD and require different (partial independence) IRT models (Carle et al., 2014; Reardon &

Raudenbush, 2006). Finally LD may result from lack of unidimensionality (Chen & Thissen, 1997; Houts & Edwards, 2013). In the analyses of physical function, depression, anxiety, and pain, despite strong evidence of unidimensionality, LD was observed. In the cases of depression and anxiety and pain, given that the Cronbach's alpha (Cronbach, 1951) estimates were lower than those of McDonald's omega, additional support for unidimensionality was observed; however, slight multidimensionality may be present, leading to a slightly elevated risk of false DIF detection.

Reeve and colleagues (this issue) found that the assumption of unidimensionality was met in the analyses of fatigue; however, some item pairs evidenced local dependency. These authors found two item pairs that showed potential LD; these items were removed from each LD pair and the discrimination parameter change examined. The result was a reduction in discrimination parameter values for the LD-paired item. Removal of the item did not affect the DIF results. Teresi and colleagues (this issue-a, this issue-b) performed sensitivity analyses removing items from pairs with high LD values and (resulting) high discrimination parameters. Discrimination parameters for remaining items were reduced for depression and DIF results did not change. However, in the analyses of anxiety, some small changes were observed in the results of DIF, none of high magnitude. In the analyses of pain, one item was removed from the analyses due to very high LD values with other items.

Model fit

IRT model fit was evaluated by several authors using the generalization of Orlando and Thissen's S-X^2 indicator for polytomous-response data (Orlando & Thissen, 2003). Reeve and colleagues (2016) found an item that did not fit because it was worded in a positive direction in contrast to other items. The IRT model was then fit to the full sample for the 13 remaining items; however, the S-X^2 indicator showed significant p-values (indicating lack of fit) for nearly all items. The authors posited that the S-X^2 statistics were inflated by the large sample size, such that item-level model fit could not be assessed adequately.

Jensen and colleagues (in press) investigated four short form models, none of which fit the data. There were positively and negatively worded items, and the authors posit that respondents were not always paying attention to the switch in the response scale from negative to positive wording, leading to a method effect. There was content overlap as well; e.g., "I tried hard to get to sleep" and "I had difficulty falling asleep." A two-step solution was applied. Items with content overlap were removed. The method effect was controlled by allowing the residual covariances of negatively worded items to covary. Ultimately, several items required removal.

Missing Data

Although little missing data were observed, most investigators treated missing data by removing subjects with 50 % or more missing items, and used individual imputation and prorating algorithms for the remainder of the items with missing data. For example, Fieo et al., (in press) used this approach. Jensen et al. (in press) used a hybrid approach. All participants with more than half of the items missing were excluded, and then a pairwise deletion approach (Asparouhov & Muthén, 2010) was applied to missing data from the remaining respondents using Mplus (available from: http://www.statmodel.com/ download/GstrucMissingRevision.pdf). Most software packages also permit estimation with inclusion of all available data by using maximum likelihood procedures and applying the expectation maximization algorithm. Respondents responding to as few as one item may be included; however, use of one item may not result in the best estimate of the underlying attribute for those individuals.

Discussion

Methods effects and model assumption violations: Moderate-level methodological challenges were encountered in the analyses of the PROMIS short forms. These included methods effects and local independence assumption violations. As reviewed in this article, local dependencies can result in inflated slope parameters and type 1 error (false DIF detection). Methods factors arose from 1) type of response category 2) combining different type of response categories, e.g., frequency and amount 3) inclusion of mixed mode of response, e.g., positively and negatively worded items 4) content overlap and possibly 5) response fatigue.

Model Assumption Violations: Local dependencies were observed for most of the short forms, often resulting in inflated *a* (discrimination) parameters. The LD was due to positively and negatively worded items, content overlap and more DIF in items measuring frequency as contrasted with amount of a symptom (e.g., Jensen et al., in press; Reeve et al., this issue). For example, Jensen and colleagues (in press), examining sleep items found that the unrestricted model resulted in very poor fit, due in part to multiple reverse-scored (positively and negatively worded items), changing response options and item overlap. The reverse-scoring was observed primarily in the sleep domain; the fatigue domain had only one reverse-scored item. Model fit improved by removing items with too much content overlap and allowing the residual variances of positively worded items to covary. Introducing a methods factor or residual correlations under these circumstances is not optimal; however, few options were available. A pair of items that are redundant, but reverse coded will have high LD values but the item set as a whole may be essentially unidimensional. Item pairs with high LD will contribute to poor fit (RMSEA too high), as was observed in the analyses of sleep.

Although Jones and colleagues (in press) found only the first eigenvalue significant ($p < 0.01$) after using the permutation test, suggesting that the physical function items were essentially and strongly unidimensional, there was substantial local dependence observed, which these authors did not model. Thus their findings of numerous items with DIF could have been affected by local dependency, resulting in false DIF detection. However, the impact was trivial for most analyses. For most comparisons, the mean differences in the estimates of the physical function latent trait (assumed normal with a mean of 0, and standard deviation of 1 in the reference group) with and without control for DIF did not exceed 0.03. The exception was for the Hispanic-reference group comparison, where larger differences were observed.

The analyses of depression, anxiety, and cognition domains, although providing evidence for essential unidimensionality, also showed evidence of high LD values. In addition, Cronbach's alpha coefficients estimated with Pearson correlations had lower values than the McDonald's omega total coefficients. However, when estimated with ordinal alpha and polychoric correlations, these values were aligned better. As discussed, alpha values will be lower than McDonald's omega under conditions of unidimensionality and heterogeneous loadings (Zinbarg et al., 2005), perhaps providing additional indirect evidence in support of the essential unidimensionality of the item sets.

As reviewed, local dependencies can result in poorly estimated and inflated discrimination parameters. In the analyses of depression and pain, higher discrimination parameters than expected were observed, and one item with a high LD was in part the cause. Similarly, Reeve and colleagues noted that the IRT item discrimination parameters for the first seven fatigue items (which use a frequency response scale) ranged from 2.45 to 3.98 and for the last six items (which use an amount response scale) ranged from 4.58 to 6.44. Removing an item from an LD pair typically resulted in a reduced discrimination parameter for the other item (e.g., from 6.3 to 5.1). Reeve and colleagues also noted that items on a frequency scale (*never* to *always*) evidenced more DIF than did those on a magnitude metric (*not at all* to *very much*). Additionally, the items that were on the magnitude metric had higher discrimination parameters (after adjusting for inflation due to local dependencies), indicating a better relationship to the underlying attribute.

The effects of contextual factors on item performance are well-known (Chen & Thissen, 1997; Steinberg, 2001; Thissen, Bender, Chen, Hayashi, & Wiesen, 1992). These problems were handled by a) deletion of items b) modeling and c) sensitivity analyses to examine the effects of violations on DIF detection. In general, DIF findings were robust to most of the violations. However, item removal was necessary for some analyses (fatigue, sleep, and pain).

Response Styles: Jones and colleagues (in press) in studying physical function short form items noted that the group comparison resulting in the greatest DIF impact (just below threshold for high impact) was for the groups: Hispanic vs. non-Hispanic White. Paz and colleagues also found DIF of high magnitude and impact for Spanish speakers vs. English speakers for the physical function item bank. They note that the *some* category was problematic in that Spanish speakers were much more likely to select response categories with the word *some*. They speculate that Spanish-speakers may prefer the middle category. This is contrary to most findings reviewed by McHorney and Fleishman (2006) in which they note that Hispanics were more likely to use extreme response categories; however, this may vary across the domains studied and the ethnic composition of the Hispanic group, which may not be homogeneous (Yang, Cazorla-Lancaster, & Jones, 2008). The extreme response style might be more prevalent in measures of psychological distress than physical function (see Teresi, Ramirez, Lai, & Silver, 2008). Response styles such as acquiescence and tendencies toward selection of extreme categories if consistent across items may not be detected easily and may affect the scale latent means and variances (Little, 2000). Such extreme response styles can render invalid cross-national comparisons and may require modeling. However, if the effect is observed only for some items, as is the case here, DIF analyses will most likely identify the problematic items. Otherwise external measures of response style may be required. However, an alternative strategy is to model the extreme responses. Bolt and Newton (2011) developed a multidimensional, nominal response model in order to model simultaneously the substantive trait, e.g., depression as well as the extreme response style trait. For the PROMIS physical function item set, there did not appear to be differences in response style from non-Hispanic White or Black groups; however, across most items, the Asians/Pacific Islanders tended to have an extreme response endorsement pattern, in that they were more likely to report the lowest level of impairment. Given that some authors identified more DIF (depression, anxiety, fatigue, and pain) among the Asians/Pacific Islanders as contrasted with other groups, the effects of extreme response style on the results cannot be ruled out.

DIF Methods: In terms of tests of DIF, as reviewed earlier there are two approaches to the Wald test for polytomous items. The first approach Wald 1 assumes that iterative purification and anchor items are used. The second approach, Wald 2 does not rely on anchor items, and uses the all-other approach for identification of DIF. Both Wald tests have the advantage of testing DIF for more than two groups. The Wald 2 test has the theoretical advantage of requiring only one model and is thus less computationally intensive than other DIF methods. However, Wald 2 is not recommended because of poor performance in comparison with Wald 1 which requires iterative purification and anchor items. In general, the evidence (Woods et al., 2013) appears to favor Wald 1 over IRTLR and Wald 2; thus the overall efficiency of Wald 2 may be compromised by elevated type 1 error. An advantage of both Wald tests is that there are fewer model comparisons that might inflate type 1 error rates. However, the Wald test relies on the IRT model and associated assumptions as well as robust estimation of item parameters and their error covariance matrix. Additionally, magnitude tests are required as separate steps.

The observed score OLR method relies on fewer assumptions than latent variable models; however, it is less efficient because it tests one item at a time, and may be less accurate. Reeve and colleagues when applying the observed OLR approach in sensitivity analyses found many items to evidence DIF, and the results were not always consistent with those observed using the Wald test. Observed score methods with short tests are less reliable, resulting in false DIF detection (DeMars, 2010). Use of such a method may require as many as eight DIF-free anchor items in order to obtain well-controlled type 1 error (Shih, Liu, & Wang, 2014). As reviewed above, carefully selected anchor items are generally required for more accurate DIF assessment, and latent variable OLR models incorporating magnitude measures are recommended.

Summary: Latent variable model state-of-the-art methods for examining measurement equivalence were introduced briefly in this paper to orient readers to the approaches adopted in this set of papers. Several methodological challenges underlying anchor item selection and model assumption violations were presented as a backdrop for the articles on measurement equivalence of PROMIS measures appearing in this and a second issue of Psychological Test and Assessment Modeling

Acknowledgements

Partial funding for these analyses was provided by the National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971(PI: Potosky, Moinpour) and by the National Institute on Aging, 1P30AG028741-01A2 (PI: Siu). The authors thank Adam Carle, Ph.D. and Katja Ocepek-Welikson, M.Phil. for reviewing earlier versions of this paper, and Stephanie Silver, MPH for editorial assistance in the preparation of this manuscript.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300). doi: 10.1111/j.1745-3984.1999.tb00558.x

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397-438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2010). Weighted least squares estimation with missing data. Mplus; (Available from: http://www.statmodel.com/download/GstrucMissingRevision.pdf)

Benjamini, Y., & Hochberg, Y. (1995). Controlling for the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289-300. doi: 10.2307/2346101

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137-143. doi: 10.1007/s11336-008-9100-1

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141. doi: 10.1207/S15324818AME1502_01

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*, 814-833. doi: 10.1177/00131644103 88411

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509-540. doi: 10.1207/s15327906mbr2704_2

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-566. doi: 10.1037/0033-2909.105.3.456

Cai, L. (2013). FlexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309-329. doi: 10.1348/000711007X249603

Cai, L., Thissen, D., & du Toit, S. H. C. (2012). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Carle, A., Jean-Pierre, P., Winters, P., Valverde, P., Wells, K., Simon, M., … & Fiscella, K. (2014). Psychometric Evaluation of the Patient Satisfaction With Logistical Aspects of Navigation (PSN-L) scale using item response theory. *Medical Care, 52*, 354-361 doi: 10.1097/MLR.0000000000000089; PMCID: PMC4149289

Carle, A. C., Cella, D., Cai, L., Choi, S. W., Crane, P. K., Curtis, S. M., ... & Hays, R. (2011). Advancing PROMIS's methodology: Results of the third PROMIS Psychometric Summit. Expert Review of *Pharmacoeconomics & Outcome Research, 11*(6), 677-684. doi: 10.1586/erp.11.74

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., & Rose, M., on behalf of the PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11. doi: 10.1097/01.mlr.0000258615.42478.55

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 39*, 391-404. doi: 10.1007/BF02296132

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289. doi: 10.2307/1165285

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2013). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear models when impact is present. *Applied Psychological Measurement, 38*, 18-36. doi: 10.1177/0146621 613488643

Cheung, G. W. & Rensvold, R. B. (2003). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. doi: 10.1207/ S15328007SEM0902_5

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: A R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulation. *Journal of Statistical Software, 39*(8), 1-30. doi: 10.18637/jss.v039.i08

Chou, C-P. & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier and Wald test. *Multivariate Behavioral Research, 25*, 115-136. doi: 10.1207/s15327906mbr2501_13

Chou, Y-T. & Wang, W-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*, 717-731. doi: 10.1177/0013164410379322

Cohen, P., Cohen, J., Teresi, J., Marchi, P., & Velez, N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14*(2), 183-196. doi: 10.1177/014662169001400207

Cohen, A. S., Kim, S-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335-350. doi:10.1177/014662169301700402

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential item functioning in a satisfaction scale. *Journal of Applied Psychology, 85*, 451-461. doi: 10.1037//0021-9010.85.3.451

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. doi: 10.1007/s11136-009-9464-4

Cook, K. F., Bombardier, C. H., Bamer, A. M., Choi, S. W., Kroenke, K., & Fann, J. R. (2011). Do Somatic and Cognitive Symptoms of Traumatic Brain Injury Confound Depression Screening? *Archives of Physical Medicine and Rehabilitation, 92*(5), 818-823. doi: 10.1016/j.apmr.2010.12.008

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: Difdetect and difwithpar. *Medical Care, 44*, S115-S123. doi: 10.1097/01.mlr.0000245183.28384.ed

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., … & Teresi, J. A. (2007b). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research, 16*, 69-84. doi: 10.1007/s11136-007-9185-5

Crane, P. K., Van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256. doi: 10.1002/sim.1713

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi: 10.1007/BF02310555

DeMars, C. E. (2010). Type 1 error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*, 961-972. doi: 10.1177/0013164410 366691

DeMars, C. (2014). An illustration of the effects of ignoring a secondary factor. *Applied Psychological Measurement*, *38*, 406-409. doi: 10.1177/0146621614529360

Edwards, M. C. & Cai, L. (2008). *A new diagnostic procedure to detect departures from local independence in item response models.* Paper presented at the Quantitative Forum at the L. L. Thurstone Psychometric Laboratory, Chapel Hill, NC.

Egberink, I. J. L., Meijer, R. R., & Tendeiro, J. N. (2015). Investigating measurement invariance in computer-based personality testing: The impact of using anchor items on effect size indices. *Educational and Psychological Measurement, 75,* 126-145. doi: doi:10.1177/ 0013164414520965

Fieo, R., Ocepek-Welikson, K., Kleinman, M., Eimicke, J., Crane, P. K., Cella, D., & Teresi, J. A. (in press). Measurement Equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS®) Applied Cognition-General Concerns short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling.*

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and the IRT likelihood ratio test. *Applied Psychological Measurement, 29*, 278-295. doi: 10.1177/0146621605275728

Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. Illinois Institute of Technology. *Dissertation Abstracts International, 54*(04B), 2266.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309-32. doi:10.1177/01466219922031437

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*, 432-457. doi: 10.1111/j.1745-3984.2010.00122.x

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393. doi: 10.1177/0013164406294781

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability of Likert-type and ordinal response data: A conceptual, empirical and practical guide. *Practical Assessment, Research and Evaluation, 17,* 1-13, ISSN 1531-7714.

Gelin, M. N. & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63*(1), 65-74. doi: 10.1177/0013164402239317

Glas, C. A. W. & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106. doi: 10.1177/0146621602250530

Glas, C. A. W. & Verhelst, N. (1995). Testing the Rasch model. In G. Fisher & I. Molennar (Eds.), Rasch Models: *Foundations, recent developments and applications* (pp 69-96). New York: Springer.

González-Betanzos, F. & Abad, F. J. (2012). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology:European Journal of Research Methods for the Behavioral and Social Sciences, 8*, 130-145. doi: 10.1027/1614-2241/a000046

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44* (11, Suppl. 3), S78-S94. doi: 10.1097/01.mlr.0000245454.12228.8f

Hahn, E. A., Kallen, M. A., Jensen, R. E., Potosky, A. L., Moinpour, C., Ramirez, M., …, Teresi, J. A. (In press). Measuring social function in ethnically diverse cancer populations: Evaluation of measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS®) Ability to Participate in Social Roles and Activities short form. *Psychological Test and Assessment Modeling.*

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(Suppl. 11)*,* S182-S188. doi: 10.1097/01.mlr.0000245443.86671.c4

Hambleton, R. K. & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45*, 153-171. doi: 10.1023/A:1006941729637

Hambleton, R. K. & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, I,* 1-30.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications, Inc.

Herrel, F. E. (2009). Design; design package. R package version 2:3.0. Retrieved from http://CRAN R-project.org/package=Design.

Hidalgo, M. D., Gomez-Benito, J., & Zumbo, B. D. (2014). Binary logistic regression analysis for detecting differential item functioning: Effectiveness of R^2 and delta log odds ratio effect size measures. *Educational and Psychological Measurement, 74*, 927-949. doi: 10.1177/0013164414523618

Holland, P.W. & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, Inc.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185. doi: 10.1007/BF02289447

Houts, C. R. & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37*, 541-562. doi: 10.1177/0146621613491456

Houts, C. R. & Cai, L. (2013). FlexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Jensen, R. E., King-Kallimanis, B. L., Sexton, E., Reeve, B. B., Moinpour, C. M., Potosky, A.L. …, Teresi, J.A. (in press). Measurement properties of the PROMIS® Sleep Disturbance short form in a large, ethnically diverse cancer cohort. *Psychological Test and Assessment Modeling.*

Jensen, R. E., Moinpour, C. M., Keegen, T. H. M., Cress, R. D., Wu, X-C., Paddock, L. A., …, Potosky, A. L. (2016). The Measuring Your Health study: Use of a community-based cancer registry to recruit a large, diverse cohort for analyses of measurement equivalence. *Psychological Test and Assessment Modeling, 58.*

Jodoin, M. G. & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349. doi: 10.1207/S15324818AME1404_2

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions for the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care, 44*(11, Suppl 3), S124-S133. doi: 10.1097/01.mlr.0000245250.50114.0f

Jones, R. N. & Gallo, J. J. (2002). Education and sex differences in the Mini-Mental Status Examination: Effects of differential item functioning. *Journal of Gerontology: Psychological Sciences, 57B,* P548-P558. doi: 10.1093/geronb/57.6.P548

Jones, R. N., Tommet, D., Ramirez, M., Jensen, R. E., & Teresi, J. A. (in press). Differential *item functioning in Patient Reported Outcomes Measurement Information System* (PROMIS®) Physical Functioning short forms: Analyses across ethnically diverse groups. *Psychological Test and Assessment Modeling.*

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 408-426. doi: 10.1007/BF02291366

Jöreskog, K. & Goldberger, A. (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 10*, 631-639. doi: 10.2307/2285946

Jöreskog, K. G. & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36* (3), 347-387. doi:10.1207/S15327906347-387

Jöreskog, K. & Sorbom, D. (1996). LISREL8: Analysis of linear structural relationships: Users Reference Guide. Scientific Software International, Inc.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56*, 255-278. doi: 10.1007/BF02294462

Kim, S. H. & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355. doi: 10.1177/014662169802200403

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44(2)* 93-116. doi: 10.1111/j.1745-3984.2007.00029.x

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. *Psychology and Aging, 17*(3), 379-391 doi:10.1037/0882-7974.17.3.379

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58.*

Kopf, J., Zeileis, A., & Stobl, C. (2015a). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement, 39*, 83-103. doi: 10.1177/0146621614544195

Kopf, J., Zeileis, A., & Stobl, C. (2015b). Anchor selection strategies for DIF analysis: Review, assessment and new approaches. *Educational and Psychological Measurement, 75*, 22-56. doi: 10.1177/0013164414529792

Langer, M. M. (2008). *A re-examination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation). University of North Carolina at Chapel Hill library, http://search.lib.unc.edu/search?R=UNC b5878458.

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*, 847-861. doi: 10.1177/0013164411432333

Liu, Y. & Maydeu-Olivares, A. (2012). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement, 73*, 254-274. doi: 10.1177/0013164412453841

Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology, 31*, 213-219. doi: 10.1177/0022022100031002004

Liu, Y. & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology, 67*, 496-513. doi: 10.1111/bmsp.12030

Liu, Y. & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement, 36*, 670-688. doi: 10.1177/0146621612458174

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251-265. doi: 10.1177/0146621608 321760

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

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138. doi: 10.1002/j.2333-8504.1977.tb01128.x

Lord, F. M. & Novick, M. R. (with contributions by A. Birnbaum) (1968). *Statistical theories of mental test scores.* Reading Massachusetts: Addison-Wesley Publishing Company, Inc.

Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: The R package Rm for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20. doi: 10.18637/ jss.v020.i09

Malida, M., Van de Vijver, F. J. R., Srinivasan, K., Tranlser, C., Sukumar, P., & Rao, K. (2008). Adapting a cognitive test for different culture: An illustration of qualitative procedures. *Psychology Science Quarterly, 50*, 453-468. doi: 10.1177/1073191109341445

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*, 357-367. doi: 10.1177/014662169802200404

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*, 99-114. doi: 10.1177/01466210022031552

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

McHorney, C. A. & Fleishman, J. A. (2006). Assessing and understanding measurement equivalence in health outcome measures. *Medical Care, 44*(11), S205-S210. doi: 10.1097/01.mlr.0000245451.67862.57

Meade, A. W. & Lautenschlager, G. J. (2004). A Comparison of IRT and CFA Methodologies for Establishing Measurement Equivalence. *Organizational Research Methods, 7*, 361-388. doi: 10.1177/1094428104268027

Meade, A., Lautenschlager, G., & Johnson, E. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430-455. doi: 10.1177/0146621606297316

Meade, A. W. & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*, 1016-1031. doi: 10.1037/a0027934

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 302-307. doi: 10.1037/0033-2909.115.2.300

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of *Educational Research, 13*, 127-143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. doi: 10.1007/BF02294825

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177-185. doi: 10.1007/BF02289699

Meredith, W. & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44, Suppl 3, S69-S77. doi: 10.1097/01.mlr.0000245438.73837.89

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement,* 17, 297-334, doi: 10.1177/014662169301700401

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195. doi: 10.1007/BF02293979

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*, 92-109. doi: 10.3102/1076998606298035

Morales, L. S., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the Differential Item and Test Functioning (DFIT) framework. *Medical Care, 44*(11, Suppl. 3), S143-S151. doi: 10.1097/01.mlr.0000245141.70946.29

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132. doi: 10.1007/BF02294210

Muthén, B. (1989a). Latent variable modeling in heterogeneous populations. Meetings of Psychometric Society (1989, Los Angeles, California and Leuven, Belgium). *Psychometrika, 54*(4), 557-585.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81-117.

Muthén, B. & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (p 16). Los Angeles: University of California.

Muthén, B. O., du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Los Angeles: UCLA.

Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133-142. doi: 10.3102/10769986010002133

Muthén, L. K. & Muthén, B. O. (1998-2013). M-PLUS Users Guide. Sixth Edition. Los Angeles, California: Authors Muthén and Muthén.

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298. doi: 10.1177/0146621603027004004

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. *Medical Care, 44*, S134-S142. doi: 10.1097/01.mlr.0000245251.83359.8c

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2011). DFIT8 for Window User's Manual: Differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance of the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17. doi: 10.1111/j.1745-3984.2006.00001.x

Paz, S. H., Spritzer, K. L., Morales, L., & Hays, R. D. (2013). Evaluation of the Patient-Reported outcomes Information System (PROMIS) Spanish-language physical functioning items. *Quality of Life Research, 22*, 1819-1830. doi: 10.1007/s11136-012-0292-6

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37. doi: 10.1177/014662169501900104

R Development Core Team (2008). *R: A language and environment for statistical computing.* Vienna, Austria (ISBN 3-900051-07-0).

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502. doi: 10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207. doi:10.1177/014662169001400208

Raju, N. S. (1999). DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*, 133-147. doi: 10.1177/0146621608319514

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-528. doi: 10.1037//0021-9010.87.3.517

Raju, N. S., Oshima, T. C., & Wolach, A. (1995a). Differential functioning of items and tests (DFIT): dichotomous and polytomous (Computer Program). Chicago: Illinois Institute of Technology.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995b). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368. doi: 10.1177/014662169501900405.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmarks Paedagogiske Institut (Danish Institute of Educational Research).

Rasch, G. (1980; original work published in 1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reardon, S. F. & Raudenbush, S. W. (2006). A partial independence item response model for surveys with filter questions. *Sociological Methodology, 36*, 257-300. doi: 10.1111/j. 1467-9531.2006.00181.x

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., …, Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl 1), S22-S31. doi: 10.1097/01.mlr.0000250483.85507.04

Reeve, B. B, Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., … & Chen, W-H. (2016). Psychometric evaluation of the PROMIS Fatigue Measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling, 58.*

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566. doi: 10.1037/0033-2909.114.3.552

Revelle, W. (2015). Psych: package Psych. Retrieved from http://CRAN.R-project.org/package=PSYCH

Revelle, W. & Zinbarg, R. E. (2009). Coefficient Alpha, Beta, Omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74*, 145-154 doi: 10.1007/s11336-008-9102-z

Rizopoulus, D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*, 1-25. doi: 10.18637/jss.v017.i05

Rizopoulus, D. (2009). Ltm: Latent Trait Models under IRT. http://cran.r project.org/web/packages/ltm/index.html.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114. doi: 10.1007/BF02290599

Setodji, C. M., Reise, S. P., Morales, L. S. Fongwam, N., & Hays, R. D. (2011). Differential Item Functioning by Survey Language Among Older Hispanics Enrolled in Medicare Managed Care A New Method for Anchor Item Selection. *Medical Care, 49*, 461-468. doi: 10.1097/MLR.0b013e318207edb5

Seybert, J. & Stark, S. (2012). Iterative linking with the differential functioning of items and tests (DFIT) Method: Comparison of testwide and item parameter replication (IPR) critical values. doi: 10.1177/0146621612445182

Shih, C. -L. & Wang, W. -C. (2009). Differential item functioning detection using multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199. doi: 10.1177/0146621608321758

Shih, C. -L., Liu, T.-H., & Wang, W. -C. (2014). Controlling type 1 error rates in assessing DIF for logistic regression method with SIBTEST regression correction procedure and DIF-free-then-DIF strategy. *Educational and Psychological Measurement, 74*, 1018-1048. doi: 10.1177/0013164413520545

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120. doi: 10.1007/s11336-008-9101-0

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247. doi: 10.1111/j.1745-3984.1991.tb00356.x

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306. doi: 10.1037/0021-9010.91.6.1292

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508. doi: 10.1037/0021-9010.89.3.497

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*, 332-342. doi: 10.1037/0022-3514.81.2.332

Steinberg, L. & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415. doi: 10.1007/s11136-011-9969-5

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408. doi: 10.1007/BF02294363

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. *Medical Care, 44*(Suppl. 11), S152-S170. doi: 10.1097/01.mlr.0000245142.74628.ab

Teresi, J. A., Golden, R. R., Cross, P., Gurland, B., Kleinman, M., & Wilder, D. (1995). Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology, 48*, 473-483. doi: 10.1016/0895-4356(94)00159-N

Teresi, J. A. & Jones, R. N. (2013). Bias in psychological assessment and other measures. In K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology: Vol 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology* (p. 139-164). American Psychological Association: Washington, DC. doi: 10.1037/14047-008.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651-1683. doi: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H

Teresi, J. A., Ocepek-Welikson, K., Cook, K. F., Kleinman, M., Ramirez, M., Reid, M. C., & Siu, A. (in press). Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) Pain short-forms in Ethnically Diverse Cancer and Palliative Care Populations. *Psychological Test and Assessment Modeling.*

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K, Gibbons, L. E., …, & Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measure of physical functioning ability and general distress. *Quality Life Research, 16*, 43-68. doi: 10.1007/s11136-007-9186-4

Teresi, J., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. E., Crane, P. K., Jones, R. N., …, & Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly. 51*:148-180. NIHMSID#136951

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016, a). Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System (PROMIS®) Depression short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58.*

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016, b). Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS®) Anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58.*

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on Differential Item Functioning (DIF) impact analyses. *Journal of Aging & Health, 24*(6), 1044-1076. Available online first, March 15, 2012. doi: 10.1177/089826412436877

Teresi, J. A., Ramirez, M., Lai, J-S., & Silver, S. (2008). Occurrences and sources of differential item functioning (DIF) in patient reported outcomes measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly, 50*, 538-612.

Thissen, D. (2001). IRTLRDIF (Version v.2.0b): University of North Carolina at Chapel Hill: L.L. Thurstone Psychometric Laboratory.

Thissen, D. (1991). MULTILOG^{TM} User's Guide Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago: Scientific Software, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In: P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. Braun (Eds.) *Test Validity* (pp. 147-169). Hillsdale, New Jersey, Lawrence Erlbaum, Associates.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false discovery rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77-83. doi: 10.3102/10769986027001077

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory dependence: A preliminary report (Research Memorandum 92-2).* Chapel Hill: Thurstone Laboratory, University of North Carolina at Chapel Hill.

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10. doi: 10.1007/BF02291170

Van de Vijver, F. J. R. & Leung, K. (1997). Methods and data analysis for cross-cultural research. Thousand Oaks, CA: Sage Publications.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. doi: 10.1177/109442810031002

Wainer, H. (1993). Model-based standardization measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.). *Differential Item Functioning* (pp. 123-135). Hillsdale NJ: Lawrence Erlbaum, Inc.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. Journal of educational measurement, 28(3), 197-219. doi: 10.1111/j.1745-3984.1991.tb00354.x

Wang, W. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261. doi: 10.3200/JEXE.72.3.221-261

Wang, W. C. & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498. doi: 10.1177/0146621603259902

Wang, W-C. & Shih, C-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*, 166-180. doi: 10.1177/0146621609355279

Wang, W-C., Shih, C-L., & Sun, G-W. (2012). The DIF-free-then DIF strategy for the assessment of differential item functioning (DIF). *Educational and Psychological Measurement, 72*, 687-708. doi: 10.1177/0013164411426157

Wang, W-C., Shih, C-L, & Yang, C-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731. doi: 10.1177/0013164409332228

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33,* 42-57. doi:10.1177/0146621607314044

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison of two group analysis. *Multivariate Behavioral Research, 44*, 1-27. doi: 10.1080/00273170802620121

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH tests and IRTLRDIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement, 35*, 145-164. doi: 10.1177/0146621610377450

Woods, C. M. & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement,* 35, 339-36. doi: 10.1177/0146621611405984

Woods, C. M. & Harpole, J. (2015). How item residual heterogeneity affects tests for differential item functioning. *Applied Psychological Measurement, 39*, 251-263. doi: 10.1177/0146621614561313

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*, 532-547. doi: 10.1177/0013164412464875

Yang, F. M., Cazorla-Lancaster, Y., & Jones, R. N. (2008). Within group differences in depression among older Hispanics living in the United States. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 63*(1), 27-32. doi: 10.1093/geronb/63.1.P27

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi: 10.1177/014662168400800201

Yang, F. M., Heslin, K. C., Mehta, K. M., Yang, C. W., Jones, R. N., Ocepek-Welikson, K., …, & Teresi, J. (2011). A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos. *Psychological Test and Assessment Modeling, 53*(4), 440-460.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurment, 39*, 291-309. doi: 10.1111/j.1745-3984.2002.tb01144.x

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. doi: 10.1007/s11336-003-0974-7

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.

Zumbo, B. D. & Thomas, D. R. (1997). *A measure of effect size of a model-based approach for studying DIF.* The Edgeworth Series in Quantitative Behavioural Science. Working Paper. Edgeworth Laboratory for Quantitative Behavioral Science. University of British Columbia. Prince George, BC.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.