Article

Journal of Educational and Behavioral Statistics
2022, Vol. 47, No. 6, pp. 666–692
DOI: 10.3102/10769986221109208
Article reuse guidelines: sagepub.com/journals-permissions

Testing Differential Item Functioning Without Predefined Anchor Items Using Robust Regression

Weimeng Wang
Yang Liu
University of Maryland

Hongyun Liu
Beijing Normal University

Differential item functioning (DIF) occurs when the probability of endorsing an item differs across groups for individuals with the same latent trait level. The presence of DIF items may jeopardize the validity of an instrument; therefore, it is crucial to identify DIF items in routine operations of educational assessment. While DIF detection procedures based on item response theory (IRT) have been widely used, a majority of IRT-based DIF tests assume predefined anchor (i.e., DIF-free) items. Not only is this assumption strong, but violations to it may also lead to erroneous inferences, for example, an inflated Type I error rate. We propose a general framework to define the effect sizes of DIF without a priori knowledge of anchor items. In particular, we quantify DIF by item-specific residuals from a regression model fitted to the true item parameters in respective groups. Moreover, the null distribution of the proposed test statistic using robust estimator can be derived analytically or approximated numerically even when there is a mix of DIF and non-DIF items, which yields asymptotically justified statistical inference. The Type I error rate and the power performance of the proposed procedure are evaluated and compared with the conventional likelihood-ratio DIF tests in a Monte Carlo experiment. Our simulation study has shown promising results in controlling Type I error rate and power of detecting DIF items. Even when there is a mix of DIF and non-DIF items, the true and false alarm rate can be well controlled when a robust regression estimator is used.

Keywords: *differential item functioning; multiple-group IRT; measurement invariance; robust regression; delta method; multiple imputation; likelihood-ratio test; implicit differentiation*

1. Introduction

In test theory, an item is said to exhibit differential item functioning (DIF) if individuals of the same ability level from different groups have unequal probabilities to select a given response to the item. The presence of DIF items may

greatly impact the validity of a test and thus jeopardize the inference drawn from the test score. For instance, if item parameters are the functions of group membership, between group differences in test scores may not reflect true difference in the trait but are merely due to the variation of its parameters across groups. Consequently, the test score may be biased for or against the examinees of specific groups. Therefore, it is crucial to identify DIF items in routine operations of the educational assessment.

While DIF detection methods based on the multiple-group item response theory (IRT) have been widely used (see, e.g., Glas, 1998; Lord, 1980; Thissen et al., 1993; Woods et al., 2013), many IRT-based DIF tests require knowing exactly which items are DIF-free items or anchor items[1] (e.g., Shih & Wang, 2009). Under the multiple-group IRT framework (Bock & Zimowski, 1997), item parameters of the underlying IRT model from different groups must be placed on a common scale before any DIF test statistics can be implemented to test the statistical significance of the difference. In doing so, differences in item response functions of different groups are solely due to item parameters irrespective of any potential differences in the underlying distributions of the latent trait across groups. For this purpose, anchor items must be explicitly chosen for the majority of the IRT DIF detection methods. Ideally, anchor items are supposed to be DIF-free to avoid inflated false alarm rate. Numerous studies have shown the critical role of DIF-free anchors in making a correct statistical inference in DIF detection. For instance, one of the most commonly used anchoring methods is the *all-other* anchor method, which uses all other items as anchors except the studied items (see, e.g., Cohen et al., 1996; Kim & Cohen, 1998). However, simulation studies have found that the all-other method only performs well when DIF is balanced[2] or when there is no or very few DIF items; otherwise, the Type I error rate of detecting DIF items can be inflated (Finch, 2005; Wang, 2004; Wang & Yeh, 2003; Woods, 2009). The reason is that if there are DIF items, the latent variable distributions are incorrectly estimated, which further leads to spurious between-group discrepancies in estimated item parameters for non-DIF items. Consequently, the false alarm rates of the DIF tests may be seriously inflated (see Kopf et al., 2013; Wang, 2004, for illustrative examples with scale shift due to anchor contamination).

Previous studies on DIF analysis without predefined anchor items can be roughly categorized into two-step and one-step approaches. Two-step approaches focus on the anchor selection strategy that firstly determines which items should be used as anchors, and then, the DIF detection is carried out at the second step (for overview, see Kopf et al., 2015; Shih & Wang, 2009). The one-step approach identifies DIF items directly without explicitly specifying anchor items (e.g., Frederickx et al., 2010; Magis et al., 2015; Strobl et al., 2015; Tutz & Berger, 2016).

In general, the anchor item selection strategy (e.g., Kopf et al., 2015; Wang et al., 2012; Wang & Su, 2004; Woods, 2009) aims to select the anchor items

empirically via an iterative procedure, which is often commonly referred to as item purification in the literature. IRT-based item purification alternates between item calibration and removal of DIF items until no further DIF items can be identified. The selected DIF-free anchor items are then used as a common scale for the final DIF detection for the rest of the items. Nevertheless, the stepwise anchor selection method is not only time inefficient but also lacks theoretical justifications. Moreover, in practice, it is difficult to implement an appropriate anchor selection strategy. Kopf et al. (2015) found that the optimal anchor selection strategy depends on the sample size, the proportion of DIF items, the direction of the DIF, and also the anchor length. On the contrary to the stepwise solution, one-step approaches have a sound statistical justification. Approaches of this type include but are not limited to the regularized DIF detection method using Lasso (e.g., Belzak & Bauer, 2020; Magis et al., 2015). However, drawing statistical inference (e.g., hypothesis testing and interval estimation) based on those methods can be challenging.

The current work develops an alternative one-step DIF detection approach that does not count on a priori knowledge of anchor items. Instead, an automatic process in finding the anchor items to link the latent scale of the two groups (i.e., a subset of DIF-free anchor items) is proposed. Specifically, it requires separate calibration of item parameters using item response data from the two groups. Then, a *reference line* is determined by regressing one set of item parameters onto the other, preferably using robust regression methods. Lastly, the test statistics can be formularized as residuals from the reference line determined by the majority of item parameters. We propose DIF test statistics with justified asymptotic properties and evaluate their finite-sample performance via a simulation study. We focus on the most commonly used IRT model—two-parameter logistic (2PL) model (Birnbaum, 1968) for notational conciseness and computational convenience. It is worth mentioning that placing a reference line through the items, preferably DIF free items, is commonly used in the Rasch modeling (see the graphical tests in Wright & Stone, 1999). Similarly, treating DIF effects as residuals of a linear regression model has been mentioned before but in a different setting. Robitzsch and Lüdtke (2020) proposed a robust linking approach based on a robust regression to estimate the mean of the latent ability of the focus group in the Rasch modeling framework. The unique contribution of the current study is the automatic process in finding the reference line determined by the subset of DIF-free items (i.e., anchor items) and extensions to models outside of the Rasch family.

The rest of this article is organized as follows. First, the test statistic is defined in a general form. Then, the asymptotic distribution of the proposed test statistic is derived analytically or approximated using the multiple imputation procedure (Yang et al., 2012). Lastly, the performance of such a test statistic is assessed by a Monte Carlo simulation.

2. A New DIF Detection Method

2.1. Multiple-Group IRT and DIF

Considering a dichotomous response to item j for examinee i from group g, denoted $Y_{ij}^{(g)} \in \{0, 1\}$, $i = 1, \ldots, N_g$, $j = 1, \ldots, J$, $g = 1, 2$. The 2PL model specifies the probability of endorsing an item ($Y_{ij}^{(g)} = 1$) given the latent ability $\theta_i^{(g)}$ as

$$P\{Y_{ij}^{(g)} = 1 | \theta_i^{(g)} = \theta\} = \frac{\exp\left[a_j^{(g)}\left(\theta - b_j^{(g)}\right)\right]}{1 + \exp\left[a_j^{(g)}\left(\theta - b_j^{(g)}\right)\right]}, \tag{1}$$

in which $a_j^{(g)}$ and $b_j^{(g)}$ denote item discrimination and difficulty parameters for group g. Assume that $\theta_i^{(1)} \sim N(0, 1)$ and $\theta_i^{(2)} \sim \mathcal{N}(\mu, \sigma^2)$. For the second group, it is possible to write $\theta_i^{(2)} = \sigma\tilde{\theta}_i^{(2)} + \mu$, where $\tilde{\theta}_i^{(2)} \sim \mathcal{N}(0, 1)$. When expressed in terms of the standard normal $\tilde{\theta}_i^{(2)}$, the transformed discrimination and difficulty parameters in the second group satisfy $\tilde{a}_j^{(2)} = \sigma a_j^{(2)}$ and $\tilde{b}_j^{(2)} = \frac{b_j^{(2)} - \mu}{\sigma}$. Define the anchor set for "a-DIF" as

$$\mathcal{A} = \{j = 1, \cdots, J : a_j^{(1)} = a_j^{(2)} = \frac{\tilde{a}_j^{(2)}}{\sigma}\}. \tag{2}$$

Similarly, "b-DIF" is

$$\mathcal{B} = \{j = 1, \cdots, J : b_j^{(1)} = b_j^{(2)} = \tilde{b}_j^{(2)}\sigma + \mu\}. \tag{3}$$

Equations 2 and 3 imply that item parameters of the two groups, when calibrated separately, fall on a straight line if items are DIF-free (see also Stocking & Lord, 1983). Specifically, item discriminations $a_j^{(1)}$ and $\tilde{a}_j^{(2)}$ fall on a line that passes through the origin; meanwhile, item difficulties $b_j^{(1)}$ and $\tilde{b}_j^{(2)}$ also fall on a line, which has the inverse slope but does not necessarily pass through the origin. Any deviation of the item from the line indicates DIF. Anchor sets $\mathcal{A}$ and $\mathcal{B}$ do not have to be exactly the same. We further discuss how the two anchor sets can help define DIF effect sizes. In the sequel, we use the term "separate calibration" to refer to item calibration based on two groups of data separately using the same measurement model with the standard normal latent variable.

Let $\boldsymbol{\xi}_j^{(1)} = (a_j^{(1)}, b_j^{(1)})'$ and $\tilde{\boldsymbol{\xi}}_j^{(2)} = (\tilde{a}_j^{(2)}, \tilde{b}_j^{(2)})'$. By Equations 2 and 3, we define item j's effect size of DIF $\delta_j = (\delta_j^a, \delta_j^b)'$ as the deviation of $(\boldsymbol{\xi}_j^{(1)}, \tilde{\boldsymbol{\xi}}_j^{(2)})'$ from the reference line determined by anchor items. As the slope of the reference line for

"a-DIF" is the inverse of the slope of the reference line for "b-DIF," we have the following two ways of defining the effect size of DIF differing by the specific item parameters used to locate the respective reference line. First, the effect size of "a-DIF" and "b-DIF" can be defined as the deviation from the reference line determined by item parameters $a_{\mathcal{A}}$ and $b_{\mathcal{B}}$, respectively. Specifically, item j's effect size of "a-DIF" (δ_j^a) quantifies the deviation of $(a_j^{(1)}, \tilde{a}_j^{(2)})'$ from the line determined by $a_{\mathcal{A}}$, whereas the item j's effect size of "b-DIF" (δ_j^b) quantifies the deviation of $(b_j^{(1)}, \tilde{b}_j^{(2)})'$ from the line determined by $b_{\mathcal{B}}$ that does not necessarily pass through the origin. Alternatively, the effect size of "a-DIF" or "b-DIF" can be defined as the deviation from the reference line determined by both the item slope and item intercept parameters $a_{\mathcal{A}}$ and $b_{\mathcal{B}}$. For example, to find the reference line for "b-DIF", the slope can be first obtained by inverting the slope of the reference line of the "a-DIF", after which the intercept can be estimated using $b_{\mathcal{B}}$ by fixing the slope. For ease of illustration, we focus on the first way of defining effect size, and the derivation of the second method is documented in the Online Supplemental Material. The pros and cons of each method are discussed in the result section.

Given the current framework of defining DIF, a critical step is to determine the reference line, which in turn can facilitate quantifying the effect size of DIF (i.e., deviation from the reference line). Ideally, the line should only be determined by item parameters of a subset of items which are DIF free, so that the effect size of the non-DIF item is zero, whereas the effect size of the DIF item is larger than zero. To illustrate our idea, Figure 1 displays three examples for the effect size of "a-DIF." Item discrimination parameters for Group 1 are plotted against those for Group 2 estimated from separate calibration. The reference line is estimated using the ordinary least square (OLS) method and the least trimmed square (LTS) method. Under the null hypothesis (the figure on the left in Figure 1), both methods consistently reach the same reference line. On the contrary, when there is a mix of DIF and non-DIF items (see figure on the right in Figure 1), the two methods are slightly different. In particular, the OLS is sensitive to "outliers," which are the DIF items in this case. Consequently, the resulting reference line tilts toward DIF items and thus creates nonzero effect sizes for non-DIF items. A more robust method—LTS method—is resistant to the influence of DIF items, which creates a reference line by only using a subset of the items (usually $\geq 50\%$). Therefore, effect sizes of non-DIF items are zero, and DIF items do not fall on the reference line. For the effect size of DIF, technically any statistics that can quantify the deviation of the two DIF items from the dashed line can be utilized as the effect sizes of DIF. Three examples are demonstrated in the graph including the vertical distance, the perpendicular distance, and the horizontal distance. In the next section, a general formulation of the test statistic and three specific examples are provided.

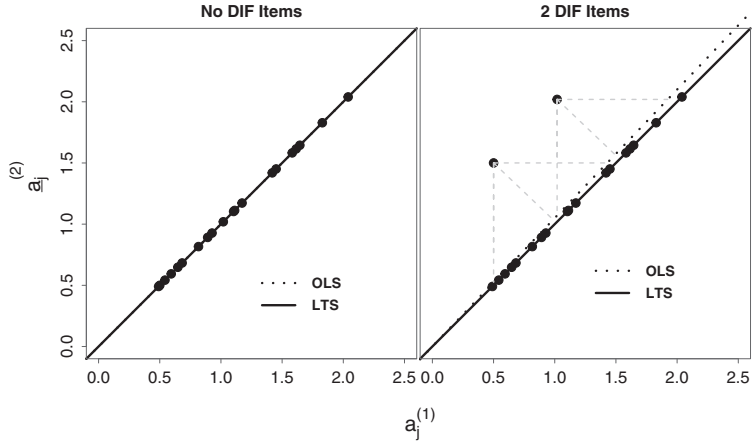

FIGURE 1. *Illustration of three examples of effect sizes of "a-DIF" for item j. The item parameters from separate calibration are plotted against each other. Two methods of obtaining the reference line are shown here: (1) OLS = the ordinary least square method (dotted line) and (2) LTS = the least trimmed square method (solid line). Left figure shows the results, when there is no DIF items, both methods end up with the same reference line. Right figure shows the results with a mix of DIF and non-DIF items. Dashed lines indicate different effect sizes of "a-DIF" vertical distance, perpendicular distance, and horizontal distance. DIF = differential item functioning.*

2.2. A General Formulation of the Test Statistics

In general, the intercept (α) and the slope (β) of the reference line can be expressed as functions of the item parameters and can be obtained by minimizing a specific objective function over α and β:

$$F(\boldsymbol{\xi}, \alpha, \beta) := \sum_{j=1}^{J} \rho\big(\boldsymbol{\tau}_j(\boldsymbol{\xi})\big), \tag{4}$$

where $\boldsymbol{\tau}_j$ is a function of $\boldsymbol{\xi}$ quantifying the deviation from the line for item j and ρ quantifies the contribution of each residual to the objective function (F). If ρ and $\boldsymbol{\tau}_j$ are differentiable, the intercept α and slope β of the reference line necessarily solve the estimating equations

$$\begin{aligned}
\frac{\partial F(\boldsymbol{\xi}, \alpha, \beta)}{\partial(\alpha, \beta)} &= \frac{\partial \sum_{j=1}^{J} \rho\big(\boldsymbol{\tau}_j(\boldsymbol{\xi})\big)}{\partial(\alpha, \beta)}, \\
&= \left(\sum_{j=1}^{J} \varphi\big(\boldsymbol{\tau}_j(\boldsymbol{\xi})\big) \frac{\partial \boldsymbol{\tau}_j(\boldsymbol{\xi})}{\partial \alpha}, \quad \sum_{j=1}^{J} \varphi\big(\boldsymbol{\tau}_j(\boldsymbol{\xi})\big) \frac{\partial \boldsymbol{\tau}_j(\boldsymbol{\xi})}{\partial \beta} \right), \\
&= \mathbf{0},
\end{aligned} \tag{5}$$

in which $\varphi = \frac{\partial \rho(\boldsymbol{\tau}_j)}{\partial \boldsymbol{\tau}_j}$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}^{(1)\prime}, \tilde{\boldsymbol{\xi}}^{(2)\prime})'$ are the true item parameters in separate calibration, and $\boldsymbol{\xi}^{(1)} = (\boldsymbol{\xi}_1^{(1)}, \ldots, \boldsymbol{\xi}_J^{(1)})'$ and $\tilde{\boldsymbol{\xi}}^{(2)} = (\tilde{\boldsymbol{\xi}}_1^{(2)}, \ldots, \tilde{\boldsymbol{\xi}}_J^{(2)})'$. Then, item j's effect size of DIF is a function of $\boldsymbol{\xi}$, denoted $\boldsymbol{\tau}_j(\boldsymbol{\xi})$, which measures the vertical distance[3] of $(\boldsymbol{\xi}_j^{(1)}, \tilde{\boldsymbol{\xi}}_j^{(2)})'$ from a reference line determined by (possibly a subset of) $\boldsymbol{\xi}$ using Equation 6:

$$\boldsymbol{\tau}_j(\boldsymbol{\xi}) = (\hat{\beta}\tilde{\boldsymbol{\xi}}^{(2)} + \hat{\alpha}) - \boldsymbol{\xi}^{(1)}, \tag{6}$$

in which $\hat{\alpha}$ and $\hat{\beta}$ denote the estimated intercept and the slope of the reference line, respectively. Notice that the intercept of the reference line for "a-DIF" is zero as it passes through the origin.

Although we use a more general notation using $\boldsymbol{\xi}$, a-DIF and b-DIF are investigated separately and contain their own objective functions. In principle, we can find the two regression lines simultaneously by defining an objective function that combines the residuals for a and b parameters and minimizing the function with respect to one slope and one intercept. However, the a and b parameters typically have different scales. An extreme example is that all the b_js are close to zero and all the a_js are close to 1. Simply adding up all the residual terms, which allows the residuals for a and b to equally contribute to the objective function, is then problematic. As a result, we recommend investigating a-DIF and b-DIF separately as described in Section 2.1. In particular, our simulation shows that the two-step approach: (1) determining the slope of the reference line of a-DIF based on the a parameters only and (2) fixing the slope of the reference line of b-DIF as the inverse of estimated slope for the reference line of a-DIF and determining the intercept of the reference line of b-DIF using only the b parameters, performs the best.

2.3. Connection With Conventional DIF Assessment

So far, we have discussed the proposed test statistic. Before introducing the statistical tests, it is critical to highlight the similarities and differences with the traditional definition of the DIF effect size. Comparing with the conventional definition of the DIF effect size, the proposed definition of effect size as a deviation from the reference line obtained by minimizing an objective function is more general (see Equation 4). The conventional multiple group IRT method sets the latent scale by the anchor items, which is similar as drawing the reference line passing through a set of equally weighted and predetermined nonempty anchor sets $\mathcal{A}_0 \subseteq \mathcal{A}$ and $\mathcal{B}_0 \subseteq \mathcal{B}$. In this case, $\rho(\boldsymbol{\tau}_j(\boldsymbol{\xi})) = (\boldsymbol{\tau}_j(\boldsymbol{\xi}))^2$ for all $j \in \mathcal{A}_0$ for a-DIF and for all $j \in \mathcal{B}_0$ for b-DIF. As a comparison, our proposed method finds the subsets of DIF-free items by using 50% of the item (e.g., LTS method) or by tuning (e.g., bisquare method), which does not necessarily weight items equally. In this regard, the traditional methods of using fixed anchors can be viewed as special cases of our general

definition of estimating equations. Despite the slight difference, we want to highlight the similarity between the proposed method and the traditional approach. The subset of DIF-free items, which locates the reference line, are anchor items because the latent scale is also set by the same subset of the DIF-free items, which further determines the reference line as a function of the latent scale (i.e., mean and variance of the second group). It is the engine to locate these anchor items that makes the difference between the new and the traditional.

With that being said, we still have to impose an additional assumption that the majority of the items should be DIF-free to ensure that the reference line goes through the anchor items. The assumption of the majority of the non-DIF items is implicitly required by the limit of the breakdown point in the robust regression. For robust regression methods in the current study, we require $> 50\%$ of the items are DIF-free. For example, the upper bound of the breakdown point of LTS in this case approximately 50% if 50% of the observation is trimmed (see Rousseeuw & Leroy, 1987, Theorem 4 in Chapter 3). Although the bisquare method could potentially have a larger break down point controlled by the tuning parameter k, without the 50% of the DIF-free item assumption, it still cannot guarantee that the reference line passes through the non-DIF items. Therefore, the assumption is imposed to ensure that the solution, even if it exists, is accurate, so that the effect sizes of the DIF-free items are zero and those for DIF items are nonzero. Violations to this assumption could result in erroneous inferences under extreme cases. Considering an extreme scenario where all DIF items perfectly fall on Line 1 and the anchor items fall on Line 2 as displayed in Figure 2, if more than 50% of the items are DIF items, Line 1 will be the solution that minimizes the objective function even when a robust ρ is chosen. As a result, DIF items can be incorrectly identified as non-DIF items and non-DIF items can be incorrectly identified as DIF-items.

2.4. Sampling Variability

As the item parameters of IRT models are usually estimated using the maximum likelihood estimator, the sampling variability of item parameters should be taken into account to derive the sampling variability of the test statistic. Under suitable regularity conditions (Birch, 1964), the maximum likelihood estimator[4] of item parameters, denoted $\hat{\boldsymbol{\xi}}$, is asymptotically normal:

$$\sqrt{N}\left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}\right), \tag{7}$$

in which $\boldsymbol{\mathcal{I}}$ is the Fisher information matrix[5] evaluated at the true item parameter $\boldsymbol{\xi}$, and $N = N_1 + N_2$. We assume that $N_1/N_2 \to c$, where $c > 0$ is a constant. Then, the sampling variability of the test statistic can be derived or approximated depending on the nature of $\mathbf{g}$. For example, function $\mathbf{g}(\boldsymbol{\xi})$ might be differentiable explicit functions with closed-form expressions (e.g., OLS), differentiable

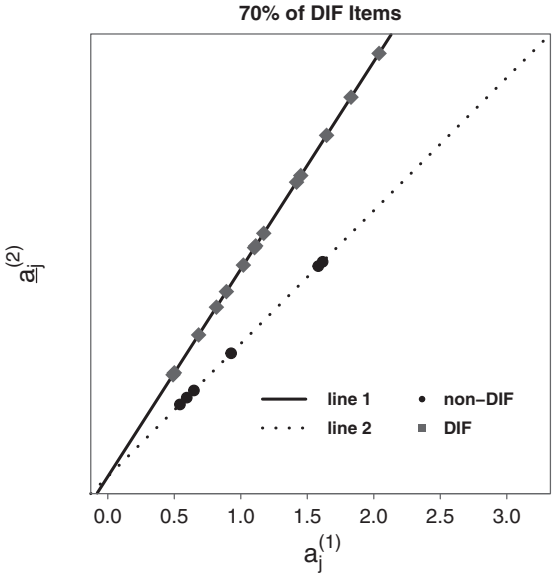

FIGURE 2. *Illustration of an extreme example where more than 50% of the items are DIF items. The item parameters from separate calibration are plotted against each other. Solid line (Line 1) indicates the reference line obtained by the least trimmed square method, which goes through the DIF items. Dotted line (Line 2) indicates the correct reference line, which goes through the anchor items. Gray squares show the DIF items and black circles show the non-DIF items. DIF = differential item functioning.*

implicit functions (e.g., bisquare loss regression), or other nonregular functions (e.g., LTS). The direct and implicit delta methods can be applied for the first two cases, respectively. The multiple imputation procedure can be applied to the last case by simulation. The asymptotic distributions of the test statistics of the three cases are derived or approximated as follows.

2.4.1. Differentiable explicit functions. If the function, $\boldsymbol{\tau}_j(\cdot)$ has nonzero first derivatives, denoted $\boldsymbol{\nabla}\boldsymbol{\tau}_j(\cdot)$, in some neighborhood of $\boldsymbol{\xi}$; then, the multivariate delta method (Bickel & Doksum, 2015, p. 319) implies that $\boldsymbol{\tau}_j(\hat{\boldsymbol{\xi}})$ is also asymptotically normal:

$$\sqrt{N}\left(\boldsymbol{\tau}_j(\hat{\boldsymbol{\xi}}) - \boldsymbol{\tau}_j(\boldsymbol{\xi})\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\nabla}\boldsymbol{\tau}_j(\boldsymbol{\xi})\mathcal{I}^{-1}\boldsymbol{\nabla}\boldsymbol{\tau}_j(\boldsymbol{\xi})'\right). \quad (8)$$

For example, the reference line can be obtained by the OLS regression. Then, $\boldsymbol{\tau}_j(\boldsymbol{\xi})$ is the perpendicular distance shown in Equation 6 and, thus, can be expressed as a closed-form function of the item parameters. In particular, $\mathbf{g}(\boldsymbol{\xi})$ and $\nabla\boldsymbol{\tau}_j(\boldsymbol{\xi})$ are documented in Online Appendix A for both "a-DIF" and "b-DIF".

2.4.2. Differentiable implicit functions. If the function $\boldsymbol{\tau}_j(\cdot)$ has nonzero first derivatives but $\mathbf{g}(\boldsymbol{\xi})$ is differentiable implicit functions, then the implicit delta method can be applied. For example, the robust estimator bisquare loss regression's ρ is defined as

$$\rho(e) = \begin{cases} \frac{k^2}{6} 1 - [1 - (\frac{e}{k})^2]^3, & \text{for } |e| \leq k, \\ \frac{k^2}{6}, & \text{for } |e| > k. \end{cases} \tag{9}$$

Then, $\varphi(e)$ is given by

$$\varphi(e) = \begin{cases} [1 - (\frac{e}{k})^2]^2 e, & \text{for } |e| \leq k, \\ 0, & \text{for } |e| > k. \end{cases} \tag{10}$$

Solving the estimating Equation 4 with φ specified in Equation 10 involves an iterative process. Therefore, $(\alpha, \beta) = (g_1(\boldsymbol{\xi}), g_2(\boldsymbol{\xi}))$ are implicit functions. Further, $\nabla \boldsymbol{\tau}_j(\boldsymbol{\xi})$ can be expressed as a function of $\frac{\partial \mathbf{g}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$, which can be then obtained by the implicit function theorem (Rudin, 1964, Chapter 9) as follows:

$$\frac{\partial \mathbf{g}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \frac{\partial [g_1(\boldsymbol{\xi}), g_2(\boldsymbol{\xi})]}{\partial \boldsymbol{\xi}} = -\left[\frac{\partial \mathbf{F}(\boldsymbol{\xi}, g_1(\boldsymbol{\xi}), g_2(\boldsymbol{\xi}))}{\partial (\alpha, \beta)}\right]^{-1} \frac{\partial \mathbf{F}(\boldsymbol{\xi}, g_1(\boldsymbol{\xi}), g_2(\boldsymbol{\xi}))}{\partial \boldsymbol{\xi}}. \tag{11}$$

Finally, the sampling variability of $\boldsymbol{\tau}_j(\hat{\boldsymbol{\xi}})$ can be obtained by the multivariate delta method as in Equation 8. The complete analytical solution is documented in the Online Appendix B.

From Equation 10, we can see that the magnitude of k controls which item is contributing to locating the reference line. Thus, items whose residuals (effect sizes) are small (i.e., $\leq k$) function similarly as the anchor items but with weights. As a result, finding the value of k is critical in locating anchor items, the reference line, and DIF items. We, therefore, recommend two ways to find the value of k. First, a relatively simple way is to use a sufficiently small value. Ideally, the value of k should be close to the smallest true effect sizes of DIF items to ensure that there is no false inclusion of DIF items in the anchor set in large samples. Our simulation study has shown that using a simple value of .2, for example or any k that is smaller than the smallest true effect size, as the sample size becomes sufficiently large, the performance in detecting DIF items is reasonably well. Alternatively, one can fix k at the median of the absolute residuals using the OLS method. Given our assumption that the majority of items are DIF-free, the empirical investigation of k using the median residual is reasonable. Our simulation study shows that the two methods yield nearly identical results (see

Figure 1 in the Online Supplemental Material for detail). The data generation and manipulating factors are documented in Section 3.1.

2.4.3. Nonregular functions. Alternatively, when $\boldsymbol{\tau}_j(\cdot)$ is not differentiable or the analytic derivatives cannot be easily computed, the sampling variation of the test statistic is approximated using a multiple imputation procedure. For example, the reference line can be estimated using the LTS—a robust estimator that minimizes the residual sum of squares using only 50% of the data. Notice that the LTS in the traditional multiple regression setting has a trimming parameter which ranges from $J/2$ to J to decide the breakdown point of the LTS estimator. However, in the current setting, we did not manipulate the trimming parameter of the LTS. It is suggested to trim 50% of the observations. To approximate the sampling variability of the test statistic, multiple sets of plausible item parameters are drawn from $N(\hat{\boldsymbol{\xi}}, (N\hat{\boldsymbol{\mathcal{I}}})^{-1})$. $\boldsymbol{\tau}_j(\cdot)$ is then evaluated at the imputed parameters, resulting in an approximation to the sampling distribution (Yang et al., 2012).

3. Simulation

The goal of the current simulation study is twofold: (1) to evaluate the finite-sample performance of the proposed DIF detection method in terms of its Type I error rate and power of detecting DIF items and (2) to compare its performance with the state-of-the-art IRT DIF detection method using the likelihood ratio test (Thissen et al., 1993). Item purification method using iterative backward scheme is also applied to the current simulation as a benchmark. Specifically, three examples of effect sizes of DIF corresponding to three ways of estimating the reference line were investigated under two conditions: (1) the null condition where all items were DIF-free and (2) the alternative condition where there was a mix of DIF and non-DIF items. Item parameters were estimated using the R package mirt (Chalmers, 2012). All other computations were performed in the statistical computing environment R (R Core Team, 2019).

3.1. Simulation Setup

3.1.1. Manipulating factors. Five factors were manipulated in the current simulation design including (1) the total sample size ($N_1 + N_2 = 1{,}000$ and $2{,}000$), (2) the ratio between two groups $\frac{N_1}{N_2} = 1$ and 3, (3) the percentage of DIF items (0%, 10%, and 40%), (4) the effect size of DIF (small and large), and (5) the direction of DIF (balanced and unbalanced DIF). Together, there were four null conditions with 0% DIF item. For the alternative conditions, all manipulated factors were fully crossed with DIF items, which ended up with a total of 32 conditions in total ($2^5 = 32$). All varying factors across conditions were selected as a result of their relevance to the performance of the proposed method. Furthermore, different levels of the factors were chosen for realistic considerations to

improve generalizability. Each of the manipulated factor and its potential impact on the performance of the proposed DIF test statistic are discussed below.

Particularly, the sample size of each group can influence the standard error estimates of the item parameter, which in turn can impact the reference line estimates and, thus, impact the test statistic. Besides, by manipulating the sample size, the finite sample performance of the proposed test statistic can be evaluated. Specifically, four levels of the sample size condition are considered, small equal ($N_1 = N_2 = 500$), large equal ($N_1 = N_2 = 1{,}000$), unequal small ($N_1 = 750$ and $N_2 = 250$), and unequal large ($N_1 = 1{,}500$ and $N_2 = 500$). The chosen magnitude of the sample size is typically observed in both the applied research and the methodology research related to DIF (e.g., Bolt et al., 2004; Chan et al., 2004; Magis et al., 2010; Rodebaugh et al., 2006; Wang et al., 2012; Woods et al., 2013). For the sample size ratio between the two groups in comparison, it mimics a real DIF detection scenario between different ethnicity groups in the United States. For example, the two groups in comparison can be Caucasians versus Hispanic, which usually takes up 60% and 20% of the population in a testing context (Woods et al., 2013). In addition, the proportion of DIF items may greatly impact the power and the false alarm rate in detecting DIF items under the alternative condition where the percentage of DIF items varies from 10% to 40%. It is anticipated that the higher the percentage is, the better the performance of the test statistic based on the robust estimator is as compared with its nonrobust estimator counterparts. By increasing the percentage of DIF items, the test statistic based on the regular regression estimator is likely to suffer from inflated false alarm rate. For the same reason, the effect size of the DIF is also manipulated. For both a-DIF and b-DIF, two realistic values were considered with $\delta_j = |b_j^{(1)} - b_j^{(2)}| = |a_j^{(1)} - a_j^{(2)}|$ centered around .4 and .8 to represent a moderate and a large effect size for both a-DIF and b-DIF, which are typically considered in the DIF simulation study (Langer, 2008; Magis & De Boeck, 2012; Wang & Yeh, 2003; Woods et al., 2013) and observed in the applied research context. For each, $\delta_j \sim \mathcal{U}(0.3, 0.4)$ or $\mathcal{U}(0.7, 0.9)$. Lastly, the direction of the DIF, the sign of the effect size of DIF (δ_j) representing whether all DIF items are favoring one group, is manipulated. Two cases are considered, the unbalanced DIF and balanced DIF. The unbalanced DIF will only, most likely, favor the reference group, indicating that $\delta_j \geq 0$. However, this is not always the case in practice (Wang & Yeh, 2003). For example, it is unlikely that all items will favor the male group instead of the female group. Consequently, in the balanced DIF case, the DIF items can favor any groups, but on average, no group has an advantage at the test level. In other words, the average of DIF effect size is 0 ($\bar{\delta}_j = 0$). The balanced DIF case was achieved by setting 50% of the DIF items favoring one group and the rest favoring the other. We would expect that test statistics based on the robust estimator will outperform those based on regular regression methods under the unbalanced DIF case. On the contrary, the

robust estimator and nonrobust estimator are expected to perform equally well under the balanced case.

3.1.2. Data generating model. The binary response for item j for examinee i from group g, denoted $Y_{ij}^{(g)} \in \{0, 1\}$, $i = 1, \ldots, N_g$, $j = 1, \ldots, J$, $g = 1, 2$, was generated from a 2PL model (Birnbaum, 1968; see Equation 1). The total number of items was fixed at 20. The data generating item parameters are visualized in Figure 3. These values were randomly sampled from the following distribution: for each j, $a_j^{(2)} = a_j^{(1)} \sim \mathcal{LN}(0, 0.3^2)$; $b_j^{(2)} = b_j^{(1)} \sim \mathcal{N}(0, 1)$, which resembles the item parameter distribution from a simulation study (Langer, 2008), which generates similar item parameters published by Lord and Novick (2008). Under the alternative condition, the item discrimination parameters of the DIF items for the second group were increased by an effect size of $\delta_j = 0.4$ or 0.8. The latent variables of Groups 1 and 2 were generated from $\mathcal{N}(0, 1)$ and $\mathcal{N}(0.5, 1.5^2)$, respectively. A total of 1,000 replications for each condition were implemented.

3.1.3. Proposed test statistics. The DIF analysis based on the proposed method consists the following steps: (1) conducting separate item calibration and obtaining the maximum likelihood estimates of item parameters for the two groups $\boldsymbol{\xi}^{(1)} = (a^{(1)\prime}; b^{(1)\prime})'$ and $\tilde{\boldsymbol{\xi}}^{(2)} = (\tilde{a}^{(2)\prime}; \tilde{b}^{(2)\prime})'$, (2) regressing $\boldsymbol{\xi}^{(1)}$ on $\tilde{\boldsymbol{\xi}}^{(2)}$ and calculating the test statistic $\boldsymbol{\tau}_j(\hat{\boldsymbol{\xi}})$, and (3) conducting hypothesis test using either the analytical solution or approximated sampling distribution of the test statistic.

For each replicated data set, the aforementioned three methods—the OLS, the LTS, and the bisquare—were implemented. For each method, separate item calibration was first conducted using the maximum likelihood estimation with the Expectation Maximization (EM) algorithm. The convergence criterion for the EM cycle was set to be 10^{-4}. In the meantime, we adopted the default maximum number of iterations (NCYCLES = 500). The marginal likelihood function was approximated by a 61-point equally spaced rectangular quadrature points ranging from -6 to 6 (theta_lim = c(-6, 6)), quadpts = 61). The Fisher information matrix $\boldsymbol{\mathcal{I}}$ was estimated by the observed information matrix using a central difference approximation method (SE.type = "Oakes"). The test statistics $\boldsymbol{\tau}_j(\boldsymbol{\xi})$ were the vertical distance of $(\boldsymbol{\xi}_j^{(1)}, \tilde{\boldsymbol{\xi}}_j^{(2)})'$ from the reference line obtained using different estimating equations. For the OLS method, the reference line was determined by minimizing the sum of squares of the vertical distances. The sampling variation was derived using the multivariate delta method as described in Section 2.4. The reference line for the LTS method was obtained by minimizing the sum of squares of the vertical distance using only 50% of the data. The sampling variability was approximated using the multiple imputation procedure. Specifically, 1,000 plausible item parameters were drawn from $\mathcal{N}(\hat{\boldsymbol{\xi}}, (N\hat{\boldsymbol{\mathcal{I}}})^{-1})$;

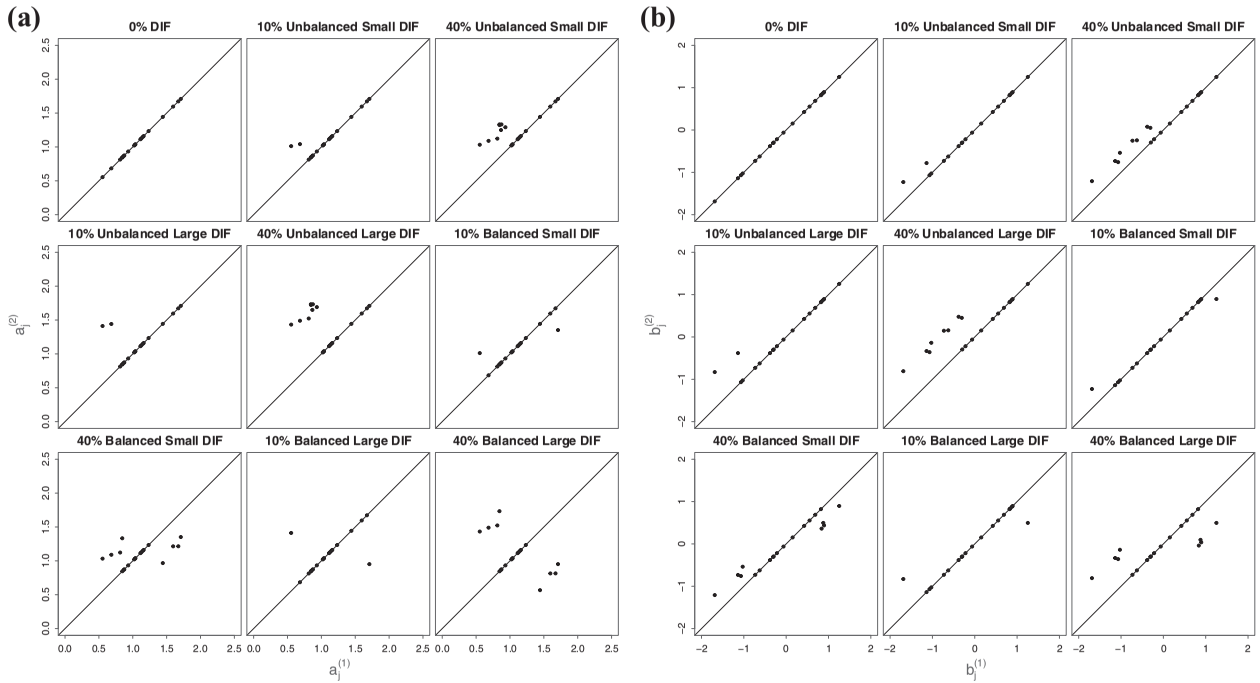

FIGURE 3. *Scatterplot of the true item parameters for 30 items under all conditions. Left: "a-DIF." Right: "b-DIF." Item parameters including the item discrimination parameter* $a_j^{(1)} = a_j^{(2)}$ *and item difficulty parameter* $b_j^{(1)} = b_j^{(2)}$. *These values are fixed across replications. The mean of* $a_j^{(1)} = a_j^{(2)} = 1.02$ *and the mean of* $b_j^{(1)} = b_j^{(2)} = 0.18$ *for Group 1. DIF = differential item functioning. (a) Item Discrimination Parameters. (b) Item Difficulty Parameters.*

1,000 sets of test statistics were then evaluated at the imputed item parameters, which approximated the sampling distribution (see Yang et al., 2012, for details). Lastly, the reference line of the bisquare loss method was obtained by minimizing the objective function of the bisquare loss function for the vertical residuals. The sampling distribution was derived using the implicit function theorem as detailed in the Online Appendix B. For the value of k, we investigate the two proposed methods to estimate k: (1) fixing k at a smaller value (i.e., 0.2 value smaller than the true effect size) and (2) fixing the value of k at the median of the absolute value of residuals of the OLS method. Our simulation study shows that the two methods yield nearly identical results (see Figure 1 in the Online Supplemental Material). For better visualization of the empirical cumulative distribution function (ECDF), we only include the results using Method 1.

3.1.4. Likelihood ratio test. The IRT LRT test is usually achieved by conducting a nested model comparison between a compacted model and an augmented model. The compact model involves the likelihood of the parameter estimates for a given item j, assuming the measurement invariance, whereas the augmented model allows for additional item parameters to be freely estimated across the two groups. Typically, anchor items are required to be constrained to be the same to link the metric of the latent abilities of the two groups. Depending on how anchor items are selected, different schemes within the IRT LRT exist (e.g., all-other method, constant anchor items; iteratively backward and forward; see Kopf et al., 2015; Wang & Yeh, 2003, for different anchor schemes comparison). For fair comparison with the proposed method without priori knowledge of anchor items, the *all-other* anchor scheme and iteratively backward is employed here. All-other anchor scheme tests a DIF item by using all items as anchors except the studied item. In the current analysis, all-other IRT LRT method was achieved by first fitting the most constrained multiple group IRT model with all item parameters constraint to be the same across groups. The mean and variance of the latent ability of the first group were set to be 0 and 1, and those of the second group were freely estimated. LRT test was then conducted by freeing one item at a time with the rest of the items constraint in the multiple group IRT model. This step was achieved by first fitting the most constrained multiple group IRT model using the mirt package in R (multipleGroup). DIF analysis was then conducted using the DIF function with the all-other scheme (scheme = "drop"). The iteratively backward method starts with the most constrained multiple group IRT model and freely estimates the mean and variance of the second group. Then, the algorithm loops sequentially over each item by treating all other items as anchors. Then, the algorithm removes items with significant likelihood ratio tests from the anchor set and tests each item remaining in the anchor set. The above procedures are repeated until no more DIF items was found. DIF analysis was conducted using the DIF function with scheme = "drop_sequential." See the helping document in mirt for details.

3.1.5. Evaluation criterion. Rejection rates were used to examine the Type I error rate and power of detecting DIF items. To evaluate the hypothesis test for each type of the test statistic under all nominal levels, we examined the ECDFs of one-tailed p values for testing the null hypothesis (e.g., $H_0 : \boldsymbol{\tau}_j(a) \leq 0$). Ideally, the p values should be uniformly distributed for non-DIF items. Deviation from the $U(0, 1)$ indicates incorrect Type I error rate. In contrast, p values for DIF items are more likely to concentrate in the vicinity of 0 or 1 depending on the sign of the test statistic being evaluated. ECDFs cannot be plotted with the item purification method, and thus, rejection rates were calculated using the two-tailed p values for the null hypothesis $H_0 : \boldsymbol{\tau}_j(a) = 0$ at the nominal level 0.05. Bar plots were used to visualize the results. Results show the comparison between the all-other anchor scheme with the proposed method using the ECDF curves. Additional bar plots in comparing the performance of the proposed method and the iteratively backward anchor scheme are presented in the Online Supplemental Materials.

3.2. Results

3.2.1. 0% DIF item. When there is no DIF item exists, the ECDF of p value should fall approximately along the diagonal line when the Type I error rate is controlled. As is shown in Figure 4, which visualizes ECDFs under all sample size and group size ratio conditions, all proposed methods and also the LRT test can well control the Type I error rate for both "a-DIF" and "b-DIF."

3.2.2. 10% DIF items. When the percentage of DIF items is small (10%), the "a-DIF" and the "b-DIF" behave similarly under all conditions and therefore are summarized together. Also, as items of the same type (DIF or non-DIF items) tend to have similar results, only one item of each kind is shown in the figures to save space. Figure 5 shows the ECDFs of the p value under all conditions with the balanced condition in Figure 5a and the unbalanced condition in Figure 5b. When DIF is balanced, the false alarm rate is acceptable by all methods regardless of the sample size and the effect size condition. However, when the DIF items are unbalanced (all items are favoring one group) as is shown in Figure 5b, LRT (dotted line) and OLS (dashed line), though a little bit better than the LRT, can under reject the null hypothesis. More importantly, increasing the sample size only worsen the underrejection rate (dotted line and dashed line deviate from the diagonal line even more under the large sample condition in Figure 5b). The reason is that as the sample size gets larger, item parameters are estimated more accurately with less uncertainty, and thus, the reference line is pulled over by DIF items, which results in the nonzero effect sizes for non-DIF items. Notice that this does not occur for the balance DIF condition since the DIF effects are cancelled out by the DIF items and the reference line can be accurately estimated by non-DIF items. On the contrary, all of our proposed robust test statistics (i.e., LTS and bisquare) perform well in terms of the rejection rate when the item is DIF-free.

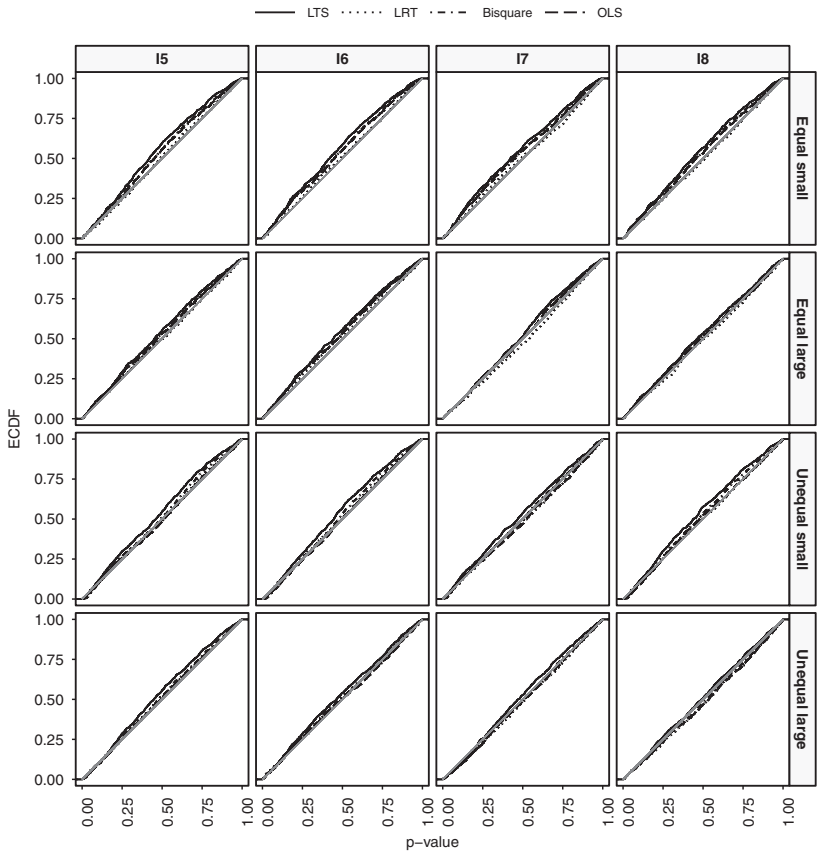

FIGURE 4. *"a-DIF" empirical cumulative distribution functions (ECDFs) of the* p *value under the null conditions: (1) equal small:* $N_1 = N_2 = 500$, *(2) equal large:* $N_1 = N_2 = 1{,}000$, *(3) unequal small:* $N_1 = 250$ *and* $N_2 = 750$, *and (4) unequal small:* $N_1 = 500$ *and* $N_2 = 1{,}500$. *Noticed that the "b-DIF" ECDFs are similar to the "a-DIF" as is shown here and thus not repeated here. LTS = least trimmed square method; LRT = likelihood ratio test; bisquare = Tukey's bisquare method; OLS = ordinary least square method; DIF = differential item functioning.*

Finally, power is decent even when the sample size and the effect size are small for all methods. Increasing the sample size or the effect size can also increase the power in detecting DIF items.

3.2.3. 40% DIF items. When the percentage of the DIF items is large (40%), "a-DIF" and "b-DIF" behaves quite differently and will be discussed separately hereafter. All methods behave similarly as when the percentage of DIF is small

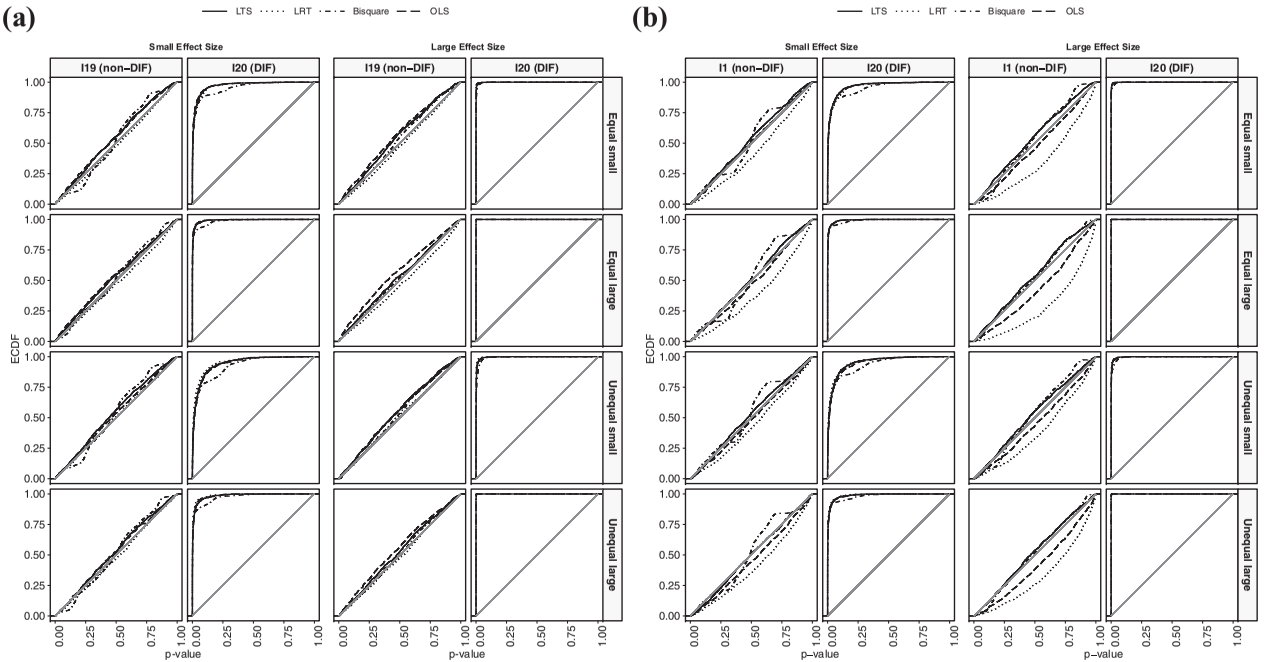

FIGURE 5. *"a-DIF" empirical cumulative distribution functions (ECDFs) of the* p *value under the alternative conditions when there is only 10% of DIF items. Left: balanced DIF. Right: unbalanced DIF. (1) equal small:* $N_1 = N_2 = 500$, *(2) equal large:* $N_1 = N_2 = 1{,}000$, *(3) unequal small:* $N_1 = 250$ *and* $N_2 = 750$, *and (4) unequal small:* $N_1 = 500$ *and* $N_2 = 1{,}500$. *Noticed that the "b-DIF" ECDFs are similar to the "a-DIF" as is shown here and thus not repeated here. LTS = least trimmed square method; LRT = likelihood ratio test; Bisquare = Tukey's bisquare method; OLS = ordinary least square method; DIF = differential item functioning. (a) Balanced DIF. (b) Unbalanced DIF.*

(see Section 3.2). Figure 6 displays the ECDFs of the p value for "a-DIF" when the percentage of DIF items is large in the same fashion as in Figure 5. However, as compared with the 10% DIF condition, larger percentage of DIF items is more challenging for all methods to maintain the rejection rate at the nominal level for the non-DIF items. For example, even when DIF is balanced (Figure 6a), the null hypothesis can be overrejected for the LTS and the bisquare method. The OLS method and the LRT method can maintain the rejection rate at the nominal level when the effect size is small but can overreject (OLS) or underreject (LRT) the null hypothesis when the effect size is large. When the DIF is unbalanced (Figure 6b), OLS and LRT can severely underreject the null hypothesis, while LTS and bisquare perform a little bit better. Especially, when the effect size is large and the sample size is large, bisquare method can maintain the rejection rate at the nominal level. Lastly, as for the power of detecting DIF, all of our proposed methods have decent power in detecting DIF items, while LRT still suffers from lack of decent power in detecting DIF items as compared with our proposed method.

For b-DIF, when the percentage of DIF items is large, all methods failed to differentiate the DIF items and the non-DIF items as can be seen from Figure 7, showing the most difficult condition with unbalanced DIF. With our proposed method, it is possible that the reference line can go through the DIF items instead of the non-DIF items with "b-DIF" because both the intercept and the slope have to be estimated from the item difficulty parameters with sampling variability. As a simple solution, the second way of quantifying the effect size as illustrated in Section 2.1 is utilized. We fix the slope of the reference line as the inverse of the slope of the reference line estimated by the "a-DIF" (i.e., $\beta_{b-dif} = 1/\beta_{a-dif} = 1/g_2(\mathbf{a})$). Then, we estimate the intercept (α) with the item difficulty parameters conditional on the fixed $\hat{\beta}$ estimated from the item discrimination parameter. The corresponding standard errors are adjusted with the fixed slope method to reflect the sampling variability of a. Analytical derivations are provided in the Online Appendix C. As can be seen in Figure 8 after applying the proposed solution to "b-DIF," our proposed methods (the bisquare method and the LTS method) outperform the LRT test in controlling the rejection rate for non-DIF items while also maintaining decent power in detecting DIF items. However, the OLS can have equally inflated false alarm rate as the LRT because of the influence of the DIF items on the reference line. It is worth to mention that only the robust method can control the false alarm rate around the nominal level when the sample size increases. The asymptotic performance of the LRT and the OLS method gets worse because the effect sizes of non-DIF items are nonzero. After the adjustment, the behavior for our approach resembles the "a-DIF" under similar conditions and thus is not repeated here. Compared with the method of finding reference line of "b-DIF" based on b only, borrowing information from

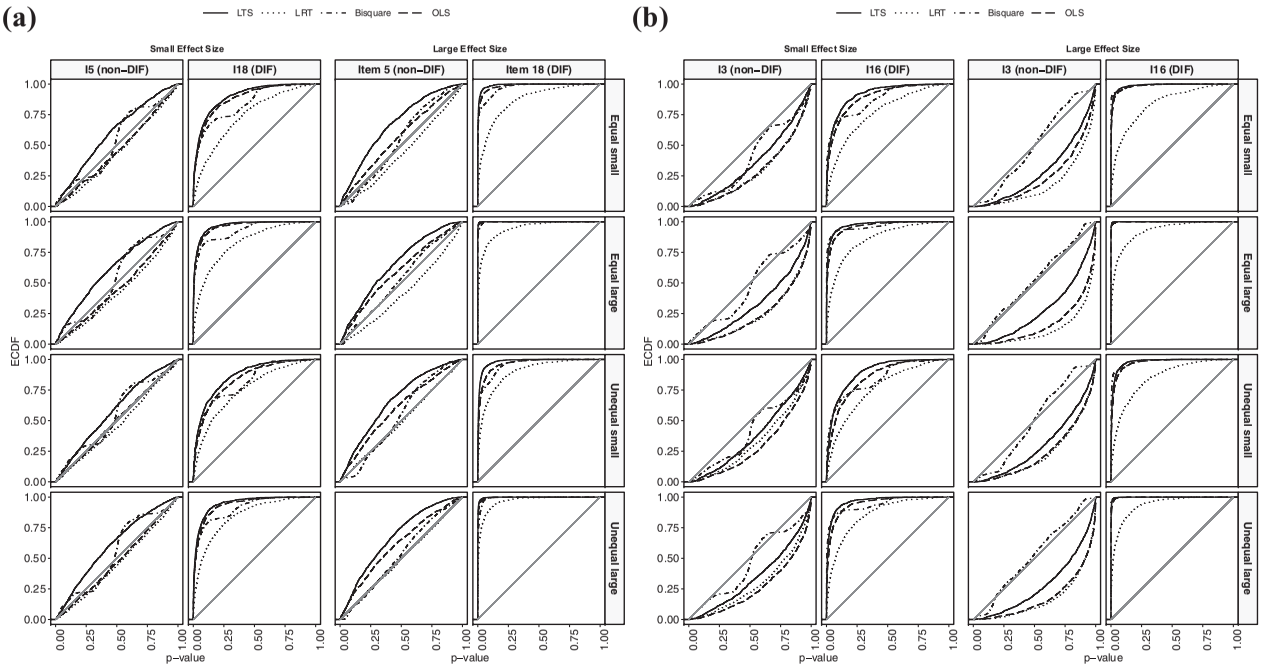

FIGURE 6. *"a-DIF" empirical cumulative distribution functions of the* p *value under the alternative conditions when there is only 40% of DIF items. Left: balanced DIF. Right: unbalanced DIF. (1) equal small:* $N_1 = N_2 = 500$*; (2) equal large:* $N_1 = N_2 = 1{,}000$*; (3) unequal small:* $N_1 = 250$ *and* $N_2 = 750$*; (4) unequal small:* $N_1 = 500$ *and* $N_2 = 1{,}500$*. LTS = least trimmed square method; LRT = likelihood ratio test; Bisquare = Tukey's bisquare method; OLS = ordinary least square method; DIF = differential item functioning. (a) Balanced DIF. (b) Unbalanced DIF.*

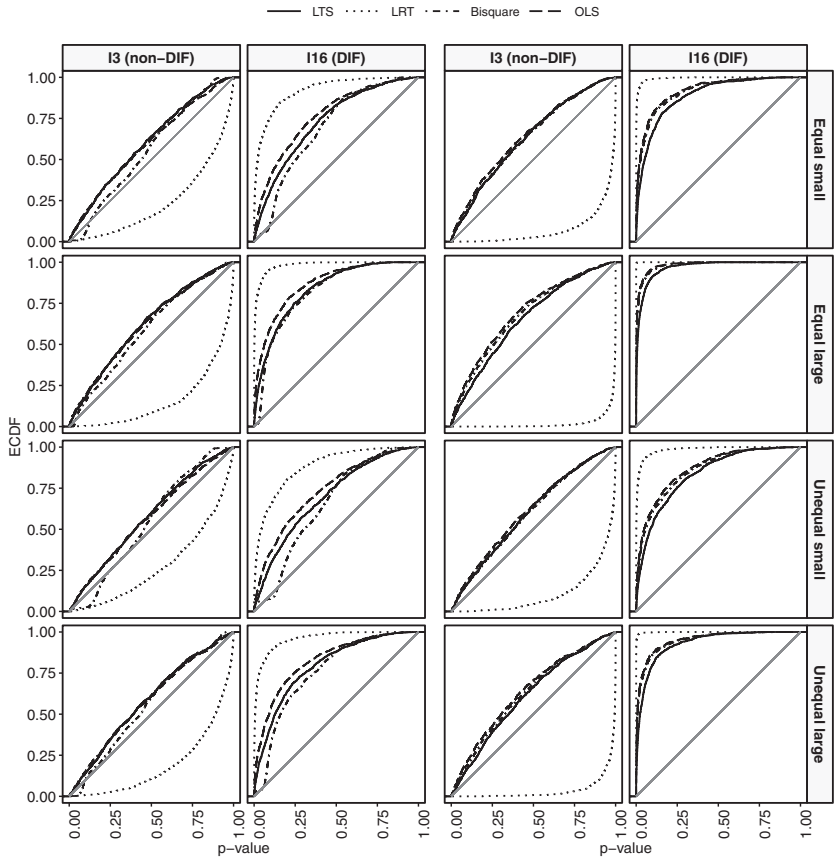

FIGURE 7. *"b-DIF" empirical cumulative distribution functions of the* p *value under the alternative conditions with 40% of unbalanced DIF. (1) equal small:* $N_1 = N_2 = 500$*; (2) equal large:* $N_1 = N_2 = 1{,}000$*; (3) unequal small:* $N_1 = 250$ *and* $N_2 = 750$*; (4) unequal small:* $N_1 = 500$ *and* $N_2 = 1{,}500$*. LTS = least trimmed square method; LRT = likelihood ratio test; bisquare = Tukey's bisquare method; OLS = ordinary least square method; DIF = differential item functioning.*

a can help estimating the slope more accurately. Thus, we recommend to check "a-DIF" first and subsequently "b-DIF" using information from *a*.

3.3. Summary and Discussion

In a nutshell, the rejection rate for the non-DIF items can be well controlled when there is no DIF items. When there is a mix of DIF and non-DIF items, all proposed methods have shown decent power in detecting DIF even under the most difficult condition where the DIF items are unbalanced and the majority of

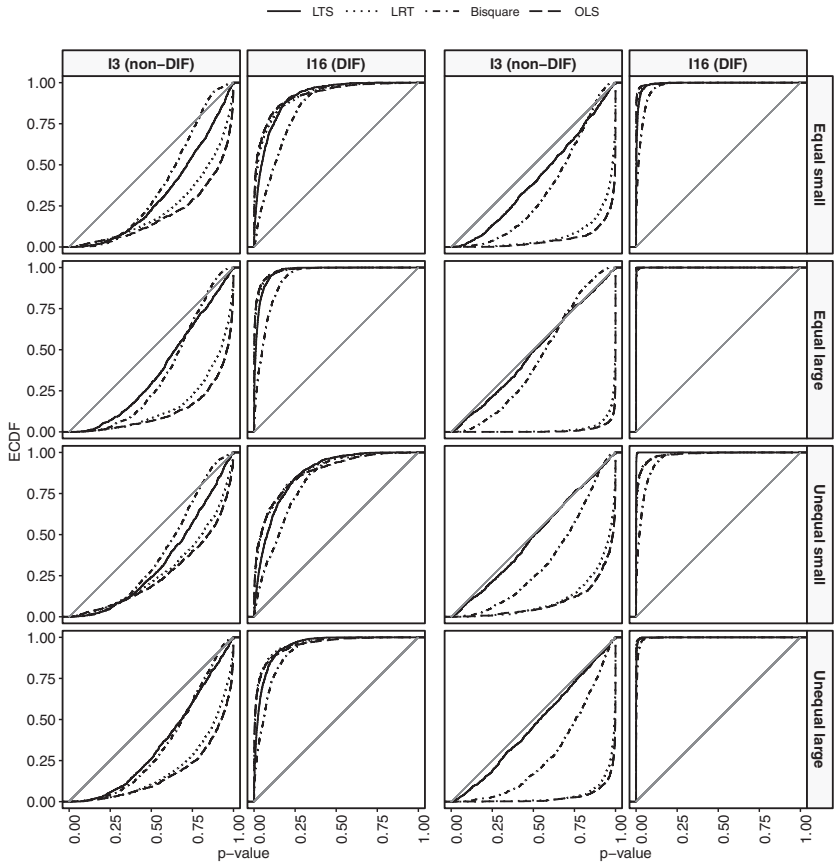

FIGURE 8. *The "b-DIF" empirical cumulative distribution functions (ECDFs) of the* p *value using the modified approach based on both item slope and item intercept under the alternative conditions with 40% of unbalanced DIF: (1) equal small:* $N_1 = N_2 = 500$*; (2) equal large:* $N_1 = N_2 = 1{,}000$*; (3) unequal small:* $N_1 = 250$ *and* $N_2 = 750$*; 4) unequal small:* $N_1 = 500$ *and* $N_2 = 1{,}500$*. LTS = least trimmed square method; LRT = likelihood ratio test; bisquare = Tukey's bisquare method; OLS = ordinary least square method.*

the items are DIF. As for the false alarm rate, a number of factors can influence the performance of the proposed method including the percentage of DIF items, the balance or the unbalanced DIF items, the sample size, and the effect size measure. Under balanced and less percentage of DIF item condition, all proposed methods can control the false alarm rate and perform equally well if not better than the LRT test. Under the most challenging condition where the percentage of DIF items is large and the unbalanced DIF exists, our robust method (LTS

and the bisquare method) outperforms the LRT in maintaining the rejection rate at the nominal level for non-DIF items. Most importantly, the asymptotic performance in controlling the rejection rate only applies to our robust method. However, if the percentage of DIF items is too large (>50%), the robust method eventually breaks down. The Type I error rate in incorrectly detecting a DIF item can be inflated and power in detecting a DIF item can be deflated. The reason is that if the majority of the items are DIF items and somehow align in the same direction, the reference line could be located by a mixed of DIF and non-DIF items or more severely by DIF items only. Effect sizes of DIF items can approach zero and thus result in deflated power. Conversely, the effect size of non-DIF items can be large, such that the Type I error rate can be inflated.

4. Limitation and Future Direction

The current study proposed an innovative DIF detection method that does not require a priori specified anchor item. It can be easily implemented in routine operations of an educational assessment. Our simulation study has shown promising results in controlling the Type I error rate and power of detecting DIF items. Even when there is a mix of DIF and non-DIF items, the false alarm rate can be well controlled using a robust estimator (i.e., the LTS and the bisquare methods in the current study). If there is balanced DIF, OLS can perform equally well with the LTS and the bisquare method.

The proposed method is advantageous over previous studies that deal with anchor contamination issues in several aspects. First of all, the proposed DIF detection method employs a one-step estimation procedure that is computationally efficient. Unlike the traditional IRT DIF detection method, the current method does not require repeatedly fitting models. Once item parameters are calibrated independently for each group, the proposed DIF test statistic can be conducted simultaneously for all items. In addition, this method can be fairly easily implemented in any psychometric software with the ability of fitting IRT models. Moreover, this general definition of the effect size of DIF can be extended to different IRT models (e.g., graded response model [GRM] Samejima, 1969) as well as comparison more than two groups.

The proposed framework of testing DIF can be extended to polytomous IRT models. Taking the GRM as an example. Separate calibration still indicates that item parameters (a_j or b_{jc} for each c) will fall on a straight line if there is no DIF items. Then, the proposed method described in Section 3.1 is applicable in the GRM setting. Specifically, the item discrimination parameter (a) will fall on a straight line that goes through the origin. Accordingly, reference lines of the item category location parameter b_{jc} for each response category c will not necessarily go through the origin. Notice that for all categories c, reference lines have the same slope, which is the inverse of the slope for the "a-DIF" reference line. Generalization to testing DIF across multiple groups is also possible. In

particular, reference lines for "a-DIF" and "b-DIF" still exist but in a g-dimensional space, where g is the number of groups. Similarly, the effect size of DIF can be any types of statistic that can quantifies the deviation from the line. Then, an omnibus hypothesis test can be constructed to test whether item functions differently across groups (i.e., $H_0 : a_j^{(1)} = a_j^{(2)} = \ldots = a_j^{(g)}$).

Given that, there are still several issues that need to be cautiously addressed by future research. First, as is observed that the false alarm rate of incorrectly detecting a DIF item using nonrobust estimators (i.e., OLS) is large especially when the percentage of DIF is large and unbalanced. As robust approaches to obtain the reference line can significantly reduce the false alarm rate (e.g., the LTS and the bisquare methods), additional robust estimators can be investigated.

Second, one of the disadvantages of the bisquare method is the reliance on the tuning parameter k, the magnitude of which can influence DIF detection results. Our investigation has shown that when k is too large, the bisquare method is no different from the OLS method and thus is no longer robust. However, too small value of k makes it sensitive to the sampling variability and thus leads to incorrect false alarm rate. An appropriate value of k is the value that is closer to the true effect size. Here, we proposed two easy methods: (1) using a relatively small value and (2) using empirical results from the OLS method to have a rough estimate of k. Although our simulation results have shown identical results of the two, future research is encouraged to explore alternative approaches to select k.

Lastly, given that our current simulation study only investigates the comparative performance with the all-other anchor method. More thorough comparison with different anchor selection strategies (e.g., item purification and regularization methods) should be conducted in the future study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Weimeng Wang https://orcid.org/0000-0001-9255-8156
Hongyun Liu https://orcid.org/0000-0002-3472-9102

Notes

1. See Belzak and Bauer (2020), Magis et al. (2015), Strobl et al. (2015), Tutz and Berger (2016) for exceptions.
2. When the differential item functioning (DIF) effects are cancelled out by DIF items.

3. Other distance can be defined in a similar fashion: For example, the perpendicular distance is given by $\tau_j(\boldsymbol{\xi}) = \frac{\beta\tilde{\xi}^{(2)} - \xi^{(1)} + \alpha}{\sqrt{1+\beta^2}}$.
4. The item discrimination and difficulty parameters are estimated within each group, assuming that the latent ability follows a standard normal distribution.
5. The Fisher information matrix is rescaled because of our assumptions on the sample sizes.

References

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673.

Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics*. Chapman and Hall/CRC.

Birch, M. W. (1964). A new proof of the Pearson-Fisher theorem. *The Annals of Mathematical Statistics*, *35*, 817–824. https://doi.org/10.1214/aoms/1177703581

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Addison-Wesley.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 433–448). Springer.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment*, *16*(2), 155.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) Scale: An item response theory analysis. *Medical Care*, 281–289.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26.

Finch, H. (2005). The mimic model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*(4), 432–457.

Glas, C. A. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, 647–667.

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*(4), 345–355.

Kopf, J., Zeileis, A., & Strobl, C. (2013). Anchor methods for DIF detection: A comparison of the iterative forward, backward, constant and all-other anchor class.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56.

Langer, M. M. (2008). *A reexamination of lord's Wald test for differential item functioning using item response theory and modern error estimation* [Unpublished doctoral dissertation]. Department of Psychology, University of North Carolina.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862.

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent type I error inflation in differential item functioning. *Educational and Psychological Measurement*, *72*(2), 291–311. https://doi.org/10.1177/0013164411416975

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135. https://doi.org/10.3102/1076998614559747

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software manual]. https://www.R-project.org/

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, *62*(2), 233–279.

Rodebaugh, T. L., Woods, C. M., Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (2006). The factor structure and screening utility of the Social Interaction Anxiety Scale. *Psychological Assessment*, *18*(2), 231.

Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection.

Rudin, W. (1964). *Principles of mathematical analysis* (Vol. *3*). McGraw Hill.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Society.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184–199.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates.

Tutz, G., & Berger, M. (2016). Item-focused trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727–750.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113–144. https://doi.org/10.1207/s15324818ame1702_2

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57.

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*(3), 532–547. https://doi.org/10.1177/0013164412464875

Wright, B., & Stone, M. (1999). Identifying item bias. In *Measurement essentials* (pp. 57–64). Wide Range, INC. https://www.rasch.org/measess/me-all.pdf

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in IRT scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290. https://doi.org/10.1177/0013164411410056

Authors

WEIMENG WANG is a PhD candidate in the Department of Human Development and Quantitative Methodology at the University of Maryland; email: weimengbonnie@gmail.com. Her research focuses on item response theory, differential item functioning, and clinical outcome assessments.

YANG LIU is an assistant professor in the Department of Human Development and Quantitative Methodology at the University of Maryland, 1230B Benjamin Building, 3942 Campus Drive, College Park, MD 20742; email: yliu87@umd.edu. His research focuses on various inferential problems in item response theory.

HONGYUN LIU is a professor, Faculty of Psychology, Beijing Normal University, No.19 Xin Jie Kou Wai Street, Hai Dian District, Beijing 100875, China; email: hyliu@bnu.edu.cn. Her research area focuses on statistical analysis methods–statistical analysis methods in the application of psychology and the large-scale assessment.

Manuscript received January 7, 2021
Revision received March 18, 2022
Accepted May 29, 2022