Article

The Mediated MIMIC Model for Understanding the Underlying Mechanism of DIF

Educational and Psychological Measurement
2016, Vol. 76(1) 43–63

DOI: 10.1177/0013164415576187
epm.sagepub.com
SAGE

Ying Cheng[1], Can Shao[1], and Quinn N. Lathrop[1]

Abstract

Due to its flexibility, the multiple-indicator, multiple-causes (MIMIC) model has become an increasingly popular method for the detection of differential item functioning (DIF). In this article, we propose the mediated MIMIC model method to uncover the underlying mechanism of DIF. This method extends the usual MIMIC model by including one variable or multiple variables that may completely or partially mediate the DIF effect. If complete mediation effect is found, the DIF effect is fully accounted for. Through our simulation study, we find that the mediated MIMIC model is very successful in detecting the mediation effect that completely or partially accounts for DIF, while keeping the Type I error rate well controlled for both balanced and unbalanced sample sizes between focal and reference groups. Because it is successful in detecting such mediation effects, the mediated MIMIC model may help explain DIF and give guidance in the revision of a DIF item.

Keywords

differential item functioning, MIMIC model, mediation effect, item response theory

When people from different groups (e.g., female vs. male, or wealthy vs. impoverished) with the same latent trait (e.g., aptitude or proficiency) level show differential probability of endorsing certain response options to an item on a scale or test, this item is referred to as an item with differential item functioning (DIF). DIF is an important topic in psychological and educational measurement. Discussion of DIF, formerly known as item bias, dates back to the Civil Rights movement in the 1960s,

[1]University of Notre Dame, Notre Dame, IN, USA

Corresponding Author:
Ying Cheng, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA.
Email: ycheng4@nd.edu

which brought the issue of fair testing to public attention (Dorans, 1989). Students with the same level of proficiency, if the testing is done fairly, should have the same probability of answering an item correctly (i.e., endorsing the correct response option), regardless of their ethnicity, gender, social economic status, and so on. Items with DIF, therefore, are considered harmful to fair testing. In spite of its root in educational testing, DIF has been linked to the vast body of literature in psychological testing on measurement invariance (e.g., Ekermans, Saklofske, Austin, & Stough, 2011; Meade, Lautenschlager, & Johnson, 2007) in recent years and has therefore been attracting attention from researchers with broader interests and backgrounds.

Many methods over the years have been proposed to identify items with DIF, including the Mantel–Haenszel (MH) approach (Holland & Thayer, 1988; Mantel & Haenszel, 1959) for dichotomous items, the generalized Mantel–Haenszel test (GMH; Mantel & Haenszel, 1959; Somes, 1986; Zwick, Donoghue, & Grima, 1993) for polytomous items, the SIBTEST (for dichotomous items; Shealy & Stout, 1993) and poly-SIBTEST (for polytomous items; Chang, Mazzeo, & Roussos, 1996) approach, the logistic regression (LR) procedure (for dichotomous items; Swaminathan & Rogers, 1990), the ordinal logistic regression approach (for polytomous items), and the IRT-model-based likelihood ratio goodness-of-fit test proposed by Thissen, Steinberg, and Gerrard (1986), which can be applied to both dichotomous and polytomous data.

Camilli and Shepard (1994) noted that confirmatory factor analysis (CFA) has potential in DIF detection since the comparison of group differences on a secondary factor is allowed. Based on that, different approaches using CFA to detect DIF have been developed, such as the multi-group CFA method (Pae & Park, 2006), the modification indices method (Chan, 2000), and the multiple-indicator, multiple-causes (MIMIC) model (Hauser & Goldberger, 1971; MacIntosh & Hashim, 2003) approach.

This article focuses on the MIMIC model approach, which is very flexible. It can handle both dichotomous and polytomous items (Finch, 2005; Wang & Shih, 2010). It can also include multiple grouping variables, for example, both gender and ethnicity, in the analysis simultaneously and allow for interactions among these variables. These grouping variables can be either observed or latent variables. It can easily control for covariates (e.g., age of the participant in a study on cognitive development in early childhood) and allow both categorical and continuous background or DIF variables (Glockner-Rist & Hoitjink, 2003; Muthén, 1988). These flexibilities make the MIMIC model approach now one of the most popular approaches for DIF detection in recent studies (e.g., Finch, 2005; Gallo, Anthony, & Muthén, 1994; Wang, Shih, & Yang, 2009; Woods, 2009; Woods & Grimm, 2011).

In spite of the paramount attention on DIF detection, there is little research on the underlying mechanism of DIF, that is, what causes DIF (Yao & Li, 2010). Among the few studies that have tapped on the issue, most focus on multidimensionality in the multidimensional item response theory (MIRT) framework. In other words, DIF is assumed to be caused by the existence of auxiliary dimensions (Ackerman, 1992; Penfield & Lee, 2010; Walker & Beretvas, 2001). In this article, we propose to

examine the underlying mechanism of DIF using mediation analysis in the framework of the MIMIC model.

Mediation analysis hypothesizes that one variable (X) affects a second variable (M), which, in turn, affects a third variable (Y) (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The intervening variable M is called a mediator, which can either be a complete or partial mediator. A complete mediator M can fully mediate the relation between X and Y. In other words, when M is taken into account, the direct relation between X and Y vanishes. Partial mediation means the mediator can only explain part of the relation between X and Y.

We posit that by including a mediator in the MIMIC model, we can obtain a clearer picture of the underlying mechanism of DIF. For example, if the response to a math item (Y) involves baseball knowledge (M), students from immigrant families (X) may be at a disadvantage in answering the question given the same math ability (θ). The baseball knowledge (M) may completely or partially mediate the DIF effect, that is, the effect of X on Y when controlling for θ. If complete mediation effect is found, baseball knowledge fully explains the DIF effect, otherwise partially. Generally, if a potential mediator (M), which can be either a manifest or latent variable, is identified, we can test if it significantly mediates the effect of X, a background variable such as gender or ethnicity, on Y, an item response, controlling for the latent trait of interest, θ. If it does, we can further test if it is a complete or partial mediator. If complete mediation is found, the mediator should be respecified as a direct background variable and the model should be refit. As mentioned earlier, the background variable X can be either a categorical or continuous variable in the MIMIC model. Therefore, if the mediator is a continuous variable there is no problem respecifying it as the DIF variable. By respecifying the mediator M to be the DIF variable X, the model becomes more parsimonious. Complete or partial mediation, either way we will have better understanding of the underlying mechanism of DIF and will be better positioned to revise the item if necessary.

The mediated MIMIC model approach is different from the MIRT approach in several important aspects. First, it does not rely on the MIRT models, which assume that the item responses are categorical and that the source of DIF is a continuous latent trait. The mediator in the mediated MIMIC model framework theoretically may well be a categorical variable, latent or manifest, and the item responses may well be continuous. Second, the mediated MIMIC model allows the detection of complete or partial mediation effect. The multidimensionality perspective, on the other hand, only tells if there is DIF caused by an auxiliary dimension or not, but does not differentiate between partial or complete DIF. For these reasons, we believe that the two approaches are conceptually distinct and the mediated MIMIC model is more versatile.

In this study, we would like to investigate the performance of the mediated MIMIC approach in its detection of the underlying cause or possible mediator of the DIF effect. Note that this is not a conventional DIF study, and the focus is not to detect the DIF effect itself. The investigation will be conducted under both balanced

and unbalanced designs. A balanced design means that the sample sizes for the reference and focal groups are equal or similar (e.g., Wang & Shih, 2010). In practice, however, we often encounter unbalanced sample sizes for reference group and focal groups. For example, the reference group is ethnic majority while the focal group is a minority group. It is suggested by researchers that we should be cautious about DIF analysis with a very small reference or focal group. Mazor, Clauser, and Hambleton (1992) suggested 200 people in each group for the MH approach. Zieky (1993) suggested (100, 400),[1] (200, 400), and (500, 500) as the minimum sample sizes used for DIF detection at Educational Testing Service with MH chi-square test, MH chi-square statistic without continuity correction, and MH delta-based *z* test, respectively. Paek and Guo (2011) carried out a simulation study using the MH approach and found that when the total sample size is fixed, the balanced design yields a higher power and more accurate DIF estimates than the unbalanced design. When the focal group sample size is fixed and is small, larger reference group leads to higher power and more accurate DIF parameter estimates. To examine the performance of the mediated MIMIC approach under various sample sizes of the focal and reference groups, we include different total sample sizes crossed with different ratios between focal versus reference sample size. This way, we are able to investigate not only the influence of total sample size but also the effect of the sample size ratio between focal and reference groups.

The rest of the article is organized as follows. First, we delineate the MIMIC model method for DIF detection, followed by the introduction to the mediated MIMIC model approach. Then a simulation study is carried out to examine the effectiveness of the proposed approach in detecting the mediation effect of interest, that is, the underlying DIF mechanism, under various sample size conditions. To illustrate the application of the mediated MIMIC approach, we included a real data example. Finally, recommendations are made regarding sample sizes and the application of the mediated MIMIC model in practice.

The MIMIC Model for the Detection of DIF

In a standard factor analytic model, an underlying factor can influence some manifest variables, which we call indicators. The assumption is that the dependence among the indicators comes from their common dependence on the underlying factor or factors. By further incorporating variables that influence the latent factor(s), the MIMIC model is derived. Thus, the multiple indicators reflect the underlying factors (often referred to as the measurement model in the structural equation modeling framework), and the multiple causes affect the underlying factors (the structural component). Take, for example, a test with 10 items measuring students' math ability and other 10 measuring verbal ability. The measurement model will include two underlying factors—verbal and math ability—each having 10 corresponding indicators. If the researcher is interested in finding out whether mother's and father's education will affect students' math and verbal ability, then mother's and father's education

can be used as cause variables to those latent factors. In other words, mother's and father's education would affect the two latent abilities, which would in turn affect the indicators.

Specified appropriately, the MIMIC model can be used for DIF detection. The measurement component takes the following form:

$$y_i^* = \lambda_i \theta + \beta_i z + \varepsilon_i, \tag{1}$$

where y_i^* is the latent response propensity variable, θ is the latent trait of interest (e.g., math ability), and λ_i is the factor loading. Here, z is the background or grouping variable (e.g., 1 indicates focal group and 0 indicates reference group). Note that in the MIMIC model approach, z can have more than two categories. It can also be a continuous variable. This may not be true for other DIF detection methods, for example, the MH approach, which relies on two-by-two contingency tables. Meanwhile, the MIMIC model approach also allows for multiple grouping variables and their interactions—another distinct advantage. To keep our illustration straightforward, here we include only one categorical grouping variable z. Then β_i indicates the relationship between the grouping variable z and item response y_i^*, when θ is controlled for. A significant β_i therefore suggests the presence of DIF[2] with item i. Last, ε_i is the random error usually assumed to be normally distributed with mean 0.

The latent propensity response y_i^* is related to the observed ordinal item responses y_i through a threshold model as follows, where τ_{ij} s are the thresholds between two adjacent score categories:

$$y_i = \begin{cases} 0 & \text{when } y_i^* \leq \tau_{i1}, \\ 1 & \text{when } \tau_{i1} < y_i^* \leq \tau_{i2}, \\ 2 & \text{when } \tau_{i2} < y_i^* \leq \tau_{i3}, \\ & \cdots \\ J & \text{when } y_i^* > \tau_{iJ}. \end{cases} \tag{2}$$

It is clear here that the MIMIC model approach is able to handle both dichotomous (when $J = 2$) and polytomous item responses (when $J > 2$).

The structural component is as follows:

$$\theta = \gamma z + \zeta, \tag{3}$$

where γ is the regression coefficient for the grouping variable z. A significant γ indicates real group difference in θ, that is, impact (Ackerman, 1992; Camilli, 1993). It is important not to confuse impact with the difference in performance between groups (e.g., focal vs. reference) on an item. The performance gap may be attributed to real difference in the latent trait θ between groups, or to the DIF effect, or a combination of both. Impact refers to the real difference in the latent trait θ. Research has shown that impact can confound with the DIF effect and make the detection of DIF more difficult (Hidalgo & López-Pina, 2002). In the simulation study below, we will consider

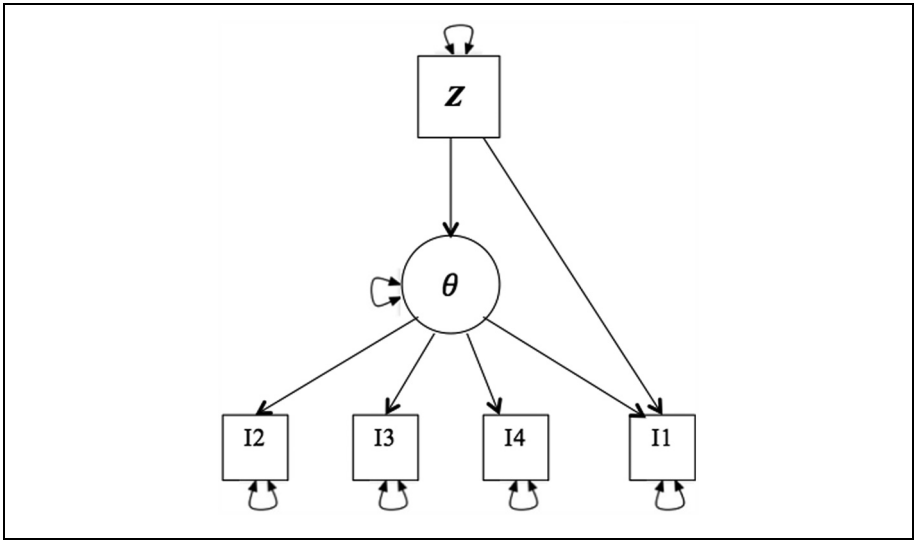

Figure 1. The MIMIC model for DIF detection.
Note. I2, I3, and I4 constitute the "anchor"; when I1 is tested for DIF, I2 to I4 are assumed DIF free.

the existence of impact. ζ is random error usually assumed to follow normal distribution with mean 0. It is also assumed independent of the grouping variable.

Figure 1 illustrates the detection of DIF of Item 1 (I1) when it is the item of interest. The detection of DIF with MIMIC models involves estimating the direct effect of the grouping variable on the item response, for example, the direct path from *z* to I1 in Figure 1. The direct effect indicates the difference in item response among groups given the same latent trait. Therefore, if the direct effect path is significant, it suggests the presence of DIF. The indirect effect ''regresses'' the latent trait onto the grouping variable *z* (the path from *z* to θ in Figure 1) and indicates whether the mean of this latent variable differs across groups, thus accounting for real group difference or impact in the latent trait.

Method

There is growing literature on using the MIMIC model for DIF detection (Fleishman, Spector, & Altman, 2002; Jones, 2006). In reality, it is unknown which items are free of DIF. Therefore, usually an iterative procedure is employed, where one item is tested for DIF and the rest of the items serve as the anchor set. This is repeated for every item on the scale. An example of such iterative procedure is proposed by Wang and Shih (2010), called the ''MIMIC model with scale purification'' (M-SP) method. M-SP finds the anchor set by first testing each item for DIF one at a time and using all other items as the anchor. Then it removes items that are classified as having DIF

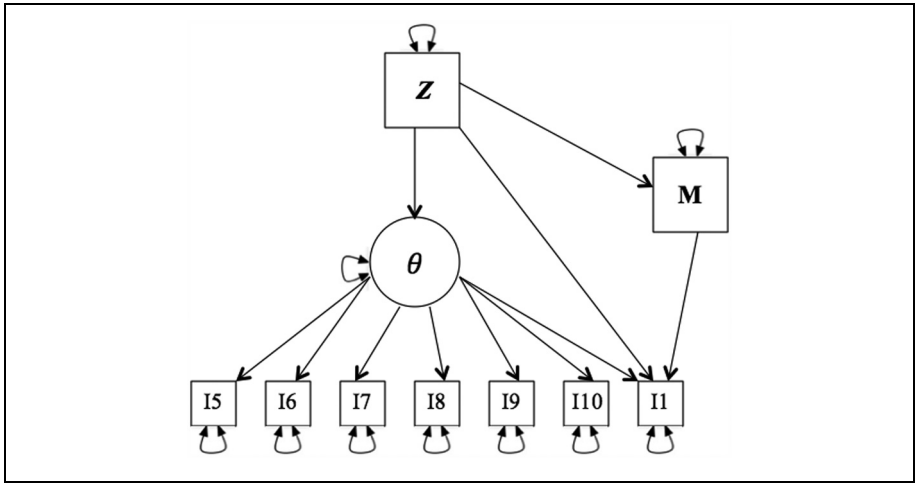

Figure 2. The mediated MIMIC model for DIF detection.

from the anchor set. This purified anchor is subsequently used to test all remaining items for DIF, and this process repeats until the same set of items are detected as showing DIF for two successive iterations.

Using the MIMIC model to detect DIF is well established in the literature. We focus here on the detection of mediation in the context of DIF. In other words, in our simulation we assume that the step of DIF detection has been completed, and items with DIF have already been detected. From a practitioner's perspective, the natural next step is to find out what causes DIF in these items. In such a case the mediated MIMIC model can be used to find out whether a variable *M* mediates the relationship between the item response and the grouping variable, while controlling for θ. If so, *M* fully or partially accounts for the DIF effect.

Figure 2 shows the mediated MIMIC model for Item 1. As discussed earlier, a variable *M* can mediate the relationship between group membership and an item response, conditioning on latent trait θ. If *M* fully explains the relationship, it is called complete mediation. For example, if Asian Americans are found to score higher on an item on social anxiety than Whites given the same underlying level of social anxiety, it means the item exhibits DIF. But when acculturation is taken into account (i.e., as a mediator), the direct path from the grouping variable (e.g., *z* in Figure 2) to the item response (e.g., I1 in Figure 2) may become nonsignificant. In this case we call acculturation a complete mediator. If the direct path (e.g., *z* to I1) and the mediation path (e.g., *z* to *M* to I1) both are significant, acculturation is a partial mediator.[3] When complete mediation is found, the mediator completely accounts for the DIF effect; when partial mediation is found, the mediator helps account for the DIF effect but not in entirety. Other possible mediators could be identified and the group of mediators together may be able to fully account for the DIF effect.

Practically, the success of using the mediated MIMIC model to understand the underlying mechanism of DIF understandably relies on the identification of possible mediators. Taking the math test item as an example again, to realize that baseball knowledge may be a mediator requires careful item review and content expertise. Once a mediator is proposed and observed, the mediated MIMIC model can be applied to check if any mediation effect exists.

The success of the mediated MIMIC model can be evaluated in simulation studies in terms of the power and Type I error in its detection of the mediation effect. Namely, the method should be able to identify a true mediator and, at the same time, screen nonmediators. In the simulations that follow, by manipulating the (relative) magnitude of the direct DIF effect (i.e., the path from z to an item response when the mediation effect is accounted for) and the mediation effect, we will compare the power and Type I error of detecting the mediation effect in each condition. We expect that when the magnitude of mediation effect increases, the power of mediation detection will go up. Also, we expect the magnitude of direct DIF will influence the power of mediation detection, conditioning on the magnitude of the mediation effect. In regard to total sample size and the ratio between focal and reference group, we expect that larger sample size leads to higher power. When the total sample size is fixed, the balanced sample size conditions are expected to yield larger power than the unbalanced conditions.

Simulation Study

In this simulation study, we consider three total sample sizes (N = 600, N = 1,000, N = 2,000), crossed with three ratios of focal versus reference sample size: 1:1, 1:2, and 1:4, where 1:1 represents the balanced condition, and 1:2 and 1:4 are two unbalanced conditions. In total, we have 9 sample size conditions. This way, we can investigate the influence of total sample size, the sample size ratio, and their interaction.

Item responses are generated with the graded response model (GRM; Samejima, 1969), where each item is associated with K parameters: a discrimination parameter (a_i) and ($K-1$) location parameters (b_{ik}), and K is the number of response categories ($K \geq 2$). In this study, the generating item parameters for the reference group came from the first 10 items of Cohen, Kim, and Baker (1993). Each item has 5 response categories. Item i has five (K = 5) item parameters: a discrimination parameter a_i and four threshold parameters b_{ik}, where k = 1, 2, 3, 4 and $b_{i1} < b_{i2} < b_{i3} < b_{i4}$. For a test taker in the reference group with latent trait θ, the probability of endorsing category k of item i is given as follows. When k = 1,

$$P_{i1}(\theta) = 1 - \frac{\exp(a_i(\theta - b_{i1}))}{1 + \exp(a_i(\theta - b_{i1}))}. \quad (4)$$

When $1 < k < K$,

Table 1. Power and Type I Error for Each DIF Item of (500, 500) Conditions.

Total DIF		Condition	Item 1	Item 2	Item 3	Item 4
0.20	Power	DE = 0.00, ME = 0.20	1	0.998	0.998	1
		DE = 0.07, ME = 0.13	0.908	0.866	0.928	0.936
		DE = 0.10, ME = 0.10	0.722	0.672	0.684	0.690
		DE = 0.13, ME = 0.07	0.394	0.384	0.448	0.380
	Type I error	DE = 0.20, ME = 0.00	0.084	0.072	0.066	0.054
0.30	Power	DE = 0.00, ME = 0.30	1	1	1	1
		DE = 0.10, ME = 0.20	0.998	0.996	1	1
		DE = 0.15, ME = 0.15	0.968	0.948	0.972	0.946
		DE = 0.20, ME = 0.10	0.692	0.654	0.714	0.706
	Type I error	DE = 0.30, ME = 0.00	0.076	0.044	0.052	0.056
0.50	Power	DE = 0.00, ME = 0.50	1	1	1	1
		DE = 0.17, ME = 0.33	1	1	1	1
		DE = 0.25, ME = 0.25	1	1	1	1
		DE = 0.33, ME = 0.17	0.986	0.988	0.988	0.978
	Type I error	DE = 0.50, ME = 0.00	0.040	0.048	0.044	0.044

Note. DIF = differential item functioning; DE = direct DIF effect; ME = mediation effect.

$$P_{ik}(\theta) = \frac{\exp\left(a_i\left(\theta - b_{i(k-1)}\right)\right)}{1 + \exp\left(a_i\left(\theta - b_{i(k-1)}\right)\right)} - \frac{\exp(a_i(\theta - b_{ik}))}{1 + \exp(a_i(\theta - b_{ik}))}. \tag{5}$$

When $k = K$,

$$P_{iK}(\theta) = \frac{\exp(a_i(\theta - b_{iK}))}{1 + \exp(a_i(\theta - b_{iK}))}. \tag{6}$$

Items 1 to 4 were the items chosen to exhibit DIF. To simulate the DIF effect, the threshold parameters of each of the four DIF items were added a constant for the focal group. Such a data generation scheme mimicked that of Cohen et al. (1993). To be specific, to generate responses for a test taker from the focal group, we replaced the b_{ik} in Equations 4 to 6 by b_{ikF}, where $b_{ikF} = b_{ik} + \text{TE}$, and TE represents the total DIF magnitude. For example, if TE is 0.2, then the item appeared more difficult to the focal group members by 0.2 logits. Note that the magnitude of effect is measured in logits, as in Wang and Shih (2010), and the total DIF effect (TE) is the sum of the direct DIF effect (DE) and the mediation effect (ME).

Simulees from the reference group had their abilities generated from N(0, 1). The focal group's abilities were drawn from N(1, 1). This means that true group difference in latent trait, or impact, exists between the reference and focal groups. Items 1 to 4 were set to have the same magnitude of uniform DIF, and all other items were DIF-free and together served as the anchor set.

The third column of Table 1 summarizes all the 12 DIF and mediation magnitude conditions included within each sample size condition. By setting the magnitude of the mediation effect at 0 (i.e., Conditions 5, 9, and 12), we obtained the Type I error of the detection of the mediation effect, because these are the null conditions. This examines the capability of the mediated MIMIC model approach in screening insubstantial mediators. On the other hand, by setting the mediation magnitude to be nonzero (see Table 1, the magnitude varies from 0.20 to 0.50), we obtained the power of the mediation effect detection. If the direct DIF effect is 0, it means that *M* completely mediates the DIF effect. We set the total DIF magnitude at 0.5, 0.3, and 0.2, which were well within the range of DIF magnitude in the literature. For example, Cohen et al. (1993) simulated DIF magnitude of 0.5; Su and Wang (2005) simulated DIF magnitude between 0.1 and 0.4; and Wang and Shih (2010) simulated DIF magnitude of 0.25. Again these are measured in logits. Given a total DIF magnitude, we manipulated the ratio of direct DIF effect and the mediation effect. The levels of ratio were 0 (complete mediation), 1:2, 1:1, 2:1, and $+\infty$ (no mediation). Crossed with the 3 levels of total DIF magnitude, there are 15 conditions (see Table 1).

The simulation was performed in **R** (R Development Core Team, 2011) and **Mplus** (Muthén & Muthén, 2011). We first used **R** to generate data and then called **Mplus** to estimate mediated MIMIC models using the *Mplus Automation* package (Hallquist & Wiley, 2012) from **R**. In total 135 conditions were simulated: 9 (sample size conditions) $\times$ 15 (DIF effect conditions). Each condition was replicated 500 times. The weighted least squares (WLS) approach was used to estimate the parameters in the mediated MIMIC model. By getting the parameter estimates of the indirect path, *Z* to *M* (denoted as $\hat{\alpha}$), and *M* to a DIF item (denoted as $\hat{\beta}$), we obtained the unstandardized estimated indirect effect, which is the product of these two estimates ($\hat{\alpha}\hat{\beta}$). The estimated standard error of indirect effect was calculated using the estimated standard errors of the indirect path ($\hat{\sigma}_{\hat{\alpha}}$ and $\hat{\sigma}_{\hat{\beta}}$), following $\hat{\sigma}_{ind} = \sqrt{\hat{\alpha}^2\hat{\sigma}^2_{\hat{\alpha}} + 2\hat{\alpha}\hat{\beta}cov(\hat{\alpha},\hat{\beta}) + \hat{\beta}^2\hat{\sigma}^2_{\hat{\beta}}}$. The standardized estimate of the indirect effect, $\hat{\alpha}\hat{\beta}/\hat{\sigma}_{\text{ind}}$, asymptotically follows N(0,1) under the null hypothesis that there is no mediation effect (MacKinnon, 2008). Therefore, the *p* value of the standardized estimate can be found by referencing the sample $\hat{\alpha}\hat{\beta}/\hat{\sigma}_{\text{ind}}$ against N(0, 1). When $p < .05$, the mediation effect is considered significant, otherwise non-significant.

Table 1 shows the Power and Type I Error for each of the four DIF items, with a balanced sample size of (500, 500), that is, 500 test takers were from the reference and focal group, respectively. In Table 1, ME represents the magnitude of the mediation effect and DE represents the magnitude of the direct effect. Again the ME and DE add up to the total DIF magnitude, TE. The four DIF items have rather similar power and Type I error under each DIF effect condition, so in the rest of the article we report the average power and Type I error across the four DIF items. Type I error rates range between 0.040 and 0.084, suggesting that Type I error is generally well controlled. Power to detect the mediation effect is at least 0.866 as long as the ME is

Table 2. Power and Type I Error for All Conditions.

			N = 600			*N* = 1,000			*N* = 2,000		
TE		Condition	(300, 300)	(200, 400)	(120, 480)	(500, 500)	(333, 667)	(200, 800)	(1,000, 1,000)	(667, 1333)	(400, 1,600)
0.20	Power	DE = 0.00, ME = 0.20	0.967	0.964	0.958	0.999	1	0.999	1	1	1
		DE = 0.07, ME = 0.13	0.723	0.710	0.723	0.910	0.918	0.900	0.994	0.995	0.999
		DE = 0.10, ME = 0.10	0.484	0.506	0.495	0.692	0.708	0.683	0.930	0.924	0.936
		DE = 0.13, ME = 0.07	0.275	0.280	0.253	0.402	0.3 80	0.376	0.060	0.609	0.644
	Type I error	DE = 0.20, ME = 0.00	0.067	0.058	0.070	0.069	0.060	0.060	0.061	0.059	0.050
0.30	Power	DE = 0.00, ME = 0.30	1	1	1	1	1	1	1	1	1
		DE = 0.10, ME = 0.20	0.972	0.957	0.967	0.999	0.999	0.997	1	1	1
		DE = 0.15, ME = 0.15	0.808	0.797	0.798	0.959	0.958	0.957	1	1	1
		DE = 0.20, ME = 0.10	0.495	0.496	0.473	0.692	0.695	0.680	0.939	0.927	0.932
	Type I error	DE = 0.30, ME = 0.00	0.070	0.051	0.057	0.057	0.059	0.063	0.051	0.055	0.054
0.50	Power	DE = 0.00, ME = 0.50	1	1	1	1	1	1	1	1	1
		DE = 0.17, ME = 0.33	1	1	1	1	1	1	1	1	1
		DE = 0.25, ME = 0.25	0.995	0.999	0.997	1	1	1	1	1	1
		DE = 0.33, ME = 0.17	0.889	0.883	0.887	0.985	0.981	0.978	1	1	1
	Type I error	DE = 0.50, ME = 0.00	0.069	0.063	0.072	0.044	0.064	0.058	0.064	0.048	0.054

Note. DIF = differential item functioning; TE = total DIF effect; DE = direct DIF effect; ME = mediation effect.

not too small (i.e., at least 0.13). Meanwhile, the psychometric properties of the items (i.e., IRT parameters) do not seem to have an effect on the detection of the mediation effect. The effect of the total DIF magnitude and the ratio between ME and DE are dominating.

Since the four items do not differ much in their performance, for the other sample size combinations only the mean power and Type I error rate of the detection of the mediation effect are reported (see Table 2). The averages were taken over the four items and 500 replications. Across the conditions, the Type I error is between 0.044 and 0.072, indicating that the Type I error is well controlled as it is close to the nominal level of 0.05.

Across all conditions, when ME $\geq$ 0.15, the power is about 0.8 or higher. Other factors being equal, when the total sample size increases, the power increases. Other factors being equal, when the total DIF magnitude increases, the power increases. Other factors being equal, when the ratio of ME to DE increases, the power increases. All these trends are consistent with our expectation. The effect of the ratio of reference to focal group sample size, however, has no clear pattern. In fact, the difference in power from different sample size combinations is very small, usually in the second decimal place. Therefore, with a total sample size of 600 or up, it does not matter much whether the sample sizes from reference and focal groups are balanced or not. When the sample size is 2,000, even when the total DIF effect is very small (i.e., 0.20) and the mediation effect is tiny (i.e., 0.10), the power is at least 0.924. In general, the effect of the total sample size on power is much larger than that of the sample size ratio. Also as long as the sample size is large (i.e., 1,000 or up) and the mediation effect is not tiny (i.e., 0.10 or up), the power is at least 0.680.

Across all conditions, the smallest power occurs when the sample size is small (i.e., 600), and the mediation effect is tiny (i.e., ME = 0.10 or below). Figure 3 visually displays the distribution of the standardized mediation effect $\hat{\alpha}\hat{\beta}/\hat{\sigma}_{ind}$ of Item 1 when the sample size combination is (120:480) and the total DIF effect is 0.20. The solid line in each plot is the distribution of $\hat{\alpha}\hat{\beta}/\hat{\sigma}_{ind}$ we obtained through 500 replications when there is mediation effect (ME = 0.07, 0.10, 0.13, and 0.20, respectively, for the four plots, with DE = 0.20 − ME, which means DE = 0.13, 0.10, 0.07, and 0, respectively). The dashed line is the distribution of $\hat{\alpha}\hat{\beta}/\hat{\sigma}_{ind}$ by setting ME = 0 with DE = 0.13, 0.10, 0.07, and 0, respectively (this is to be consistent with the simulation conditions that produce the solid line). Because ME = 0, the dashed line provides the null distribution where there is no mediation effect. For each plot, the only cause of difference between the solid and dashed line is the ME magnitude. As the ME increases, the distribution of the empirical distribution of the standardized mediation effect, that is, $\hat{\alpha}\hat{\beta}/\hat{\sigma}_{ind}$, pulls further away from the null distribution. When the ME is small, the two distributions have a substantial amount of overlap and it is therefore difficult to tell them apart. That is why the power is low when the ME is tiny. As the two distributions get further and further separated, the power increases.

In summary, the mediated MIMIC model is successful in detecting existing mediation effect when DIF occurs and screening nonmediators. Type I error rate is

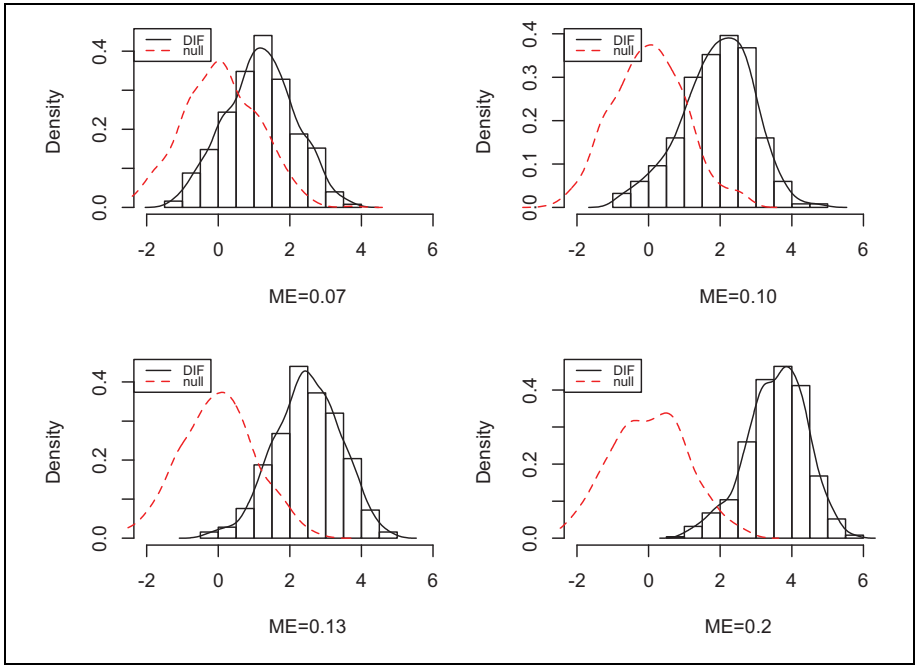

Figure 3. The distribution of mediation effect (with ME vs. no ME) estimates/SE for Item 1 in (120, 480) condition.

generally well controlled and power is very high for most conditions. Power only drops when the mediation effect is tiny and when the sample size is as small as 600. The mediated MIMIC model performs very well when total sample size is big enough (e.g., N = 1,000 or up) even if we have unbalanced groups and fairly small mediation effect.

Real Data Example

In this section, we illustrate the application of the mediated MIMIC model to help understand the cause of the DIF effect. We used 2007 U.S. data from the Trends in International Mathematics and Science Study (TIMSS). The data set is available at https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010024. The scale of 8th-grade students' enjoyment of science with 8 items was examined. Each item uses a 4-point Likert-type scale where 1 = *agree a lot*, 2 = *agree a little*, 3 = *disagree a little*, 4 = *disagree a lot*. The data set includes responses from 7,095 students after list-wise deletion of entries with missing data. Based on our simulation study, such a large sample size should allow us to identify the mediation effect if it exists. Some of the

Table 3. The Scale of 8th-Grade Students' Enjoyment of Science.

1	I usually do well in science
2	I would like to take more science in school
3	Science is more difficult for me than for many of my classmates
4	I enjoy learning science
5	Science is not one of my strengths
6	I learn things quickly in science
7	Science is boring
8	I like science

Table 4. DIF Items and the Corresponding *z* Values.

DIF item	*z* Value	*p* Value
5	−3.015	.003
6	−3.282	.001
7	2.789	.005

Note. DIF = differential item functioning.

items are reverse coded so that for all items, larger value indicates higher level of enjoyment of science. See Table 3 for the items.

In practice, it is unknown if there is any DIF item on the scale; and if yes, how many. Therefore, we first examined whether there was any DIF item on this scale, when gender was used as grouping variable. To answer this, the M-SP method (Wang & Shih, 2010) was used to detect DIF items. Male students were treated as the reference group (coded as 0) and female students were treated as the focal group (coded as 1). After this step, Items 5, 6, and 7 were identified as DIF items. Bonferroni correction was adopted to control for the Type I error ($\alpha = 0.05/8 = 0.006$). The *z* values of these three items in the last iteration are shown in Table 4. They all led to *p* values that were below .006 and were therefore identified as DIF items. The negative *z* values indicate that a male student at the same enjoyment level of science as a female student is more likely to think science is more of their strength (Item 5), and is more likely to think that they learn science quickly (Item 6). The positive *z* values indicate that a female student at the same enjoyment level of science as a male student is more likely to endorse that science is interesting (Item 7).

The next step is to find a mediator to help explain the DIF effect. By examination of the variables included in the data set, we considered the sum score of the scale of self-confidence in learning science and math. Note that higher score indicates less self-confidence. By fitting each DIF item in the mediation model with the mediator, we obtain the *z* values of the direct effects and indirect effects as shown in Table 5. It can be found that the direct path from gender to Item 5 is not significant, and the direct path from gender to Items 6 and 7 are significant; the indirect path from

Table 5. Direct Effect and Mediation Effect for DIF Items.

Item	Direct	*p* Value	Mediation	*p* Value
5	−0.403	.687	−3.715	.000
6	−3.580	.000	−2.295	.022
7	3.888	.000	−0.608	.543

Note. DIF = differential item functioning.

gender to the mediator then to Items 5 and 6 are significant, and the indirect path from gender to the mediator then to Item 7 is not significant. This means that for Items 5 and 6, the proposed mediator helps account for the DIF effect for these items. Furthermore, for Items 5 and 6, the path from grouping variable to the mediator is positive, and the path from mediator to DIF item is negative, meaning that females at the same enjoyment level of science report lower level on these two items because they feel less confident on science and math than their male counterparts do. For Item 5, the mediator is a complete mediator since the direct path is no longer significant. For Item 6, since the direct path is still significant, the mediator is a partial mediator. It is possible that other mediators may be identified and added to the model and the DIF effect can be completely mediated by a group of mediators collectively. This real data example illustrates how the mediated MIMIC model can be used in practice to help understand the underlying DIF mechanism. **Mplus** codes of the mediated MIMIC model are provided in the Appendix for interested readers.

We would like to reiterate that it requires substantial content knowledge and familiarity with the items to identify possible mediators. In this empirical example, confidence in learning science and math turns out a significant mediator for Items 5 and 6, but not for Item 7. So confidence does not explain why a female student at the same enjoyment level of science as a male student is more likely to endorse that science is interesting. Further examination of the data set may help us identify another variable as the mediator for the DIF effect of Item 7. It is not uncommon that different causes will be identified for the DIF effects on different items. For example, baseball knowledge may be the cause of DIF for one item, and reading level may be the cause for another. In fact, if a common cause can be identified for all DIF items, then we should consider modeling that cause as a second dimension.

Discussion and Conclusion

DIF detection has received tremendous attention in educational and psychological testing. The motivation of our work is to help better understand the underlying mechanism of DIF. For example, a student with English as his or her second language may have lower score than a native speaker on a math item even if they have the same math ability. Language proficiency may mediate the DIF effect. It is very important in practice to understand the underlying cause for DIF, because otherwise

a DIF item may simply be dropped from an exam. However, ''a fine-grained analysis would likely uncover all items as systematically biased (with respect to certain socially defined groups)'' (Lubinski, 1996). Shall we drop all items requiring reading comprehension ability on a math test? New item development is costly and time-consuming. It is desirable to understand the underlying mechanism of DIF and revise an item accordingly.

The mediated MIMIC model allows us to test any variable that we believe might help explain DIF, and therefore can offer valuable information on how to revise a DIF item or even to implement targeted intervention. This way the mediated MIMIC model allows practitioners to go a step further than simply identifying DIF. Practitioners can find and then act on the mediators of DIF to not only improve tests but also to intervene and empower test takers with the skills required to test at their ability.

In simulation studies we show that for various total sample sizes and group sample size ratios, the mediated MIMIC model is successful in detecting the mediation effect when it actually occurs in the context of DIF, while keeping the Type I error rate well controlled in most practical conditions. The power only drops when the mediation effect is tiny and when the total sample size is 600 or smaller. When we have a large total sample size, the mediated MIMIC model is efficient in detecting even very small mediation effect.

We would like to emphasize here again that multiple mediators can be tested using the mediated MIMIC model framework. One mediator may be identified as a partial mediator. Combined with other mediators, they may completely account for the DIF effect. With that said, in practice, proposing good mediators requires considerable knowledge about the scale of interest and about the possible mediator variables. It is also true that to fit the mediated MIMIC model it requires data on both the scale of interest and the mediators. Data of possible mediators may not be easy to come by. This is one challenge of using the mediated MIMIC approach to account for the DIF effect. But this challenge is equally applicable to the MIRT approach, which was proposed to help account for the DIF effect. Additionally, the MIMIC model approach as described in this article assumes that the means of the latent trait between two groups differ but the variances across groups are equal. There is no such restriction when the MIRT approach is used. It warrants further study to examine the influence of such a constraint on the power and Type I error in detecting the mediation effect.

In the future, we would like to extend this study in several aspects. First, we would like to relax the assumption that the DIF items are known in our simulations. For that, we need to test which items have DIF, using for example the MIMIC model with scale purification method. Once the DIF items are identified, we will then proceed to test the mediation effect. Relaxing this assumption might have an effect on the conditions where the total DIF effect is small. If the total DIF effect is very small, the power of DIF detection may be low, and the item may not even be flagged for DIF. Consequently, no mediation effect will be tested even if it does exist. The other danger is inflated Type I error in DIF analysis. In other words, an item free of DIF may

be flagged for DIF because of multiple testing and an unnecessary test for the mediation effect will follow. Such an issue is inherent in DIF analysis when multiple items are tested for DIF. When the scale is iteratively purified, the large number of hypothesis tests inhibits the use of simple methods such as Bonferroni correction for Type I error control, because the power would be prohibitively low. Raykov, Marcoulides, Lee, and Chang (2013) addressed this issue by introducing the Benjamini–Hochberg procedure for multiple hypothesis testing through controlling the false discovery rate, or the rate of incorrect rejection of item-specific hypotheses concerning DIF.

Second, the DIF effect investigated in this study is uniform DIF, meaning that one group is biased in a certain direction over the entire latent trait range. It will be imperative to extend the mediation analysis to nonuniform DIF analysis in the future. The detection of nonuniform DIF using the MIMIC model was delineated in Woods and Grimm (2011). We expect to explore the mechanism of both uniform and nonuniform DIF using the mediated MIMIC model in the future. Another issue could emerge in the context of detecting DIF in polytomous items, namely, differential step functioning (DSF; Penfield, 2007, 2010), which means that the magnitude and/or direction of the DIF effect changes across the steps or categories underlying the polytomous response process. In our simulation study, the DIF effect is simulated as a difference between the threshold parameters of the focal and reference groups (i.e., $b_{ikF} = b_{ikR} + \text{TE}$, where TE represents the total magnitude of the DIF effect). Note that here the difference is a constant, meaning the magnitude and direction of the DIF effect does not vary across the response categories. It is yet to be examined how DSF affects the utility of the mediated MIMIC model approach.

Appendix

The Format of the Data Set

The figure below shows how the data is organized*:

```
0   3   4   3   3   3   3   4   4   4
1   4   1   4   1   4   2   4   3   2
0   3   4   2   4   3   3   4   4   5
0   4   4   3   4   4   3   4   4   2
0   4   1   4   4   4   4   1   4   2
1   4   3   3   2   3   3   2   2   2
0   4   3   3   3   2   3   3   3   2
1   4   1   4   2   3   4   1   2   2
1   3   4   4   3   3   4   2   3   2
1   2   1   2   3   2   2   2   2   5
1   3   3   3   2   3   3   2   2   2
1   4   3   4   2   4   4   2   3   2
1   4   2   4   3   4   4   2   3   2
```

Each row represents the response of one student. The first column is the grouping variable, 1 means focal group (female students), 0 means reference group (male). Columns 2 to 9 are the 8 items of the 8th-grade students' enjoyment of science scale; Column 10 is the sum score of self-confidence in learning science and math.

*Only the interested variables are exported from the original data set, and students with missing responses are removed.

Mplus *Code for Real Data Analysis*

```
TITLE: Mediated MIMIC DIF;

DATA:
  FILE = C:\desk\data.txt;

VARIABLE:
  NAMES =Gen SCI1-SCI8 Sum;
  USEVAR =Gen SCI1 SCI2 SCI3 SCI4 SCI5 SCI8 Sum;
  CATEGORICAL = SCI1 SCI2 SCI3 SCI4 SCI5 SCI8;

ANALYSIS: ESTIMATOR IS WLS;
PARAMETERIZATION=THETA;

MODEL:
  R by SCI1 SCI2 SCI3 SCI4 SCI5 SCI8;
  R on Gen;
  SCI5 on Sum;
  Sum on Gen;
  SCI5 on Gen;

MODEL INDIRECT:
  SCI5 ind Gen;
```

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: NSF 1350787.

Notes

1. The first number is the minimum sample size for the focal group, and the second number indicates the minimum sample size for the reference group.

2. The DIF here is uniform DIF. For nonuniform DIF testing using MIMIC, please see Woods and Grimm (2011).
3. Researchers may continue to find other possible mediators. The MIMIC model framework can easily accommodate multiple mediators at the same time. The collective mediation effect can also be tested in a straightforward manner.

References

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. *Journal of Educational Measurement*, *24*, 67-91.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research*, *35*, 169-199.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*, 333-353.

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.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel–Haenszel method. *Applied Measurement in Education*, *2*, 217-233.

Ekermans, G., Saklofske, D. H., Austin, E., & Stough, C. (2011). Measurement invariance and differential item functioning of the Bar-On EQ-i:S measure over Canadian, Scottish, South African and Australian samples. *Personality and Individual Differences*, *50*, 286-290.

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*, 278-295.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*, 275-284.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences*, *49*, 251-264.

Glockner-Rist, A., & Hoitjink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*, 544-565.

Hallquist, M., & Wiley, J. (2012). *MplusAutomation, version 0.5-1: An R package for automating Mplus model estimation and interpretation*. Retrieved from http://cran.r-project.org/web/packages/MplusAutomation/index.html

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, *2*, 81-117.

Hidalgo, M. D., & López-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement*, *62*, 32-44.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care*, *44*(11), 124-133.

Lubinski, D. (1996). Applied individual differences research and its quantitative methods. *Psychology, Public Policy, and Law*, *2*, 187-203.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*, 372-379.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. London, England: Routledge.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83-104.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*, 443-451.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte-Carlo examination of the sensitivity of the DFIT framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement*, *31*, 430-455.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus: Statistical analysis with latent variables* (Version 6.12). Los Angeles, CA: Muthén & Muthén.

Pae, T. I., & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*, 475-496.

Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel–Haenszel DIF detection procedure. *Applied Psychological Measurement*, *35*, 518-535.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*, 187-210.

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement*, *47*, 129-149.

Penfield, R. D., & Lee, O. (2010). Test based accountability: Potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*, *47*(1), 6-24.

R Development Core Team. (2011). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: Author. Retrieved from http://www.R-project.org.

Raykov, T., Marcoulides, G. A., Lee, C.-L., & Chang, C. (2013). Studying differential item functioning via latent variable modeling: A note on a multiple-testing procedure. *Educational and Psychological Measurement*, *73*, 898-908.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

Shealy, R., & Stout, W. (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.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician*, *40*, 106-108.

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel–Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*, 313-350.

Swaminathan, H., & Rogers, H. J. (1990). Detecting item bias using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*, 147-163.

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

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.

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

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.

Yao, L., & Li, F. (2010, May). *A DIF detection procedure in multidimensional item response theory framework and its applications*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–348). Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.