MIMIC Methods for Detecting DIF Among Multiple Groups: Exploring a New Sequential-Free Baseline Procedure

Applied Psychological Measurement
2016, Vol. 40(7) 486–499
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616659738
apm.sagepub.com

Seokjoon Chun¹, Stephen Stark¹, Eun Sook Kim¹,
and Oleksandr S. Chernyshenko²

Abstract

A simulation study was conducted to investigate the efficacy of multiple indicators multiple causes (MIMIC) methods for multi-group uniform and non-uniform differential item functioning (DIF) detection. DIF was simulated to originate from one or more sources involving combinations of two background variables, *gender* and *ethnicity*. Three implementations of MIMIC DIF methods were compared: constrained baseline, free baseline, and a new sequential-free baseline. When the MIMIC assumption of equal factor variance across comparison groups was satisfied, the sequential-free baseline method provided excellent Type I error and power, with results similar to an idealized free baseline method that used a designated DIF-free anchor, and results much better than a constrained baseline method, which used all items other than the studied item as an anchor. However, when the equal factor variance assumption was violated, all methods showed inflated Type I error. Finally, despite the efficacy of the two free baseline methods for detecting DIF, identifying the source(s) of DIF was problematic, especially when background variables interacted.

Keywords

differential item functioning, latent variable models, simulation

In applied psychology and in educational measurement, there is significant interest in identifying underlying “sources” of differential item functioning (DIF). DIF is said to occur when the psychometric properties of an item, such as discrimination and difficulty, differ for individuals selected from subpopulations that have equal standing on the trait that is being measured (Drasgow, 1987).¹ Many authors have noted that DIF results from items measuring secondary factors or dimensions on which comparison groups systematically differ (Camilli, 1992; Lopez-Rivas, Stark, & Chernyshenko, 2009; Shealy & Stout, 1993). For example, DIF on mathematical reasoning items might result from differences among comparison groups in English proficiency.

¹University of South Florida, Tampa, USA

²Nanyang Technological University, Singapore

Corresponding Author:

Seokjoon Chun, Department of Psychology, University of South Florida, 4202 East Fowler Ave., PCD 3120, Tampa, FL 33620, USA.

Email: seokjoon@mail.usf.edu

Alternatively, in personality assessment, DIF on conscientiousness items might result from socially desirable responding that is more prevalent among job applicants than non-applicants or from cultural differences (Chernyshenko, Stark, & Guenole, 2007; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001).

A related and, perhaps, more perplexing practical problem in DIF analysis is how to parse examinee groups for comparisons and what to do if different items are flagged as problematic in separate analyses involving background variables that co-occur, such as gender and ethnicity or gender and language. For example, what remedies might be suggested if one subset of items in a math test exhibits DIF in comparisons of males and females, because they tap spatial abilities, and another subset exhibits DIF in comparisons of Whites and English-as-second-language (ESL) Hispanics because they are influenced by English proficiency? Preferably, one could compare item properties using both background variables simultaneously to determine their relative contributions to DIF and inform revisions that would minimize the overall untoward effects of bias.

The purpose of this research was to propose and evaluate a method that might be suitable for detecting DIF due to two or more background variables and their potential interactions. More specifically, a Monte Carlo simulation was conducted to explore the efficacy of DIF detection using three implementations of Multiple Indicators Multiple Causes (MIMIC; Jöreskog & Goldberger, 1975) methodology, which is rooted in the confirmatory factor analysis (CFA) tradition.

CFA Methods for DIF Detection

In the last decade alone, many papers have compared and contrasted CFA and item response theory (IRT) approaches with DIF detection (e.g., Kim & Yoon, 2011; Stark, Chernyshenko, & Drasgow, 2006). Some have suggested that IRT methods are advantageous because they fit non-linear models directly to item responses, rather than linear models to inter-item correlation or covariance matrices, which is especially problematic with dichotomous variables. On the contrary, the unidimensionality assumptions of many IRT DIF methods and software packages limit the scope of invariance testing relative to CFA software, which can easily accommodate multidimensionality and multiple groups. Moreover, with the advent of categorical CFA methods, which incorporate the threshold (τ) structure with latent response process (Y^*), there is no longer a fundamental distinction between CFA and IRT (McDonald, 1999), so one could argue that the more general CFA approaches to DIF detection should be preferred.

There are at least two general ways to pursue DIF detection in the categorical CFA framework: Multi-group CFA (MGCFA) and MIMIC. MGCFA compares the fit of increasingly constrained measurement models across two or more groups. Essentially, it compares a model with one or more parameters estimated freely across groups to a model in which those parameters are constrained to be equal. Although the efficacy of MGCFA methods has been well established (Kim, Yoon, & Lee, 2012), there are some noteworthy limitations. In MGCFA, only one categorical variable can be used to define the comparison groups, so comparisons involving more than one background variable are cumbersome, and it is difficult to isolate the source of DIF when detected. In addition, because parameters are estimated for each group separately, each group must be large enough to adequately estimate model parameters, which leads to large overall sample size requirements. In contrast, MIMIC methods estimate just one set of model parameters using the total sample of respondents and test for DIF by adding or deleting paths from variables associated with group membership to the items under investigation. Because the full sample is used for parameter estimation, the total sample size needed for effective MIMIC DIF detection can be considerably smaller than with MGCFA (B. O. Muthén, 1989) and does not

increase when more than two groups must be compared. Also, by allowing for the inclusion of more than one background variable and its interactions, MIMIC models can be used to explore why DIF occurs, which makes it an attractive alternative to MGCFA methods.

MIMIC DIF Detection for Ordered Categorical Variables

Like some other CFA and IRT DIF methods, MIMIC DIF analysis involves comparing the fit of a series of full and reduced models with the goal of determining whether items measuring a latent variable are equally discriminating and difficult across comparison groups. What makes MIMIC unique is the way this is accomplished. Rather than fixing and freeing parameters reflecting item discrimination (loadings) and difficulty (thresholds) across groups, MIMIC tests for DIF by adding or deleting direct paths to items emanating from the background variables associated with group membership, and impact is accounted for by paths from grouping variables to the common factor. Essentially, MIMIC tells us how grouping variables affect item properties and factor means.

The first step in building a baseline model for MIMIC DIF analysis is to select a categorical background variable z that defines group membership. If two-group characteristics are of interest, for example, then two background variables can be selected, such as gender and ethnicity, and called $z1$ and $z2$. To test for a potential interaction of these variables, a third background variable $z1z2$ can be created. The next step in building a baseline model is to draw direct paths (γ) from each background variable to the common factor (θ) and paths (α) from the common factor to the items that will be tested for DIF, as shown in Figure 1a. Here, each categorical observed item response (Y) is construed as the manifestation of the underlying continuous response process (Y^*). And thresholds (τ s) on the response continuum determine whether a person endorses in one category or the next.

The next step depends on whether one wishes to use a *constrained baseline* (“all-others”; Wang, 2004) method or a *free baseline* (“constant anchor”; Wang, 2004) method to perform model comparisons (Stark et al., 2006). Most MIMIC DIF research has been conducted using the constrained baseline method in which items are tested for uniform DIF, associated with group differences in item thresholds (τ), by using the model shown in Figure 1a as a baseline and adding paths from each grouping variable (β) to individual items in a sequence of compact versus augmented model comparisons. If an augmented model fits significantly better than the compact baseline, then the item under investigation is flagged as a DIF item (Kim et al., 2012). Figure 1b shows the augmented model that would be used to test for threshold (uniform) DIF on Item 2 of a scale containing k items.

If one also wants to test for loading (non-uniform) DIF on Item 2, then additional variables can be created to reflect the moderating effect of background variables on common factors scores (Woods & Grimm, 2011). This is illustrated in Figure 1c, which shows three new variables $\theta z1$, $\theta z2$, and $\theta z1z2$ having direct paths (ω) to Item 2. Note that this augmented model performs an omnibus hypothesis test for DIF on item loadings (α) and thresholds (τ).

Although the constrained baseline approach to testing for DIF is convenient because it allows every item to be evaluated, it often leads to high Type I error rates because the baseline model is misspecified when DIF is present (Stark et al., 2006). To deal with this problem, researchers have explored using stricter p values for DIF detection or corrections to chi-square statistics on which model comparisons may be based (Oort, 1998). However, evidence is mounting that alternative free-baseline approaches are not only logically justified but also more effective (Lopez-Rivas et al., 2009; Stark et al., 2006; Woods & Grimm, 2011).

Free-baseline approaches to DIF detection begin by forming a baseline model that has only the necessary constraints for identification. In MGCFA DIF analysis, this might involve

Figure 1. Constrained baseline approach to test for DIF on Item 2 of a scale containing $i = 1, 2, \dots, k$ items with c categories per item: (a) The baseline model, (b) the augmented model to test for threshold (uniform) DIF, and (c) the augmented model to test for threshold and loading (non-uniform) DIF.

Note. j = group membership, γ_j = latent mean difference associated with background variable (z_j), θ = the common factor measured by the items, α_i = item loading, τ_i = item threshold, ω_{ij} = loading DIF, and β_{ij} = threshold DIF. DIF = differential item functioning.

Figure 2. Free baseline approach to test for DIF on Item 2 of a scale containing $i = 1, 2, \dots, k$ items with c categories per item: (a) The baseline model, (b) the reduced model to test for threshold and loading DIF.

Note. j = group membership, γ_j = latent mean difference associated with background variable (Z_j), θ = the common factor measured by the items, α_i = item loading, τ_i = item threshold, ω_{ij} = loading DIF, and β_{ij} = threshold DIF. DIF = differential item functioning.

constraining the loadings and thresholds for one item to be equal across comparison groups. Reduced models are then formed by constraining the loading and threshold parameters simultaneously for one additional item at a time and examining the change in goodness of fit for each reduced model relative to the baseline. If the fit worsens significantly, then the item under investigation is flagged as DIF. Stark et al. (2006) showed that this method yielded high power and low Type I error for IRT and mean and covariance structure (MACS; Sörbom, 1974) DIF detection, and Woods and Grimm (2011) showed that this general approach is effective with MIMIC.

Figure 2a shows a single-anchor free-baseline model for MIMIC DIF detection. Note that the model contains the same latent and observed variables as shown in Figure 1c, except that there

are paths from the variables associated with group membership to all items except Item 1, which was conveniently chosen as an anchor for model identification. To perform an omnibus test for DIF due to group differences in loadings and thresholds on Item 2, the paths to Item 2 from the variables associated with group membership are deleted, as shown in Figure 2b. This process is repeated for the other items in the scale and, in each case, a statistically significant decrease in goodness of fit relative to the baseline model indicates DIF.

As discussed by Stark et al. (2006), this single-anchor omnibus free-baseline approach to DIF detection has some desirable features. For example, by keeping the baseline model consistent across comparisons, error does not propagate as it would if items found to be free of DIF were subsequently added to the baseline to increase the size of the anchor group in the hope of increasing power, as is sometimes done in MGCFA invariance testing. In addition, by using just a single-anchor item, there is less chance of contaminating the baseline model by including a DIF item. However, as noted by Lopez-Rivas et al. (2009), the effectiveness of this method relies on choosing a satisfactory anchor—one that is unbiased and adequately discriminating. Otherwise, performance could be as bad as or worse than the constrained baseline approach, which works reasonably well when contamination due to DIF is not severe.

To choose a suitable anchor, Lopez-Rivas et al. (2009) suggested a *sequential-free baseline* approach similar to Thissen, Steinberg, and Wainer (1988). Specifically, they suggested performing DIF analysis in two steps. First, conduct constrained baseline tests to identify items that appear to be free of DIF. Then, choose the most discriminating non-DIF item as the anchor for subsequent free baseline tests of the other items in the scale. A study by Meade and Wright (2012) found that this two-step method, sequential-free baseline method for DIF analysis was the most effective of several that were considered. But, there has been no research exploring its efficacy with MIMIC, so studies are needed to see whether that finding generalizes.

Finally, because MIMIC DIF detection methods are relatively new, no published simulation studies have examined their efficacy with two or more grouping variables and interactions. Second, there have been no large-scale evaluations of free and constrained baseline tests for non-uniform DIF when an unbiased anchor item is not chosen a priori; choosing a problematic anchor item could completely undermine the benefits of the free baseline process. Third, very little research has examined the robustness of MIMIC methods to violations of the equal variance assumption for the common factor (θ) across groups. Violations of that assumption could inflate Type I error regardless of how the baseline model is specified. These unexplored issues motivated the following simulation study.

Method

Unlike most DIF simulations which have focused on pairwise group comparisons, MIMIC DIF detection was investigated here using *four* comparison groups, resulting from the co-occurrence of gender (male, female) and ethnic group status (majority, minority). “Male-Majority (MMA)” served as a reference group. “Male-Minority (MMI),” “Female-Majority (FMA),” and “Female-Minority (FMI)” served as focal groups whose item parameters were manipulated to create DIF associated with gender (G), ethnicity (E), and gender by ethnicity interactions ($G \times E$).

For comparability with previous research, and because many variables were examined, scale length was fixed at 15 items, and generating dichotomous and polytomous item parameters were based on the non-DIF and small DIF conditions of Stark et al. (2006).² In non-DIF conditions, reference group item parameters were used to generate both reference and focal group item responses, based on a categorical MGCFA model (B. O. Muthén & Asparouhov, 2002), via Mplus 7.11 (Muthén & Muthén, 1998–2013) scripts run using SAS PROC IML (SAS

Institute, 2010). In DIF conditions, DIF was simulated on Items 3, 8, 11, and 15 by decreasing focal group loadings by 0.15 (non-uniform DIF) or by increasing focal group thresholds by 0.25 (uniform DIF). (A detailed description of the data generation process, tables of reference and focal group generating item parameters, and plots showing the resulting main, marginal, and interactive effects associated with the grouping variables can be obtained from the first author upon request).

Independent Variables

176 experimental conditions were created by manipulating seven independent variables:

1. Response categories: dichotomous (i.e., two options), polytomous (i.e., five options).
2. Sample size per group (MMA = MMI = FMA = FMI): 125, 250.
3. Impact: none ($\mu_{Male} = \mu_{Female} = 0$), 0.5 SD ($\mu_{Male} = 0$, $\mu_{Female} = -0.5$).
4. Factor variance: equal ($\psi_{Male} = \psi_{Female} = 1$), unequal ($\psi_{Male} = 1$, $\psi_{Female} = 0.7$).
5. Type of DIF: none, threshold (τ), loading (λ).
6. Source of DIF: G, G E, $G \times E$, $G G \times E$, $G E G \times E$.
7. Baseline model: constrained, free, sequential-free.

Variable 6 was nested within the last two levels of Variable 5. Variable 7 was completely nested. That is, the same data sets were used for the constrained, free, and sequential-free baseline conditions. In each condition, 100 data sets were generated for the comparison groups using the designated item parameters and, assuming normality, the factor means and variances associated with independent Variables 3 and 4, respectively.

Analysis Details

MIMIC analyses were performed using Mplus 7.11 (Muthén & Muthén, 1998–2013). The robust maximum likelihood estimation (MLR) option was chosen so that the “XWITH” command could be used to model interactions between latent and grouping variables for non-uniform DIF tests. As was noted by Woods and Grimm (2011), modeling interactions involving binary grouping variables violates the normality assumption, which may inflate Type I error. However, as recommended by L. K. Muthén and Muthén (1998–2013, p. 694), the “XWITH” option was used with the MLR estimator, which produces “maximum likelihood parameter estimates with standard errors and a chi-square test statistic that are robust to non-normality and non-independence of observations” (L. K. Muthén & Muthén, 1998–2013, p. 603).

Model identification and standardization were accomplished by fixing the intercept of the common factor (θ) to 0 and fixing the variance of the common factor to 1 in the reference and focal groups. Because all items other than the one under investigation could be used to anchor the metric in constrained baseline analyses, every item was tested for DIF. However, free-baseline tests required an explicit referent for the model comparisons, so one item had to be left out of the DIF analyses.

For comparability with previous research by Stark et al. (2006), and to explore a recommendation by Lopez-Rivas et al. (2009) concerning the choice of anchor items for free baseline DIF analyses, this simulation examined MIMIC DIF detection using the following procedure. First, *constrained* baseline DIF analyses were performed on Items 1 to 15. In the *free* baseline conditions, Item 1, a discriminating non-DIF item, served as the designated referent for exploring DIF detection under the desirable, uncontaminated-one-item-anchor scenario. In the *sequential-free baseline* conditions, the most discriminating item identified as non-DIF in the constrained

baseline analyses was chosen as an anchor for subsequent free baseline tests on each replication, and the 14 remaining items were analyzed. The sequential-free baseline conditions thus provided a more realistic picture of MIMIC performance than the free baseline conditions, because a contaminated anchor could have been chosen.

The constrained, free, and sequential-free baseline analyses were performed in accordance with the procedures described in connection with Figures 1 and 2. Omnibus likelihood ratio (LR) tests were conducted for uniform and non-uniform DIF based on the Satorra and Bentler (2001) method for nested model testing with scaled chi-squares, which adjusts for potential bias due to multivariate non-normality (Bryant & Satorra, 2012):

$$\chi^2_{\text{DIFF}} = \frac{-2 \times (\text{loglikelihood}_{\text{reduced}} - \text{loglikelihood}_{\text{full}})}{C_{\text{LR}}},$$

where

$$C_{\text{LR}} = \frac{(q_{\text{reduced}})(c_{\text{reduced}}) - (q_{\text{full}})(c_{\text{full}})}{(q_{\text{reduced}}) - (q_{\text{full}})}, \quad (1)$$

q is the number of model parameters, c is a scaling correction factor for MLR chi-squares reported in the Mplus output, C_{LR} is the scaling factor for the chi-square difference statistic, and subscripts *reduced* and *full* refer to the reduced (compact) and full (augmented) models. For every studied item, the observed χ^2_{DIFF} was compared with a critical chi-square (12.59) corresponding to critical $p = .05$ and 6 df , because there were three covariates (G , E , $G \times E$) and three interactions with the common factor (θ^*G , θ^*E , $\theta^*G \times E$). If the observed χ^2_{DIFF} exceeded the critical chi-square, the item was flagged as DIF.

In addition, to see whether the MIMIC methods could accurately identify the source(s) of DIF following a hit (correct detection of a DIF item), outcomes known as Type III errors (Mosteller, 1948) were recorded. Here, Type III error was defined as the number of times the source of DIF was misidentified, divided by the number of hits, averaged over replications. Beginning with a baseline model that contained paths from all grouping variables to a studied item, three reduced models were formed by successively deleting paths from the $G \times E$, G , and E grouping variables in that order, and evaluating the χ^2_{DIFF} statistics with respect to a critical chi-square of 5.99 based on 2 df and critical $p = .05$. If a statistically significant result occurred for any grouping variable other than the true source of DIF, a Type III error was recorded. (Note that the errors were not calculated for the “ $G \ E \ G \times E$ ” conditions, because DIF was simulated based on all of the sources, that is, gender, ethnicity, and their interaction.)

To facilitate the interpretations of Type I error, power, and Type III error results, separate ANOVAs and a limited number of planned pairwise comparisons were performed. The ANOVAs tested for main effects and interactions involving up to three independent variables, and omega-square (ω^2) effect size statistics were reported, where .01, .06, and .14 represent small, medium, and large effects, respectively (Cohen, 1998).

Results

Type I error is defined as the proportion of items erroneously flagged as DIF (i.e., false positives) averaged over replications. Power is defined as the proportion of items correctly identified as DIF (i.e., hits) averaged over replications. It is important to note that there were no differences in power or Type I error across impact and no-impact conditions. This result was

expected, because MIMIC DIF methods account for latent mean differences explicitly in the baseline model, and impact is usually not a significant factor in CFA or IRT DIF studies.

It was also expected that Type I error rates near the nominal (.05) level would be observed in the equal factor variance conditions, but inflated Type I error would be observed in the unequal variance conditions, which violate MIMIC assumptions. These expectations were confirmed in the “No DIF” conditions. In the equal factor variance conditions, Type I error was generally near .05, with the highest value being .06. However, in the unequal factor variance conditions, several values exceeded .09, and one reached .14. Interestingly, somewhat better results were observed in the constrained baseline conditions, but this finding did not generalize to conditions where DIF was simulated.

Due to space limitations, power and Type I error results for polytomous conditions only are presented in Table 1 (Equal Variance) and Table 2 (Unequal Variance); results for the dichotomous conditions can be obtained from the authors. Overall, it can be seen that Type I error was lower in the equal variance conditions, Type I error was highest in the constrained baseline conditions, power improved as sample size increased, and power was higher for detecting DIF on thresholds than loadings. All of these findings were consistent with expectations. Importantly, similar power was observed in the sequential-free and free baseline conditions, indicating the viability of the sequential-free method in the absence of a priori information for choosing a referent. In fact, a detailed examination of individual simulation runs indicated that a DIF item was inappropriately chosen as an anchor only 1% of the time.

Further inspection of the equal variance results revealed that, in accordance with expectations, Type I error rates were markedly lower for the free and sequential-free baseline methods. Whereas only a few values reached .08 in those conditions, values were frequently high in constrained baseline conditions and reached a maximum of .36 in the polytomous $n = 250$ conditions with DIF due to main effects and interactions. Also consistent with expectations, power to detect threshold DIF was higher in polytomous conditions than in dichotomous conditions, with values averaging .85 and .62, respectively. Detailed inspection of the unequal variance results reveals a similar pattern. The most noteworthy finding is the substantially higher Type I error that occurred in all unequal factor variance conditions. Clearly, violating the equal variance assumption of MIMIC methodology led to spurious DIF detection, so a stricter statistical criterion for flagging DIF items is needed when model violations are suspected.

To further examine the results in Tables 1 and 2 and estimate effect sizes for the manipulations, ANOVAs were performed using Type I error and power as the dependent variables. For Type I error, baseline method, factor variance, and their interaction were statistically significant. As expected, Type I error was much higher for the constrained baseline method than the free and sequential-free baseline methods ($\omega^2 = .13$), and Type I error was significantly higher in the unequal factor variance conditions than in the equal variance conditions ($\omega^2 = .11$). There was also a statistically significant interaction between factor variance and baseline model ($\omega^2 = .05$).

Table 3 presents the ANOVA results for power. For power, as is usually the case, sample size had the largest effect ($\omega^2 = .32$) followed by the type and source of DIF. Threshold DIF was easier to detect than loading DIF ($\omega^2 = .21$), and DIF detection was better with polytomous data ($\omega^2 = .04$) than with dichotomous data. Only small effect sizes ($\omega^2 = .01$) were observed for two key variables of interest, baseline model and factor variance. Due to inflated Type I error in the constrained baseline and unequal variance conditions, the power differences across conditions were not as large as they might have been if Type I were controlled.

Finally, ANOVAs were conducted to examine the capacity of the MIMIC DIF methods to detect the underlying sources of DIF. These analyses used Type III error as the dependent variable. Overall, very high Type III error rates were observed, meaning that the true sources of DIF could not be reliably identified. Surprisingly, baseline model had the largest effect ($\omega^2 = .68$),

Table 1. Power and Type I Error in Polytomous, Equal Variance Conditions.

Source of DIF	<i>n</i>	Type of DIF	Impact	Baseline					
				Constrained		Free		Sequential-free	
				Power	Type I	Power	Type I	Power	Type I
G	125	Threshold	none	0.61	.12	0.63	.06	0.59	.05
			0.5 SD	0.56	.12	0.63	.05	0.57	.06
		Loading	none	0.22	.07	0.25	.06	0.23	.06
	250	Threshold	0.5 SD	0.25	.07	0.28	.06	0.26	.05
			none	0.90	.18	0.95	.05	0.93	.04
		Loading	0.5 SD	0.88	.16	0.93	.04	0.90	.04
GE	125	Threshold	none	0.46	.06	0.49	.05	0.49	.05
			0.5 SD	0.47	.06	0.49	.05	0.48	.04
		Loading	none	0.91	.18	0.94	.06	0.89	.05
	250	Threshold	0.5 SD	0.90	.20	0.94	.06	0.90	.05
			none	0.41	.08	0.45	.06	0.44	.05
		Loading	0.5 SD	0.44	.07	0.46	.06	0.46	.05
	250	Threshold	none	1.00	.32	1.00	.05	0.99	.04
			0.5 SD	0.99	.32	1.00	.04	1.00	.05
		Loading	none	0.76	.07	0.79	.05	0.77	.05
G × E	125	Threshold	0.5 SD	0.77	.08	0.81	.04	0.80	.04
			none	0.67	.13	0.69	.06	0.63	.06
		Loading	0.5 SD	0.68	.13	0.71	.05	0.65	.06
	250	Threshold	none	0.20	.07	0.24	.06	0.23	.05
			0.5 SD	0.26	.07	0.27	.06	0.26	.05
		Loading	none	0.93	.17	0.97	.05	0.94	.04
G G × E	125	Threshold	0.5 SD	0.93	.18	0.96	.04	0.91	.03
			none	0.45	.06	0.50	.05	0.49	.04
		Loading	0.5 SD	0.53	.06	0.57	.05	0.57	.04
	250	Threshold	none	0.74	.14	0.77	.05	0.70	.06
			0.5 SD	0.73	.15	0.78	.05	0.69	.06
		Loading	none	0.28	.07	0.29	.06	0.29	.05
	250	Threshold	0.5 SD	0.66	.07	0.32	.05	0.31	.05
			none	0.96	.22	0.97	.05	0.96	.04
		Loading	0.5 SD	0.94	.21	0.97	.04	0.95	.04
GE G × E	125	Threshold	none	0.52	.06	0.57	.06	0.56	.05
			0.5 SD	0.56	.06	0.58	.04	0.58	.04
		Loading	none	0.93	.20	0.97	.05	0.93	.05
	250	Threshold	0.5 SD	0.94	.21	0.96	.05	0.93	.04
			none	0.44	.08	0.51	.06	0.48	.05
		Loading	0.5 SD	0.49	.08	0.53	.06	0.52	.05
	250	Threshold	none	1.00	.36	1.00	.05	1.00	.04
			0.5 SD	1.00	.37	1.00	.04	1.00	.04
		Loading	none	0.79	.08	0.84	.05	0.84	.05
			0.5 SD	0.83	.10	0.88	.04	0.88	.04

Note. DIF = differential item functioning.

followed by source of DIF ($\omega^2 = .15$), and their interaction ($\omega^2 = .13$). Despite showing substantially better power and Type I error, the free and sequential-free baseline methods performed much worse in this respect, with Type III error near .95 in several instances. In contrast, the constrained baseline methods averaged a somewhat better .33. Interestingly, the same general pattern of results was observed across methods, with errors being the highest when interactions between sources of DIF were present: GE conditions (.08 constrained, .57 free), G (.16

Table 2. Power and Type I Error in Polytomous, Unequal Variance Conditions.

Source	n	DIF	Impact	Baseline					
				Constrained		Free		Sequential-free	
				Power	Type I	Power	Type I	Power	Type I
G	125	Threshold	none	0.62	.12	0.70	.11	0.67	.11
			0.5 SD	0.58	.13	0.68	.11	0.62	.12
		Loading	none	0.26	.06	0.47	.11	0.45	.10
	250	Threshold	0.5 SD	0.26	.06	0.46	.10	0.45	.10
			none	0.93	.17	0.97	.13	0.96	.12
		Loading	0.5 SD	0.88	.18	0.94	.12	0.89	.14
GE	125	Threshold	none	0.47	.04	0.75	.14	0.74	.13
			0.5 SD	0.52	.05	0.82	.13	0.82	.13
		Loading	none	0.91	.20	0.95	.10	0.91	.09
	250	Threshold	0.5 SD	0.89	.20	0.94	.11	0.90	.11
			none	0.38	.07	0.56	.10	0.55	.09
		Loading	0.5 SD	0.45	.06	0.56	.11	0.56	.11
G × E	125	Threshold	none	1.00	.33	1.00	.13	1.00	.13
			0.5 SD	0.99	.34	1.00	.12	0.99	.14
		Loading	none	0.74	.06	0.89	.13	0.88	.12
	250	Threshold	0.5 SD	0.74	.06	0.90	.13	0.91	.12
			none	0.68	.14	0.74	.10	0.67	.10
		Loading	0.5 SD	0.68	.15	0.75	.11	0.70	.11
G G × E	125	Threshold	none	0.19	.06	0.24	.10	0.23	.09
			0.5 SD	0.21	.06	0.27	.11	0.27	.10
		Loading	none	0.93	.17	0.97	.13	0.96	.13
	250	Threshold	0.5 SD	0.92	.18	0.97	.12	0.96	.11
			none	0.41	.06	0.53	.13	0.52	.13
		Loading	0.5 SD	0.49	.06	0.60	.12	0.36	.08
GE G × E	125	Threshold	none	0.75	.15	0.82	.11	0.77	.11
			0.5 SD	0.73	.15	0.81	.11	0.76	.12
		Loading	none	0.27	.06	0.45	.11	0.45	.10
	250	Threshold	0.5 SD	0.28	.05	0.49	.11	0.48	.10
			none	0.96	.21	0.99	.13	0.99	.13
		Loading	0.5 SD	0.93	.23	0.98	.12	0.95	.13
GE G × E	125	Threshold	none	0.53	.05	0.77	.14	0.76	.14
			0.5 SD	0.58	.06	0.81	.13	0.81	.13
		Loading	none	0.94	.21	0.97	.10	0.95	.09
	250	Threshold	0.5 SD	0.93	.22	0.96	.11	0.93	.11
			none	0.37	.06	0.56	.10	0.56	.09
		Loading	0.5 SD	0.47	.06	0.63	.10	0.62	.10
GE G × E	125	Threshold	none	1.00	.37	1.00	.13	1.00	.14
			0.5 SD	1.00	.39	1.00	.13	1.00	.14
		Loading	none	0.71	.07	0.89	.13	0.89	.12
	250	Threshold	0.5 SD	0.78	.07	0.92	.12	0.92	.12
			none						
		Loading	0.5 SD						

Note. DIF = differential item functioning.

constrained, .70 free), G, G × E (.21 constrained, .73 free), and G × E (.85 constrained, .95 free). The differences between the free and sequential-free baseline methods were minimal.

Discussion

In recent years, MIMIC methods have been suggested for DIF detection in situations involving multiple groups (Kim et al., 2012; Woods, 2009). However, until now, there has been no hard

Table 3. ANOVA Results for Power.

Source	df_B	F	ω^2
Sample size per group (N)	1	5,458.65	.32
Type of DIF (D)	1	4,013.43	.23
Source of DIF (S)	4	921.94	.21
Response categories (R)	1	678.33	.04
Factor variance (V)	1	229.45	.01
Baseline model (B)	2	112.82	.01
$D \times R$	1	1,219.38	.07

Note. Results are shown only for the factors that accounted for at least 1% of the variance in power. All effects shown were significant at $p < .05$. ω^2 = proportion of variance accounted for. df_B = degree of freedom between; for all effects, degrees of freedom within = 479. DIF = differential item functioning.

evidence to support that practice because the simulations showing MIMIC efficacy have used two-group designs (Finch, 2005; Kim et al., 2012; Woods, 2009). In addition, there has been little, if any, research examining MIMIC accuracy for detecting DIF due to combinations of background variables and interactions. To our knowledge, this is also the first study that examined the performance of MIMIC DIF methods across the three implementation approaches (i.e., sequential-free, free, and constrained baselines) and under violations of the equal factor variance assumption.

As expected, the authors found that the free and sequential-free baseline methods provided better overall DIF detection than the constrained baseline method. Furthermore, the more practical sequential-free baseline method performed nearly as well as the free baseline method, which indicates that the proposed empirically driven process for choosing an anchor item for DIF detection was highly effective.

Although using just one anchor item has been shown to yield slightly lower power than using a larger DIF-free anchor subset (Lopez-Rivas et al., 2009; Meade & Wright, 2012), the two-step approach proposed in this article provides a way to screen the maximum number of items for DIF, while reducing the risk of Type I error caused by inadvertently choosing a contaminated anchor. The latter has been shown to be much more of a concern (Stark et al., 2006). However, as was noted by a reviewer, the performance of DIF detection under the MIMIC methods might be improved as the number of DIF-free anchor items is increased, for example, from one to five anchor items (Lopez-Rivas et al., 2009; Meade & Wright, 2012). Therefore, further research is required to investigate the increased efficacy of MIMIC methods afforded by more anchor items.

Also, as expected, all three methods performed markedly better when the MIMIC equal factor variance assumption was satisfied. When this assumption was violated, Type I error was above .05 regardless of the baseline model specification. Therefore, a correction factor or stricter statistical criterion for flagging DIF items is recommended when violations are suspected.

Despite the intuitive appeal of MIMIC methods for identifying the source(s) of DIF when more than one background variable is considered, the high Type III error rates indicate that caution must be exercised. None of the methods examined here were effective overall, and Type III errors rose sharply when background variables interacted. Research is therefore needed to improve efficacy before MIMIC methods can be recommended for this purpose.

We hope that future research will also examine how MIMIC DIF detection efficacy varies as a function of the difference in factor variances across comparison groups. This study explored only one possibility, a 0.3 standard deviation difference, due to the large number of conditions.

However, smaller variance differences may be more realistic and have less adverse effect on performance. It would also be interesting to explore MIMIC efficacy with measures involving more than one dimension by design.

Altogether, this research conclusively shows that MIMIC methods provide a worthwhile alternative to MGCFA and traditional IRT approaches to detecting DIF. Of the MIMIC implementations considered here, the sequential-free baseline method is recommended because it is both practical and effective. When just one item is used as an anchor for DIF detection, the choice becomes especially important. And, in most practical situations, the authors believe the empirically driven sequential-free process will perform best in the long run.

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.

Notes

1. The null differential item functioning (DIF) hypothesis is $P(Y_{ij} = y | \theta_i, G) = P(Y_{ij} = y | \theta_i)$. Given the trait score (θ_i), the conditional probability of the response of the i th person on the j th variable (Y_{ij}) is independent of variable G .
2. We used Multi-group confirmatory factor analysis (MGCFA) to simulate dichotomous data consistent with the item response theory (IRT) two-parameter normal (2PN) model (without guessing) and polytomous data consistent with a graded response model.

References

- Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 19, 372-398.
- Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, 16, 129-147.
- Chernyshenko, O. S., Stark, S., & Guenole, N. (2007). Can the discretionary nature of certain criteria lead to differential prediction across cultural groups? *International Journal of Selection and Assessment*, 15, 175-184.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19-29.
- 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.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631-639.
- Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, 18, 212-228.
- Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, 72, 469-492.
- Lopez-Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters upon DIF detection using the free-baseline likelihood ratio test. *Applied Psychological Measurement*, 33, 251-265.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

- Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*, 1016-1031.
- Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics, 19*, 58-65.
- Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.
- Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes, 4*(5), 1-22. Retrieved from <http://www.statmodel.com/mplus/examples/webnote.html#web4>
- Muthén, L. K., & Muthén, B. O. (1998-2013). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.
- Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107-124.
- SAS Institute. (2010). *SAS 9.3 user's guide*. Cary, NC: Author.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moments structure analysis. *Psychometrika, 66*, 507-514.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects bias/DIF as well as item bias/DIF. *Psychometrika, 59*, 159-194.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.
- Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied psychology, 86*, 943-953.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 19*, 1292-1306.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.
- Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.
- 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.