PSYCHOMETRIKA—VOL. 86, NO. 3, 674–711
SEPTEMBER 2021
https://doi.org/10.1007/s11336-021-09775-0

DIFFERENTIAL ITEM FUNCTIONING ANALYSES OF THE PATIENT-REPORTED OUTCOMES MEASUREMENT INFORMATION SYSTEM (PROMIS®) MEASURES: METHODS, CHALLENGES, ADVANCES, AND FUTURE DIRECTIONS

JEANNE A. TERESI

COLUMBIA UNIVERSITY STROUD CENTER

HEBREW HOME AT RIVERDALE; RIVERSPRING HEALTH

WEILL CORNELL MEDICAL CENTER

NEW YORK STATE PSYCHIATRIC INSTITUTE

CHUN WANG

UNIVERSITY OF WASHINGTON COLLEGE OF EDUCATION

MARJORIE KLEINMAN

NEW YORK STATE PSYCHIATRIC INSTITUTE

RICHARD N. JONES

BROWN UNIVERSITY

DAVID J. WEISS

UNIVERSITY OF MINNESOTA

Several methods used to examine differential item functioning (DIF) in Patient-Reported Outcomes Measurement Information System (PROMIS®) measures are presented, including effect size estimation. A summary of factors that may affect DIF detection and challenges encountered in PROMIS DIF analyses, e.g., anchor item selection, is provided. An issue in PROMIS was the potential for inadequately modeled multidimensionality to result in false DIF detection. Section 1 is a presentation of the unidimensional models used by most PROMIS investigators for DIF detection, as well as their multidimensional expansions. Section 2 is an illustration that builds on previous unidimensional analyses of depression and anxiety short-forms to examine DIF detection using a multidimensional item response theory (MIRT) model. The Item Response Theory-Log-likelihood Ratio Test (IRT-LRT) method was used for a real data illustration with gender as the grouping variable. The IRT-LRT DIF detection method is a flexible approach to handle group differences in trait distributions, known as impact in the DIF literature, and was studied with both real data and in simulations to compare the performance of the IRT-LRT method within the unidimensional IRT (UIRT) and MIRT contexts. Additionally, different effect size measures were compared for the data presented in Section 2. A finding from the real data illustration was that using the IRT-LRT method within a MIRT context resulted in more flagged items as compared to using the IRT-LRT method within a UIRT context. The simulations provided some evidence that while unidimensional and multidimensional approaches were similar in terms of Type I error rates, power for DIF detection was greater for the multidimensional approach. Effect size measures presented in Section 1 and applied in Section 2 varied in terms of estimation methods, choice of density function, methods of equating, and anchor item selection. Despite these differences, there was considerable consistency in results, especially for the items showing the largest values. Future work is needed to examine DIF detection in the context of polytomous, multidimensional data. PROMIS standards included incorporation of effect size measures in determining salient DIF. Integrated methods for examining effect size measures in the context of IRT-based DIF detection procedures are still in early stages of development.

Correspondence should be made to Jeanne A. Teresi, Columbia University Stroud Center, New York, NY, USA. Email: teresimeas@aol.com; jat61@cumc.columbia.edu

Key words: PROMIS, measurement, differential item functioning, multidimensional IRT, effect size estimates.

Numerous articles on differential item functioning (DIF) have been published in *Psychometrika* recently (Chalmers, 2018; Chang et al., 2017; Liu et al., 2016; Strobl et al., 2015; Wang et al., 2018). The importance of examining DIF in cross-national surveys (Zwitzer et al., 2017) has been emphasized; however, few articles have described the measurement statistics work of the Patient-Reported Outcomes Measurement Information System (PROMIS®) international effort to standardize measurements for use in research and clinical assessment. One such article (Yu et al., 2018) presents an analysis of the PROMIS anxiety and depression short-forms using a large ethnically diverse cohort from the MyHealth survey (Jensen et al., 2016a). Depression and anxiety were two of the domains examined for DIF in the PROMIS item banks (Teresi et al., 2009; Choi et al., 2011) and short-forms (Teresi et al., 2016a, 2016b); however, only unidimensional approaches were used to examine DIF in these two domains, specifically, and across most PROMIS domains in general. An important issue discussed in detail in Section 1.3.1 is that inadequately modeled multidimensionality may lead to false DIF detection. Additionally, few studies have compared the performance of unidimensional and multidimensional DIF models in the context of two correlated traits. In this article, we present the results of a multidimensional approach to examining PROMIS depression and anxiety short-forms.

Discussed in this article are several factors that PROMIS investigators identified as important in the context of DIF detection. These include the need to: adjust for multiple comparisons, incorporate magnitude effect size measures, select DIF-free anchor items carefully, examine differences in group trait distributions, and check unidimensionality assumptions. PROMIS investigators also promoted the practice of generating DIF hypotheses by content experts to guide interpretation of findings. This topic is not included in this review, but is referenced in the example in Section 2. This article is organized as follows. In Section 1, a basic presentation of several unidimensional methods used to examine DIF in PROMIS measures is given, followed by a presentation of their multidimensional expansions. Methods for estimation of effect size are also presented. A summary of factors that may affect DIF detection, and challenges encountered in PROMIS DIF analyses are provided. Advances made in PROMIS DIF detection methodology and future directions are also discussed. Section 2 of the article builds on previous unidimensional analyses of depression and anxiety short-forms by PROMIS investigators (Teresi et al., 2016a, 2016b) to present an illustration of DIF detection using a multidimensional model. Effect size estimates are presented and compared. A simulation study is also included to compare the performance of a specific DIF approach described below in the context of unidimensional versus multidimensional item response theory (IRT) models.

Although early PROMIS studies of DIF included a number of non-parametric as well as parametric methods such as those reviewed in Millsap and Everson (1993), this presentation will focus on IRT-based or related approaches, including a brief discussion of confirmatory factor analyses (CFA) approaches to perform hierarchical tests of measurement invariance. Not included will be prediction invariance, nor the relationship between measurement and prediction invariance as presented recently in this journal (Culpepper et al., 2019). An important point is that while formulas are presented to orient the reader to the topic, a detailed explication is not provided in this review; rather, the reader is referred to the original work for details. It is emphasized that the motivation for this article is to present in Section 1 a broad review for the general reader of DIF detection methods used by PROMIS investigators, with an illustration of newer multidimensional approaches given in Section 2.

1. Overview of DIF Detection Methods

1.1. General Definition of Measurement Invariance

This Section provides a conceptual orientation to examining measurement invariance, including differential item functioning. The relationship between factor analysis and item response theory approaches is discussed in Section 1.3.4.2. A general formulation of DIF in the context of measurement invariance (Mellenbergh, 1989; Meredith, 1964, 1993; Meredith & Teresi, 2006) is that the conditional distribution of observed score, *X*, given the latent trait (θ) is independent of the group *G*, and can be expressed as:

$$f(X|\theta, G) = f(X|\theta), \tag{1-1}$$

which for polytomous items modeled with a graded response model can be formulated as:

$$P(X_{\mathrm{i}} = k|\theta, G) = P(X_{\mathrm{i}} = k|\theta) \tag{1-2}$$

for an item with K categories (see Kim & Yoon, 2011; Chang et al., 2017).

DIF is observed when the probability of item response differs across comparison groups such as gender, language or race/ethnicity, after conditioning on level of the state or trait measured, such as depression or anxiety. When the probability of response is consistently higher (or lower) for one of the comparison groups across all levels of the trait, uniform DIF is observed; in contrast, DIF is non-uniform when the probability of response is in a different direction for groups at different levels of the state or trait. Formal definitions are presented below.

Most analyses of PROMIS data have relied on unidimensional latent variable models, using item response theory (Hambleton et al., 1991; Lord, 1980; Lord & Novick, 1968; Rasch, 1960), specifically the log-likelihood ratio test (Orlando-Edelen et al., 2006; Thissen et al., 1988, 1993) or Wald tests based on Lord's chi-square (Lord, 1980; Teresi et al., 2000; Woods et al., 2013). Other main methods used were ordinal logistic regression (OLR; Zumbo, 1999) using latent variable models (Choi et al., 2011; Crane et al., 2004), and multiple indicators, multiple causes (MIMIC; Jöreskog & Goldberger, 1975; Jones, 2006; Muthén, 1984). Another approach used was structural equation models (SEM; Jöreskog & Goldberger, 1975; Jöreskog & Sorbom, 1996), specifically multiple group confirmatory factor analyses (MGCFA; Jöreskog, 1971; Meredith, 1964).

PROMIS guidelines and standards recommended for DIF assessment have been provided (Reeve et al., 2007; http://www.nihpromis.org/science/publications), and their use was illustrated by Carle et al. (2011). These methods were used in a two-part series summarizing findings of DIF in PROMIS short-forms (Reeve & Teresi, 2016; Teresi & Reeve, 2016). The guidelines were developed to promote the best practices for DIF detection, anchor item selection, effect size estimation and tests of dimensionality, discussed below. A detailed discussion of these guidelines is beyond the scope of this paper. Most of the methods used incorporated the basic guidelines, which in brief included DIF hypothesis generation (not discussed in this manuscript), tests of model assumptions, iterative anchor item selection, application of adjustments for multiple comparisons, use of a second method in sensitivity analyses, and use of magnitude (effect size) measures at the item and scale level rather than reliance only on significance tests.

1.2. Unidimensional Approaches to DIF Detection Applied by PROMIS Investigators

This Section describes unidimensional IRT-based approaches to DIF detection used by PROMIS investigators, statisticians, or affiliates.

1.2.1. IRT-based DIF Tests PROMIS items were usually polytomous with ordered categories; thus, the graded response model (GRM; Samejima, 1969) was used as the basis for DIF detection for three approaches applied: the log likelihood ratio test, the Wald test, and latent variable ordinal logistic regression. Given ordered responses, $x = k$ and $k = 1, 2, \ldots m$, a_i is the discrimination for item i and b_{ik} the difficulty parameters for response category k :

$$P(x = k) = P^*(k) - P^*(k + 1) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{ik+1})}} \tag{1-3}$$

An equivalent formulation for Eq. (1-3) in slope intercept form was used in the analyses described in Section 2:

$$P(x = k) = P^*(k) - P^*(k + 1) = \frac{1}{1 + e^{-\mathrm{D}(a_i\theta + d_k)}} - \frac{1}{1 + e^{-\mathrm{D}(a_i\theta + d_{k+1})}} \tag{1-3b}$$

In Eq. 1-3b, d_k is an intercept parameter, and D is a scaling constant (Chalmers, 2012). In both Eqs. (1-3) and (1-3b), $P^*(k)$ is the item response function describing the probability that a response is in category k or higher, for each value of the latent trait, θ (see Orlando-Edelen et al., 2006; Thissen, 1991). There are $k - 1$ boundary response functions describing the cumulative probability of responding in category k or higher. For the slope-intercept form, the probability of responding in category k is the difference in probabilities of responding in category k or higher and $k + 1$ or higher. Given m response categories for item i, multiple d_{ik} parameters from $k = 1$ to $m - 1$ are estimated, and the DIF effects are estimated for a_i and $d_\mathbf{i}$, where d is related to item location (see also Eqs. 1-8 and 2-1).

The item response theory likelihood ratio (IRT-LRT) method tests a series of IRT models established by fixing and freeing parameters. A typical IRT-LRT approach begins with an omnibus test of both the a and b parameters. 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. The Wald statistic, equivalent to Lord's Chi-square (Lord, 1980) and extended for polytomous data by Cohen et al. (1993), is asymptotically equivalent to the likelihood ratio test (Thissen, 1991; Thissen et al., 1993). As summarized in Teresi et al. (2000), Lord (1980, p. 223) proposed a Chi-square statistic, the Wald test for DIF, in which vectors of IRT item parameters are compared.

$$\chi^2 = \underset{\sim i}{\mathbf{v}'} \sum_i^{-1} \mathbf{v}_{\boldsymbol{i}}. \tag{1-4}$$

The hypotheses is a simultaneous test 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 $\widehat{a}_{i1} - \widehat{a}_{i2}$. The extension to m_k categories for the graded response model produces a vector of item parameters for each group, e.g., for the studied group, this is: $\left\{\hat{a}_{k\mathrm{S}}, \hat{b}_{k1\mathrm{S}}, \ldots, \hat{b}_{k(m_k-1)\mathrm{S}}\right\}'$. More advanced estimation procedures for the covariance matrix (Cai, 2008) introduced by Langer (2008) and simultaneous equating procedures have been incorporated into IRT software used for DIF detection, including Flexible Multilevel Multidimensional Item Analysis and Test Scoring (FlexMIRT; Cai, 2013; Houts & Cai, 2013) and Item Response Theory for Patient-Reported Outcomes (IRTPRO; Cai et al., 2011).

Vectors of IRT parameters can be tested for DIF using two approaches to the Wald test. The Wald 1 method uses anchor items in DIF detection that may be pre-selected. Anchor items are presumed DIF-free items used to set the metric for group comparisons. Each test item is examined by freeing the item parameters for group comparisons, while fixing the remaining item parameters as group equivalent. The Wald 2 method does not select for anchor items; within a single model the scale is identified by fixing the reference group mean to 0 and the standard deviation to 1, and estimating the studied group mean and standard deviation. The second step of the Wald-2 method tests all items simultaneously, with mean and standard deviation fixed at values estimated in the first step. An advantage of both Wald tests over IRT-LRT is that there are fewer model comparisons that might inflate Type I error rates because DIF testing can be performed across multiple groups rather than two at a time as with IRT-LRT. The Wald procedure requires at most two model fittings, while IRT-LRT will require one more than the number of studied items due to the nested model comparison approach. Evidence (Woods et al., 2013) supports the use of Wald 1 over IRT-LRT and Wald 2 in terms of Type I error inflation. The final p values are adjusted using Benjamini–Hochberg (B–H; Benjamini & Hochberg, 1995; Thissen et al., 2002) methods. Additionally, magnitude (effect size) tests described in detail in Section 1.4 have been performed as separate steps. A variant of this methodology is illustrated in Section 2.

1.2.2. IRT-Ordinal Logistic Regression (IRT-OLR) The method used as a primary or secondary approach to DIF analyses in many studies of PROMIS measures and item banks was logistic regression (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.

$$\text{logit}\,[\mathrm{P}(Y \leq k)] = a_k + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) \tag{1-5}$$

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 (Zumbo, 1999). Three nested models are examined: (1): $\alpha + \beta_1 x_1$; (2): $\alpha + \beta_1 x_1 + \beta_2 x_2$; (3): $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_{1*} x_2)$, where x_1 is the trait variable, and x_2 the group or studied covariate. 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 by trait. The main effect of the group variable is tested for uniform DIF in the threshold parameters, and a significant interaction term $\beta_3 (x_{1*} x_2)$ is indicative of non-uniform DIF. Specific criteria are used to identify salient DIF using OLR, e.g., comparing the R^2 values between the second and first steps in order to measure the unique DIF effect (Gelin & Zumbo, 2003). Such effect size measures can reduce Type I error inflation (Hidalgo et al., 2014; Jodoin, & Gierl, 2001). One limitation of the logistic regression approach is that there is not direct modeling of the group differences in the latent trait.

IRT-OLR (Crane et al., 2004, 2006, 2007; Mukherjee et al., 2013) substitutes latent trait estimates from an IRT model for the observed score conditioning variable and incorporates effect sizes into the uniform DIF detection procedure. An iteratively purified IRT trait is estimated as the matching criterion. A program, lordif, was developed by PROMIS investigators (Choi et al., 2011) to perform the analyses. Ltm in R (Rizopoulus, 2006, 2009) is used to obtain IRT item parameter estimates for the GRM (Samejima, 1969), and the Design package is used for the OLR procedure (Herrel, 2009). Lordif software includes a number of DIF effect size measures: the change in Beta and pseudo-R^2 from models with and without DIF terms, as well as magnitude and impact indices based on IRT parameters (Kim et al., 2007), described in Section 1.4.

1.2.3. MIMIC Unidimensional Models

1.2.3.1 The MIMIC Model The MIMIC model is a variant of the factor analytic structural equation model (SEM) and, assuming all items load on a single underlying latent trait, is equivalent to a unidimensional IRT model, but with different parameterization (Muthén & Muthén, 1998–2019). A measure of DIF is the direct effect of a studied variable on the item response estimated from a model that includes the trait variable. As reviewed in Teresi and Jones (2013, p. 152, 2016), the measurement model (shown in Eq. 1-6) is expanded to include direct effects of background variables, and the SEM (Eq. 1-7) includes Γ (regressions of the underlying trait) and describes the effects of covariates (studied group) on the underlying trait (θ, referenced in the factor analyses literature as η).

$$\boldsymbol{y}^{*} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{K}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1-6}$$

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}. \tag{1-7}$$

Direct effects ($\boldsymbol{K}$) are estimated from a regression of individual items' latent response variables (y^{*}) on studied group covariates (x). A significant value for $\boldsymbol{K}$ in model 1-6 is indicative of an item difficulty shift for members of the group, $\boldsymbol{x}$ or uniform DIF. In the context of uniform DIF and small studied group sample sizes, MIMIC has evidenced superior performance in DIF detection compared with IRT-LRT methods (Woods, 2009b). Traditional MIMIC models assess only uniform DIF; however, Woods and Grimm (2011) introduced interaction terms to model non-uniform DIF; this work has been extended to multidimensional models (Lee et al., 2017). As expected, traditional MIMIC models without interaction terms to detect non-uniform DIF performed poorly in the presence of DIF in factor loadings (Kim et al., 2012).

Several factors that can affect DIF detection have been examined in the context of MIMIC. High Type I error rates have been observed (e.g., Finch, 2005; Kim et al., 2012; Wang et al., 2009); however, MIMIC with iterative scale purification for polytomous data resulted in less Type I error (Wang & Shih, 2010). Another factor resulting in Type I error inflation is the multiple comparisons associated with DIF testing. Kim et al. (2012) studied MIMIC models with categorical, polytomous, and continuous variables in terms of sample and group size, number of anchors, location, and magnitude of DIF. They used the Oort (1998) critical value adjustment to control Type I error beyond what was achieved with standard methods for multiple comparison adjustment. Power remained adequate, and the reduction in Type I error inflation in MIMIC likelihood ratio tests, and particularly with a contaminated (with DIF) baseline model, was achieved with use of the Oort adjustment.

PROMIS investigators (Jensen et al., 2016b; Jones, 2006; Jones et al., 2016) applied unidimensional MIMIC models that did not include recent advances in modeling non-uniform DIF using interaction terms (Woods & Grimm, 2011) and/or a multidimensional framework (Lee et al., 2017). The latter model assumes that multidimensional constructs were intended, as contrasted with a bifactor model in which the auxiliary factor is considered an unintended nuisance variable to be modeled. The bifactor model approach has been used in many PROMIS DIF studies to provide evidence in support of essential unidimensionality. However, because multidimensionality can masquerade as DIF, it is possible that Type I error inflation (excess false positive DIF) could result from not modeling multidimensionality in DIF analyses adequately. Methods have been proposed that theoretically could accommodate multidimensionality using MIMIC in the context of understanding DIF mechanisms. These include the use of mediation (Cheng et al., 2016, 2020) and moderated mediation (Montoya & Jeon, 2020). (See also Jones, 2019 for a discussion of DIF in the context of effect modification.) The applications, however, have focused on binary data, with limited analyses of multidimensional data. Presented in Section 1.3.3 are recent advances in multidimensional MIMIC models.

1.3. Multidimensional Models

In this Section, multidimensionality and DIF is discussed, followed by a presentation of multidimensional latent variable models used in DIF detection, including IRT, MIMIC, and multiple group CFA.

1.3.1. Multidimensionality and DIF An argument has been advanced that multidimensionality is a cause of DIF; however, others have argued that not all DIF is due to multidimensionality, but could be due to the complexity of loadings across groups (McDonald, 2000), relative bias (Boorsboom et al., 2002) or translation, such that the item performs well within groups and is related well to the construct measured, but shows DIF due to factors such as poor translation (Boorsboom, 2006). Another formulation is that other dimensions represent intentional or unintentional traits. Unintentional dimensions are considered as nuisance dimensions (Shealy & Stout, 1993), which produce adverse DIF. An unintended trait results in members of one group experiencing "systematic disadvantage" (Furlow et al., 2009). Benign DIF is considered to exist if the multidimensionality is intentional with an auxiliary trait measured.

IRT-type models used in educational testing, psychological assessments, and in PROMIS health and mental health assessments typically assume unidimensionality. Because a major concern of PROMIS investigators was that violations of the unidimensionality assumptions of IRT models used could lead to Type I error inflation and false DIF detection (Ackerman, 1992; Mazor et al., 1998), they focused on methods for detecting whether essential unidimensionality existed. The notion of essential unidimensionality, in which one dominant domain exists together with minor unintended dimensions, was introduced by Stout (1987, 1990). Tests of essential unidimensionality have been developed to evaluate if minor dimensionality is ignorable. Numerous reviews of methods to assess dimensionality exist [e.g., Junker (1991), Stout (1987), Reise (2012)], and this topic is not presented here. PROMIS investigators, e.g., Reise (2012), promoted use of the bifactor model to inform dimensionality assessments in analyses of PROMIS item banks and short-forms. Few analyses of PROMIS measures incorporated the possibility of multidimensionality. However, some investigators used models that could theoretically be applied to multidimensional data, including MIMIC (Muthén, 1984) and confirmatory factor analyses (e.g., McDonald, 2000; Meredith, 1964), presented below. Early methods to accommodate multidimensionality in DIF testing included MULTI-Simultaneous Item Bias (MULTI-SIBTEST; Stout et al., 1997), an extension of SIBTEST (Shealy & Stout, 1993); as mentioned above, these nonparametric methods were not used widely in PROMIS and are not discussed here.

1.3.2. Multidimensional IRT Extensions of unidimensional DIF methods to multidimensional approaches have been advanced in the context of logistic regression and item response theory (Kahraman et al., 2009). In early formulations, the logistic regression model was used with an observed score with Wald tests of parameters in the context of differences in the log likelihood functions of full and reduced models. These authors extended this model to a multidimensional IRT model for binary data incorporating the latent trait θ instead of the observed score using the compensatory two-parameter logistic (2PL) multidimensional model (MIRT; Reckase & McKinley, 1991). The probability that person j responds affirmatively to item i is:

$$P\left(y_{ij}=1|\theta_j\right)=\frac{\exp\left(\sum_{k=1}^{K}a_{ik}\theta_{jk}-d_i\right)}{1+\exp\left(\sum_{k=1}^{K}a_{ik}\theta_{jk}-d_i\right)} \tag{1-8}$$

where y_{ij} is the score on item i ($i = 1,\ldots,$ I) by person j($j = 1,\ldots,$ J), $\boldsymbol{\theta}_j$ is a vector of trait parameters for person j on K dimensions (k $= 1,\ldots,$ K), $\theta_j = (\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}, \ldots, \theta_{jk})$, $\boldsymbol{a}_i$ is a vector of item discrimination parameters $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{ik})$, d_i is a scalar related to item location. The $\boldsymbol{\theta}$ are multivariate normal with mean 0 and variances 1, and with covariances as free parameters (see Kahraman et al., 2009, p. 156). A multidimensional model for polytomous data is presented in Section 2 of this paper. As stated earlier, the analyses of anxiety and depression short-forms in PROMIS applied a unidimensional graded response model; for the analyses in Section 2 of this manuscript a multidimensional IRT DIF model (e.g., Suh & Cho, 2014) was used. In the prior analyses, evidence supported essential unidimensionality; however, depression and anxiety could be considered as two correlated constructs that are multidimensional.

1.3.3. MIMIC Multidimensional Models

1.3.3.1 Multidimensional MIMIC Models MIMIC models permit items to load on multiple traits and can thus be used to model multidimensional data. A recent simulation study (Lee et al., 2017) examined an extended multidimensional MIMIC-interaction model in which an interaction term between the latent variable, e.g., depression and anxiety states, and a group variable, e.g., race/ethnicity or gender, is examined. The model can also be parameterized as a multidimensional IRT model. The model presented (equation 3, Lee et al., 2017) measures k latent traits:

$$y_i^* = \lambda_{1i}\theta_1 + \ldots + \lambda_{ki}\theta_k + \beta_i z + \omega_{1i}\theta_1 z + \ldots + \omega_{ki}\theta_k z + \varepsilon_i, \tag{1-9}$$

where the factor loadings ($\lambda_{1i}, , , , \lambda_{ki}$) link item i to the latent traits $\theta_1 \ldots .\theta_k$. The interaction terms (ω_{1i} through ω_{ki}) represent nonuniform DIF effects for item i, and z is the categorical studied group variable. An important feature of this model is that an effect size for uniform DIF can be estimated, and anchor items included. As reviewed below, anchor item selection methods may have an impact on the degree of accuracy of DIF detection. The finding of a simulation study by Lee et al. was that unlike that reported by Woods and Grimm (2011), elevated Type I error for the interaction model was not observed because Lee et al. used a procedure in MPlus to adjust for the violation of the normality assumptions in the interaction of a categorical group variable with a continuous latent variable. Additionally, these authors adjusted for multiple comparisons using the Benjamini–Hochberg (Benjamini & Hochberg, 1995) method. Lee et al. (2017) observed that the MIMIC interaction model was not always powerful for detection of non-uniform DIF. As would be expected, power for DIF detection was greater for larger sample sizes; however, longer scales mitigated to some extent the loss of power due to smaller sample sizes under conditions of non-uniform DIF. Of interest is that, like many other simulation studies, the more anchor items, the greater the power. The effect of purification was unknown as it was not examined, and results apply only to binary data.

Bulut and Suh (2017) compared MIMIC, multidimensional IRT-LRT, and logistic regression in the context of multidimensional data. IRT-LRT was more powerful and generally evidenced less Type 1 error inflation than the MIMIC interaction model and the logistic regression approach. Thus, the IRT-LRT approach was applied in the illustration in Section 2. Logistic regression in particular evidenced greater Type I error than the other approaches; however, the conditioning variable was an observed rather than a latent variable. In the PROMIS applications using lordif (Choi et al., 2011), a latent variable was used in the ordinal logistic regression model as the conditioning variable.

1.3.4. Multiple Group Confirmatory Factor Analysis (MGCFA)

1.3.4.1 Overview of the Approach MGCFA is an approach used by some PROMIS investigators. 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). Measurement invariance tests are typically based on evaluations of nested model Chi-square differences and changes in model fit indices. Jensen et al. (2016b) applied this approach to examination of the PROMIS sleep disturbance short-form. A measurement model can be estimated separately, but simultaneously. Model identification and measurement model calibration are achieved by imposing equality constraints on the measurement model parameters and variance parameters for the latent variable across groups. With categorical dependent variables and a least squares parameter estimation approach, model modification indices (Chi-square scaled derivatives from the model fit function) are generated for all constrained or fixed parameters, and provide an estimate of the expected change in model Chi-square if the parameter was freely estimated. Formal testing for DIF requires imposing implied constraints and testing improvement with robust Chi-square difference tests (Jones, 2006; Muthén, 1989). With categorical variables and a robust maximum likelihood parameter estimation approach, modification indices are not available, but a likelihood ratio test procedure can be used to similar effect (see Thissen, 2001). 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, 1989). A robust parameter estimation procedure, based on a mean and variance adjusted weighted least squares procedure (WLSMV; Muthén et al., 1997) with adjusted critical values (Oort, 1998; Kim & Yoon, 2011), can be implemented in Mplus (Muthén & Muthén, 1998–2019).

Different levels of equality constraints (subject to model identification) across these models constitute a hierarchy of factorial invariance presented in several reviews (Byrne et al., 1989; Cheung & Rensvold, 2003; Gregorich, 2006; Mellenbergh, 1989; Meredith, 1993; Meredith & Teresi, 2006; Vandenberg & Lance, 2000). Strong factorial invariance is assumed if groups have equivalent τ (threshold/difficulty) and λ (factor loading) values (see Meredith, 1993). 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.

1.3.4.2 Relationship Between IRT and Factor Analysis 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 et al., 2002; Reise et al., 1993; Takane & de Leeuw, 1987). DIF detection using these approaches has been compared (Kim & Yoon, 2011; Stark et al., 2006). 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). Many comparisons of MGCFA to IRT-based approaches used an ordinal linear regression (CFA) method, e.g., Raju et al. (2002), Stark et al. (2006); however, a more comparable method is to use an ordered categorical approach with thresholds, e.g., Kim and Yoon (2011).

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. A common default is to fix the first loading to 1 to set the metric, while permitting the variance of the factor to be estimated freely. 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. Parameterization of the measurement model to link to IRT parameters is discussed in Muthén and Asparouhov (2002).

Recently, Chang et al. (2017) performed a unification of the GRM and categorical CFA by discretizing the underlying normal item variable and setting the uniqueness to 1.0 across subgroups. They show that the models differ primarily in terms of the identifiability constraints. As is well known, the MGCFA approach permits uniqueness as well as thresholds and loadings to vary across groups, in contrast with the IRT assumption of homogeneous error variances (Woods & Harpole, 2015). Chang et al. (2017) compared GRM "with a usual one anchor item" method to MGCFA that identifies the model with one fixed loading. They found that the GRM-type parameterization was more powerful than MGCFA for DIF detection under conditions of heterogeneous DIF sizes across the latent variable continuum. GRM in practice is not usually applied with one anchor item because simulations have shown that DIF detection is improved in terms of power and Type I error inflation reduction when multiple anchor items were used together with purification (e.g., Finch, 2005; Shih & Yang; Wang et al., 2009; Woods, 2011). However, the analyses in Section 2 were conducted using one carefully selected anchor without DIF for each subscale examined, because in analyses of PROMIS short-forms, identification of multiple DIF-free anchor items was a challenge (see Reeve & Teresi, 2016; Teresi & Reeve, 2016). Anchor item selection is discussed in Section 1.5.2.

1.4. DIF Magnitude (Effect Size) Measures at the Item and Scale Level

This Section describes the state of the art in examining magnitude of DIF, also referred to as effect size estimation. Because a goal in PROMIS DIF analyses was to retain items if possible, given the considerable effort to design the item pools, and the limited number of items available, PROMIS investigators advanced the field of DIF assessment by recommending the inclusion of several magnitude (effect size) measures at the item and scale level. Significant DIF will be observed for most items if sample sizes are large (Boorsboom, 2006). Incorporation of magnitude measures such as R square change in the ordinal logistic regression DIF detection method, for example, can help to reduce flagging of non-salient DIF (Gomez-Benito et al., 2013). Thus, examination of DIF effect sizes is an integral part of DIF analyses (Kleinman & Teresi, 2016; Rouquette et al., 2019; Stark et al., 2004; Steinberg & Thissen, 2006; Teresi, 2006; Teresi et al., 2012).

1.4.1. DIF Item Level Magnitude IRT-based approaches to effect size estimation have included examination of differences in parameter estimates, e.g., Steinberg and Thissen (2006), or methods based on differences in the expected item and scale score functions (e.g., Wainer, 1993; Raju et al., 1995). Effect size measures based on the expected item score differences were proposed by Wainer (1993) and extended for polytomous data by Kim et al. (2007). The expected score is the sum of the weighted probabilities of scoring in each of the possible categories for a polytomous item, taking a graded response form. The boundary response function can be defined as follows, where a_i is the item discrimination and β_{ik} are location parameters (see Kim et al., 2007 for a detailed explication)

$$P_{ik}^{*}(\theta) = \{1 + \exp[-\alpha_i(\theta - \beta_{ik})]\}^{-1}. \tag{1-10}$$

For item i, with k categories, with y category values, the true (expected) score for item i can be expressed as:

$$T_i(\theta) = \sum_{k=1}^{K_i} y_{ik} P_{ik}(\theta). \tag{1-11}$$

A method used for quantification of the difference in the average expected (true) item scores is the non-compensatory DIF (NCDIF) index (Raju, 1999; Raju et al., 1995; Flowers et al., 1999; Oshima et al., 2006; Oshima et al., 2009; Raju et al., 2009). Additionally, PROMIS affiliates advanced methods for quantifying the difference in expected item scores, e.g., Woods (2011), which were programmed into PROMIS DIF software, lordif (Choi et al., 2011), as well as other PROMIS effect size software (Kleinman & Teresi, 2016). Graphic displays of group differences in expected item and scale score functions are available in IRTPRO (Cai et al., 2011). More recently, Orlando-Edelen presented an effect size measure based on her earlier work (Orlando-Edelen et al., 2015) at the 2019 PROMIS Psychometric Summit at a session on measuring DIF effect sizes (Teresi, 2019). Chalmers et al. (2016) and Chalmers (2018) extended the work of Wainer (1993), Raju (1988; 1990) and Raju et al. (1995) with more advanced estimation and a stronger statistical approach.

1.4.1.1 Equivalent Formulations

1.4.1.1.1 Wainer Standardized Impact Indices Four indices of DIF magnitude, labeled T(1) to T(4) were introduced by Wainer (1993); two variants used by PROMIS investigators for polytomous items are:

$$T\,(1) = \int_{-\infty}^{\infty} [T_R\,(\theta) - T_S\,(\theta)]\, dG_S\,(\theta) \tag{1-12}$$

$$T\,(3) = \int_{-\infty}^{\infty} [T_R\,(\theta) - T_S\,(\theta)]^2\, dG_S\,(\theta) \tag{1-13}$$

where $T_R\,(\theta)$ and $T_S\,(\theta)$ are the true score functions for the reference and studied comparison groups, respectively, and $G_S\,(\theta)$ is the studied group distribution (see Kim et al., 2007, p. 105).

1.4.1.1.2 NCDIF Raju et al. (1995) proposed an NCDIF statistic equivalent to Wainer's T(3) statistic. NCDIF for item i is defined as the average squared difference between the true or expected scores for an individual j as a member of the studied group (S) and as a member of the reference group (R). Two estimated scores are computed for each, one based on the subject's trait estimate and the estimated item parameters for the studied group and the other based on the trait estimate and the estimated item parameters for the reference group. Each subject's difference score is squared and summed for all subjects ($j = 1, N_S$) to obtain NCDIF. Similar to T(3), NCDIF is weighted by the actual distribution of θs in the studied group. PROMIS investigators

(Kleinman & Teresi, 2016; Teresi et al., 2007) used equivalent formulations to that of Wainer's (1993) T(3) index $\sum_{j=1}^{N^S} \left[T_S(\theta) - T_R(\theta)\right]^2 / N_S$ to calculate:

$$\mathrm{NCDIF}_i = \left[\sum_{j=1}^{N_S} \left(T_{\mathrm{ijS}} - T_{\mathrm{ijR}}\right)^2\right] \Big/ \mathrm{N_S}, \tag{1-14}$$

where N_S is the number of subjects in the studied group; T_{ijS} is the expected (true) score for subject j in studied group S; and T_{ijR} is the expected (true) score for subject j as if a member of reference group R. Choi et al. (2011) used an equivalent formula from Raju et al. (1995, equation 10) to compute NCDIF in lordif.

Different equating methods and densities are used in the calculations. For example, unlike methods described below, for NCDIF the estimates of the latent trait (θ) are calculated separately for each group and equated together with the item parameters using the Stocking and Lord (1983) procedures. Baker's (1995) EQUATE program has been used in an iterative fashion to equate the θ and item parameter estimates for the two groups and place them on a common metric. If DIF is detected, the item showing DIF is excluded from the equating algorithm, and new DIF-free equating constants are computed, and purified iteratively. Iterative purification of equating constants has been shown to reduce Type I error (Seybert & Stark, 2012). As shown above, NCDIF is calculated squaring differences between expected response functions; the three methods described below are not based on squared differences.

1.4.1.1.3 Average Unsigned Difference (AUD; Woods, 2011) Non-uniform DIF occurs when the probability of response is in a different direction for the reference and studied groups, at different levels of the latent ability, θ. Both the AUD and NCDIF measure the magnitude of both uniform and non-uniform DIF. However, instead of the squared difference used in NCDIF, the AUD is constructed by calculating the absolute value of the difference between the expected item response functions, weighted by the presumed normal focal (studied) group distribution. Instead of the actual studied group distribution, 81 quadrature points at 0.1 intervals from -4 to $+4$ are used in the calculations. (The illustration in Section 2.5 uses the actual estimates of θ for the studied group instead of assuming a normal distribution for the studied group.)

$$\mathrm{AUD} = \sum_{i=1}^{N_S} \left|\left[T_R(\theta) - T_S(\theta)\right]\right| / N_S. \tag{1-15}$$

The AUD is the same as Wainer's (1993) *T (1)* (if there is no crossing), except that it is the absolute value of the differences across subjects that is summed and divided by $\mathrm{N_S}$. When the AUD is close to the value of *T(1)*, this is an indication of uniform DIF, that is, the probability of response is consistently higher for either the reference or studied group across all levels of the latent ability (θ). It is helpful to report both $T(1)$ and AUD to investigate instances of non-uniform DIF, in which case, $T(1)$ and AUD could differ substantially. Wainer's *T(1)* can be negative. The AUD is included in PROMIS DIF software (e.g., Choi et al., 2011; Kleinman & Teresi, 2016); however, in this software AUD is weighted by the actual studied group density, as contrasted with the method proposed by Woods (2011), for which the presumed normal studied group density is used.

1.4.1.1.4 wABC (Orlando-Edelen et al., 2015) The wABC is the average of the area between the expected item score curves, weighted by the normal distribution. Because there is no equating between groups on θ, there are two different distributions. Simultaneous linking is performed, holding θ constant. The wABC is computed twice (once for the studied and once for the reference group), where $\phi^{(R)}$ denotes the reference group normal distribution with mean $= 0$ and SD $= 1$, and $\phi^{(S)}$ accounts for the mean shift in the normal distribution estimated from the generating IRT DIF model (see Orlando-Edelen et al., pp. 97–98).

$$\text{wABC}_{R_i} = \int_{\theta} \left| T_i^{(R)}(\theta) - T_i^{(S)}(\theta) \right| \phi^{(R)}(\theta)\, d\theta, \tag{1-16}$$

$$\text{wABC}_{S_i} = \int_{\theta} \left| T_i^{(R)}(\theta) - T_i^{(S)}(\theta) \right| \phi^{(S)}(\theta)\, d\theta. \tag{1-17}$$

Estimation is achieved through calculating weights, at different quadrature points along the presumed normal θ distribution. A product of the difference in the absolute expected (true) scores for the studied and reference groups and the normal underlying distribution is computed. The weight, based on the normal distribution, is the proportion in the reference group in the 0.25 interval of θ. Then, the two group-specific wABC estimates are averaged based on the proportion of the sample in each group. The wABC is a non-compensatory statistic in that there is no DIF cancellation (differences in one direction favoring one group do not cancel those in the opposite direction). The wABC is similar to the AUD in PROMIS software; however, simultaneous linking is performed and the two presumed normal densities from -4 to $+4$ are used for weighting instead of the studied group distribution. Depending on the similarities of the distributions between the studied and reference groups, the wABC will equal the AUD. NCDIF (Raju et al., 1995) as calculated by PROMIS investigators (Kleinman & Teresi, 2016) used the actual distribution for the studied group, and did not assume a normal distribution. Woods (2011) also used the studied group distribution in the calculations of the AUD, but assumed a normal density.

1.4.1.1.5 Differential Response Function (DRF) Chalmers (2018) calculated compensatory and non-compensatory bias measures based on Wainer's (1993) formulation of standardized impact and Raju's et al. (1995) operationalization of these measures in DFIT.

The DRF approach, like the DFIT (Raju, 1999; Wainer, 1993) standardized effect size method, examines group differences in expected response functions, averaged and weighted; however, the marginal density $[f(\theta)]$ is used rather than the studied group density function.

Chalmers (2018, equation 4) defines non-compensatory response bias as

$$\beta_{NC} = \int \left| S\left(C|\boldsymbol{\Psi}^{(R)}, \theta\right) - S\left(C|\boldsymbol{\Psi}^{(S)}, \theta\right) \right| f(\theta)\, w(\theta)\, d\theta \tag{1-18}$$

where $S\left(C|\boldsymbol{\Psi}^{(R)}, \theta\right)$ and $S\left(C|\boldsymbol{\Psi}^{(S)}, \theta\right)$ represents the scoring function for the reference and studied group, respectively, $f(\theta)$ represents the marginal density, $w(\theta)$ is a weight function used if focusing on specific regions of θ, and $\boldsymbol{\Psi}$ is the vector of item parameters. C refers to a total scale score or a 'bundle' of items within a scale; thus, the formula for β_{NC} is the general form if one is examining the whole measure or some subset. It is made up of expected scores on the items which then each has the expected scoring function of $S(c)$, given the parameters for that item and level of θ. If one is examining only one item at a time, as is the case with most PROMIS applications, $S(c)$ is the scoring function.

The illustration of these magnitude measures in Table 8 in Section 2 of this paper presents a modified version of the Chalmers (2018) β_{NC} statistic. For comparison with the other statistics, the β_{NC} was calculated at the item level instead of the total scale or 'item bundle' level. Thus, for any given value of θ, the reference group $S(c)$ is the expected item level score for item i, given the estimated parameters for the reference group, and the studied group $S(c)$ is the expected item level score for the same item given the estimated parameters for the focal (studied) group. As recommended by Chalmers, the estimates in Table 8 are based on the combined density of the estimated θs for both groups to calculate β_{NC}, but instead of estimating the density function at Q quadrature points along the θ axis and assigning weights, the actual distribution of estimated θs for both reference and studied groups combined was used.

Chalmers' (2018) compensatory bias (β_C) statistic is related to Wainer's *T(1)*, while the non-compensatory bias statistic (given above) is related to Wainer's *T(3)*. Note that *T(3)* is the squared difference between groups, whereas β_{NC} is the absolute value (the average absolute difference) and is the average (weighted) score group difference between the response functions. While only binary data were examined, future work proposed by Chalmers (2018) includes simulations for polytomous and multidimensional data and the development of methods to obtain the asymptotic sampling distributions for these statistics.

1.4.1.2 Summary of Item Level Magnitude Measures The statistics used to estimate magnitude of DIF share similarities, but vary in terms of whether the index is squared or the absolute value taken, the weighting density, the linking methodology, and the estimation approach.

1.4.1.2.1 Indices Wainer's (1993) *T(3)* is the same as Raju's et al. (1995) NCDIF, which is the sum of the squared differences divided by the number of subjects. The AUD is different because it is the sum of the absolute values of the differences across all subjects of the group, and weighted by the presumed normal distribution of the studied group. It will be the same as *T(1)* if there is no crossing DIF so that all differences are in the same direction, either favoring the studied or the reference group. The absolute averaged differences in expected scores are calculated for the AUD, wABC, and β_{NC}; in contrast, NCDIF is calculated based on the average squared difference between an individual's estimate as members of the studied group versus the reference group.

1.4.1.2.2 Linking Raju et al. (1995) equated the parameters so that the two distributions are on the same scale; this procedure from Stocking and Lord (1983) was also used in two PROMIS-related software packages: lordif (Choi et al., 2011) and in the software developed by Kleinman and Teresi (2016). There may be more error with this equating procedure than with the use of simultaneous estimation and linking. Another factor affecting estimates relates to the selection and use of anchor items to set the metric. This topic is discussed briefly in Section 1.5 and in reference to the illustration in Section 2.5.

1.4.1.2.3 Estimation Orlando-Edelen and colleagues performed two estimation procedures to calculate wABC, one for the studied and one for the reference group, using an approximation of the integral and calculating the midpoint of small intervals at quadrature points of 0.25 θ, using a normal density from -4 to $+4$. This procedure was followed in the illustration in Section 2.5. Woods (2011) also assumed a normal distribution for the studied group, using weights at 81 quadrature points along the scale from θ =-4 to $+4$. (.10 intervals). Chalmers (2018) used quadrature points for estimation (61 quadrature nodes across the θ range -6 to $+6$).

1.4.1.2.4 Weighting Density The studied group distribution is used for weighting by Wainer (1993), Raju et al. (1995), and PROMIS investigators because it has often been observed to be non-normal, extending from −2 to +2 or less. In practice, extending calculations beyond ±3.5 results in weights close to 0 and thus has little effect on the estimated effect size. The AUD (Woods, 2011), in contrast, presumes a normal distribution for the studied group. This is similar to the procedures of Orlando-Edelen et al. (2015) in the calculation of wABC, except these authors used the presumed normal distribution for both the studied and reference groups. wABC is similar to the AUD estimated by PROMIS investigators; however, simultaneous linking is performed and the normal density used for weighting instead of the actual studied group distribution. β_{NC} is based on the overall (combined) marginal density. A normal density function is not assumed, and a weight function can be used if the investigator is interested in a particular area of the curve.

1.4.1.3 Cutoff Values Cutoff values for NCDIF were established based on simulations (Fleer, 1993; Flowers et al., 1999; Raju, 1999), and effect sizes estimated (see Meade et al., 2007). Simulations by Meade et al. 2007 resulted in the recommendation to use empirically derived DIF cutoff values. Choi et al. (2011) incorporated this approach in the PROMIS DIF detection software (lordif) used widely in PROMIS. Item parameter replication methods have been recommended to derive sample-specific estimates (Seybert & Stark, 2012) in the context of power for DIF detection. However, given the goal of flagging only items with large DIF magnitude, the practical meaning of group differences in expected scores reflected in NCDIF is also of consideration; thus, the higher threshold values were used by PROMIS investigators (see Kleinman & Teresi, 2016). Chalmers (2018) provides the sampling distribution for the non-compensatory differential response function and a method for calculating bootstrap confidence intervals for the estimated β_{NC} statistic. This is an area that requires more work.

1.4.2. Scale-level DIF Impact Raju et al. (1995) introduced the concept of differential functioning at the level of the scale. PROMIS investigators adopted the term scale level impact to describe this effect size; however, as noted earlier, in much of the literature on DIF, the term impact has been used to refer to group difference in the trait distributions. The expected scale scores are summed to obtain expected total scale scores, based on estimated parameters for the reference and studied groups, respectively. Stark et al. (2004) extended this work to assume a normal density for the studied group in the development of an effect size measure.

$$DTFR = \int \left[TCC_R(\theta) - TCC_S(\theta)\right] f_S(\theta)\, d\theta \tag{1-19}$$

where $f_S(\theta)$ is the trait density for the studied (focal) group, which is assumed to be normally distributed. The test characteristic curve [TCC$_s(\theta)$] represents a person's expected total test score based on the studied group parameter estimates of a's and b's for all items in the measure and TCC$_R(\theta)$ represents the same person's expected total test score based on the reference group a and b parameter estimates. A DTFR value represents measurement bias in terms of the raw score point difference. A positive value for DTFR indicates bias against the studied group, and a negative value indicates a bias against the reference group.

Recent improvements in the calculation of the DTF statistics have been advanced (Chalmers et al., 2016); Stark et al. (2004), the assumed normal studied group density is not used in their calculations. Two statistics, signed DTF (sDTF) and unsigned DTF (uDTF), were recommended that are similar to Raju's DTF statistics, except that weights at selected quadrature points corresponding to levels of θ (assumed normal) are used instead of the actual distribution of estimated θs in the studied group to calculate the estimated total test scores.

The sDTF measure from Chalmers et al. (2016, equation 5) is

$$sDTF = \int \left[T\left(\theta, \psi_R\right) - T\left(\theta, \psi_S\right)\right] g\left(\theta\right) d\theta, \tag{1-20}$$

where g (θ) is a weighting function with numerical evaluation at discrete quadrature points.

For each value of θ, the two estimated total scale scores are calculated and the studied group estimated total scale score is subtracted from that of the reference group and multiplied by the weight associated with that level of θ. These values are added over the quadrature points (from θ of −6 to +6). Unlike Raju's DTF, differences are not squared so the result is positive if the scale favors the reference group and negative if it favors the studied group. Based on this formulation, the item-level signed DIF (sDIF) is given by Chalmers (2016, equation 2.18).

$$sDIF = \int \left(S\left(c|\theta, \psi_R\right) - S\left(c|\theta, \psi_S\right)\right) g\left(\theta\right) d\theta = \int \left(T\left(C|\theta, \Psi_R\right) - T\left(C|\theta, \Psi_S\right)\right) g\left(\theta\right) d\theta, \tag{1-21}$$

Item-level signed and unsigned DIF statistics (Chalmers, 2016) are included in the "mirt" software (Chalmers, 2012). Note that this is similar to the formula for the item-level bias statistic given above. A comparison of these item-level statistics and estimation methods is presented in Section 2, Table 8.

The uDTF is similar to the sDTF, except that the absolute value of the difference between the two estimated total scale scores is calculated. The result is always positive and is calculated similar to the way the AUD is calculated at the item level.

Among the advantages to this approach described by Chalmers (2018) is that because they are not squared, the statistics are in the same metric as the expected scale scores, and the result does not depend on the distribution of any of the comparison groups or specific populations sampled. Thus, it is argued that because the result is not dependent on either distribution, results will not be as affected as other estimation methods (e.g., Raju, 1999) if the groups are of very different sizes and the studied group is smaller.

1.5. Challenges to DIF Detection

Factors that may affect DIF detection and that are discussed in Sections 1.3 and 1.4 include violations of unidimensionality assumptions, and failure to adjust for multiple comparisons and for magnitude of DIF at the item and scale level. Briefly reviewed in this Section are two topics that affect DIF detection and are examined in Section 2: group differences in latent trait distributions and anchor item selection.

1.5.1. Distributions Differences in θ distributions between groups as was often observed in PROMIS DIF detection applications (e.g., Paz et al., 2013) can result in inflated Type I error rates (DeMars, 2010; Li et al., 2012), and non-normal latent distributions can impact the performance of DIF detection methods. For example, comparison of Mantel, Generalized Mantel, and PolySIBTEST (non-parametric tests) to IRT-LRT DIF using ordinal items showed that while all procedures were affected by non-normal subgroup latent distributions, IRT-LRT DIF was more robust to latent nonnormality than the nonparametric approaches (Woods, 2011). Increasing the number of anchors had a mitigating effect on Type I error inflation across methods; additionally, the AUD (averaged unsigned difference presented above) was observed to be the most consistently accurate effect size. These methods are discussed in Section 1.4 and are illustrated in Section 2. The effect of impact (different group distributions) is also demonstrated in Section 2.

1.5.2. Anchor Item Selection Method Recently, there has been considerable work on the topic of anchor item selection (e.g., Kopf et al., 2015a, 2015b; Meade & Wright, 2012; Setodji et al., 2011; Shih & Wang, 2009; Shih et al., 2014; Wang, 2004; Wang, & Yeh, 2003; Wang & Woods, 2017). Best methods for selecting DIF-free anchor items have been reviewed (e.g., Kopf et al., 2015a, 2015b; Teresi & Jones, 2016; Wang & Shih, 2010; Woods, 2009a), and several methods for anchor item selection have been advanced. One approach is the so-called all-other or all-other with purification, often used in IRT-LRT (Bolt, 2002; Kim & Cohen, 1998), 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. Another method is the constant anchor approach (Thissen et al., 1993), based on the assumption that the anchor set is known from other studies or procedures. This approach, used in the analyses of many PROMIS DIF studies, relies on iterative purification, using the Wald procedure or that of Wood's rank order method (Woods, 2009a) to select anchor items to avoid Type I error inflation (Rikis & Oshima, 2017). The rank order method was used in sensitivity analyses to examine the convergence of identified anchor items and in cases in which not enough anchor items were identified with the standard approach (all-others-as-anchors). However, it is noted that to the extent that there are group differences in the trait estimates and greater percent DIF, power for DIF detection is reduced.

As reviewed, use of a reference anchor item or set has been found to improve Type I error rates in DIF detection for several models, e.g., MIMIC (Wang & Shih, 2010), IRT-LRT (Woods, 2009a), and hierarchical generalized linear models (Chen et al., 2013). Stark et al. (2006) suggested selecting a single anchor from among those tested with the highest factor loading. Often this approach is used in factor analyses in selection of the most discriminating item to set the metric for the latent construct. Because IRT-LRT tests have greater power when the discrimination parameter is larger (Ankenmann et al., 1999; Lopez Rivas et al., 2009), it has been recommended that anchor items be those with the highest discrimination parameters (González-Betanzos & Abad, 2012). Selection of items with higher discrimination parameters has recently been recommended by Wang and Woods (2017) in the context of the Wald test. A variant of this approach was used in the illustration in Section 2 of this paper.

2. Illustration

The following Section provides an illustration of a multidimensional approach to examining two short-form measures and is motivated by the goal of moving beyond unidimensional approaches to DIF detection.

2.1. Background

The Patient-Reported Outcomes Measurement Information System (PROMIS®) "Roadmap Initiative" was funded by the National Institutes of Health in 2004 to improve and standardize the measurement of symptoms and health outcomes by constructing item banks using item response theory (Cella et al., 2007; Reeve et al., 2007). Although the original anxiety and depression item banks were evaluated for DIF using the unidimensional IRT-based methods described earlier (Choi et al., 2011; Teresi et al., 2009), little data existed that permitted evaluation of the performance of PROMIS measures across ethnically diverse groups. The Measuring Your Health (MYHealth; Jensen et al., 2016a) study of PROMIS short-form measures in a stratified random sample of 5506 ethnically diverse patients with cancer was thus initiated in 2010 to partially redress this gap.

PROMIS depression and anxiety short-forms scales continue to be evaluated for clinical validity (Schalet et al., 2016) and to determine minimally important differences (Yost et al., 2011). However, less analysis of differential item functioning has been performed. For example, the 8-item PROMIS physical function, fatigue, and depression short-forms have been examined among a sample of adults with chronic obstructive pulmonary disease (Bjorner et al., 2014). Thresholds and factor loadings were compared across mode of administration (interactive voice, paper, digital assistant, and personal computer on the Internet) using multigroup CFA. DIF effects (impact) were examined by fixing and freeing IRT threshold and slope parameters. The authors concluded that there was no salient DIF by mode.

More recently, the short-form Depression and Anxiety scales were examined for DIF with unidimensional IRT models (Teresi et al., 2016a, 2016b) described in Section 1 of this paper. Specifically, the primary analyses used IRT Wald tests (Cai et al., 2011); sensitivity analyses were performed using an IRT ordinal logistic regression (IRT-OLR) approach (Choi et al., 2011). Both methods adjusted for multiple comparisons and magnitude (effect size) measures using the methods of Wainer (1993), Raju et al. (1995) and Kim et al. (2007) described above and integrated into the interpretation of DIF findings. To our knowledge, no studies have examined the short-forms using multidimensional models. Using unidimensional IRT models, many short-form depression items tested positive for DIF, particularly among Asians/Pacific Islanders as contrasted with the White reference group (Teresi et al., 2016a). Because this illustration focuses on gender, findings are summarized for that variable. After adjustment for multiple comparisons, one item (4, sad) was flagged for DIF using the primary IRT-Wald test approach and four with the IRT-OLR approach (items 3,4,6,7). No items evidenced high magnitude DIF for gender, as shown by overlapping item response functions and low effect size estimates. The aggregate impact was negligible. Although there was a correspondence of the findings to the DIF hypotheses in a number of instances, the magnitude and impact of DIF were negligible.

Fewer items with significant DIF were observed for anxiety as contrasted with depression (Teresi et al., 2016b). Contrary to the hypotheses, the findings were of very little DIF by gender group. One item was identified with DIF after correction for multiple comparisons for the primary IRT-Wald test—item 21 (difficulty calming down). The IRT-OLR method flagged most of the items with DIF. As with depression, very little DIF of high magnitude was evidenced in the PROMIS short-form items and none of high scale-level aggregate impact. A recent study (Taple et al., 2019) examining the depression and anxiety PROMIS short-forms using lordif (Choi et al., 2011) also identified negligible DIF impact.

Recently, the MyHealth PROMIS Depression and Anxiety short-forms were used as an illustration in this journal (Yu et al., 2018) of new methods for mediation analyses. The analyses examined health disparities in the context of observed clinically important mean differences (of 3 or more points; Yost et al., 2011) between non-Hispanic White and Hispanic White respondents. Hispanic White respondents reported greater depression and anxiety. An obvious question is whether these differences reflect true differences or could be due to DIF.

2.2. *Aims*

The aim of this illustration was to use multidimensional models to examine the PROMIS anxiety and depression subscales and to use simulations to compare unidimensional and multidimensional approaches.

2.3. *Methods*

2.3.1. Sample Patients for the MyHealth study were recruited from four Surveillance, Epidemiology, and End Results (SEER) cancer registries located in California (2), Louisiana, and New Jersey between 2010 and 2012 (Jensen et al., 2016a). The population-based sample of recently diagnosed cancer patients was oversampled to include larger samples of racial/ethnic minorities and younger patients. Eligible participants were 21-84 years old at the time of initial diagnosis of their first primary cancer. Eligibility was restricted to persons diagnosed with one of seven cancers, and able to read English, Spanish, or Mandarin. Sampling was stratified by four race-ethnicity groups (Non-Hispanic White, Hispanic, Non-Hispanic Black, Non-Hispanic Asian) and three age groups at diagnosis (21-49, 50-64, 65-84 years). This study was approved by Institutional Review Boards at Georgetown University, the State of California, and each research site.

2.3.2. Measure Items for inclusion in the MyHealth survey were selected from the PROMIS short-forms (as of 2010) or their high frequency of selection when administered online using the PROMIS computerized adaptive testing (CAT) assessment center. The frequency assessment was based on a prior sample of cancer patients scoring at least one-half standard deviation above (i.e., higher symptoms) the US general population mean (see Jensen et al., 2016a).

The PROMIS depression and anxiety short-forms were developed based on item bank parameters generated from an IRT graded response model (Samejima, 1969). The final bank contained 28 depression and 22 anxiety items. Depressive symptoms assessment and the 11 anxiety items were subdomains of emotional distress (Choi et al., 2010). The PROMIS short-form depression scale was developed by selecting items that maximized measurement precision and were most informative regardless of their location on the trait (Choi et al., 2010; Pilkonis et al., 2011). Short-form items were selected from the item bank based on the rank-order of IRT information provided and higher frequency of administration in the CAT. The depression short-form was almost as precise as the CAT in the middle and upper part of the distribution, and less so at the extremes of the distribution. In addition to the short-form items, three items were selected for this study based on their rank-ordering in terms of overall IRT information. Items were administered using a five-point response scale: 'never,' 'rarely,' 'sometimes,' 'often' and 'always.' The timeframe was the past 7 days. Anxiety items were administered using the same five-point response scale and one-week timeframe as for the depression items. In addition to the eight-item short-form, three other anxiety items were selected for inclusion based on high overall information estimated from a graded response IRT model and coverage across the latent attribute continuum, or their inclusion in other short-form measures. (Details of the origins of the short-form items used in the MyHealth survey can be found in Jensen et al. 2016a).

2.3.3. Statistical Approach Several DIF detection methods described in Section 1 have been extended to the multidimensional context, including the multiple indicators multiple causes (MIMIC) models (Lee et al., 2017), the IRT likelihood ratio test (IRT-LRT; Thissen et al., 1993), and logistic regression (Bulut & Suh, 2017). Among these three methods, the logistic regression approach does not permit direct modeling of latent trait differences (i.e., impact) across groups because the primary idea of the method is to compare a null model (assuming no DIF) for an item to two nested models formed in hierarchy with an explanatory group variable and group-by-θ interaction variable. That said, individual trait scores are estimated (subject to identification constraints such as fixing the mean and variance of trait to 0 and 1 in the reference group), and therefore, group mean differences can be estimated from the sample mean difference. In contrast, in the MIMIC model, θ is regressed on an observed grouping variable to allow for a mean difference (Woods, 2009b), i.e., $\theta = \beta \times \text{group} + \varepsilon$, where the variance of the residual, ε, is fixed to 1 to identify the scale. As a result, the MIMIC approach can only model impact in terms of the group mean difference, but not group covariance differences. The IRT-LRT (Suh & Cho, 2014) is by far the most general approach, and its main idea is to compare two hierarchically nested models and evaluate if one model fits better than the other. A statistical significance of the χ^2 difference statistic indicates the presence of DIF. This approach could easily accommodate group differences in both mean and covariance, although it usually requires a larger sample size than the MIMIC approach (Woods, 2009b).

When the test or scale displays a simple structure, no prior study has evaluated whether using a MIRT model for DIF detection will perform differently from the traditional DIF detection approach by treating the subscales separately. Therefore, we first present a real data illustration using the PROMIS Depression and Anxiety scales, followed by a simulation study to provide further evidence. The item response function for the MIRT model considered in this subSection is

$$P\left(x=k\right)=\frac{1}{1+e^{-(\boldsymbol{a}_i\boldsymbol{\theta}+d_{ik})}}-\frac{1}{1+e^{-(\boldsymbol{a}_i\boldsymbol{\theta}+d_{ik+1})}}. \tag{2-1}$$

Compared with Eq. 1-3, here both $\boldsymbol{a}_i$ and $\boldsymbol{\theta}$ are multidimensional vectors, and for a simple structure MIRT model, only one element in $\boldsymbol{a}_i$ is nonzero, implying the specific factor that item i measures. By MIRT convention, the scaling factor D is not included in the equation. Again, note that compared to Eq. 1-3, here we considered a slope-intercept notation instead, so that the DIF effect is on $\boldsymbol{a}_i$ and $\boldsymbol{d}_i$ (i.e., a m-dimensional vector).

2.3.3.1 Real Data Illustration For the real data illustration, gender was considered as the grouping variable and the IRT-LRT method implemented in the "mirt" package (Chalmers, 2012) was used. The "mirt" package permits specification of anchor items and estimation of the mean and variance of θ for the focal (studied) group. Two approaches to DIF detection are permitted: the likelihood ratio test and Wald tests. The LRT approach was used in this illustration because preliminary study showed no appreciable difference between the two methods under considered conditions.

In the unidimensional IRT context, two different approaches were taken for the IRT-LRT method: the first approach assumes no impact (i.e., no group difference in θ distributions) with θ assumed to have a mean of 0 and variance of 1 in both groups; hence, anchor items are not needed to link the groups. The second approach assumes that there is impact and θ is assumed to have a mean of 0 and variance of 1 only in the reference group, whereas the mean and variance of θ are estimated for the focal (studied) group. In this regard, at least one anchor item per subscale is needed to establish a linking between the two groups. In the MIRT context, the same two approaches were used, except that when there is impact, the correlation between two θs (i.e., depression and anxiety in this example) is estimated for the reference group, whereas both the mean and covariance of $\boldsymbol{\theta}$ are estimated for the focal group. In this case, item 1 (worthless) from the Depression scale and item 5 (nervous) from the Anxiety scale were used as anchor items because they were among the items that exhibited the least amount of DIF in prior studies (Teresi et al., 2016a, 2016b). In the "mirt" package, the "DIF" function was used along with an "add" scheme, which means that a most unconstrained model is fitted first assuming all non-anchor items have DIF. Then, the constraints on the item parameters for the target (studied) item are added one item at a time for each of the tested (studied) items.

2.3.3.2 Simulation Study Design A simulation study was conducted to further compare the performance of the IRT-LRT method within the unidimensional IRT (UIRT) and MIRT contexts. The two-dimensional two-parameter logistic model was used and both uniform and non-uniform DIF were considered. To be consistent with the PROMIS data structure, the test length was set at 21 with the first 10 items loading on factor 1 and the remaining items loading on factor 2. The true item parameters for the focal (studied) group were also obtained from the PROMIS items (but with only the first threshold parameter per item); they are presented in Table 1. The uniform DIF magnitude between 0.6 and 1.7 was randomly added to the designated DIF items (i.e., d parameter) to mimic the simulation design of the latest research (e.g., Belzak & Bauer, 2020). The values shown in parentheses in Table 1 are the threshold parameters for the focal (studied) group. Five out of 21 items were simulated to have uniform DIF, and they were items 6, 11, 12, 13, and 14. That is, one item from the first factor and four items from the second factor exhibited DIF. This unbalanced DIF proportion was intentionally created so that the results are more generalizable. Table 2 presents the non-uniform DIF added on the discrimination parameters. These DIF values were chosen to produce wABC (weighted area between the expected score curves described in Section 1.4.1.1.4) values between 0.46 and 0.97, the range used in Edelen et al. (2015). Sample size was fixed at 1000 per group. Both no-impact and impact conditions were included. When there was no impact, $\boldsymbol{\theta}s$ from both groups were simulated from a bivariate normal distribution with means of 0 and covariance matrix of $\begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$, where 0.85 is close to the correlation between depression and anxiety. When there was an impact, the means were still fixed at 0 in both groups, whereas the two covariance matrices for the focal and reference groups were $\begin{bmatrix} 1 & 0.15 \\ 0.15 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$, respectively. The small correlation chosen in the focal group was intentional to produce a large impact, which is considered a more challenging scenario. Most previous DIF research either did not consider impact or considered low impact or impact only on the mean difference (e.g., Bulut & Suh, 2017). One-third of the items were assumed to be DIF-free and they served as anchor items. Evaluation criteria include Type I error rate and power for DIF detection, as well as item parameter recovery in terms of root mean squared error (RMSE) and average bias. Fifty replications were conducted per condition, and the results were averaged over the replications. In sensitivity analyses, one condition (no impact, uniform DIF) was selected for which 100 replications were run. A plot showing how Type I error and power from both UIRT and MIRT approaches changed as a function of the number of replications showed that 50 replications were sufficient. The final results after increasing the number of replications to 100 were almost the same as with 50.

2.4. Results

2.4.1. Real Data Illustration Table 3 presents the DIF analysis results from IRT-LRT implemented in the UIRT and MIRT contexts. The false discovery rate (FDR) control (Benjamini & Hochberg, 1995) was used to correct for the family-wise error rate. Items bolded are significant at $\alpha = .05$ (i.e., false discovery rate), whereas the remaining items are significant at $\alpha = .01$. As shown, when assuming no impact, every item in the UIRT context was flagged as having DIF, whereas 14 out of 21 items in the MIRT context were flagged. On the other hand, when assuming there is an impact, the DIF results become more consistent with prior studies in which the mean and variance of θ were estimated for the focal groups (e.g., Teresi et al., 2016a, 2016b), and the estimated population parameters of the θ distribution are presented in Table 4. It appears from Table 4 that there is a group difference between the mean and covariance of θ, with females (group 2) showing higher levels of depression and anxiety than males. This reinforces that the results from the first approach may not be valid due to the fallible assumption of no impact.

2.4.2. Simulation Study Table 5 presents the Type I error and power from UIRT and MIRT approaches under both no-impact and impact conditions. As shown, the MIRT approach generated slightly higher power, but the two approaches produced the same Type I error rate. The power from non-uniform DIF conditions was consistently smaller because (1) the manipulated DIF size in the uniform conditions was larger (the wABC values were 0.190–0.425) and they are similar to those reported in Narayanan and Swaminathan (1996), and (2) detecting non-uniform DIF with dichotomous items may be difficult because when the two item response functions cross, it is difficult to clearly attribute the difference in those functions to *a*- or *b*- (or *d*-) parameters. Table 6 presents the item parameter recovery results, which show that the MIRT approach slightly outperformed the UIRT approach by producing more accurate item parameter estimates, but the

Table 1.
True item parameters in the simulation study (the values in parentheses are the true item parameters for the focal (studied) group; DIF magnitude is thus the difference between item parameters from the studied and reference groups.)

Item	a_1	a_2	d
1. Worthless	3.46		− 1.45
2. Nothing to look forward to	4.17		− 1.77
3. Helpless	3.84		− 1.36
4. Sad	3.95		0.88
5. Felt like a failure	4.22		− 2.07
6. Depressed	4.49		− 0.23 (0.35)
7. Unhappy	4.38		0.60
8. Hopeless	5.93		− 2.60
9. Discouraged about the future	3.79		− 0.37
10. Disappointed in myself	3.32		− 0.82
11. Fearful		2.98	− 0.36 (0.26)
12. Anxious		3.63	0.35 (1.20)
13. Worried		3.72	1.59 (2.60)
14. Hard to focus		4.44	− 1.50 (− 0.20)
15. Nervous		4.33	− 0.02
16. Uneasy		4.92	0.02
17. Tense		4.15	0.37
18. Worries overwhelmed me		4.18	− 1.08
19. Needed help for anxiety		3.89	− 1.83
20. Many situations made me worry		3.41	0.28
21. Difficulty calming down		3.57	− 1.28

difference was not substantial. Table 7 presents the mean absolute bias of the DIF magnitude. It appears that for some DIF items, the MIRT approach produced smaller bias, whereas for some items, the UIRT approach performed slightly better. There is no appreciable difference between the impact and no-impact conditions.

2.5. *Magnitude (Effect Size) Measures*

2.5.1. Estimation Approach DIF effect size measures are reviewed in Section 1.4. Shown in Table 8 are the results of their application to anxiety and depression data for unidimensional and multidimensional models. When reviewing these results, it is important to recall that methods for estimating the parameters may differ that assumptions about the underlying density functions may vary, and the resulting statistics can be a squared value, an unsigned absolute value, or a signed value; thus, not all are directly comparable. It is noted that different software was used to produce these effect sizes, and thus, different algorithms were used in the estimation of the IRT parameters. For example, for T1, AUD, wABC, modified B_{NC}, and NCDIF from Magnits (Kleinman & Teresi, 2016), parameters were generated from IRTPRO (Cai et al., 2011) and significance was determined using the Wald tests. For calculation of the effect size measures, the studied group parameters were equated to be on the same scale as the reference group using Baker's EQUATE (Baker, 1995) program. For the lordif analyses, uniform and non-uniform DIF were determined using the likelihood ratio Chi-square test. Uniform DIF was obtained by comparing the log likelihood values from models one and two. Non-uniform DIF was obtained by comparing the log likelihood values from models two and three, described in Section 1.2.2. Because DIF

Table 2.
Non-uniform DIF condition: True item parameters in the simulation study (the values in parentheses are the true item parameters for the focal (studied) group; DIF magnitude is thus the difference between item parameters from the studied and reference groups).

Item	a_1	a_2	d
Var1_dep	3.460	0	−1.446
Var2_dep	4.170	0	−1.771
Var3_dep	3.840	0	−1.355
Var4_dep	3.950	0	0.880
Var5_dep	4.215	0	−2.069
Var6_dep	4.488 (3.488)	0	−0.226
Var7_dep	4.377	0	0.599
Var8_dep	5.929	0	−2.604
Var9_dep	3.787	0	−0.365
Var10_dep	3.324	0	−0.821
Var1_anx	0	2.981 (2.181)	−0.356
Var2_anx	0	3.628 (2.628)	0.354
Var3_anx	0	3.721 (2.621)	1.593
Var4_anx	0	4.444 (3.244)	−1.503
Var5_anx	0	4.331	−0.021
Var6_anx	0	4.917	0.018
Var7_anx	0	4.152	0.371
Var8_anx	0	4.177	−1.076
Var9_anx	0	3.894	−1.825
Var10_anx	0	3.407	0.283
Var11_anx	0	3.556	−1.275

Table 3.
Items flagged as having DIF (in italics) by IRT-LRT methods implemented in the UIRT and MIRT contexts.

	Approach 1: no impact	Approach 2: with impact
UIRT model Gender effect on Depression	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)	(***3***, 4, 6)[a]
UIRT model Gender effect on Anxiety	(11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21)	(11, ***14, 16***, 21)
MIRT model Gender effect on Depression and Anxiety	(3, 4, 6, 7, 11, 12, 13, ***14***, 15, 16, 17, 18, 19, 20)	(3, 4, ***5***, 6, ***7***, 11, 12, 13, ***17***, ***18***, 21)

[a] Values in bold italic are significant at the false discovery rate of .05, whereas the remaining items are significant at .01.

TABLE 4.
Estimated population mean and covariance of θ from two groups.

	Group1 (Males)					Group2 (Females; studied group)				
	Mean1	Mean2	Cov11	Cov12	Cov22	Mean1	Mean2	Cov11	Cov12	Cov22
UIRT model Sex effect on Depression	0		1			0.215		0.891		
UIRT model Sex effect on Anxiety		0			1		0.298			0.915
MIRT model Sex effect on Depression and Anxiety	0	0	1	0.931	1	0.094	0.160	1.223	1.091	1.193

TABLE 5.
Type I error rate and power from the simulation study.

		Uniform DIF		Non-uniform DIF	
		UIRT	MIRT	UIRT	MIRT
No-impact	Type I	.013	.016	.024	.026
	Power	.908	.936	.616	.648
Impact	Type I	.024	.018	.018	.013
	Power	.924	.944	.676	.684

was not detected using the pseudo-R^2 measures or the change in Beta criterion, no anchor items were specified. NCDIF was calculated in lordif and compared with those obtained from Magnits.

For the comparison of unidimensional and multidimensional models, the multipleGroup function within “mirt” (Chalmers, 2012) was used to model the data as unidimensional or multidimensional, and then, the parameters from these models were used to calculate the uDIF and sDIF statistics using the DRF function. For uDIF and sDIF, anchor items were specified based on prior results using two specifications: the single anchor item method used in the real data illustration above as well as multiple anchors identified from earlier analyses.

2.5.2. Magnitude (Effect Size) Results As shown in Table 8, the default effect size for anchor items is close to zero. As shown, the impact on the effect size of use of different sets of anchor items (only one per scale in brackets or multiple anchors) is minimal, with values generally within 0.01 of each other. The values with only one anchor (in brackets) tend to be slightly larger. Values for unidimensional versus multidimensional models were similar. Additionally, values for T1, AUD, wABC and β_{NC} were similar (except for the sign of T1). Sensitivity analyses using simultaneous parameter estimation instead of separate group estimation and equating revealed a close correspondence between values. For example, the AUD and uDIF values were for the most part within 0.005 of one another. The differences were larger for uDIF using different numbers of anchors (about 0.02 between use of several anchors vs. one anchor). With simultaneous linking, the values of NCDIF estimated by the two software packages were almost identical.

TABLE 6.
Item parameter recovery for reference group from both approaches.

		Uniform DIF						Non-uniform DIF					
		UIRT			MIRT			UIRT			MIRT		
		a_1	a_2	d	a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
No-impact	Bias	.049	.006	− .040	− .012	− .049	−.053	− .004	.015	− .008	− .026	− .019	− .022
	RMSE	.328	.246	.177	.302	.233	.206	.335	.324	.164	.327	.306	.186
Impact	Bias	.011	− .001	.001	− .011	− .017	− .035	− .013	.027	− .014	− .037	− .004	− .008
	RMSE	.272	.245	.153	.260	.239	.156	.327	.322	.153	.322	.321	.158

RMSE root mean squared error.

TABLE 7.
Mean absolute bias of DIF magnitude estimates.

		UIRT					MIRT				
	Item	6	11	12	13	14	6	11	12	13	14
Uniform DIF	No-impact	0.299	0.102	0.134	0.137	0.143	0.231	0.121	0.152	0.162	0.164
	Impact	0.271	0.112	0.103	0.166	0.152	0.425	0.249	0.204	0.135	0.149
Non-uniform DIF	No-impact	0.855	0.336	0.633	0.384	0.452	0.800	0.317	0.622	0.352	0.430
	Impact	0.870	0.295	0.475	0.352	0.336	0.892	0.280	0.462	0.337	0.335

While no items in these examples evidenced salient DIF, it is interesting to note that in the Depression scale, the item that showed the largest DIF across all the unidimensional methods was item 4 ‘I felt sad.’ For the Anxiety scale, the item that consistently showed the largest DIF was item 21 ‘I had difficulty calming down.’ These were also the items with the highest effect size for the estimates from the ‘mirt’ unidimensional and multidimensional approaches. It is noted, however, that the effect size even for the item with the largest value was indicative of a group difference of a fraction (0.19) of a point on the original metric of a five-point response scale.

2.6. *Summary and Conclusions (from the Illustration)*

2.6.1. Real Data Example and Simulation The IRT-LRT DIF detection method is a flexible method to handle group differences of θ distributions, known as impact, and was studied with both real data and in simulation. From the real data illustration, it was shown that when impact exists and the method assumes no impact, well-behaving items may be incorrectly flagged as having DIF. As a result, the IRT-LRT method is recommended. However, the successful performance of the IRT-LRT method when impact exists hinges on the correct selection of the anchor (i.e., DIF-free) items as discussed in Section 1.5.2 and in Section 3. At least one anchor item per subscale needs to be specified in advance for the unconstrained model to be identifiable. An iterative purification method may be an option (e.g., Candell & Drasgow, 1988; Clauser et al., 1993), but it is computationally slow and a more efficient method is needed, such as regularization methods (Bauer et al., 2019; Belzak & Bauer, 2020). These methods impose a penalty function on DIF parameters during estimation, so that non-DIF items will have their DIF parameters automatically suppressed to 0 without impacting the fit of the model.

Another observation from the real data illustration is that using the IRT-LRT method within a MIRT context results in more flagged items as compared to using the IRT-LRT method within a UIRT context. This result is, to some extent, consistent with the simulation evidence that MIRT DIF detection is more powerful. However, this conclusion needs to be supported by more thorough simulation studies, including other manipulated conditions such as non-uniform DIF. Additionally, the simulated data were binary; future research could examine polytomous data. Finally, the method used was the likelihood ratio test. The Wald test was used in some of the earlier studies of PROMIS measures. Although the two are asymptotically equivalent, in the context of DIF testing, use of the Wald test permits testing DIF across multiple groups rather than two groups at a time, as is the case with IRT-LRT. Future work could also compare DIF testing using IRT-LRT to IRT-Wald tests; such efforts are beginning to emerge in the context of unidimensional data (Chalmers, 2018; Woods et al., 2013). More such work with multidimensional models is needed. An additional point, as reviewed earlier in this article, is that PROMIS standards included incorporation of effect size measures in determining salient DIF. Integrated methods for examining effect size

TABLE 8.
PROMIS depression and anxiety short-form items: Magnitude (effect size) measures for gender subgroup comparisons.

Item description	NCDIF (Magnits)	NCDIF (lordif)	T1	AUD	wABC	*Modified* B_{NC}	Uni-dimensional uDIF (mirt)	Multi-dimensional uDIF (mirt)	Uni-dimensional sDIF (mirt)	Multi-dimensional sDIF (mirt)
1. I felt worthless	0.0020	0.0002	−0.0317	0.0317	0.0380	0.0353	$< 0.0001^{a}$ $[< 0.0001^{a}]^{*}$	$< 0.0001^{a}$ $[< 0.0001^{a}]$	$< 0.0001^{a}$ $[< 0.0001^{a}]$	$< 0.0001^{a}$ $[< 0.0001^{a}]$
2. I felt that I had nothing to look forward to	0.0017	0.0004	−0.0262	0.0263	0.0320	0.0295	$< 0.0001^{a}$ [0.0098]	$< 0.0001^{a}$ [0.0208]	$< 0.0001^{a}$ [0.0081]	$< 0.0001^{a}$ [0.0208]
3. I felt helpless	0.0023 (0.0065)^{b}	0.0093	0.0302	0.0345 (0.0579)	0.0393	0.0374	0.0633 [0.0733]	0.0589 [0.0836]	0.0609 [0.0724]	0.0586 [0.0836]
4. I felt sad	0.0266 (0.0391)	0.0485	0.1321†	0.1321 (0.1611)	0.1569	0.1441	0.1850 [0.1969]	0.1756 [0.1984]	0.1850 [0.1969]	0.1756 [0.1984]
5. I felt like a failure	0.0055 (0.0019)	0.0016	−0.0471	0.0471 (0.0266)	0.0598	0.0538	0.0308 [0.0217]	0.0307 [0.0101]	−0.0291 [−0.0179]	−0.0292 [−0.0039]
6. I felt depressed	0.0040 (0.0098)	0.0142	0.0497	0.0497 (0.0780)	0.0592	0.0544	0.0883 [0.1005]	0.0839 [0.1090]	0.0883 [0.1005]	0.0839 [0.1090]
7. I felt unhappy	0.0009 (0.0036)	0.0062	0.0187	0.0187 (0.0473)	0.0226	0.0205	0.0527 [0.0644]	0.0447 [0.0675]	0.0527 [0.0644]	0.0447 [0.0674]
8. I felt hopeless	0.0009	0.0008	−0.0175	0.0209	0.0262	0.0235	$< 0.0001^{a}$ [0.0195]	$< 0.0001^{a}$ [0.0341]	$< 0.0001^{a}$ [0.0186]	$< 0.0001^{a}$ [0.0334]
9. I felt discouraged about the future	0.0049 (0.0015)	0.0007	−0.0472	0.0540 (0.0287)	0.0628	0.0587	0.0334 [0.0273]	0.0417 [0.0400]	−0.0220 [−0.0102]	−0.0275 [−0.0031]
10. I felt disappointed in myself	0.0025 (0.0008)	0.0003	−0.0328	0.0381 (0.0191)	0.0444	0.0414	0.0195 [0.0171]	0.0270 [0.0300]	−0.0082 [0.0030]	−0.0121 [0.0116]
11. I felt fearful	0.0087 (0.0057)^{b}	0.0064	0.0770	0.0774 (0.0627)	0.0885	0.0840	0.0673 [0.0855]	0.0874 [0.1175]	0.0649 [0.0842]	0.0871 [0.1175]
12. I felt anxious	0.0040 (0.0016)	0.0020	0.0448	0.0471 (0.0300)	0.0590	0.0537	0.0332 [0.0527]	0.0400 [0.0732]	0.0309 [0.0519]	0.0368 [0.0698]
13. I felt worried	0.0030	0.0018	0.0473	0.0473	0.0548	0.0512	$< 0.0001^{a}$ [0.0551]	$< 0.0001^{a}$ [0.0778]	$< 0.0001^{a}$ [0.0549]	$< 0.0001^{a}$ [0.0778]
14. I found it hard to focus on anything other than my anxiety	0.0068 (0.0131)	0.0107	−0.0564	0.0566 (0.0786)	0.0716	0.0655	0.0890 [0.0689]	0.0711 [0.0381]	−0.0890 [−0.0686]	−0.0700 [−0.0348]
15. I felt nervous	0.0001	0.0003	0.0007	0.0079	0.0100	0.0089	$< 0.0001^{a}$ $[< 0.0001^{a}]$	$< 0.0001^{a}$ $[< 0.0001^{a}]$	$< 0.0001^{a}$ $[< 0.0001^{a}]$	$< 0.0001^{a}$ $[< 0.0001^{a}]$
16. I felt uneasy	0.0049 (0.0078)	0.0069	−0.0464	0.0480 (0.0655)	0.0521	0.0504	0.0718 [0.0505]	0.0761 [0.0515]	−0.0718 [−0.0504]	−0.0760 [−0.0399]

TABLE 8.
continued

Item description	NCDIF (Magnits)	NCDIF (lordif)	T1	AUD	wABC	*Modified* B_{NC}	Uni-dimensional uDIF (mirt)	Multi-dimensional uDIF (mirt)	Uni-dimensional sDIF (mirt)	Multi-dimensional sDIF (mirt)
17. I felt tense	0.0016	0.0005	0.0213	0.0255	0.0347	0.0301	$< 0.0001^a$ [0.0258]	$< 0.0001^a$ [0.0430]	$< 0.0001^a$ [0.0256]	$< 0.0001^a$ [0.0312]
18. My worries overwhelmed me	0.0011	0.0003	0.0187	0.0214	0.0261	0.0245	$< 0.0001^a$ [0.0214]	$< 0.0001^a$ [0.0621]	$< 0.0001^a$ [0.0188]	$< 0.0001^a$ [0.0599]
19. I felt like I needed help for my anxiety	0.0017 (0.0055)	0.0039	−0.0282	0.0282 (0.0504)	0.0339	0.0320	0.0561 [0.0356]	0.0402 [0.0161]	−0.0561 [−0.0355]	−0.0402 [−0.0030]
20. Many situations made me worry	0.0006	0.0002	0.0108	0.0187	0.0244	0.0215	$< 0.0001^a$ [0.0166]	$< 0.0001^a$ [0.0490]	$< 0.0001^a$ [0.0130]	$< 0.0001^a$ [0.0418]
21. I had difficulty calming down	0.0126 (0.0185)	0.0174	−0.0810	0.0810 (0.0961)	0.1030	0.0935	0.1161 [0.0967]	0.1079 [0.0731]	−0.1161 [−0.0967]	−0.1079 [−0.0731]

All non-compensatory differential item functioning (NCDIF; Raju et al., 1995) values were smaller than the threshold (0.0960); all wABC values are below the threshold of 0.3099; † indicates value above threshold of 0.10.

T1 (Wainer, 1993; Kim et al., 2007); Average Unsigned Difference (AUD; Woods, 2011); wABC (Orlando-Edelen et al., (2015); B_{NC} = non-compensatory bias; uDIF= unsigned differential item functioning; sDIF= signed differential item functioning (Chalmers, 2016, 2018).

For the lordif analyses, uniform and non-uniform DIF were determined using the likelihood ratio Chi-square test. Uniform DIF was designated after comparing the log likelihood values from models one and two. Non-uniform DIF was designated after comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R2 measures or the change in Beta criterion; thus, no anchor items were specified. Items with salient DIF were identified and NCDIF calculated.

For T1, AUD, wABC, modified B_{NC}, and NCDIF from Magnits (Kleinman & Teresi, 2016), parameters were generated from IRTPRO (Cai et al., 2011) and significance was determined using the Wald tests. For calculation of the effect size measures, the studied group parameters were equated to be on the same scale as the reference group using Baker's EQUATE (Baker, 1995) program.

For the uDIF and sDIF the statistics were calculated as separate functions. First, the multipleGroup function within "mirt" (Chalmers, 2012) was used to model the data as unidimensional or multidimensional, and then, the parameters from these models were used to calculate the uDIF and sDIF statistics using the DRF function.

[a] Anchor items identified in the original analyses (Teresi et al., 2016a, 2016b) treating depression and anxiety as unidimensional using IRTPRO (Cai et al., 2011)

[b] Elements in parentheses were based on group parameters estimated simultaneously rather than using the Stocking and Lord (1983) separate group equating algorithms; the five anchor items identified originally were included in this estimation procedure.

*Elements in brackets in the last four columns are from the unidimensional model without anchor items except for the first item for depression and the fifth item for anxiety, consistent with the analyses presented earlier in Section 2.

measures in the context of IRT DIF detection valuation are still in early stages of development (Chalmers, 2018).

2.6.2. Effect Size Measures Several effect size measures were examined in the illustration. Some methods (the AUD) are based on unsigned differences between groups, and some assess signed differences between groups, like the T1 statistic. Both can prove valuable because to the extent that they differ, one can assess instances of crossing DIF, where different areas of the curve may 'favor' one or the other group. Some statistics such as NCDIF are based on average squared differences between the groups and cannot be compared directly to results from statistics that are not squared. Another difference involves the choice of a density function. Some methods (NCDIF, T1, and AUD) use the studied group density; others (the modified β_{NC}) use the density function of both groups combined. Some methods assume a normal density function for the studied group and assign weights based on the normal distribution, or, in the case of the wABC, assume a normal density function for both the reference group and the studied group.

For comparison purposes, the statistics T1, AUD, wABC, β_{NC} and NCDIF were computed using the same software (MAGNITS), and the parameters and θs were estimated separately for the two groups, and then, the studied group parameters and θs were equated to be on the same scale as the reference group. Equating was performed in an iterative fashion, eliminating from the anchor set any items exhibiting DIF. In the case of these two scales, Depression and Anxiety, all items were retained in the anchor set. In the case of uDIF and sDIF analyses, specified items were used as anchors and set to be equal in both groups. An alternative method used in sensitivity analyses (results shown in Table 8 in parentheses) was simultaneous estimation and linking, with inclusion of previously identified anchor items. The rank order across methods was (in order from lowest to highest magnitude) the same: items 10, 9, 5, 7, 3, 6, 4. Despite differences in approach and estimation, there was considerable consistency in results, especially in identification of items showing the largest values.

3. Discussion

3.1. PROMIS DIF Analyses: Challenges and Advances

3.1.1. Anchor Item Selection Generally, it has been found that more than one anchor item should be used to ensure that linking is accurate. 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 et al., 1990). However, selecting the DIF-free anchor is a challenge. The anchor item selection method used in some PROMIS studies resembled a constant anchor approach with iterative procedures to identify a minimum of four DIF-free items. The all-other approach was used as a first step to select items; however, if very few items were DIF-free, the iterative rank order method (Woods, 2009a; Wang et al., 2012) was used to select a starting anchor set with the least DIF. If the amount of DIF is not large, DIF detection is fairly accurate with the all-other approach; however, simulations have shown that assuming that only one referent item is free of DIF was superior to a constrained model assuming all other items are DIF-free anchors (Stark et al., 2006). Although research, e.g., Woods (2009a), 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. Group sample sizes for the analyses presented for illustration using real data in Section 2 of this paper were relatively large, with 500 or more in the studied groups; however, such sizes are less common in analyses of patient-reported outcome measures.

In the case of smaller sample sizes and when most items were identified with DIF, as was the case with many PROMIS DIF analyses (see Reeve & Teresi, 2016), 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 et al., 2015), which are central to detection of salient DIF. A more recently introduced method (Wang and Woods, 2017) for anchor item selection that includes incorporation of the discrimination ability of items appeared to favor selection of more anchor items. Too few anchor items compromise power for DIF detection; however, including items with DIF in the anchor results in inflated Type I error (false DIF detection). As illustrated in the articles on DIF in PROMIS depression and anxiety short-forms (Teresi et al., 2016a, 2016b), it was not possible to obtain the desired number of anchor items for some subgroup comparisons. Thus, in Section 2, one anchor item was used in the illustration. As discussed in Section 2.6, there is a need for more efficient anchor item selection methods beyond iterative purification.

3.1.2. Adjustment for Multiple Comparisons and Effect Size Estimation PROMIS investigators incorporated adjustments for multiple comparisons and effect size estimates into DIF detection procedures. The rationale was that the item banks were relatively small, and considerable qualitative and quantitative analyses were performed in the selection of the items. The practice of adjusting for multiple comparisons has been recommended, particularly with large sample sizes because as discussed by Boorsboom (2006), with large enough sample sizes all items may evidence DIF. Effect size estimation was also a practice adopted by PROMIS investigators to identify items with salient DIF and mitigate against flagging items that may not be of a magnitude that will result in impact at the scale level.

A key difference among the approaches to effect size estimation is the sampling distribution of the density used in the calculation: the studied group, the reference group, or the marginal density; the latter choice contains distributional information from both groups and may mitigate against the influence of sample size and distribution shape on response bias measures (Chalmers, 2018). Most methods used the studied group or reference group densities separately; however, some used the joint sampling distribution of both groups by applying multiple-group IRT with simultaneous equating and marginal empirical density estimates of θ (Chalmers, 2018) or the normal studied group density (Woods, 2011). The non-compensatory DIF wABC (Orlando-Edelen et al., 2015) is the average of the area between the expected item score curves, weighted by the normal distribution. The methods proposed by Chalmers et al. (2016) and Chalmers (2018) build on the work of Wainer (1993) and Raju et al. (1995), but are more similar to Woods (2011) and Orlando-Edelen et al. (2015) and use the unsquared versions of Wainer's (1993) statistics.

Chalmers (2018) has advanced this area of research through better integration of the DIF and magnitude assessment and providing a new DIF magnitude statistic, incorporating the marginal density, $f(\theta)$ rather than a two-step approach of estimating studied and reference group densities separately. Anchor items can be included in this method. In a simulation study (Chalmers, 2018) with a 2PL generating model for binary data, the authors compared the Wald test, likelihood ratio test, and a novel non-compensatory differential response function (NCDRF) test. It was observed that Type I error rates were similar across the Wald, LR, and NCDRF methods and superior to the non-compensatory SIBTEST method. Power was relatively low across methods for the lower sample sizes of 250 per group and adequate for most conditions with 500 per group or more; however, differences in the θ distribution and number of anchor items (5 vs. 10) affected power for some items and methods. Future directions for such work include examination of polytomous and multidimensional data.

Chalmers et al. (2016) point out that the method of Raju et al. (1995) is generalizable only to similar samples as that investigated and that very skewed distributions could result in inability to detect DIF in that sample. Additionally, group differences in the latent distributions must be adjusted through linking to rescale parameter estimates rather than through use of simultaneous estimation. Chalmers (2016) scale-level statistics pertain to all potential abilities in the population and capture evidence of differential test functioning for trait levels that have not yet been observed; standard errors and confidence intervals can be estimated for these statistics. A caveat is that although differences can be evaluated at various quadrature (θ) points, in many instances evaluating the differences at θ points that reflect the observed densities in the current sample may provide a more accurate reflection of the magnitude of DIF for the data investigated.

Variations among the effect size indicator estimates were observed in Section 2.5. Such differences could occur if estimation assumes a normal density rather than the actual distribution of a studied and/or reference group. Software developers may use approximations of the integral examining midpoints of quadrature points, and the specific methods for parameter estimation could vary, e.g., IRTPRO (Cai et al., 2011); "mirt" (Chalmers, 2012) or R modules (Rizopoulus, 2006, 2009). Also, whether or not simultaneous estimation is assumed or the parameters are equated can make a difference. Finally, the way in which purification is conducted and anchor items selected could influence the final result. These differences are shown in Section 2, Table 8; however, as noted, the overall results were consistent in the rank order of the effect size values across methods, and in identification of salient DIF.

Selecting the studied group density if associated with smaller sample sizes could result in lower power. However, it is also the case that with large enough groups and in the context of patient-reported outcomes where the focus is on retaining items unless large DIF is observed and in health disparities assessment, in which the emphasis is on the studied or targeted minority group, there may be reasons to focus on the studied group density. Studied group distributions are often skewed for patient-reported outcomes. If large enough, and more representative groups of minority respondents are sampled, use of the studied group density may be justifiable. In the context of comparing multiple groups, the reference group density might be selected.

New ways of examining DIF effect sizes and associated assumptions include an odds ratio approach (Jin et al., 2018) for binary data and of mixture models for polytomous and continuous indicators, in which instead of group variables, so-called members of reference and studied groups on a variable of interest may be distributed across latent classes (DeMars, 2015; Raykov et al., 2019). Limitations include that latent class analyses introduce other sources of error, e.g., misclassification error.

3.2. *Summary, Conclusions, and Future Directions*

The illustration in Section 2 showed that many items evidenced DIF even after correction for multiple comparisons, resulting in few candidate anchor items. In these analyses, one anchor item per subscale was used for linking in the real data analysis example (Section 2.4.1) and for the effect size estimation (Table 8). However, three items for depression and five for anxiety identified in the original unidimensional analyses of these data (Teresi et al., 2016a, 2016b) were also included as anchors in sensitivity analyses. As shown in this illustration, it was important to model the group differences in latent distributions.

An issue in PROMIS was the potential for inadequately modeled multidimensionality to result in false DIF detection. Nearly all PROMIS efforts used unidimensional models. Little work has been performed comparing unidimensional and multidimensional models in the context of DIF assessment and correlated traits. The simulations presented in Section 2.4.2 provided some evidence that while unidimensional and multidimensional approaches were similar in terms of in Type I error rates, power for DIF detection was greater for the multidimensional approach. Future work is needed to examine DIF detection and effect size integration in the context of polytomous, multidimensional data.

Funding U01AR057971 (PIs: Potosky, Moinpour), NCI P30CA051008, UL1TR000101 (previously UL1RR031975) from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through the Clinical and Translational Science Awards Program (CTSA). Analyses of these data were supported by the Mount Sinai Claude D. Pepper Older Americans Independence Center (National

Institute on Aging, 1P30AG028741, Siu) and the Columbia University Alzheimer's Disease Resource Center for Minority Aging Research (National Institute on Aging, 1P30AG059303, Manly, Luchsinger). This research was also supported by the Eunice Kennedy Shriver National Institutes of Child Health and Human Development of the National Institutes of Health under Award Number R01HD079439 to the Mayo Clinic in Rochester Minnesota through subcontracts to the University of Minnesota and the University of Washington. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Katja Ocepek-Welikson, M.Phil., for analytic assistance and Ruoyi Zhu, a doctoral student in the College of Education, University of Washington for assistance in conducting the simulation study.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

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. https://doi.org/10.1111/j.1745-3984.1999.tb00558.x.

Baker, F. B. (1995). *EQUATE 2.1: Computer program for equating two metrics in item response theory*. Madison: University of Wisconsin, Laboratory of Experimental Design.

Bauer, D., Belzak, W., & Cole, V. (2019). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling A: Multidisciplinary Journal*,. https://doi.org/10.1080/10705511.2019.1642754.

Belzak, W., & Bauer, D. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*,. https://doi.org/10.1027/met0000253.

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. https://doi.org/10.2307/2346101.

Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. (2014). Difference in method of administration did not significantly impact item response: An IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. *Quality of Life Research*, *23*, 217–227.

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

Boorsboom, D. (2006). Commentary: When does measurement invariance matter? *Medical Care*, *44*(11), S176–81.

Boorsboom, D., Mellenbergh, G. J., & van Heerdon, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, *26*, 433–450.

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*, 51. https://doi.org/10.3389/feduc.2017.00051.

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. https://doi.org/10.1037/0033-2909.105.3.456.

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. https://doi.org/10.1348/000711007X249603.

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

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]*. Lincolnwood, IL: Scientific Software International Inc.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253–260.

Carle, A. C., Cella, D., Cai, L., Choi, S. W., Crane, P. K., Curtis, S. M., et al. (2011). Advancing PROMIS's methodology: Results of the third PROMIS Psychometric Summit. *Expert Review of Pharmacoeconomics & Outcome Research*, *11*(6), 677–684. https://doi.org/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. https://doi.org/10.1097/01.mlr.0000258615.42478.55.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical software*, *48*(6), 1–29.

Chalmers, R. P. (2016). A differential response functioning framework for understanding item, bundle, and test bias. Doctoral Dissertation, York University, Toronto, Ontario. https://pdfs.semanticscholar.org

Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika*, *83*, 696–732. https://doi.org/10.1007/s11336-018-9626-9.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*, 114–140.

Chang, Y.-W., Hsu, N.-J., & Tsai, R.-C. (2017). Unifying differential item functioning in factor analysis for categorical data under a discretization of a normal variant. *Psychometrika*, *82*(2), 382–406. https://doi.org/10.1007/s11336-017-9562-0.

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. https://doi.org/10.1177/0146621613488643.

Cheng, C.-P., Chen, C.-C., & Shih, C.-L. (2020). An exploratory strategy to identify and define sources of differential item functioning. *Applied Psychological Measurement*, *4*, 548–560. https://doi.org/10.1177/014662/620931/90.

Cheng, Y., Shao, C., & Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanisms of DIF. *Educational and Psychological Measurement*, *76*(1), 43–63.

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

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

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*, 125–136.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel–Haenszel procedure. *Applied Measurement in Education*, *6*, 269–279.

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. https://doi.org/10.1177/014662169301700402.

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. https://doi.org/10.1177/014662169001400207.

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. https://doi.org/10.1097/01.mlr.0000245183.28384.ed.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., & Teresi, J. A. (2007). 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. https://doi.org/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. https://doi.org/10.1002/sim.1713.

Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*, *84*, 285–309. https://doi.org/10.1007/s11336-018-9549-2.

DeMars, C. E. (2010). Type 1 error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*, 961–972. https://doi.org/10.1177/0013164410366691.

DeMars, C. E. (2015). Modeling DIF for simulations: Continuous or categorical secondary trait? *Psychological Test and Assessment Modeling*, *57*, 279–300.

Edelen, M., Stucky, B., & Chandra, A. (2015). Quantifying "problematic" DIF within an IRT framework: Application to a cancer stigma index. *Quality of Life Research*, *24*, 95–103. https://doi.org/10.1007/s11136-013-0540-4.

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. https://doi.org/10.1177/0013164414520965.

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. https://doi.org/10.1177/0146621605275728.

Fleer, P. F. (1993). *A Monte Carlo assessment of a new measure of item and test bias* (p. 2266, Vol. 54, No. 04B), Illinois Institute of Technology, Dissertation Abstracts International.

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. https://doi.org/10.1177/01466219922031437.

Furlow, C. F., Ross, T. R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*(6), 441–464. https://doi.org/10.1177/0146621609331959.

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. https://doi.org/10.1177/0013164402239317.

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. https://doi.org/10.1027/1614-2241/a000046.

Gómez-Benito, J., Dolores-Hidalgo, M., & Zumbo, B. D. (2013). Effectiveness of combining statistical tests and effect sizes when using logistic discriminant function regression to detect differential item functioning for polytomous items.

Educational and Psychological Measurement, *73*, 875–897. https://doi.org/10.1177/0013164413492419.

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), S78–S94.

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://CRANR-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. https://doi.org/10.1177/0013164414523618.

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., Moinpour, C. M., Keegan, T. H. M., Cress, R. D., Wu, X.-C., Paddock, L. A., et al. (2016a). The Measuring Your Health Study: Leveraging community-based cancer registry recruitment to establish a large, diverse cohort of cancer survivors for analyses of measurement equivalence and validity of thepatient-reported Outcomes Measurement Information System®(PROMIS®) short form items. *Psychological Test and Assessment Modeling*, *58*(1), 99–117.

Jensen, R. E., King-Kallimanis, B. L., Sexton, E., Reeve, B. B., Moinpour, C. M., Potosky, A. L., et al. (2016b). Measurement properties of the PROMIS® Sleep Disturbance short form in a large, ethnically diverse cancer cohort. *Psychological Test and Assessment Modeling*, *58*(2), 353–370.

Jin, K. Y., Chen, H. F., & Wang, W. C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement*, *42*, 613–29.

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. https://doi.org/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. https://doi.org/10.1097/01.mlr.0000245250.50114.0f.

Jones, R. N. (2019). Differential item functioning and its relevance to epidemiology. *Current Epidemiology Reports*,. https://doi.org/10.1007/s40471-019-00194-5.

Jones, R. N., Tommet, D., Ramirez, M., Jensen, R. E., & Teresi, J. A. (2016). 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*, *58*(2), 371–402.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 408–426. https://doi.org/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. https://doi.org/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. https://doi.org/10.1207/S15327906347-387.

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

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

Kahraman, N., DeBoeck, P., & Janssen, R. (2009). Modeling DIF in complex response data using test design strategies. *International Journal of Testing*, *8*, 151–166. https://doi.org/10.1080/15305050902880744.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple group categorical CFA and IRT. *Structural Equation Modeling*, *18*, 212–228. https://doi.org/10.1080/10705511-2011.557337.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*, 469–492. https://doi.org/10.1177/0013164411427395.

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. https://doi.org/10.1177/014662169802200403.

Kim, S.-H., 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. https://doi.org/10.1111/j.1745-3984.2007.00029.x.

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling*, *58*, 79–98.

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. https://doi.org/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. https://doi.org/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=UNCb5878458.

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*, *77*(4), 545–569.

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. https://doi.org/10.1177/

0013164411432333.

Liu, Y., Magnus, B. E., & Thissen, D. (2016). Modeling and testing differential item functioning in unidimensional binary item response models with a single continuous covariate: A functional data analysis approach. *Psychometrika*, *81*, 371–398.

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. https://doi.org/10.1177/0146621608321760.

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

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.

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. https://doi.org/10.1177/014662169802200404.

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

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of IRT and CFA methodologies for establishing measurement equivalence. *Organizational Research Methods*, *7*, 361–388. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/a0027934.

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

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

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

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(Suppl 3), S69–S77. https://doi.org/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. https://doi.org/10.1177/014662169301700401.

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

Montoya, A. K., & Jeon, M. (2020). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, *44*(2), 118–136.

Mukherjee, S., Gibbons, L. E., Kristjansson, E., & Crane, P. K. (2013). Extension of an iterative hybrid ordinal logistic regression/item response theory approach to detect and account for differential item functioning in longitudinal data. *Psychological Test and Assessment Modeling*, *55*(2), 127–147.

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

Muthén, B. (1989). 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, L. K. & Muthén, B. O. (1998–2019). *M-PLUS Users Guide*. Sixth Edition. Los Angeles, California: Authors Muthén and Muthén.

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished Technical Report. Available at https://www.statmodel.com/wlscv.shtml.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257–274.

Oort, E. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107–124.

Orlando-Edelen, M., Stuckey, B. D., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: Application to a cancer stigma index. *Quality of Life Research*, *24*, 95–103. https://doi.org/10.1007/s11136-013-0540-4.

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

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2009). *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. https://doi.org/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. https://doi.org/10.1007/s11136-012-0292-6.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, Anxiety and Anger. *Assessment*, *18*, 263–283.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037//0021-9010.87.3.517.

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

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

Raykov, T., Marcoulides, G. A., Menold, N., & Harrison, M. (2019). Revisiting the bi-factor model: Can mixture modeling help assess its applicability? *Structural Equation Modeling*, *26*, 110–118.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (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. https://doi.org/10.1097/01.mlr.0000250483.85507.04.

Reeve, B. B., & Teresi, J. A. (2016). Overview to the two-part series: Measurement equivalence of the Patient Reported Outcomes Measurement Information System$^{@}$ (PROMIS)$^{@}$ short forms. *Psychological Test and Assessment Modeling*, *58*(1), 31–35.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. https://doi.org/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. https://doi.org/10.1037/0033-2909.114.3.552.

Rikis, D. R. J., & Oshima, T. C. (2017). Effect of purification procedures on DIF analysis in IRTPRO. *Educational and Psychological Measurement*, *77*, 415–428.

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

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

Rouquette, A., Hardouin, J. B., Vanhaesebrouck, A., Véronique Sébille, V., & Coste, J. (2019). Differential item functioning (DIF) in composite health measurement scale: Recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. *PLoS ONE*, *14*(4), e0215073. https://doi.org/10.1371/journal.pone.0215073.

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

Schalet, B. D., Pilkonis, P. A., Yu, L., Dodds, N., Johnston, K. L., Yount, S., et al. (2016). Clinical validity of PROMIS depression, anxiety and anger across diverse clinical groups. *Journal of Clinical Epidemiology*, *73*, 119–127. https://doi.org/10.1016/j.jclinepi2015.08.036.

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. https://doi.org/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. *Applied Psychological Measurement, 36*, 494–515. https://doi.org/10.1177/0146621612445182.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

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. https://doi.org/10.1177/0013164413520545.

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. https://doi.org/10.1177/0146621608321758.

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. https://doi.org/10.1037/0021-9010.89.3.497.

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. https://doi.org/10.1037/0021-9010.91.6.1292.

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. https://doi.org/10.1007/s11136-011-9969-5.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293–326.

Stout, W., Li, H., Nandakumar, R., & Bolt, D. (1997). MULTISIB—A procedure to investigate DIF when a test is intentionally multidimensional. *Applied Psychological Measurement*, *21*, 195–213.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316. https://doi.org/10.1007/s11366-013-9388-3.

Suh, Y., & Cho, S.-J. (2014). Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement*, *38*(5), 359–375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370. https://doi.org/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. https://doi.org/10.1007/BF02294363.

Taple, B. J., Griffith, J. W., & Wolf, M. S. (2019). Interview administration of PROMIS depression and anxiety short forms. *Health Literacy Research Practice*, *6*, e196–e204. https://doi.org/10.3928/24748307-20190626-01.

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. https://doi.org/10.1097/01.mlr.0000245142.74628.ab.

Teresi, J. A. (2019). *Applying and Acting on DIF*. Moderator at the 2019 PROMIS Psychometric Summit, Northwestern University, Chicago, IL.

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* (pp. 139–164). American Psychological Association: Washington, DC. https://doi.org/10.1037/14047-008.

Teresi, J. A., & Jones, R. N. (2016). 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 Form Measures". *Psychological Test and Assessment Modeling*, *58*(1), 37–78.

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.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., et al. (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. https://doi.org/10.1007/s11136-007-9186-4.

Teresi, J., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. E., Crane, P. K., Jones, R. N., et al. (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*(2), 148–180. PMCID: PMC2844669. NIHMSID: 136951.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). 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*(1), 141–181.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016b). Measurement equivalence of the Patient Reported Outcomes Measurement Information System®(PROMIS®) anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 183–219.

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. https://doi.org/10.1177/089826412436877.

Teresi, J. A., & Reeve, B. B. (2016). Epilogue to the two-part series: Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) short forms. *Psychological Tests and Assessment Modeling*, *58*(2), 423–433.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood Ratio Tests for Differential Item Functioning*. Unpublished manual from the L.L. Thurstone Psychometric

Laboratory: University of North Carolina at Chapel Hill.

Thissen, D. (1991). *MULTILOG$^{\mathrm{TM}}$ user's guide multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software Inc.

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. https://doi.org/10.3102/10769986027001077.

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

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.

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. https://doi.org/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.

Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2018). Score-based test of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*, *83*, 132–135. https://doi.org/10.1007/s11336-017-9591-8.

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. https://doi.org/10.3200/JEXE.72.3.221-261.

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*, 166–180. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0013164409332228.

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. https://doi.org/10.1177/0146621603259902.

Wang, M., & Woods, C. M. (2017). Anchor selection using the Wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*, 17–29. https://doi.org/10.1177/014662161666680l4.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42–57. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0146621610377450.

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. https://doi.org/10.1177/0013164412464875.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*, 339–361. https://doi.org/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. https://doi.org/10.1177/0146621614561313.

Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six PROMIS cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology*, *64*(5), 507–516.

Yu, Q., Medeiros, K. L., Wu, X., & Jensen, R. E. (2018). Nonlinear predictive models for multiple mediation analysis with an application to explore ethnic disparities in anxiety and depression among cancer survivors. *Psychometrika*, *83*, 991–1006.

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.

Zwitser, R. J., Glaser, S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8.

Manuscript Received: 11 JUN 2020
Final Version Received: 2 MAR 2021
Accepted: 19 MAY 2021
Published Online Date: 12 JUL 2021