Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models

Carol M. Woods and Kevin J. Grimm

Applied Psychological Measurement 2011 35: 339
DOI: 10.1177/0146621611405984

Article

Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models

Applied Psychological Measurement
35(5) 339–361

DOI: 10.1177/0146621611405984
http://apm.sagepub.com

Carol M. Woods[1] and Kevin J. Grimm[2]

Abstract

In extant literature, multiple indicator multiple cause (MIMIC) models have been presented for identifying items that display uniform differential item functioning (DIF) only, not nonuniform DIF. This article addresses, for apparently the first time, the use of MIMIC models for testing both uniform and nonuniform DIF with categorical indicators. A latent variable interaction is added to the MIMIC model to test for nonuniform DIF. The approach is tested in simulations with small focal-group *N* and illustrated with an empirical example using a scale about agoraphobic cognitions. MIMIC-interaction models are compared with MIMIC models without the interaction as well as likelihood ratio DIF testing using item response theory (IRT-LR-DIF). The most important finding is that when the latent moderated structural equations approach is used to estimate the interaction, the Type I error in MIMIC-interaction DIF models is severely inflated.

Keywords

differential item functioning, MIMIC, item bias, item response theory, structural equation modeling

It is desirable to create items on tests, questionnaires, and interviews with measurement properties that are invariant to irrelevant characteristics of the persons being measured, such as gender and race, when one is measuring constructs like depression or antisocial personality. Differential item functioning (DIF) occurs when an item has different measurement properties for one group of people versus another, irrespective of true mean differences on the construct. To test for DIF, a reference group is compared with a focal group (or more than one, but the authors deal with a single focal group here). DIF may be uniform or nonuniform (Mellenbergh, 1989), where the group difference in endorsement probability (or item difficulty, if applicable) is constant over the latent continuum for uniform DIF but not for nonuniform DIF.

[1]University of Kansas, Lawrence, KS, USA
[2]University of California at Davis, CA, USA

Corresponding Author:
Carol M. Woods, University of Kansas, 1415 Jayhawk Blvd., Room 426, Lawrence, KS 66045-7556, USA
Email: cmw@ku.edu

There are many methods used to test for DIF. For example, the Mantel–Haenszel (MH; Holland & Thayer, 1988; Mantel & Haenszel, 1959) test, including its variants for polytomous items (Mantel, 1963; Somes, 1986), is very popular. However, MH tests have limitations, including an absence of latent variables (i.e., no adjustment for measurement error) and low power to detect nonuniform DIF (by design). Also, an emerging literature suggests that MH tests are sensitive to differences in latent-variable variances between the focal and reference groups (Pei & Li, in press) and that they lack robustness to nonnormality despite being nonparametric procedures (Woods, in press).

Two popular methods for DIF testing with latent variables use multiple-group models or multiple indicator multiple cause (MIMIC) models. The distinguishing feature of a MIMIC model (Jöreskog & Goldberger, 1975) is that at least one observed variable, called a causal indicator, predicts a latent variable. The two-group item response theory (IRT) approach for DIF testing using likelihood ratio (LR) tests (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993) is often called IRT-LR-DIF.

In a previous article that compares IRT-LR-DIF and MIMIC models with small focal-group *N* (Woods, 2009a), MIMIC models were set up so that only uniform, not nonuniform, DIF could be detected. This is consistent with the way MIMIC models are described for testing DIF in extant literature (Chen & Anthony, 2003; Christensen et al., 1999; Finch, 2005; Fleishman, Spector, & Altman, 2002; Gelin, 2005; Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Hagtvet & Sipos, 2004; MacIntosh & Hashim, 2003; Mast & Lichtenberg, 2000; B. O. Muthén, 1985, 1988, 1989; B. O. Muthén, Kao, & Burstein, 1991; Oishi, 2006; Schroeder & Moolchan, 2007; Shih & Wang, 2009; Wang & Shih, 2010).

Woods (2009a) found that the sample size needed for adequate power and reasonably accurate estimates of most item parameters was smaller for MIMIC models than for IRT-LR-DIF. With varying focal (*F*) and reference (*R*) group *N*s and scale lengths, Type I error was well controlled, and estimates of the *F* group mean were quite accurate. However, IRT-LR-DIF always had greater power to detect nonuniform DIF than MIMIC models, and bias was elevated for some MIMIC-model item parameter estimates.

The current article clarifies, for apparently the first time, how a MIMIC model with categorical indicators can be set up to test for uniform and nonuniform DIF simultaneously. The fundamental idea for testing nonuniform DIF is to add an interaction to the model. Although this may seem simple, it does not appear to be widely known. Extant literature describes MIMIC models without the interaction, thus, as being capable of testing uniform DIF only. MIMIC-interaction models for testing nonuniform DIF are explained, studied in simulations, and illustrated with empirical data. Comparisons to IRT-LR-DIF and MIMIC models without the interaction are included.

IRT-LR-DIF

To carry out IRT-LR-DIF, nested two-group item response models are compared using LR tests. Here, the IRT models are estimated using Bock and Aitkin's (1981) scheme for marginal maximum likelihood implemented with an expectation maximization algorithm (EM MML). The mean and variance of the latent variable, θ, are fixed to 0 and 1 (respectively) for the *R* group to identify the scale and estimated for the *F* group as part of the DIF analysis. A subset of items, called designated anchors, are presumed to have group invariant parameters, not tested for DIF, and used to link the metric of θ for the two groups.

Item parameters for designated anchors are constrained equal between groups, although each studied item (nonanchor) is tested individually for DIF. For a particular studied item, an analysis begins with a test of the null hypothesis that all parameters for studied item *i* are group invariant.

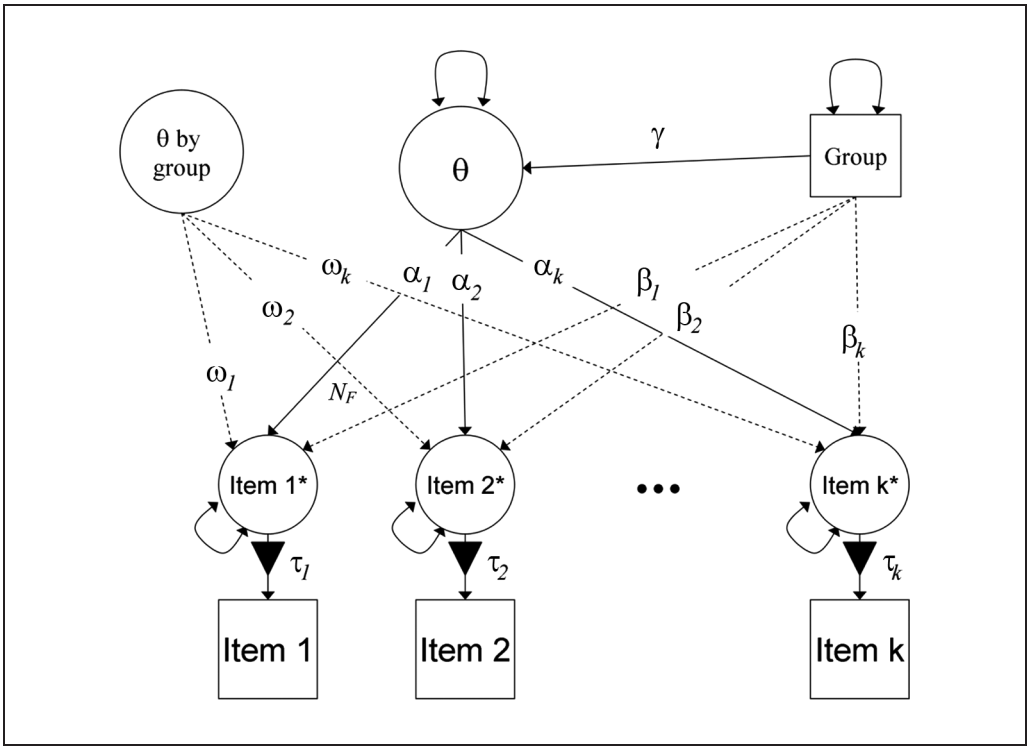

Figure 1. A MIMIC model for testing nonuniform DIF with interaction between group and θ
Note: MIMIC = multiple indicator multiple cause; DIF = differential item functioning; γ = mean difference on the latent variable, θ; items $i = 1, 2, \ldots, k$; α_i = discrimination; ω_i = nonuniform DIF effect; τ_i = threshold; β_i = group difference in the threshold.

A model with all parameters for the studied item constrained equal between groups is compared with a model with all parameters for the studied item permitted to vary between groups. In both models, parameters for all anchors are constrained equal between groups. A significant omnibus test (i.e., test for uniform and nonuniform DIF simultaneously) for an item indicates the potential presence of DIF, and follow-up tests can easily be carried out to distinguish between uniform and nonuniform DIF.

IRT-LR-DIF can be carried out with any software that is able to fit IRT models, but is particularly convenient with Thissen's (2001) free IRT-LR-DIF software, which has performed well in simulations with binary and ordinal data (Ankenmann, Witt, & Dunbar, 1999; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Sweeney, 1996; Wang & Yeh, 2003). Important assumptions of IRT-LR-DIF here are independent observations and groups, locally independent items, logistic item response functions (with an underlying continuous item response process), a normal latent variable for each group (the mean and variance may differ between groups), and DIF-free anchor items.

MIMIC Models for DIF Testing

Figure 1 depicts a unidimensional MIMIC-interaction model.[1] A continuous response process is assumed to underlie each categorical observed item response, and thresholds (τs) on the response

continuum determine whether a person responds in one category or the next (Figure 1 shows the case of binary items; generally, the number of thresholds is one less than the number of responses). The MIMIC model is parameterized as an IRT model, and EM MML is used here to fit the model to the item data directly. The latent scale is identified by fixing the mean and variance of θ to 0 and 1, respectively, and the grouping variable is dummy (0-1) coded.

The MIMIC model for DIF testing that was popularized by B. O. Muthén (e.g., 1985, 1988, 1989; see also MacIntosh & Hashim, 2003; B. O. Muthén et al., 1991) tests for uniform DIF only and does not include the interaction that is shown in Figure 1. To test for uniform DIF, item *i* is regressed on the latent variable (θ), and group, z: $y_i^* = \alpha_i\theta + \beta_i z + \varepsilon_i$, where y_i^* = continuous response process that underlies a discrete y_i, α = discrimination parameter, β = regression coefficient showing the group difference in the threshold, and ε = unique factor (error). There is evidence of DIF if z significantly predicts item responses, controlling for any mean differences on θ. Discrimination parameters are implicitly invariant; thus, the MIMIC model without an interaction tests for uniform DIF.

In MIMIC models, θ is regressed on z to allow for a mean difference. Some researchers may prefer, for conceptual reasons, to have z correlate with, instead of predict, θ. For a single covariate, results are identical either way. Setting it up so that z predicts θ permits generalization for additional predictors of θ. For example, researchers may wish to consider both gender and ethnic groups as potential sources of DIF.

The interaction between z and θ is the key to testing for nonuniform DIF. Item *i* is regressed on the interaction ($y_i^* = \alpha_i\theta + \alpha_i z + \omega_i\theta z + \varepsilon_i$), which specifies that the relationship between item response and z depends on the level of θ. To carry out DIF testing, models are compared using LR tests. An omnibus test evaluates DIF for one item at a time (called ''omnibus'' because uniform and nonuniform DIF are tested simultaneously). First, a full model is fitted in which all studied items are regressed on z and the interaction. Next, one constrained model is fitted per studied item. In these models, the studied item is regressed on neither z nor the interaction. A statistically significant difference between the full and constrained models suggests that the item potentially functions differently between groups. Statistically significant effects from the full model indicate whether the item shows uniform DIF (β_i), nonuniform DIF (ω_i), or both.

The interaction term complicates the computations because the latent variable (θ) cannot be simply multiplied by z. Various methods have been proposed to compute interactions involving latent variables (Klein & Muthén, 2007, include a recent review). However, the research is focused on interactions between continuous latent variables, and latent interactions are neither well studied nor well implemented at this time for the case of categorical indicators.

Mplus currently provides convenient estimation options for MIMIC models with categorical indicators. In the current Mplus user's guide (version 6), it is recommended that an interaction between an observed categorical variable and a continuous latent variable be obtained using either a multiple group approach or using ''XWITH'' (p. 612). XWITH invokes a maximum likelihood approach that is the same as the latent moderated structural equations (LMS) method of Klein and Moosbrugger (2000), except using a different algorithm. With LMS, the joint distribution of indicators for predictors and indicators for outcomes is represented as a mixture of normals, and mixture parameters are estimated simultaneously with the parameters of the model.

LMS assumes that the latent variables involved in the interaction are both normal, and Klein and Moosbrugger (2000) observed Type I error inflation when this assumption was violated. Although LMS is not appropriate when one variable is categorical, Mplus is a widely used program for models with categorical indicators, with a user's guide that recommends using LMS with models like the present MIMIC-interaction DIF model. For this reason, simulations are carried out using LMS (with Mplus) for MIMIC-interaction models despite the assumption violation.

Important assumptions of MIMIC-interaction models here are independent observations and groups, locally independent items, logistic item response functions (with an underlying continuous item response process), group-equivalent θ variance (the θ mean may differ between groups), and DIF-free anchor items. The EM MML methods used here assume θ is normally distributed, and the LMS procedure assumes the two variables interacting are normal (the binary ''group'' variable violates this assumption).

Relationships With Other Methods

The rationale for DIF testing using MIMIC-interaction models is analogous to that for logistic regression (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990), where item response is regressed on group, observed summed score (as a proxy for the latent variable) and their interaction. For binary logistic, a significant interaction indicates that the odds of responding correctly are greater for the *R* group at certain summed scores but greater for the *F* group at other summed scores. Logistic regression for DIF seems to work fairly well when assumptions (e.g., anchor items are actually DIF free) are met, but MIMIC models improve on logistic regression because they model measurement error.

MIMIC-interaction models are closely related to restricted factor analysis (RFA) models with an interaction (Barendse, Oort, & Garst, 2010). RFA models (Oort, 1992, 1998) and MIMIC models are set up identically to test for DIF, except that the relationship between group and θ is a correlation in RFA versus a regression in MIMIC. Both models can test for nonuniform DIF with the addition of an interaction between group and θ. The approach of Barendse et al. (2010) differs from the present methodology because it uses continuous indicators, represents group as a latent variable,[2] and treats all other items as anchors for model comparisons. Consistent with what is usually found when the anchor set is contaminated by differentially functioning (D-F) items, Barendse et al. observed inflated Type I error rates in conditions with D-F items.

Simulation Methods

Simulations were carried out to test the performance of the MIMIC-interaction model. The same conditions used by Woods (2009a) were replicated; details are given below. A C++ program[3] generated the data and wrote command files for, executed, and processed output from Mplus (L. K. Muthén & Muthén, 2007). Both binary and 5-category Likert-type item data were generated for independent conditions varying according to the *F*-group sample size ($N_F = 25$, 50, 100, 200, or 400), number of items ($k = 6$, 12, or 24), *R*-group sample size ($N_R = 500$ or 1,000), and presence versus absence of DIF. With Likert-type items, $N_F = 25$ was not used because results with binary data were poor and N_R was fixed at 1,000 because most outcomes for binary items were virtually identical with $N_R = 500$ versus 1,000. There were 100 replications for all conditions.

Binary Item Data

Binary responses were generated from the two-parameter logistic model (2PL; Birnbaum, 1968):

$$T_{ij} = \Pr(u_{ij} = 1|\theta_j) = \frac{1}{1 + \exp[-a_i(\theta_j - b_i)]}, \quad (1)$$

where a_i = discrimination and b_i = threshold for item i, and θ_j is the value of the latent variable for the jth simulee; θ_j was drawn from $N(\mu = 0, \sigma = 1)$ for the *R* group and $N(-0.4, 1)$ for the *F* group.

Item parameters for the R group were randomly drawn from distributions: $N(\mu = 1.7, \sigma = 0.3)$ for a_{iR} and $N(\mu = 0, \sigma = 1)$ for b_{iR}, chosen based on an empirical examination of item parameter estimates from various psychological scales (Hill, 2004). To avoid unrealistic extreme values, the distribution of a_{iR} was truncated on the upper end at 4 and on the lower end at 0.5 (items without nonuniform DIF) or 1.2 (items with nonuniform DIF). The maximum amount of DIF was 0.7, so truncation at 1.2 ensured that $a_{iF} \geq 0.5$. The distribution of b_{iR} was truncated at ± 2 to avoid items with all responses in a single category.

F-group parameters were defined in relation to R-group parameters. In DIF-free conditions, $a_{iF} = a_{iR}$ and $b_{iF} = b_{iR}$ for all i, and k / 3 items were used as anchors. In conditions with DIF, $2k$ / 3 items functioned differently in favor of the R group (i.e., $a_{iF} < a_{iR}$ and $b_{iF} > b_{iR}$), and k / 3 items were anchors. Half of the D-F items were variant in both a_i and b_i (nonuniform DIF) and the other half were variant in just b_i (uniform DIF). For example, when $k = 12$, there were 4 items with nonuniform DIF, 4 items with uniform DIF, and 4 DIF-free anchors.

In applications, the amount of DIF usually varies over items within a study, and typical differences: $|a_{iF} - a_{iR}|$ or $|b_{iF} - b_{iR}|$ are between .3 and .7. In this simulation, $a_{iF} = a_{iR} - \varphi$ and $b_{iF} = b_{iR} + \delta$, where φ and δ were equal to one of five equally likely values (.3, .4, .5, .6, or .7) and $\varphi \neq \delta$ (except by chance). A random uniformly distributed number determined φ (or δ) for a given item.

Likert-Type Item Data

Five-category ordinal data were generated from Samejima's (1997) graded model, a generalization of Birnbaum's 2PL model for more than two ordered categories. There is one discrimination parameter per item (a_i) and $c - 1$ threshold parameters ($b_{i1}, b_{i2}, \ldots, b_{ic-1}$), with c = total number of response categories. For each item response (u_i), the graded model trace line describes the probability that the response is in category u or higher minus the probability that it is higher:

$$T(u_i) = \frac{1}{1 + e^{[-a_i(\theta - b_{i,(u-1)})]}} - \frac{1}{1 + e^{[-a_i(\theta - b_{i,u})]}}.$$

For each item, a_{iR} was drawn from $N(\mu = 1.7, \sigma = 0.6)$ with truncation at 4.0 and 0.5 (items without nonuniform DIF) or 1.2 (items with nonuniform DIF). As with binary items, a_{iF} was either equal to a_{iR} or $a_{iR} - \varphi$, depending on whether nonuniform DIF was present.

The first R-group threshold, b_{i1R}, was drawn from $N(\mu = -0.4, \sigma = 0.9)$ with truncation at -2.5 and 1.5. Subsequent thresholds were created by adding a randomly drawn value, d_{imR}, to the immediately previous threshold (m counts differences between consecutive b_{ihR}s, where $h = 1, 2, \ldots, c - 1$). The difference between adjacent b_{ihR}s was drawn from $N(\mu = 0.9, \sigma = 0.4)$, with truncation at 0.1 and 1.5. To preserve ordering of the thresholds ($b_{i1F} < b_{i2F} < b_{i3F} < b_{i4F}$), the amount of DIF was held constant over thresholds for each item. F-group thresholds were defined as $b_{i1F} = b_{i1R} + \delta$, $b_{i2F} = b_{i2R} + \delta$, $b_{i3F} = b_{i3R} + \delta$, and $b_{i4F} = b_{i4R} + \delta$. For example, if the amount of DIF in the thresholds was $\delta = .4$ for a particular item, then every F-group threshold was equal to the corresponding R-group threshold plus .4. When an item was simulated such that 0 simulees responded in one or more of the 5 categories for either the R or F group, the categories for the item were collapsed for both groups.

MIMIC Model Omnibus DIF Tests

MIMIC models were fitted using Mplus (version 5.2; L. K. Muthén & Muthén, 2007). For each data set, $2 / 3k + 2$ different MIMIC models were fitted. In every model, the intercept of θ was

fixed to 0, θ was regressed on the grouping variable, and the variance of the residual from this regression was fixed to 1. The XWITH command was used for interactions. All models were parameterized as IRT models. The Mplus parameterization of the 2PL is

$$\Pr(u_{ij}=1|\theta_j)=\frac{1}{1+\exp[\tau_i-\alpha_i\theta_j]},$$

where τ_i is not equivalent to b_i in Equation 1, but $\tau_i = \alpha_i b_i$. Mplus parameterizes the graded model analogously: $\tau_{i1} = \alpha_i b_{i1}$, $\tau_{i2} = \alpha_i b_{i2}$, . . . , $\tau_{ic\text{-}1} = \alpha_i b_{ic-1}$.

Models were fitted to the data directly using the robust maximum likelihood estimator ('MLR'), which uses an EM MML algorithm. With MLR, the individual model χ^2 values are scaled and cannot be used (without adjustment) for nested model comparisons because a difference between two scaled χ^2 values is not distributed χ^2. To compare models, the LR statistic was divided by a scaling correction term:

$$c=\frac{df_0\frac{T_0}{S_0}-df_1\frac{T_1}{S_1}}{df_0-df_1},$$

where $df =$ degrees of freedom, $T =$ regular χ^2 value, $S =$ scaled χ^2 value, and subscripts 0 and 1 identify the bigger and smaller models, respectively. This is the Satorra and Bentler (2001) method for nested model testing with scaled χ^2s. (See also www.statmodel.com/chidiff.shtml.)

Tests of omnibus DIF (designed to detect uniform DIF, nonuniform DIF, or both) were carried out for every studied item. No follow-up tests specifically for uniform or nonuniform DIF were carried out. In conditions with DIF, a final model was fitted, in which only items with significant DIF tests were regressed on group. Final estimates of τ_i were converted to b_i so they could be compared with the true parameters.

Outcomes

DIF-free conditions. The false-positive rate (i.e., proportion of studied items with significant tests) was recorded for conditions without any D-F items. In Woods (2009a), p values were corrected using the Benjamini–Hochberg (BH; Benjamini & Hochberg, 1995) false discovery rate adjustment (Thissen, Steinberg, & Kuang, 2002; Williams, Jones, & Tukey, 1999), but this seemed to be an overcorrection for MIMIC models and IRT-LR-DIF. Here, results for all three methods are reported without any p-value correction ($\alpha = .05$).

Conditions with DIF. One study-level outcome was recorded for conditions with D-F items: The coefficient for the regression of θ on group from the MIMIC-interaction models (γ; i.e., mean difference) from the final model averaged over replications.

Item-level outcomes differed for anchors versus D-F items. For anchors, the absolute value of the mean bias was calculated for the item parameters, where b_{ij} is compared with $\hat{\tau}_{ij}/\hat{\alpha}_i$ from the MIMIC model. For D-F items, the hit rate was the proportion of items with significant tests. Hit rates were calculated separately for items with uniform versus nonuniform DIF. Woods (2009a) reported hit rates based on BH-corrected p values; here, the authors report them based on raw p values for all methods.

Four additional outcomes were used to judge how well item parameters were estimated for D-F items when the significance tests were correct. For example, if $k = 12$ and 565 of the 800 D-F items were detected, bias was evaluated for only those 565 items (800 = 8 D-F items per test times 100 replications). Bias in a_{iR}, a_{iF}, b_{ijR}, and b_{ijF} was computed using estimates from the final

model for only the items that were correctly identified as D-F (b_{ijR} and b_{ijF} were compared with $\hat{\tau}_{ijR}/\hat{\alpha}_{iR}$ and $\hat{\tau}_{ijF}/\hat{\alpha}_{iF}$).

Results for MIMIC-interaction models will be presented. None of the false alarm or hit rates have been reported previously. Otherwise, if there was a nontrivial change (arbitrarily defined as more than about .02 or .03 units) to a particular outcome with versus without the interaction, the new results are reported alongside Woods' (2009a) results.[4]

Simulation Results

Estimation Difficulty With Smaller N_F

An estimation problem that did not occur in Woods' (2009a) simulations occasionally occurred with MIMIC-interaction models. The Mplus error message was, ''The model estimation did not terminate normally due to an ill-conditioned Fisher information matrix. Change your model and/or starting values.'' The problem was extremely rare when $N_F \geq 100$ (i.e., 5 out of 5,400 replications) and occurred less frequently for Likert-type versus binary responses. It happened most often for conditions with $N_F = 25$, binary responses, and some D-F items. For conditions with 6, 12, and 24 items, respectively, failure frequencies were 5, 11, and 12 with $N = 500$ and 9, 14, and 17 with $N = 1{,}000$. The error was likely related to sparse responding in each category, which necessarily occurs with small samples. When the error occurred, the replication could not be used because full results were not produced.

False-Positive Rates

Figure 2 shows false-positive rates for IRT-LR-DIF and MIMIC models without the interaction, compared with MIMIC-interaction models. Rates for MIMIC-interaction models were unacceptably high, whereas the rates for MIMIC and IRT-LR-DIF were near the nominal level (.05) and similar to one another.

Hit Rates

As shown in Figures 3 (binary responses) and 4 (ordinal responses), MIMIC-interaction models often had the most power but recall that Type I error was unacceptably high for these models. One exception was ordinal items and nonuniform DIF—power was almost always greatest for IRT-LR-DIF. The effect of using raw, versus BH-corrected, p values was that current hit rates were less attenuated for MIMIC models (without the interaction) and IRT-LR-DIF than in Woods (2009a); the overall pattern of results was the same.

Although power to detect nonuniform DIF was quite high for IRT-LR-DIF under most conditions with ordinal responses, the authors report below that bias was also elevated for b_{ijF} in these conditions (for IRT-LR-DIF). Perhaps the power advantage of IRT-LR-DIF was related to the larger bias: If $\hat{b}_{ijF}$s were inaccurate such that the amount of DIF was overestimated for items with DIF, then the chance of rejecting H_0 increases, thereby increasing power.

Estimates of Mean Difference

For binary responses, estimated mean differences were almost exactly the same with and without the interaction; thus, results are not repeated here (see Woods, 2009a). For ordinal responses, the MIMIC-model estimate of the mean difference improved for $k = 12$ or 24 and $N_F = 50$ or 100 when the interaction was added and became extremely accurate (see Figure 5).

(Text continues on p. 351.)

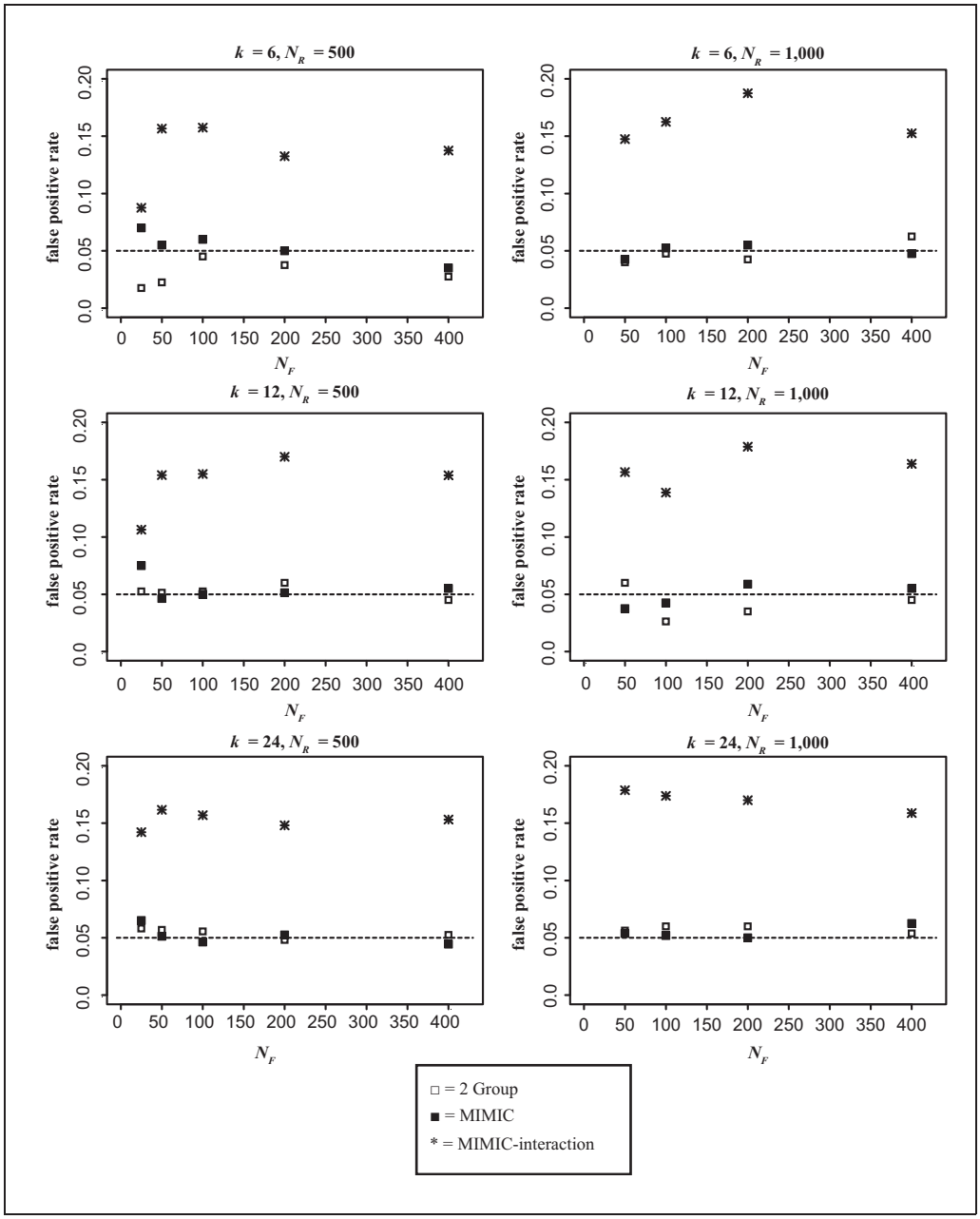

Figure 2. False positive rates based on raw *p* values for binary (left) and ordinal (right) responses
Note: MMIC = multiple indicator multiple cause; k = total items; N_R = reference group sample size; N_F = focal-group sample size; Type I error = .05.
□ = 2 Group.
■ = MIMIC.
∗ = MIMIC-interaction.

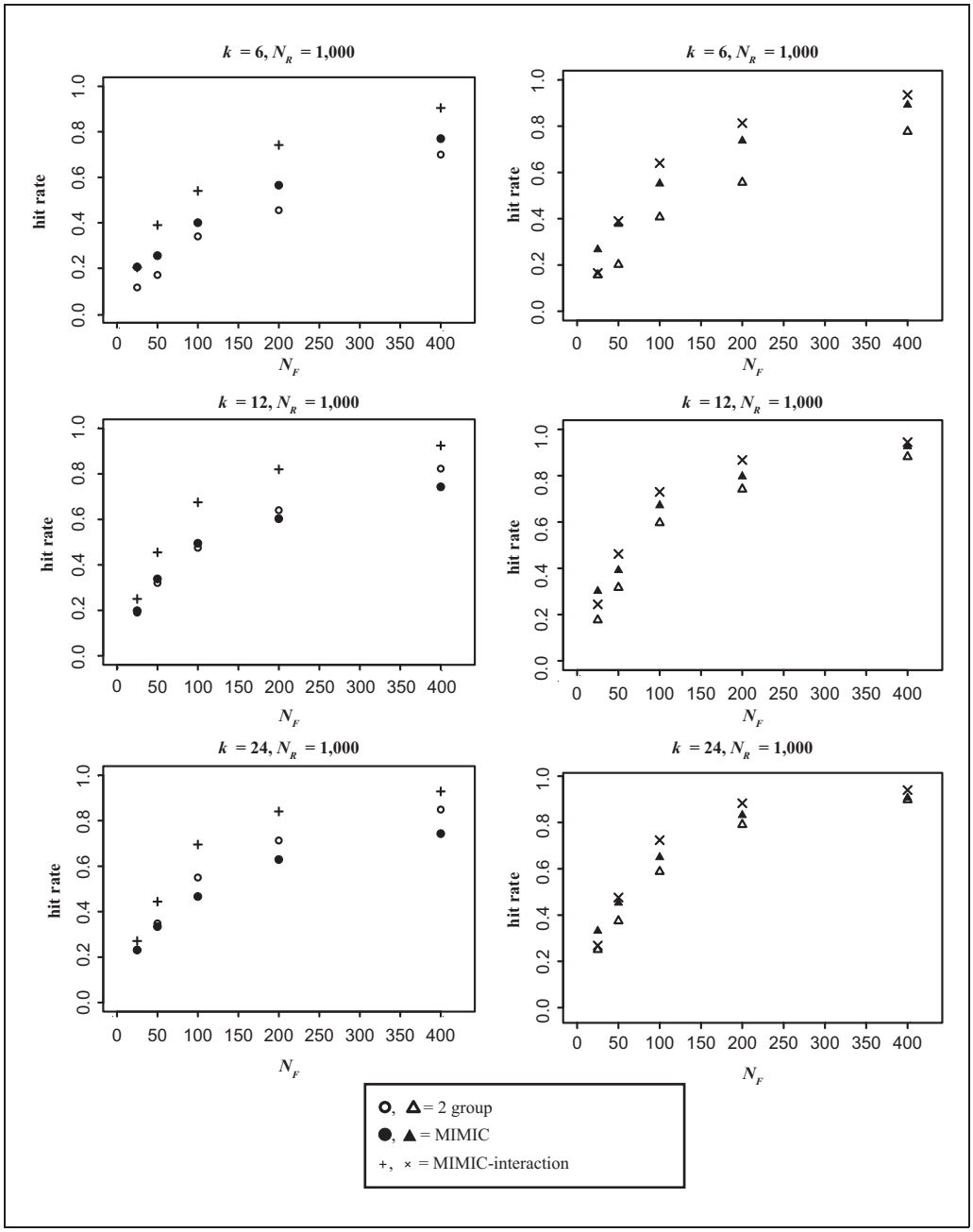

Figure 3. Hit rates for binary responses using raw *p* values for items with nonuniform (left) or uniform (right) DIF

Note: DIF = differential item functioning; MMIC = multiple indicator multiple cause; k = total items; N_R = reference group sample size; N_F = focal-group sample size.

O, △ = 2 group.

●, ▲ = MIMIC.

+, × = MIMIC-interaction.

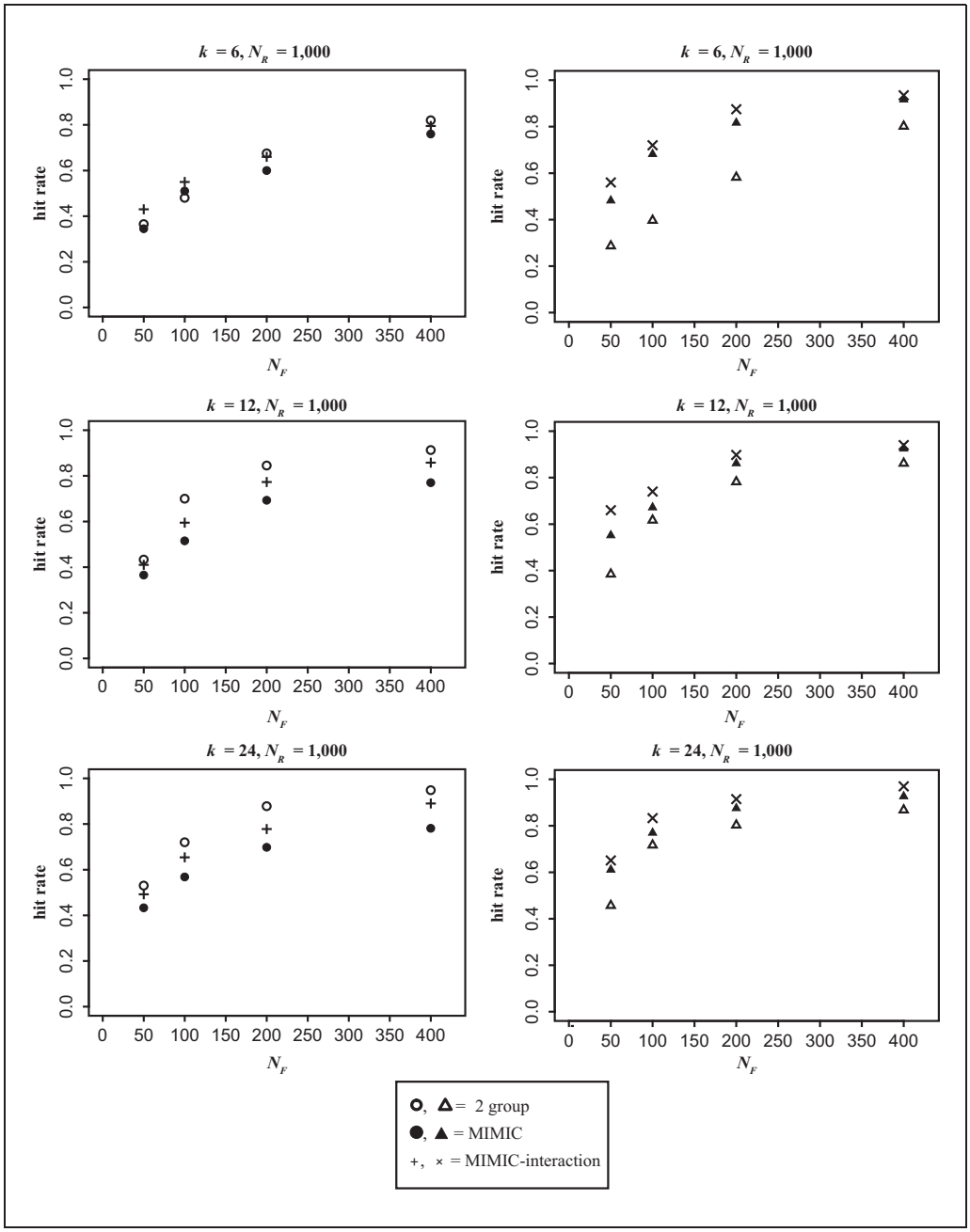

Figure 4. Hit rates for ordinal responses using raw *p* values for items with nonuniform (left) or uniform (right) DIF

Note: DIF = differential item functioning; MMIC = multiple indicator multiple cause; k = total items; N_R = reference group sample size; N_F = focal-group sample size.

○, △ = 2 group.

●, ▲ = MIMIC.

+, × = MIMIC-interaction.

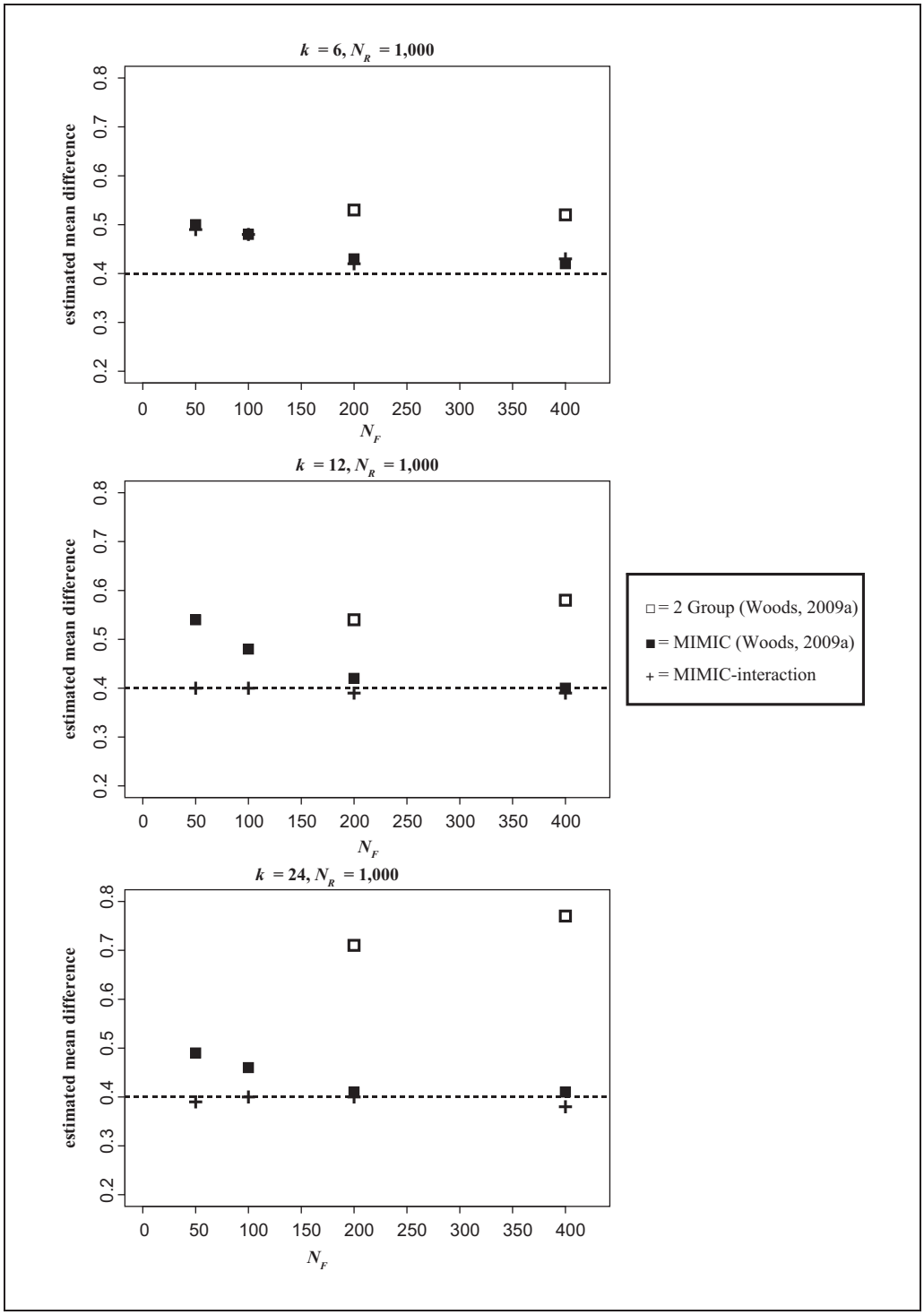

Figure 5. Items with ordinal responses

Note: MMIC = multiple indicator multiple cause.

□ = 2 Group (Woods, 2009a).

■ = MIMIC (Woods, 2009a).

+ = MIMIC-interaction.

Item Parameter Estimates

To control outliers in all analyses involving item parameters, $\hat{\alpha}_{iR}$ and $\hat{\alpha}_{iF} > 4$ were recoded to 4, and $\hat{\tau}_{ijR}/\hat{\alpha}_{iR}$ and $\hat{\tau}_{ijF}/\hat{\alpha}_{iF}$ more extreme than ± 4 were recoded to ± 4 (Woods, 2009a, did this also). For all three models, most of the recoding was needed for estimates of b_{i3} and b_{i4}, and recoding increased with smaller N_F; thus, there was less overall recoding for IRT-LR-DIF because final models were not fitted for $N_F < 200$. With IRT-LR-DIF and MIMIC models without an interaction, the proportion of recoded estimates was about equal for the R and F groups, but for MIMIC-interaction models, more F group than R-group estimates were recoded. When results were repeated with outliers left in, bias in the F-group parameters was larger than what is reported here for MIMIC-interaction models. Details are available on request.

Anchors. For both binary and ordinal responses, bias in item parameter estimates for anchor items was similar with and without the interaction in the model. For binary responses, bias in a_i for MIMIC models without the interaction ranged (more than 30 conditions) from .000 to .027 (Woods, 2009a); with the interaction, it ranged from .000 to .034. Bias in b_i without the interaction ranged from .000 to .037 (Woods, 2009a) and from .000 to .030 with the interaction.

For ordinal responses, bias ranged (over conditions) from .001 to .016 for a_i (with and without interaction), from .000 to .012 for b_{i1} (with and without interaction), from .002 to .032 (Woods, 2009a) or from .000 to .034 (with interaction) for b_{i2}, from .001 to .072 for b_{i3} (with and without interaction), and from .000 to .021 (Woods, 2009a) or .022 (with interaction) for b_{i4}.

Discrimination parameters for D-F items. For both binary and ordinal responses, bias in a_{iF} and a_{iR} for items with uniform DIF was previously low for IRT-LR-DIF and MIMIC models without the interaction (Woods, 2009a). The same was true with the MIMIC-interaction model; bias was .20 or lower for all conditions.

Of greater interest is bias in a_{iF} and a_{iR} for items with nonuniform DIF, which was elevated from MIMIC models without the interaction, but not for IRT-LR-DIF. Figure 6 shows the results for binary responses (for $N_R = 500$ but they are similar for $N_R = 1{,}000$). When the interaction was added, bias dropped dramatically in all conditions and was nearly identical to that for IRT-LR-DIF for a_{iR} and for a_{iF} when $N_F = 400$.

The improvement was not as dramatic for ordinal responses. Bias in a_{iR} (for items with nonuniform DIF) was low and nearly the same as without the interaction. Figure 7 shows bias in a_{iF} for items with nonuniform DIF. Although bias decreased with the addition of the interaction, it was still somewhat large and clearly larger than that for IRT-LR-DIF.

Reference group b_{ijR} for D-F items. Absolute bias in b_{ijR}s for items with uniform DIF was near 0 and about the same for all methods (both binary and ordinal). These results are available on request.

In contrast, adding an interaction to the MIMIC models tended to decrease the bias in b_{ijR} for items with nonuniform DIF, such that bias was similar to that for IRT-LR-DIF. Figure 8 (left) gives the absolute bias in b_{iR} with $k = 6$, 12, and 24 for binary items. N_F and k moderated each other such that with smaller k, larger N_F was needed for the improvement to manifest but with larger N_F, the improvement was apparent even with $N_F = 25$. For ordinal items with nonuniform DIF, absolute bias in b_{ijR} had been near 0 with MIMIC models and was very similar with MIMIC-interaction models (details available on request from the first author).

Focal group b_{ij}s for D-F items. Absolute bias in b_{ijF}s for items with uniform DIF improved when the interaction was added in conditions where the bias was not already near 0 (conditions with binary responses and $N_F = 25$ or 50). Otherwise, the bias remained low and similar to that observed previously (details available on request).

Adding the interaction had more impact on items with nonuniform DIF. Figure 8 (right) gives the absolute bias in b_{iF} for binary responses and $k = 6$, 12, and 24. Results depended on N_F; with $N_F = 25$, adding the interaction always improved bias. With $N_F = 50$ or 100, the bias in b_{iF}

(Text continues on p. 356.)

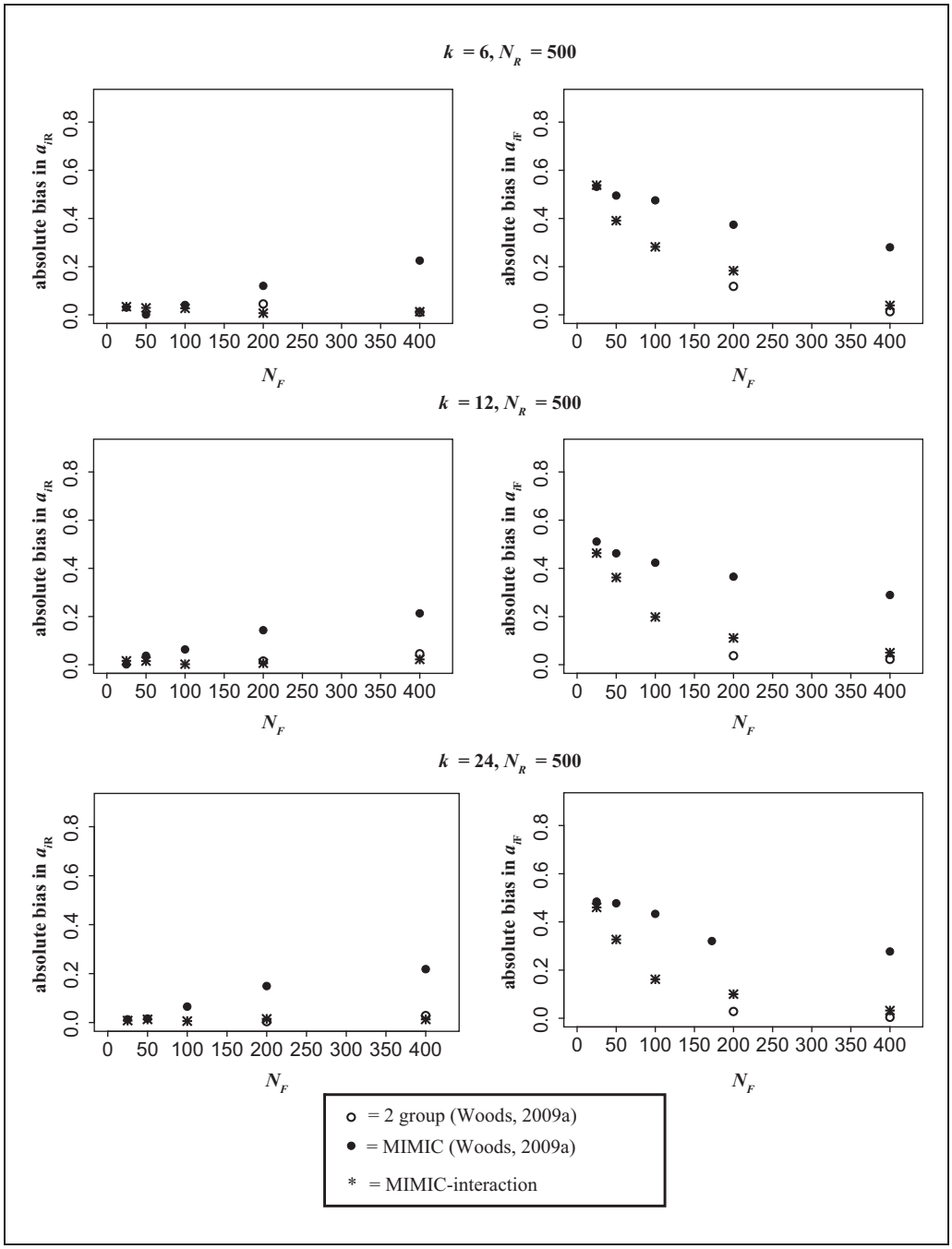

Figure 6. Items with binary responses and nonuniform DIF

Note: DIF = differential item functioning; MMIC = multiple indicator multiple cause.

o = 2 group (Woods, 2009a).

● = MIMIC (Woods, 2009a).

∗= MIMIC-interaction.

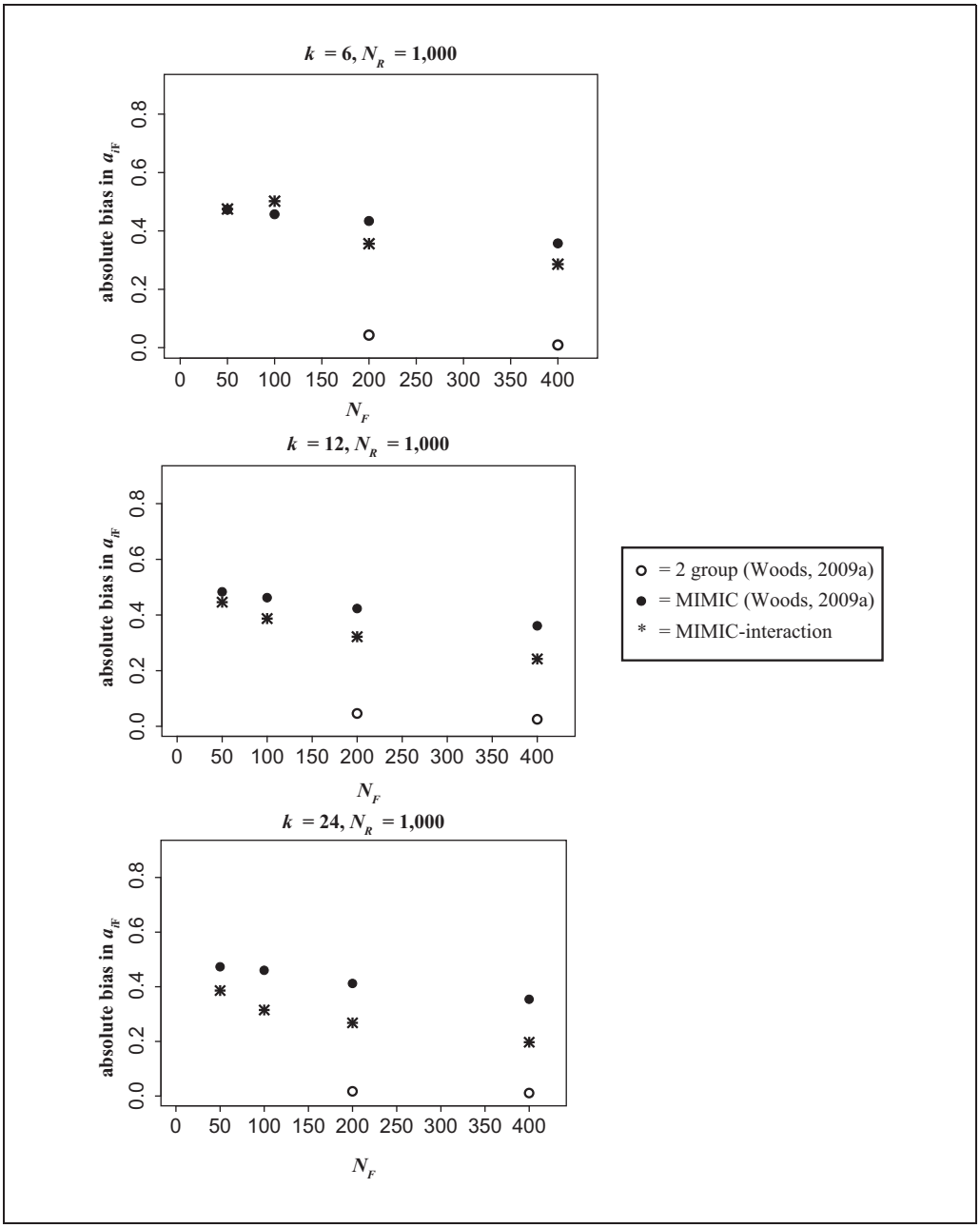

Figure 7. Items with ordinal responses and nonuniform DIF

Note: DIF = differential item functioning; MMIC = multiple indicator multiple cause.

O = 2 group (Woods, 2009a).

● = MIMIC (Woods, 2009a).

∗= MIMIC-interaction.

from interaction models was either about the same as, or greater than, that from models without the interaction. With $N_F = 200$ or 400, bias was around the same low level it had been before (with MIMIC models as well as with IRT-LR-DIF).

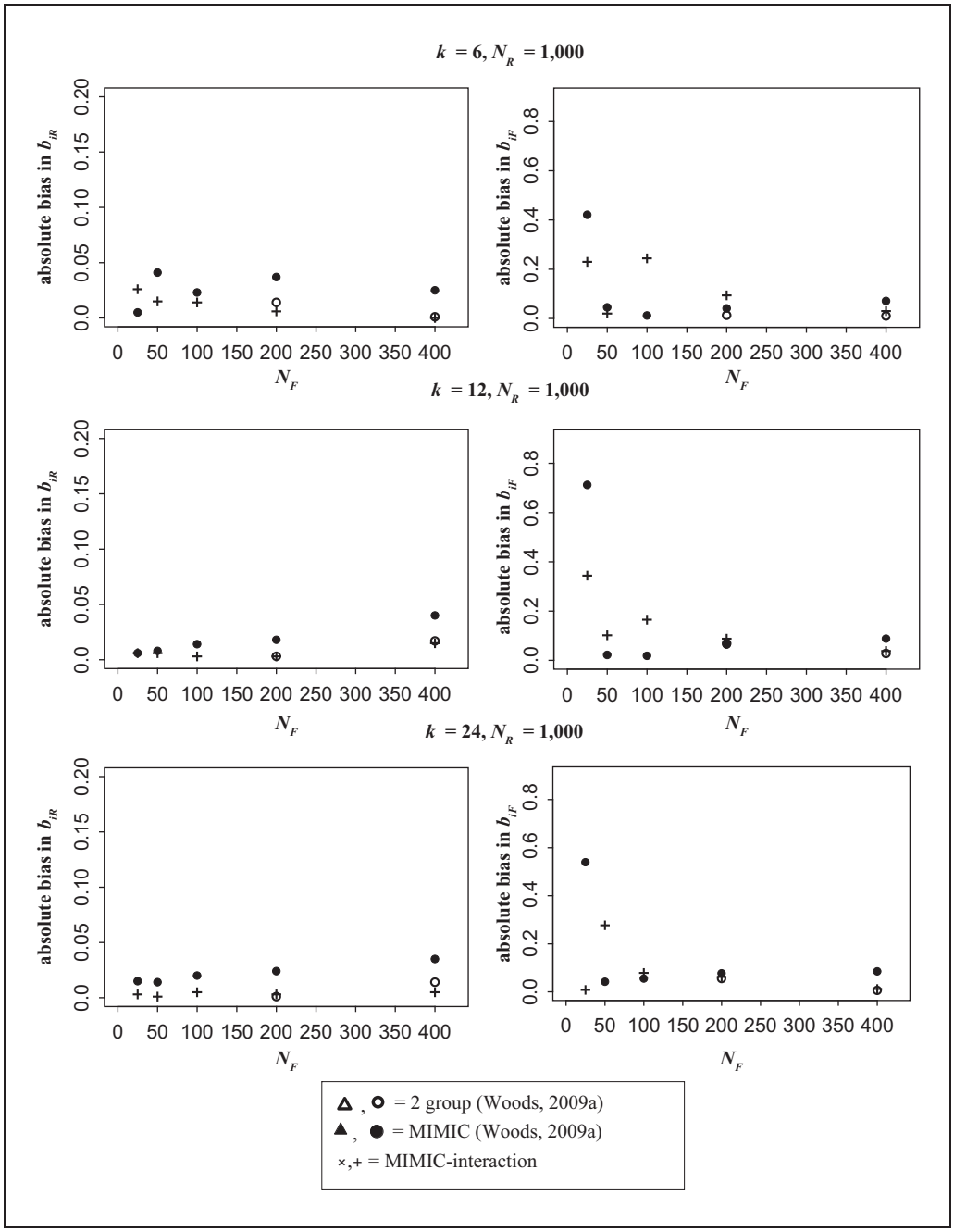

Figure 8. Items with binary responses and nonuniform DIF

Note: DIF = differential item functioning; MMIC = multiple indicator multiple cause.

○, △ = 2 group.

●, ▲ = MIMIC.

× ,+ = MIMIC-interaction.

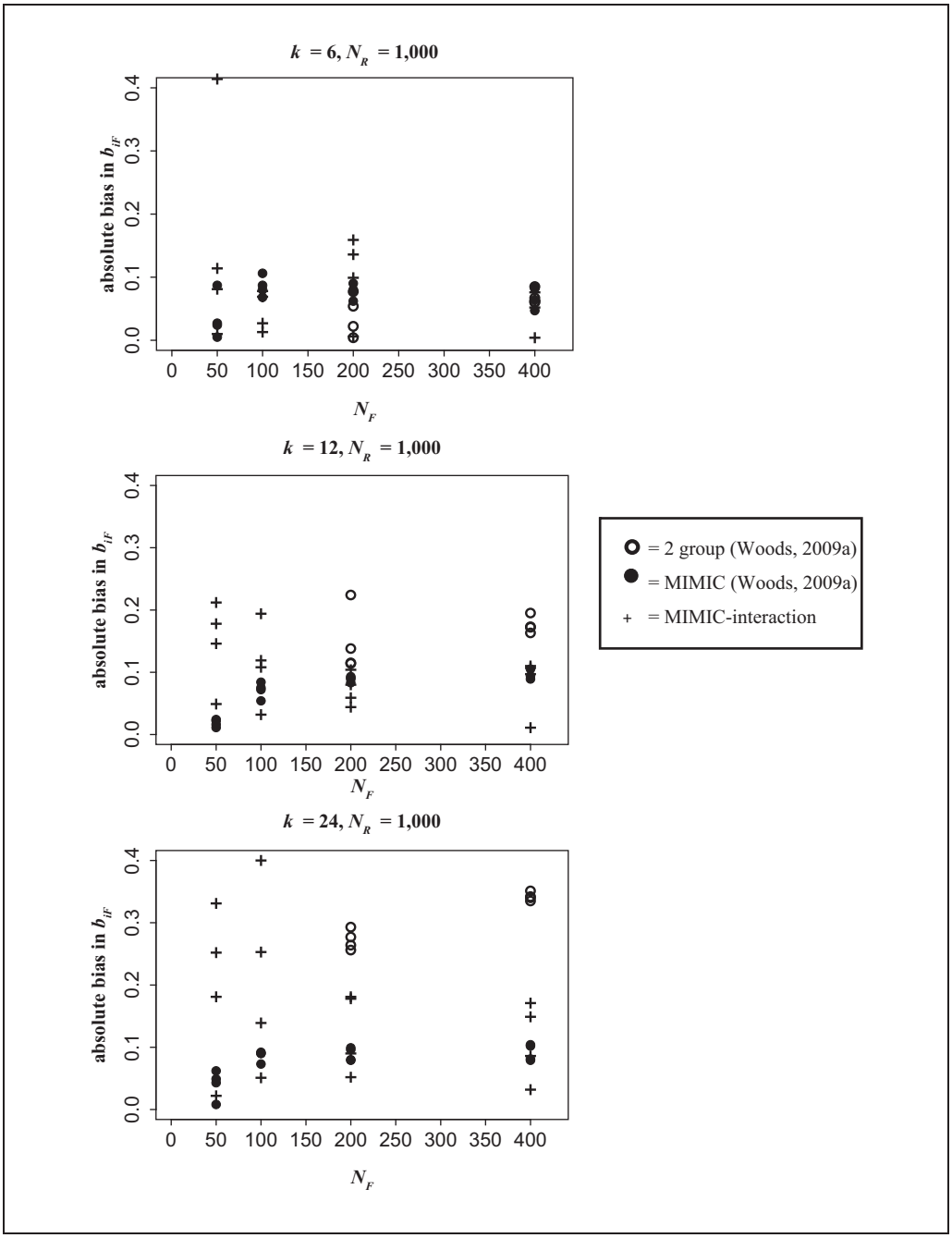

Figure 9. Absolute bias in the focal-group thresholds for ordinal responses and nonuniform DIF

Note: DIF = differential item functioning; MMIC = multiple indicator multiple cause; k = total items; N_R = reference group sample size; N_F = focal-group sample size.

O = 2 group (Woods, 2009a).

● = MIMIC (Woods, 2009a).

\+ = MIMIC-interaction.

Table 1. Multiple Indicator Multiple Cause–Model Differential Item Functioning Testing for Items on the Loss of Control Scale (139 African Americans vs. 757 Whites)

Item	$\chi^2(2)$	p_{BH}	α (*SE*)	τ_1 (*SE*)	τ_2(*SE*)	τ_3 (*SE*)	τ_4 (*SE*)	β (*SE*)	ω (*SE*)
6	19.86	<.01	1.39 (.15)[a]	–2.20 (.14)[a]	–0.37 (.10)[a]	1.38 (.12)[a]	3.55 (.21)[a]	–0.81 (.24)	0.85 (.26)
9	7.23	.05	0.85 (.13)	1.35 (.10)	2.90 (.16)	4.18 (.25)	—	—	—
12	5.04	.08	1.38 (.13)	–0.47 (.10)	1.02 (.10)	2.42 (.14)	4.32 (.23)	—	—
13	5.93	.07	1.38 (.14)	–1.23 (.11)	0.20 (.09)	1.66 (.12)	3.45 (.19)	—	—
8	Anchor		2.12 (.19)	–0.19 (.12)	2.21 (.17)	4.21 (.25)	6.59 (.45)	—	—
11	Anchor		1.61 (.18)	0.78 (.11)	2.37 (.15)	3.68 (.21)	5.34 (.34)	—	—
14	Anchor		1.40 (.14)	0.89 (.10)	2.56 (.15)	3.53 (.20)	4.71 (.29)	—	—

Note: p_{BH} = Benjamini–Hochberg adjusted *p* value, α = estimated discrimination; *SE* = standard error; τ_j = estimated threshold; β = regression coefficient showing the group difference in τ_js for this item; ω = regression coefficient showing the group difference in α for this item; — = parameter not estimated in the final model. Type I error rate α = .05. Item with significant test: "6. I am going to act foolish."
a. Applies to Whites only.

Figure 9 shows the absolute bias in b_{ijF} for ordinal responses with k = 6, 12, and 24. Although there were a few exceptions, bias increased when the interaction was added. However, with k = 12 or 24, b_{ijF} was estimated more accurately with either MIMIC model than with IRT-LR-DIF.

Empirical Example

The same data Woods (2009a) analyzed using MIMIC models (without the interaction) were analyzed here using MIMIC-interaction models, implemented as described above. Because the simulations indicated that Type I error inflation is expected, *p* values were BH corrected. This should help because when the present simulation results for MIMIC-interaction models were analyzed based on BH-corrected instead of raw *p* values, the Type I error was near the nominal rate (Woods & Grimm, 2010).

Participants rated how often each of seven thoughts or ideas about loss of control (e.g., ''I am going to be paralyzed by fear'') typically occur to them when they are nervous: *never* (1), *rarely* (2), *half the time* (3), *usually* (4), or *always* (5). Prior to DIF testing, designated anchor items were selected empirically following the rationale described by Woods (2009b). To select anchors, each item was first tested with all other items treated as anchors. Each item was tested individually in a separate MIMIC model in which the item response was regressed on z and the interaction. Considering the β / standard error (*SE*) ratios for both z and the interaction, the three items with the smallest ratios were designated anchors: 8, 11, and 14. Thus, 6, 9, 12, and 13 were studied items.

DIF testing was carried out as described for the simulations; the 4 studied items were tested individually and then a final model was fitted.[5] Table 1 lists the omnibus χ^2 test results, and the discrimination (α) and threshold (τ_j) parameter estimates from the final model. Results from the interaction model differed from previous results. Woods (2009a) found that the tests for uniform DIF were significant (Type I error α = .05) for items 6, 9, 12, and 13. In the present analysis, the omnibus test was significant only for Item 6.

The estimated mean difference on the latent variable (from the final model) was smaller here (γ = −.32, *SE* = 11) than previously (γ = −.52, *SE* = .13, Woods, 2009a) but still negative, indicating that loss of control was higher for the group coded 0 (Whites).

Because the final model treated Item 6 as functioning differently between groups, the α and τ_js given in Table 1 for this item apply only to Whites. The estimates for African Americans were obtained by adding the corresponding DIF effect. The discrimination estimates for African

Americans were equal to $\alpha + \omega$, where ω is the nonuniform DIF effect (i.e., the coefficient reflecting the relationship between the interaction and item response). Item discrimination is larger for African Americans versus Whites if ω is positive; it is larger for Whites if ω is negative. Thus, Item 6 was more discriminating for the African American versus White sample. The threshold estimates for African Americans were equal to $\tau_j + \beta$, where β is the uniform DIF effect (i.e., the coefficient reflecting the relationship between group membership and item response). A positive β indicates that the τ_js are larger for African Americans, whereas a negative β means that the τ_js are smaller for African Americans. Thus, Whites had to possess more loss of control than African Americans before endorsing a particular response category for Item 6.

Discussion

The purpose of this article was to clarify, for apparently the first time, how a MIMIC-interaction model with categorical indicators can be used to test for uniform and nonuniform DIF simultaneously. Simulations evaluated currently available methods implemented in Mplus that are likely to be used frequently in practice for fitting MIMIC-interaction models. Power to detect nonuniform DIF and bias in the item parameters were important foci of the simulations because MIMIC models without the interaction performed suboptimally on these outcomes previously (Woods, 2009a).

As expected, MIMIC-interaction models showed greater power than MIMIC models without the interaction to detect nonuniform DIF. Although this was intuitively reasonable because interaction models include a nonuniform DIF effect, Type I error was inflated for MIMIC-interaction models. Type I error inflation was probably caused by the procedures Mplus currently uses for interactions involving latent variables (the authors used version 5.2, these methods are unchanged in version 6). The XWITH code implements LMS (Klein & Moosbrugger, 2000), which assumes that both interacting latent variables are normal. When this assumption is violated, inflated Type I error has been observed previously (Klein & Moosbrugger, 2000). Because the Mplus user's guide explicitly recommends XWITH when an interaction involves an observed categorical variable, it is important for users to be aware that Type I error inflation can result. Although accompanying LMS with a BH *p*-value correction is likely to reduce the Type I error problem (Woods & Grimm, 2010), it would be better to find a more theoretically defensible method for estimating the latent interaction for these models.

Future research is needed to select a good alternative to LMS for MIMIC-interaction DIF models. The majority of extant research on latent variable interactions applies to continuous indicators and continuous latent variables. Possibilities include descendants of Kenny and Judd's (1984) seminal product indicator approach that do not make normality assumptions, including an unconstrained method (Marsh, Wen, & Hau, 2004) and the generalized appended product indicator approach (GAPI; Wall & Amemiya, 2001). Product indicator methods can be implemented in many structural equation modeling (SEM) programs and have performed well in simulations with continuous indicators (Marsh et al., 2004; Wall & Amemiya, 2001). However, they are more complicated for DIF testing because it is unclear which items should form the interaction: only the studied item, only anchors, anchors and the studied item, all items, and so on. This needs attention.

Another alternative is to specify the MIMIC-interaction model as an equivalent nonlinear mixed model and estimate the parameters using full information maximum likelihood estimation. This can be done with, for example, SAS PROC NLMIXED (SAS Institute Inc., 2004). This approach does not require product indicators, and all thresholds can be predicted from the grouping variable. Disadvantages (with NLMIXED) are that users have to specify the likelihood and start values, and it is extremely slow; it takes several hours or days to converge, whereas Mplus produces results in minutes.

As for item parameter estimates, the authors observed that adding the interaction to the MIMIC model reduced bias in the types of item parameters that had been estimated somewhat poorly from models with no interaction, without reducing accuracy for the types of item parameters that were previously estimated well. However, there were exceptions. For items with ordinal responses and nonuniform DIF, bias in discrimination was not improved by adding the interaction, and accuracy was higher from IRT-LR-DIF than MIMIC-interaction models. Also, bias in thresholds (for these types of items) got worse when the interaction was added. Parameter estimation is no doubt influenced by the procedures used to estimate the latent interaction, so finding an alternative to LMS is the pressing future problem.

MIMIC-interaction models are promising and worth pursuing. More than two groups can be more easily compared with MIMIC models, and they should perform better with smaller N_F than multiple group models because the latent variable model does not have to be fitted to data for each group separately. Woods (2009a) observed that, without an interaction in the model, MIMIC models performed better than IRT-LR-DIF with small N_F, except for some outcomes related to nonuniform DIF. Nevertheless, disadvantages of smaller N_F are increased risk of estimation failure and higher frequency of outlying item parameter estimates. In the present simulations, both of these problems were most frequent for $N_F = 25$ or 50; thus $N_F = 100$ may be more in the range of N_F where the advantage of MIMIC models over multiple group models will be apparent. Advantages of multiple group models are that they do not require estimation of a latent interaction for testing nonuniform DIF, and the latent variable variance can differ across groups.

Many additional features of these simulations can be manipulated in future work. For example, running more than 100 simulations per condition will decrease sampling variability, and the capabilities of MIMIC-interaction models can be further explored with additional types of simulated data. It will be interesting to study the performance of MIMIC-interaction models with nominal indicators, and under assumption violations, such as unequal variances and data generated from a 3PL model (a lower asymptote parameter is not currently available in the SEM framework so there would be a misspecification in the item response function). The degree of DIF could be varied as an independent variable so that its influence on study outcomes could be evaluated, statistical properties of the specific tests of uniform and nonuniform DIF could be examined, and MIMIC-interaction models could be compared with various other methods for DIF testing. The current study evaluated only the constant pattern of DIF for the thresholds, but other patterns are realistic and worthy of study. With the current Mplus method, the interaction cannot covary with its constituent main effects because the interaction is not a new distinct variable; this is not true of all other methods.

The present simulations were carried out with correctly specified (i.e., DIF-free) anchors. It is well established that Type I error is inflated when the anchors are contaminated with DIF, so it is important to attend to the correct specification of anchors with real data. Empirical methods for identifying DIF-free anchors are often used, such as the approach used in the present empirical example (introduced for IRT-LR-DIF by Woods, 2009b). For MIMIC models without interactions, Wang, Shih, and Yang (in press) and Shih and Wang (2009) have described and studied approaches for empirically selecting anchor items that take longer than the method proposed by Woods (2009b). It would be interesting to compare these three purification approaches in the future.

Acknowledgment

The authors are grateful to Daniel Serrano and Wenjing Huang for information pertaining to latent variable interactions, to Kristopher Preacher and Ryne Estabrook for helpful comments on a draft of this manuscript, and to Graham Rifenbark and Andy Aschenbrenner for assistance with computing.

Declaration of Conflicting Interests

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

Funding

This research was partially supported by NSF Grant Number SES-0818722 awarded to Carol M. Woods and NSF REECE Program Grant—DRL-0815787 awarded to David Grissmer and Kevin J. Grimm.

Notes

1. In the authors' notation for MIMIC models, Greek is used for parameters (α, β, ω, γ, τ) and latent variables (θ, Σ), and roman is used for observed variables (x, z) or index counters (k = total number of items).
2. In the present MIMIC-interaction models, the grouping variable could be treated as a categorical indicator of an underlying continuous normal latent variable, which would presumably improve the performance of the LMS estimator. However, the assumption that a normally distributed latent variable underlies observed ''group'' does not make sense for very many grouping variables used for DIF so this strategy has limited applicability.
3. The C++ code used to generate the data and write command files, execute, and process output from Mplus is available by request from the first author.
4. To understand Woods's (2009a) results a final model was fitted for two-group models only when N_F = 200 or 400 because two-group item response theory (IRT) is not a small-sample method and hit rates were so low with $N_F \leq 100$ that it was clear that accuracy would be poor for the final models. In the final model, parameters for items with nonsignificant DIF tests were constrained equal between groups, and parameters for items with significant tests were estimated separately for the two groups. The mean of θ was fixed to 0 for the *R* group and estimated for the *F* group, and the *SD* of θ was 1 for both groups.
5. The Mplus code for the final model used for the empirical example is available by request from the first author.

References

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

Barendse, M. T., Oort, F., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *Advances in Statistical Analysis, 94*, 117-127.

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Chen, C., & Anthony, J. C. (2003). Possible age-associated bias in reporting of clinical features of drug dependence: Epidemiological evidence on adolescent-onset marijuana use. *Addiction, 98*, 71-82.

Christensen, H., Jorm, A. F., MacKinnon, A. J., Korten, A. E., Jacomb, P. A., Henderson, A. S., & Rodgers, B. (1999). Age differences in depression and anxiety symptoms: A structural equation modeling analysis of data from a general population sample. *Psychological Medicine, 29*, 325-339.

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

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 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. *Journal of Gerontology: Social Sciences, 57B*, S275-S284.

Gelin, M. N. (2005). *Type I error rates of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation* (Unpublished doctoral dissertation). University of British Columbia, Canada.

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiological Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences, 55B*, P273-P282.

Hagtvet, K. A., & Sipos, K. (2004). Measuring anxiety by ordered categorical items in data with subgroup structure: The case of the Hungarian version of the trait anxiety scale of the state-trait anxiety inventory for children (STAIC-H). *Anxiety, Stress, & Coping, 17*, 49-67.

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model*. Unpublished masters thesis. University of North Carolina at Chapel Hill.

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: Erlbaum.

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.

Kenny, D., & Judd, C. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*, 201-210.

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

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*, 457-474.

Klein, A., & Muthén, B. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research, 42*, 647-673.

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

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

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.

Marsh, H. W., Wen, Z., & Hau, K. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*, 275-300.

Mast, B. T., & Lichtenberg, P. A. (2000). Assessment of functional abilities among geriatric patients: A MIMIC model of the functional independence measure. *Rehabilitation Psychology, 45*, 49-64.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*, 121-132.

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

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

Muthén, L. K., & Muthén, B. O. (2007). Mplus: Statistical Analysis with Latent Variables (Version 4.21) [Computer software]. Los Angeles, CA: Author.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*, 411-423.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*, 150-166.

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

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement. 34*, 453-456.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

SAS Institute Inc. (2004). *SAS/STAT 9.1*. Cary, NC: Author.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moments structure analysis. *Psychometrika, 66*, 507-514.

Schroeder, J. R., & Moolchan, E. T. (2007). Ethnic differences among adolescents seeking smoking cessation treatment: A structural analysis of responses on the Fagerström test for nicotine dependence. *Nicotine & Tobacco Research, 9*, 137-145.

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

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

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

Sweeney, K. P. (1996). *A Monte-Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning*. Unpublished doctoral dissertation, Fordham University, New York.

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning* (Documentation for computer program). L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

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

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

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-111). Hillsdale, NJ: Lawrence Erlbaum.

Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics, 26*, 1-29.

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

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 procedure for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*, 42-69.

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

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

Woods, C. M. (2011). DIF testing for ordinal items with poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement, 35, 145-164.*

Woods, C. M., & Grimm, K. (2010, July). *Identification of nonuniform differential item functioning using multiple indicator multiple cause models*. Presented at the International Meeting of the Psychometric Society, Athens, GA.