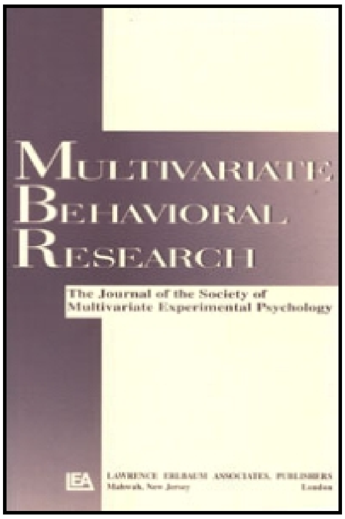

Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis

Carol M. Woods [a]
[a] Washington University in St. Louis,

Online Publication Date: 01 January 2009

To cite this Article Woods, Carol M.(2009)'Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis',Multivariate Behavioral Research,44:1,1 — 27

To link to this Article: DOI: 10.1080/00273170802620121

URL: http://dx.doi.org/10.1080/00273170802620121

Multivariate Behavioral Research, 44:1–27, 2009
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170802620121

Psychology Press
Taylor & Francis Group

Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis

Carol M. Woods
Washington University in St. Louis

Differential item functioning (DIF) occurs when an item on a test or questionnaire has different measurement properties for 1 group of people versus another, irrespective of mean differences on the construct. This study focuses on the use of multiple-indicator multiple-cause (MIMIC) structural equation models for DIF testing, parameterized as item response models. The accuracy of these methods, and the sample size requirements, are not well established. This study examines the accuracy of MIMIC methods for DIF testing when the focal group is small and compares results with those obtained using 2-group item response theory (IRT). Results support the utility of the MIMIC approach. With small focal-group samples, tests of uniform DIF with binary or 5-category ordinal responses were more accurate with MIMIC models than 2-group IRT. Recommendations are offered for the application of MIMIC methods for DIF testing.

It is usually desirable for items on psychological tests or questionnaires to measure the same construct and possess the same measurement properties for all respondents. For example, persons with the same level of depression should have the same probability of endorsing items on a depression symptom inventory regardless of their sex, ethnicity, or other group memberships. Mean depression levels may differ between groups, but in order to validly determine whether they do, items must function the same way for individuals matched on the un-

Correspondence concerning this article should be addressed to Carol M. Woods, Psychology Department, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130. E-mail: cwoods@artsci.wustl.edu

derlying construct of depression. Substantive research findings based on invalid measurement are, at best, misleading.

Differential item functioning (DIF) occurs when an item on a test or questionnaire has different measurement properties for one group of people versus another, irrespective of mean differences on the construct. In methods for DIF testing (see reviews by Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993), the construct is approximated by observed scores (e.g., sums of item scores) or a latent variable. Latent-variable methods are likely to be more accurate because they account for measurement error in the items.

Latent-variable methods for DIF testing use either multiple-group models or multiple-indicator multiple-cause (MIMIC) models, which are both a type of structural equation model (SEM). 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. For binary or ordinal data, both MIMIC and multiple-group models may be parameterized either as an item response model fitted to the data directly or as a confirmatory factor analysis model fitted to polychoric (or tetrachoric) correlations. The equivalence between item response theory (IRT) and categorical factor analysis is well established (e.g., Takane & de Leeuw, 1987). The IRT parameterization is used in the present article.

Multiple-group analysis is often preferable to the MIMIC approach because more types of hypotheses can be tested (e.g., DIF with respect to item discrimination). However, the sample size requirement is probably greater for multiple-group analysis because the latent-variable model is fitted to data for each group separately (B. O. Muthén, 1989). Small samples are common in psychological research, particularly for minority or focal groups. If MIMIC methods accurately test DIF using latent variables when the focal group is small, then they are unique and potentially quite valuable.

However, the accuracy of the MIMIC approach to DIF testing is incompletely verified, and little is known about the sample size requirements. The present research examines the accuracy of MIMIC methods for DIF testing when the focal group is small and compares results with those obtained using two-group IRT.

DIFFERENTIAL ITEM FUNCTIONING

In an IRT context, an item with DIF has a different category response function (CRF) for one group of people versus another. In other words, even when members of two different groups are matched on the latent variable, θ, their probability of giving the same response to the item is not the same. Groups are defined by, for example, sex, ethnicity, or experimental condition, with one assigned to be the reference group and the other assigned to be the focal group. The majority, or group with which a test was originally developed, is typically treated as the reference group.

DIF may be uniform or nonuniform (Camilli & Shepard, 1994; Mellenbergh, 1989). Uniform DIF occurs when the CRFs for the two groups are different and do not cross; one group is more or less likely to endorse a higher response over the entire range of θ. If the CRFs cross, then the DIF is nonuniform. Multiple-group models can test for both uniform and nonuniform DIF, but MIMIC models are sensitive to uniform DIF only.

TWO-GROUP IRT FOR TESTING DIF: IRT-LR-DIF

Multiple-group approaches for testing DIF have been described in the context of factor analysis (Jöreskog, 1971; Millsap & Yun-Tein, 2004; B. O. Muthén & Christoffersson, 1981; Sörbom, 1974) or IRT (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), and several authors have addressed similarities and differences between the two contexts (Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006). Here we focus on IRT-LR-DIF (IRT-based likelihood-ratio testing for DIF), which can be carried out with most IRT software but is particularly convenient using Thissen's (2001) IRTLRDIF program.

IRT-LR-DIF involves statistically comparing nested two-group item response models with varying constraints to evaluate whether the response function(s) for a particular item differs for the reference and focal groups. No explicit estimation of θ is needed; θ is a random latent variable treated as missing using Bock and Aitkin's (1981) scheme for marginal maximum likelihood. The mean and variance of θ are fixed to 0 and 1 (respectively) for the reference group to identify the scale and estimated for the focal group as part of the DIF analysis. A subset of items called designated anchors are presumed invariant and used to link the metric of θ for the two groups.

Item parameters for designated anchors are constrained equal between groups whereas 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. 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. The LR test statistic is -2 times the difference between the optimized log likelihoods, which is approximately χ^2-distributed with *df* equal to the difference in free parameters. Statistical significance indicates the presence of DIF. If this general test is significant, follow-up tests are easily carried out to establish whether the DIF is uniform or nonuniform.

It is also possible to perform IRT-LR-DIF tests using all other items as anchors instead of a designated anchor set. However, Type I error is inflated when the anchor set is contaminated by DIF (Stark et al., 2006; Wang & Yeh, 2003) so it is better to designate anchors. Various strategies for empirically identifying anchors have been applied or suggested (Bolt, Hare, Vitale, & Newman, 2004; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Kim & Cohen, 1995; Rensvold & Cheung, 2001; Stark et al., 2006; Wang, 2004; Woods, in press). Many of them are iterative purification procedures wherein the analysis is repeated several times with apparently differentially functioning items removed one by one. With invariant anchors, IRT-LR-DIF has performed well in simulations with binary and ordinal data (Ankenmann, Witt, & Dunbar, 1999; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Wang & Yeh, 2003).

MIMIC MODELS FOR DIF TESTING

B. O. Muthén (e.g., 1985, 1988, 1989) popularized the use of MIMIC models to test for DIF using estimation methods appropriate for categorical data (see also MacIntosh & Hashim, 2003; B. O. Muthén, Kao, & Burstein, 1991). A simple MIMIC model for DIF testing is illustrated in Figure 1. The corresponding equations are readily available in print (MacIntosh & Hashim, 2003; B. O. Muthén, 1985, 1988, 1989; B. O. Muthén et al., 1991). Figure 1 displays a standard unidimensional item response model (or factor model) with θ regressed on an observed grouping variable to allow for a mean difference. Item responses are regressed on the grouping variable to test for DIF. There is evidence of DIF if group membership significantly predicts item response, controlling for any mean differences on θ. Discrimination parameters are implicitly invariant; thus this is a model of uniform DIF.

Details of the Testing Procedure

No consensus is apparent from the published literature concerning exactly how MIMIC models should be used to test for DIF. Although it might seem initially appealing to regress responses for all items on the grouping variable to test all items for DIF (i.e., to include dashed-line paths as in Figure 1 for all items), such a model would not be identified. Thus, a different approach is required.

The strategy described by B. O. Muthén (1988, 1989) and Oort (1992, 1998) is to first fit a baseline model presuming no DIF in any item (i.e., Figure 1 without any dashed-line paths) and then to inspect modification indices (MIs),

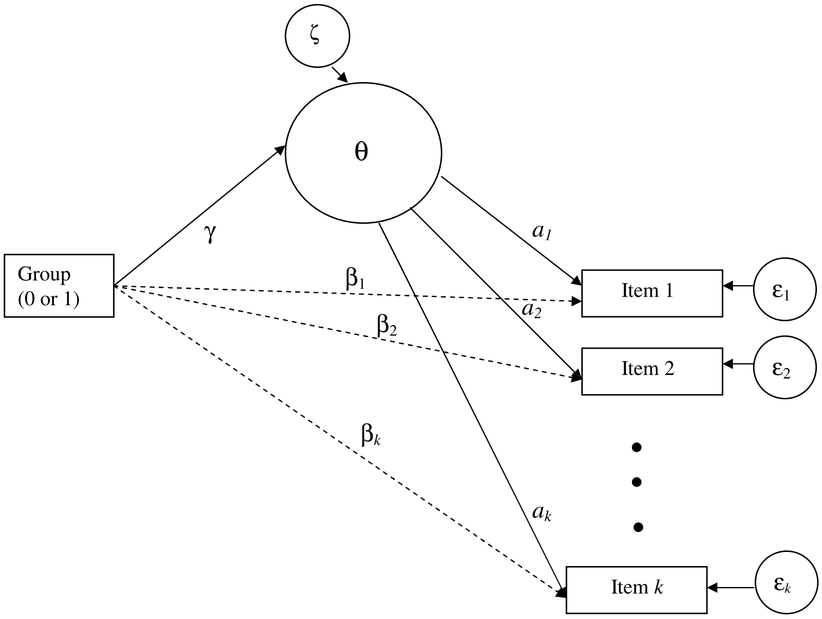

FIGURE 1 A basic MIMIC model for DIF testing; γ = regression coefficient showing the mean difference on the latent variable, θ; β_i = regression coefficient showing the group difference in the threshold for item i and the grouping variable ($i = 1, 2, \ldots k$); a_i = discrimination parameter; ε_i = measurement error for item i; ζ = residual for θ.

which reflect the expected improvement in model fit if a fixed parameter were freely estimated. Items with a "large" MI for the β_i parameter are flagged and may be regressed on the grouping variable in a subsequent model. Some practitioners have used similar MI-based approaches (Hagtvet & Sipos, 2004; Mast & Lichtenberg, 2000).

MIs might produce misleading results. One dilemma is in judging when to flag an item: The definition of a "large" index is unclear, and statistical significance depends on sample size. B. O. Muthén (1988) recognized numerous limitations of MIs: "The information from the various indexes for a certain model can be misleading since they may be highly correlated, the information really only pertains to freeing up one parameter at a time, the indexes are only good approximations for models that are close to a well-fitting one, and we may capitalize on chance in our data" (p. 228). Use of MIs does not seem advisable.

A second approach is to test each item for DIF assuming all other items are invariant (Chen & Anthony, 2003; Finch, 2005; Gelin, 2005; Oishi, 2006; Schroeder & Moolchan, 2007). The DIF-free baseline model is statistically compared with each of k models (let k = total number of items), each with a path from the grouping variable to one item. This procedure is analogous to IRT-LR-DIF with all other items as anchors. The test statistic is treated as if it is χ^2 distributed, but for that to be true, the baseline model should fit the data (Maydeu-Olivares & Cai, 2006). Fit of the baseline model worsens as the amount of DIF in the data increases. Simulations have shown that when some items function differently, Type I error is inflated with this approach to MIMIC modeling (Finch, 2005) as it is when IRT-LRT-DIF is used with all others as anchors.

Better approaches use designated anchors (Christensen et al., 1999; Fleishman, Spector, & Altman, 2002; Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000). Anchors may be chosen based on preliminary tests using all other items as anchors; Woods (in press) proposes and tests a specific strategy that can produce invariant anchor sets much of the time. In one model, all studied items can be regressed on the grouping variable with individual tests of these regression parameters interpreted as DIF tests (Christensen et al., 1999; Grayson et al., 2000). Alternatively, each studied item can be tested individually by comparing a full model that presumes DIF in all studied items with a model with the DIF path removed for one studied item (Fleishman et al., 2002; Woods, Oltmanns, & Turkheimer, 2008). This latter strategy (the free-baseline designated-anchor approach) is followed in the present study. It is appealing because it is most similar to the well-tested IRT-LR-DIF approach.

Previous Simulation Studies

Empirical evaluations of MIMIC methods for DIF are scant. Oort (1998) evaluated methods designed for continuous responses using MIs, Gelin (2005) examined DIF-free ordinal data using all other items as anchors, and Finch (2005) focused on binary data (with and without DIF) using all others as anchors. Both Gelin and Finch used the categorical factor analysis parameterization. None of these studies evaluated estimates from a final model (defined later).

The Current Study

The primary aim of the present research is to empirically evaluate the MIMIC approach for DIF testing using (a) the free-baseline designated-anchor model-testing strategy, (b) items with both binary and ordinal responses, (c) the IRT model parameterization, and (d) final models. A comparison to two-group IRT is included to investigate the hypothesis that, with other study characteristics held constant, the sample size needed for adequate power and reasonably accurate parameter estimation is smaller for MIMIC models than two-group analysis.

METHODS FOR STUDY 1: BINARY RESPONSES

One hundred sets of binary item data were generated under 60 independent conditions, varying according to the focal-group sample size ($N_F = 25$, 50, 100, 200, or 400), number of items ($k = 6$, 12, or 24), reference-group sample size ($N_R = 500$ or 1,000), and presence versus absence of DIF. A C++ program generated the data and wrote command files for, executed, and processed output from Mplus (L. K. Muthén & Muthén, 2007), IRTLRDIF (Thissen, 2001), and MULTILOG (Thissen, Chen, & Bock, 2003).

Data Generation

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

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

where a_i and b_i are the discrimination and threshold parameters, respectively, for item i, and θ_j is the value of the latent variable for the j^{th} simulee. In all conditions, θ_j was drawn from N($\mu = 0$, $\sigma = 1$) for the reference group and N(−0.4, 1) for the focal group.

Item parameters for the reference (R) group were randomly drawn from certain distributions: N($\mu = 1.7$, $\sigma = 0.3$) for $a_{i\text{R}}$ and N($\mu = 0$, $\sigma = 1$) for $b_{i\text{R}}$, chosen based on an empirical examination of item parameters estimated from an assortment of psychological scales (Hill, 2004). To avoid unrealistic extreme values, the distribution of $a_{i\text{R}}$ 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 $a_{i\text{F}} \geq 0.5$ (F = focal group). The distribution of $b_{i\text{R}}$ was truncated at ± 2 to avoid items with all responses in a single category.

Parameters for the F group were defined in relation to R-group parameters. In DIF-free conditions, $a_{i\text{F}} = a_{i\text{R}}$ and $b_{i\text{F}} = b_{i\text{R}}$ 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., $a_{i\text{F}} < a_{i\text{R}}$ and $b_{i\text{F}} > b_{i\text{R}}$), and $k/3$ items were DIF-free anchors (not tested for DIF). Half of the differentially functioning (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). As an example, when $k = 12$, there were 4 items with nonuniform DIF, 4 items with uniform DIF, and 4 DIF-free anchors.

In applications of IRT-LR-DIF, the amount of DIF usually varies over items within a study, and typical differences $|a_{i\text{F}} - a_{i\text{R}}|$ or $|b_{i\text{F}} - b_{i\text{R}}|$ are between .3 and .7. In this simulation, $a_{i\text{F}} = a_{i\text{R}} - \delta$ and $b_{i\text{F}} = b_{i\text{R}} + \gamma$, where δ and γ were equal to one of five equally likely values (.3, .4, .5, .6, or .7) and $\delta \neq \gamma$ (except by chance). A random number from a uniform distribution determined δ (or γ) for a given item.

MIMIC Model DIF Tests

The MIMIC approach was applied using Mplus (Version 4.21; L. K. Muthén & Muthén, 2007). For each data set, $2/3k + 2$ different MIMIC models were fitted. In every model, θ was regressed on the grouping variable and the variance of the residual from this regression was fixed to 1 to identify the scale. All models were parameterized as two-parameter logistic IRT models; however, the Mplus IRT model differs from Birnbaum's 2PL model given in Equation (1). The Mplus parameterization is

$$\Pr(u_{ij} = 1|\theta_j) = \frac{1}{1 + \exp[\tau_i - a_i\theta_j]}, \tag{2}$$

where τ_i is not equivalent to b_i in Equation (1), but $\tau_i = a_i b_i$.

Models were fitted to the data directly using the robust maximum likelihood estimator "MLR." With this estimator, the LR statistic must be divided by a term that is a function of the number of estimated parameters (p) and the scaling correction factor given by Mplus (c) for each model: $\chi^2_{difference} = \frac{LR}{\frac{(p1c1 - p0c0)}{(p1 - p0)}}$, where 0 and 1 label the models being compared (it does not matter which model is assigned 0 versus 1 for this formula as long as the assignment is kept consistent for all calculations—χ^2 is nonnegative). Further details are given on the Mplus Web site (http://www.statmodel.com/chidiff.shtml).

The *full model* was fitted first, in which all studied items were regressed on group. The log likelihood, *df*, and scaling correction factor from this model were used to test each studied item for DIF. Next, $2k/3$ *constrained models* were fitted: one for each studied item. In these models, b_i for studied item i was constrained equal between groups (the item was not regressed on group). The log likelihood, *df*, and scaling correction factor from each constrained model were used for DIF testing. A significant ($\alpha = .05$) difference between the full model and the constrained model for item i indicated that item i functioned differently. Decisions about statistical significance were based on the adjusted p value (described later) from the χ^2 nested model test, corrected to control the false discovery rate (FDR).

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.

FDR-Corrected *p* Values

The Benjamini-Hochberg (1995) procedure has been recommended for controlling the FDR for DIF testing (Thissen, Steinberg, & Kuang, 2002; Williams, Jones, & Tukey, 1999). In the present study, all p values were adjusted using a C++ implementation of the method described in the documentation for the SAS MULTTEST procedure (Version 9.1). The adjusted, rather than raw, p values were compared with $\alpha = .05$ to determine statistical significance.

Two-Group Analysis

The two-group approach was applied using IRTLRDIF (Version 2.0b; Thissen, 2001) and for some conditions, MULTILOG (Version 7; Thissen et al., 2003). The IRTLRDIF program permits designation of anchors and tests each studied item for DIF as described earlier. The p value corresponding to the χ^2 statistic from each of these tests, adjusted to control the FDR, determined statistical significance. The IRTLRDIF program provides as many estimates of the item parameters and the mean difference as there are studied items. In the present study, parameter estimates from IRTLRDIF were not used.

In conditions with DIF and $N_F = 200$ or 400, MULTILOG was used to fit a *final model*, analogous to the final MIMIC model described previously. 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; the SD of θ was 1 for both groups. Final models were fitted only for the larger values of N_F because two-group 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.

Outcomes

DIF-free conditions. One outcome was computed for each of the 30 DIF-free conditions: the false positive rate (i.e., proportion of studied items with significant tests).

Conditions with DIF. One study-level outcome was used for conditions with DIF: the mean difference from the final model, averaged over replications. The mean difference is the coefficient for the regression of θ on group with the MIMIC approach and the absolute value of the focal-group mean with IRT-LR-DIF.

Item-level outcomes differed for anchors versus D-F items. For anchors, the absolute value of the mean bias was calculated for a_i and for b_i. For D-F items, the hit rate was computed as the proportion of items with significant tests. Hit rates were calculated separately for items with uniform versus nonuniform DIF.

Six additional outcomes were used to judge how well item parameters were estimated for D-F items when the significance tests were correct. Bias in $a_{i\mathrm{R}}$, $a_{i\mathrm{F}}$, $b_{i\mathrm{R}}$, and $b_{i\mathrm{F}}$, and the absolute difference between the mean true and estimated amount of DIF in a_i and b_i were computed using estimates from the final model for only the items that were correctly identified as D-F. 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). To control outliers, estimates of $a_{i\mathrm{R}}$ and $a_{i\mathrm{F}}$ greater than 4 were recoded to 4, and estimates of $b_{i\mathrm{R}}$ and $b_{i\mathrm{F}}$ more extreme than ± 4 were recoded to ± 4.

RESULTS FOR STUDY 1

DIF-Free Conditions

Type I error was well below 5% for both methods in all conditions. This was not surprising given that all false positive rates were based on FDR-adjusted p values. Over the 30 DIF-free conditions, the false positive rate ranged from .002 to .035 (or 0.2% to 3.5%) for MIMIC models and from .000 to .015 (or 0% to 1.5%) for two-group models. The false positive rate tended to be higher for MIMIC models than for two-group models, but differences between methods diminished as N_F or k increased. There was little difference between $N_R = 500$ and 1,000. Figure 2 displays the false positive rates with $N_R = 500$ and $k = 6$ (upper), 12 (middle), and 24 (lower).

Conditions With DIF

Hit rates. Consistent with what is known about statistical power, hit rates improved as N_F increased, but the improvement rate decreased so there was little difference between $N_F = 200$ and 400. This principle of diminishing returns is probably why rates for $N_R = 1{,}000$ and 500 were very similar. Hit rates were nearly identical for all k but always greater for items with uniform versus nonuniform DIF.

The left side of Figure 3 displays hit rates for binary items with $N_R = 1{,}000$ and $k = 6$ (upper), 12 (middle), and 24 (lower). For both methods, power was very low when $N_F \leq 100$ but was in a more acceptable range when $N_F = 200$ or 400. In all conditions (except $k = 24$, $N_F = 400$), the MIMIC method had greater power to detect uniform DIF. The MIMIC method also had greater power to detect nonuniform DIF for short scales ($k = 6$) or very small N_F (25

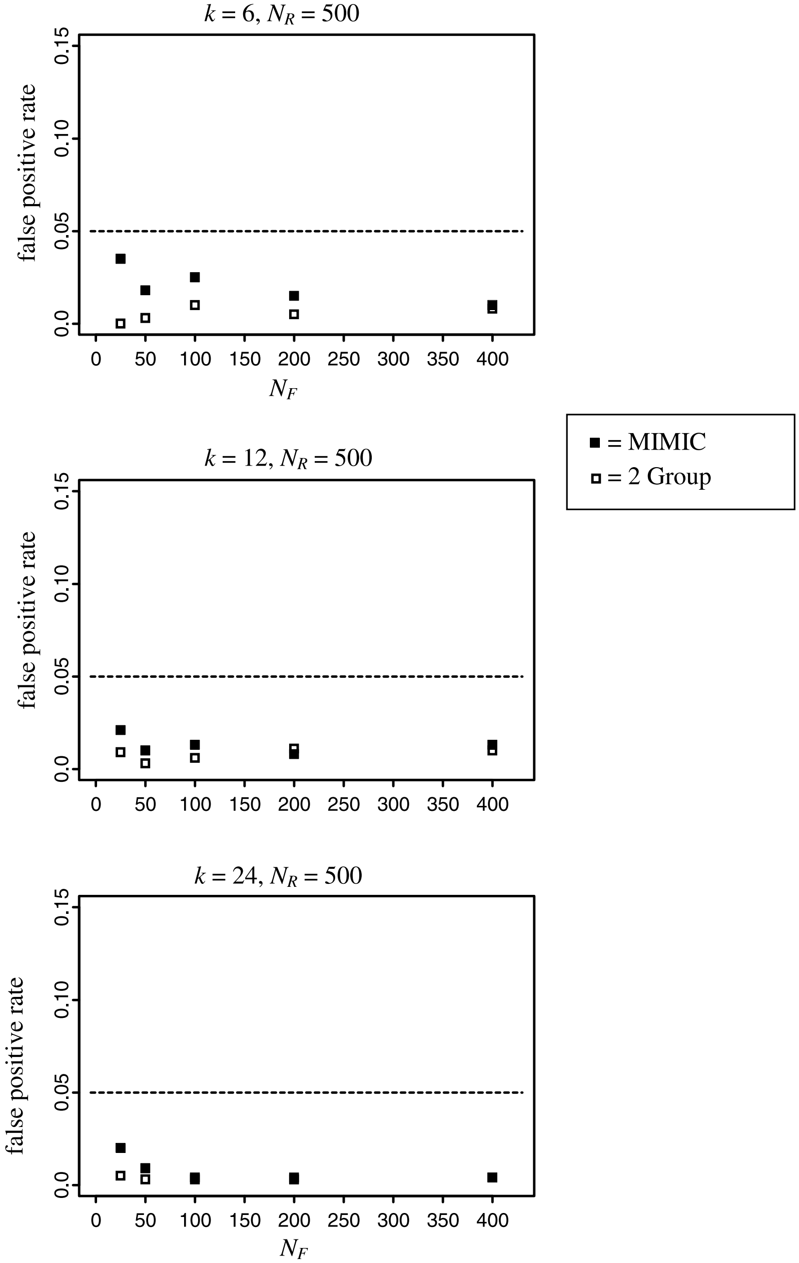

FIGURE 2 The false positive rate is the proportion of group-invariant studied items with significant DIF tests; k = number of items, N_R = reference group sample size, N_F = focal group sample size. A dashed line is drawn at .05 because $\alpha = .05$. These plots are for binary responses but the pattern was nearly identical with five-category ordinal responses.

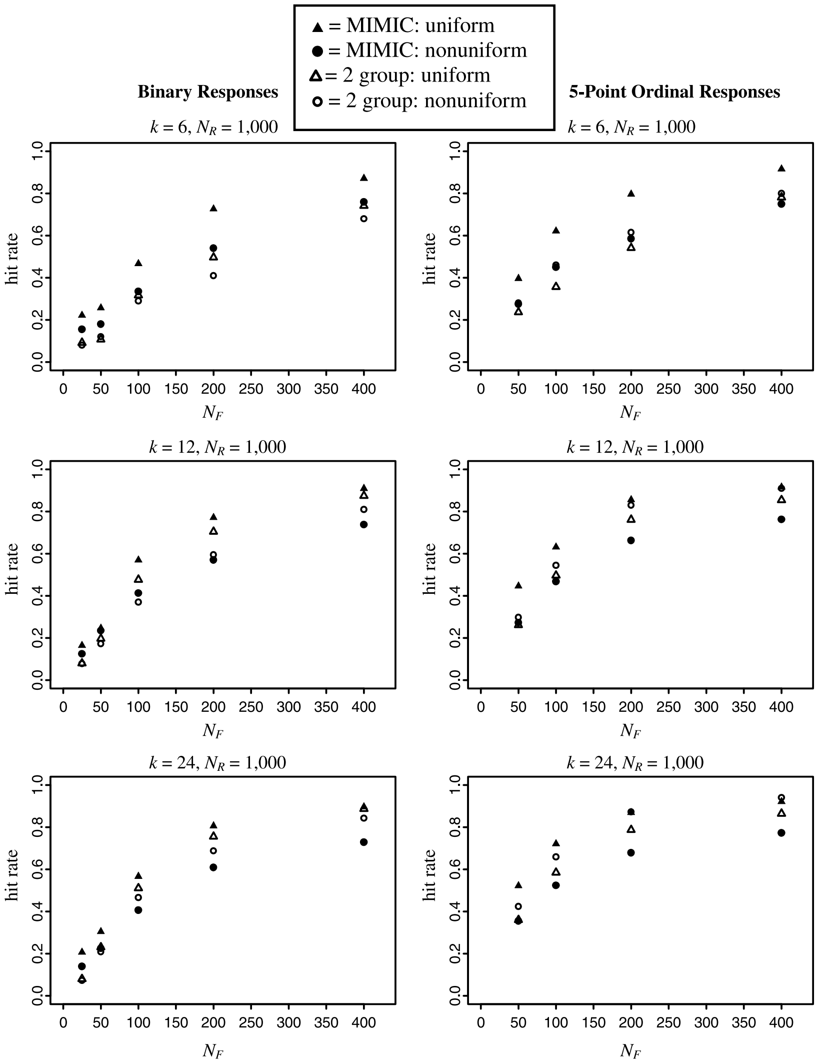

FIGURE 3 The hit rate is the proportion of differentially functioning studied items with significant DIF tests; k = number of items, N_R = reference group sample size, N_F = focal group sample size.

or 50). However, power to detect nonuniform DIF became greater for two-group analysis as k and N_F increased.

Mean difference. Estimates of the mean difference were minimally affected by N_R but clearly more accurate as N_F increased. When $k = 6$ and $N_F = 200$ or 400, estimates from MIMIC models were more accurate than those from two-group models. The two approaches produced nearly identical results for $k = 12$ and 24. Plots showing results for $N_R = 1{,}000$ and $k = 6$ (upper), $k = 12$ (middle), and $k = 24$ (lower) are displayed on the left side of Figure 4.

Item parameters for anchors. For anchors, bias in a_i and b_i was small in all conditions and similar for the two methods. For a_i, the bias ranged (over the 30 conditions) from .000 to .027 for the MIMIC method and ranged (over the 12 conditions) from .001 to .024 for two-group analysis. For b_i, the bias ranged (over conditions) from .000 to .037 for the MIMIC method and from .003 to .042 for two-group analysis.

Discrimination parameters for D-F items. The bias in $a_{i\mathrm{R}}$ and $a_{i\mathrm{F}}$ was minimally affected by N_R and k. For $N_R = 500$, Figure 5 shows the bias in $a_{i\mathrm{R}}$ (left) and $a_{i\mathrm{F}}$ (right). For items with uniform DIF, bias in $a_{i\mathrm{R}}$ was small for all N_F and nearly identical for the MIMIC and two-group methods. Bias in $a_{i\mathrm{F}}$ was also small for all N_F with the MIMIC method but a little elevated for $N_F = 200$ with the two-group approach.

For items with nonuniform DIF, bias in $a_{i\mathrm{R}}$ was low for two-group analysis but worsened as N_F increased for MIMIC models. Bias in $a_{i\mathrm{F}}$ was quite elevated for all N_F with the MIMIC method but small for two-group analysis. This makes sense because the MIMIC method is insensitive to DIF with respect to a_i.

For all conditions, bias in the amount of DIF in a_i was elevated to about .5 for items with nonuniform DIF tested with the MIMIC approach because the estimated DIF was always 0 and the true amount of DIF was about .5 (the mean of .3, .4, .5, .6, and .7). For items with uniform DIF tested with the MIMIC approach, bias in the amount of DIF was 0 because both the true and estimated amount of DIF were always 0. With two-group analysis, the amount of DIF was estimated fairly well for all items (range of bias in a_i DIF, over conditions .012 to .173).

Threshold parameters for D-F items. Bias in $b_{i\mathrm{R}}$ was low and similar to that observed for anchors. Bias was greatest for items with nonuniform DIF tested with the MIMIC approach: Over the 30 conditions, bias ranged from .000 to .066. For items with uniform DIF tested with MIMIC models, and for all D-F items tested with two-group analysis, bias in $b_{i\mathrm{R}}$ was less than .05.

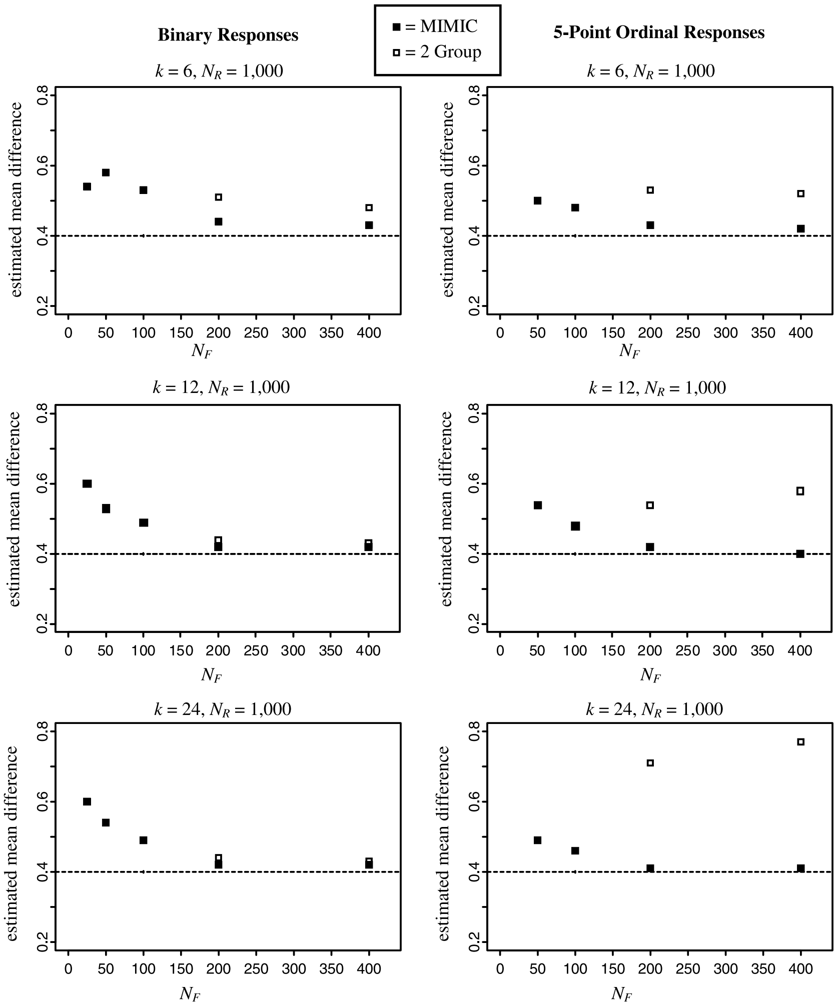

FIGURE 4 Estimates of the group-mean difference on θ, averaged over 100 replications in each condition; k = number of items, N_R = reference group sample size, N_F = focal group sample size. A dashed line is drawn at the true population value of 0.4.

Bias in $b_{i\mathrm{F}}$ from MIMIC models was quite large when $N_F < 100$ but improved dramatically when $N_F \geq 100$. The bias was minimally affected by N_R and k. Figure 6 plots the bias in $b_{i\mathrm{F}}$ for $N_R = 500$ and $k = 6$ (upper), 12 (middle), and 24 (lower). With $N_F = 200$ or 400, estimation of $b_{i\mathrm{F}}$ for items with uniform DIF was virtually identical for the two methods. However, for

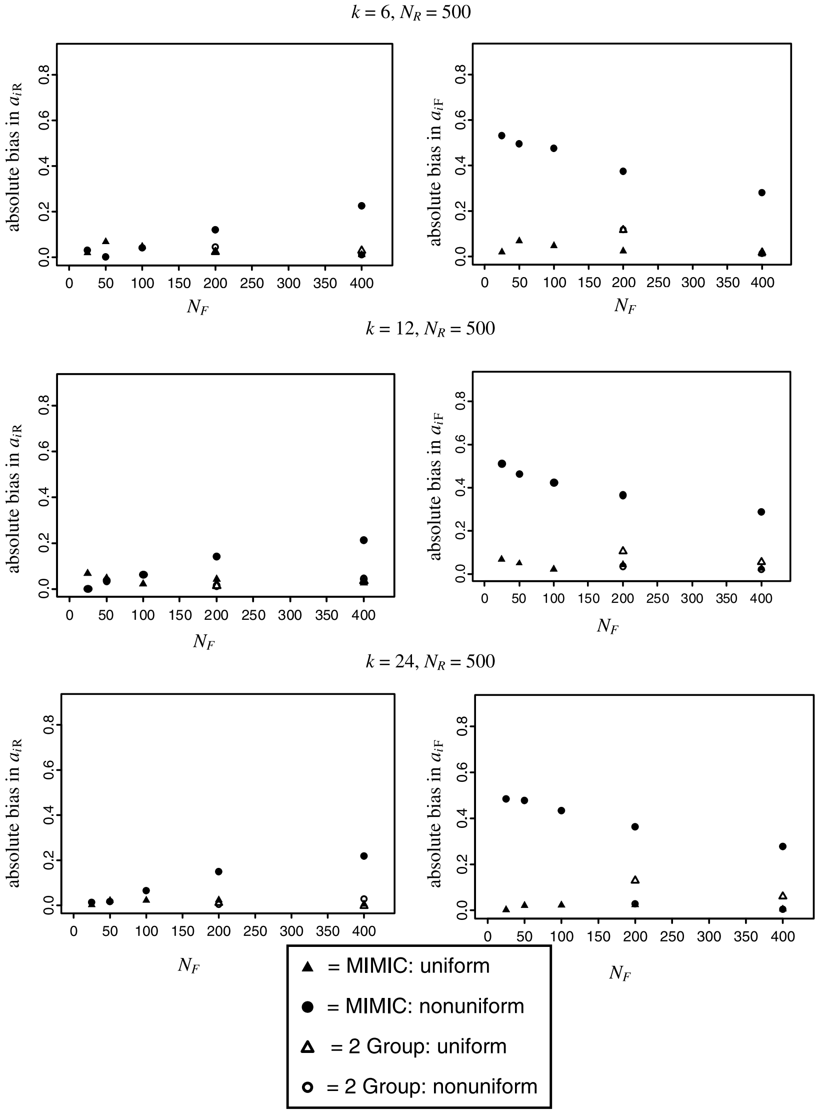

FIGURE 5 The absolute value of the mean bias in the discrimination parameter for the reference group, $a_{i\text{R}}$ (left side) or the focal group, $a_{i\text{F}}$ (right side); k = number of items, N_R = reference group sample size, N_F = focal group sample size. These plots are for binary responses but the pattern was nearly identical for five-category ordinal responses.

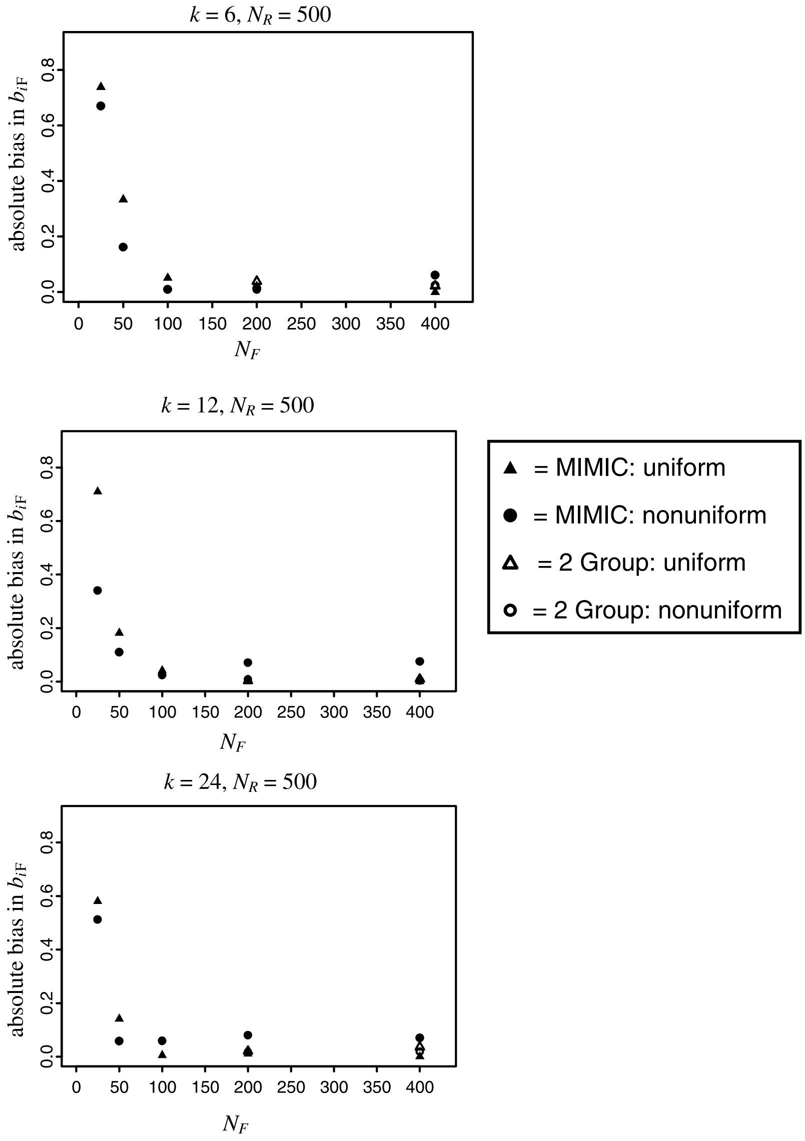

FIGURE 6 For binary responses, the absolute value of the mean bias in the reference group threshold parameter ($b_{i\mathrm{R}}$); k = the number of items, N_R = reference group sample size, N_F = focal group sample size.

items with nonuniform DIF, bias was a little greater with MIMIC models than two-group analysis ($k = 6$ with $N_F = 200$ is an exception). Because estimation inaccuracy occurred for $b_{i\mathrm{F}}$ and not $b_{i\mathrm{R}}$, the pattern of bias in the estimated amount of DIF in b_i was virtually identical to that for $b_{i\mathrm{F}}$.

METHODS FOR STUDY 2: ORDINAL RESPONSES

One hundred sets of ordinal (five-point) item data were generated for Study 2. The methodology used in Study 1 was replicated with a few differences in the design and the data generation procedure, as described subsequently.

Design

There were 24 independent simulation conditions varying according to the focal-group sample size ($N_F = 50$, 100, 200, or 400), number of items ($k = 6$, 12, or 24), and presence versus absence of DIF. No samples with $N_F = 25$ were generated because results with binary data were poor, and N_R was not varied because most outcomes for binary items were virtually identical with $N_R = 500$ versus 1,000. In Study 2, $N_R = 1{,}000$ for all conditions.

Data Generation

Five-category ordinal data were generated from Samejima's (1997) graded model, which is a generalization of Birnbaum's 2PL model for more than two ordered categories. There is one discrimination parameter per item (a_i) and one fewer threshold parameter (b_{ij}) than total number of response categories. Mplus parameterizes the graded model analogously to the 2PL such that $\tau_{ij} = a_i b_{ij}$. A common parameterization of the graded model, for the probability of responding in category v, is $\Pr(u_{ij} = v|\theta_j) = \frac{1}{1+\exp[-a_i(\theta_j - b_{i,v})]} - \frac{1}{1+\exp[-a_i(\theta_j - b_{i,v+1})]}$, where $\frac{1}{1+\exp[-a_i(\theta_j - b_{i,0})]} = 1$, $\frac{1}{1+\exp[-a_i(\theta_j - b_{i,c+1})]} = 0$, and $c =$ highest response category. In Mplus, the parameterization is $\Pr(u_{ij} = v|\theta_j) = \frac{1}{1+\exp[\tau_{i,v} - a_i\theta_j]} - \frac{1}{1+\exp[\tau_{i,v+1} - a_i\theta_j]}$ (with the analogous endpoints equal to 0 and 1).

For each item, a_{iR} was drawn from N($\mu = 1.7$, $\sigma = 0.6$) with truncation on the upper end at 4.0 and on the lower end at 0.5 (items without nonuniform DIF) or 1.2 (items with nonuniform DIF). As in Study 1, $a_{i\mathrm{F}}$ was either equal to $a_{i\mathrm{R}}$ or $a_{i\mathrm{R}} - \delta$, depending 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 $b_{i1F} = b_{i1R} + \gamma$, $b_{i2F} = b_{i2R} + \gamma$, $b_{i3F} = b_{i3R} + \gamma$, and $b_{i4F} = b_{i4R} + \gamma$.

When an item was simulated with a 0 cell frequency for either the R or F group (i.e., 0 simulees responded in one or more of the five categories), the categories for this item were collapsed (for both groups).

RESULTS FOR STUDY 2

DIF-Free Conditions

False positive rates were well under .5 for all conditions and similar for the two methods. The pattern of results was similar to that shown in Figure 2 for binary data. Across the 12 conditions, false positive rates ranged from .0019 to .0175 (0.19 to 1.75%) for the MIMIC method and from .0013 to .0200 (0.13% to 2.00%) for two-group analysis.

Conditions With DIF

Hit rates. Hit rates for ordinal items are displayed on the right side of Figure 3 for $k = 6$ (upper), 12 (middle), and 24 (lower). Power was somewhat low with $N_F = 50$, though not as low as with binary items. Power to detect uniform DIF was always greater for the MIMIC method than two-group analysis, but power to detect nonuniform DIF was either about the same for the two methods (with $k = 6$) or better for two-group analysis.

Mean difference. Estimated mean differences are shown on the right side of Figure 4. The MIMIC method was quite accurate, especially as N_F increased. As k increased, estimates from two-group analysis worsened, with rather large overestimation apparent when $k = 24$.

Item parameters for anchors. Anchor item parameters were estimated well from MIMIC models. Bias ranged (over conditions) from .001 to .016 for a_i, from .000 to .012 for b_{i1}, from .002 to .032 for b_{i2}, from .001 to .072 for b_{i3}, and from .000 to .021 for b_{i4}.

Table 1 compares bias in anchor parameter estimates from MIMIC and two-group models. Values for the two methods are comparable for a_i with all k and for b_{ih}s when $k = 6$. However, for the longer scales, accuracy in b_{ih}s estimated using two-group analysis systematically declined as k and N_F increased. The

TABLE 1
Absolute Bias in Item Parameter Estimates for Ordinal Anchors

Condition	a_i	b_{i1}	b_{i2}	b_{i3}	b_{i4}	*Free Pars*
$k = 6$			*MIMIC Models*			
$N_F = 200$	.004	.005	.008	.038	.019	33
$N_F = 400$	.001	.009	.000	.006	.010	34
$k = 12$						
$N_F = 200$	.012	.001	.002	.001	.013	65
$N_F = 400$	.003	.000	.004	.005	.004	66
$k = 24$						
$N_F = 200$	.009	.000	.011	.032	.020	128
$N_F = 400$	.004	.003	.003	.014	.002	132
$k = 6$			*Two-Group Analysis*			
$N_F = 200$	.015	.036	.025	.012	.000	41
$N_F = 400$	.015	.053	.039	.034	.026	46
$k = 12$						
$N_F = 200$	.018	.070	.091	.104	.092	89
$N_F = 400$	.003	.132	.128	.126	.127	94
$k = 24$						
$N_F = 200$	.002	.261	.255	.246	.238	180
$N_F = 400$	.005	.318	.317	.310	.305	188

Note. k = number of items; N_F = focal group sample size; a_i = discrimination parameter; b_{i1}–b_{i4} = threshold parameters; *free pars* = average (rounded to integer) number of free parameters in the final model.

number of free parameters in the final models also grew with increases in k and N_F. Greater N_F leads to more free parameters because power is higher, and when more DIF is detected, the parameters of more studied items are estimated separately for the two groups in the final model. The difference in the number of free parameters used for MIMIC versus two-group final models is larger for $k = 12$ or 24 than for $k = 6$; this may explain why bias was worse for two-group models with the longer scales.

Discrimination parameters for D-F items. In Study 2, the pattern of results for bias in $a_{i\mathrm{R}}$ and $a_{i\mathrm{F}}$ was the same as that observed in Study 1 (see Figure 5 and the earlier narrative).

Threshold parameters for D-F items. Bias in thresholds is given in Table 2 (nonuniform DIF) and Table 3 (uniform DIF) for estimates from MIMIC models (regular type) and two-group analysis (bold type). In all cases, R-group thresholds tended to be better estimated than F-group thresholds.

TABLE 2
Absolute Bias in Threshold Estimates for Ordinal Studied Items With Nonuniform DIF

	Reference Group				*Focal Group*			
Condition	b_{i1R}	b_{i2R}	b_{i3R}	b_{i4R}	b_{i1F}	b_{i2F}	b_{i3F}	b_{i4F}
	MIMIC Models							
$k = 6$								
$N_F = 50$	.012	.039	.092	.013	.024	.027	.087	.005
$N_F = 100$	.018	.011	.033	.007	.080	.087	.068	.106
$N_F = 200$	.004	.008	.001	.020	.078	.090	.080	.062
2 group	**.005**	**.013**	**.001**	**.005**	**.054**	**.022**	**.004**	**.077**
$N_F = 400$	.006	.008	.028	.039	.085	.082	.064	.047
2 group	**.031**	**.026**	**.016**	**.013**	**.085**	**.067**	**.061**	**.060**
$k = 12$								
$N_F = 50$	.002	.036	.046	.033	.011	.022	.024	.016
$N_F = 100$	.001	.010	.030	.012	.084	.072	.054	.075
$N_F = 200$	.001	.001	.003	.008	.091	.093	.089	.083
2 group	**.083**	**.083**	**.080**	**.078**	**.114**	**.115**	**.138**	**.224**
$N_F = 400$	.007	.008	.025	.023	.105	.105	.089	.094
2 group	**.120**	**.123**	**.119**	**.131**	**.172**	**.173**	**.163**	**.195**
$k = 24$								
$N_F = 50$	.004	.016	.057	.014	.062	.049	.008	.043
$N_F = 100$	.003	.002	.019	.007	.090	.092	.073	.090
$N_F = 200$	.006	.009	.028	.026	.099	.096	.079	.080
2 group	**.245**	**.245**	**.235**	**.243**	**.277**	**.256**	**.264**	**.293**
$N_F = 400$	.008	.011	.034	.036	.104	.102	.079	.081
2 group	**.300**	**.299**	**.289**	**.297**	**.351**	**.342**	**.335**	**.340**

Note. k = number of items; N_F = focal group sample size. All values are from the MIMIC approach except those in bold type, which are from two-group analysis.

For items with nonuniform DIF (Table 2), the pattern of bias in R-group thresholds was similar to that observed for anchor items. Bias was low for all N_F and both methods when $k = 6$ but increased with k and N_F (thus the number of free parameters) for two-group estimates. This pattern of increasing bias also emerged for F-group thresholds estimated using two-group analysis. For a given value of k (above 6), focal-group thresholds estimated using MIMIC models were increasingly biased as N_F increased, but accuracy was greater than with two-group analysis. For items with uniform DIF (Table 3), bias was consistently low for MIMIC-model estimates but tended to increase with increasing k and N_F for two-group estimates.

Because thresholds from two-group analysis were underestimated about the same amount for both R and F groups, the amount of bias in the estimated DIF

TABLE 3
Absolute Bias in Threshold Estimates for Ordinal Studied Items With Uniform DIF

	Reference Group				*Focal Group*			
Condition	b_{i1R}	b_{i2R}	b_{i3R}	b_{i4R}	b_{i1F}	b_{i2F}	b_{i3F}	b_{i4F}
	MIMIC Models							
$k = 6$								
$N_F = 50$	.017	.012	.087	.011	.174	.169	.266	.169
$N_F = 100$	.004	.021	.042	.005	.027	.027	.042	.017
$N_F = 200$	.001	.012	.057	.025	.009	.001	.039	.012
2 group	**.010**	**.002**	**.041**	**.023**	**.015**	**.103**	**.071**	**.022**
$N_F = 400$	.006	.003	.004	.002	.000	.003	.004	.001
2 group	**.033**	**.033**	**.033**	**.025**	**.044**	**.038**	**.028**	**.037**
$k = 12$								
$N_F = 50$	.004	.038	.081	.004	.035	.059	.096	.030
$N_F = 100$	.003	.013	.035	.020	.008	.022	.035	.005
$N_F = 200$	.009	.000	.020	.003	.008	.003	.014	.005
2 group	**.071**	**.083**	**.065**	**.083**	**.129**	**.119**	**.136**	**.172**
$N_F = 400$	.005	.010	.003	.002	.006	.010	.002	.015
2 group	**.113**	**.111**	**.122**	**.123**	**.158**	**.167**	**.175**	**.209**
$k = 24$								
$N_F = 50$	.001	.006	.048	.010	.027	.023	.051	.008
$N_F = 100$	.006	.004	.032	.003	.008	.014	.033	.002
$N_F = 200$	.004	.004	.023	.016	.003	.010	.027	.012
2 group	**.250**	**.251**	**.235**	**.243**	**.275**	**.270**	**.275**	**.290**
$N_F = 400$	.005	.003	.016	.007	.005	.003	.014	.012
2 group	**.307**	**.305**	**.299**	**.305**	**.356**	**.341**	**.319**	**.333**

Note. k = number of items; N_F = focal group sample size. All values are from the MIMIC approach except those in bold type, which are from two-group analysis.

tended to be low: it ranged, over conditions, from .002 to .151. For MIMIC models, bias in the amount of DIF ranged from .000 to .121.

EMPIRICAL EXAMPLE OF MIMIC MODELS FOR DIF TESTING

Data[1] were item responses to the Loss of Control (LOC) subscale of the Agoraphobic Cognitions Questionnaire (ACQ; Chambless, Caputo, Bright, & Gal-

[1] The data for the empirical example were originally collected by Carol M. Woods as part of a collaborative project with Jonathan S. Abramowitz and David F. Tolin. Participants constitute a subsample of the samples published previously in research by Deacon, Abramowitz, Woods, & Tolin (2003) and Woods (2006).

lagher, 1984) given via the World Wide Web by undergraduates at the University of North Carolina at Chapel Hill in exchange for course credit. The ACQ consists of 14 items; here we used only the items that constitute the LOC subscale (Items 6, 8, 9, 11, 12, 13, and 14). Participants rated how often each of 7 thoughts or ideas (e.g., "14. 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). Participants who self-identified as Black/African American ($N_F = 139$) versus White ($N_R = 757$) were compared. The majority of these participants were women ($n = 677$), and the mean age was 19 ($SD = 1.96$).

Because N_F is rather small, IRT-LR-DIF might not perform well for these data, but the present simulations suggest that MIMIC models are a good alternative. First, preliminary testing was carried out to empirically select designated anchor items. Second, DIF testing was carried out using the same procedures described for the simulation study.

Following the rationale described by Woods (in press), anchor items were selected empirically by testing each item for DIF 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 the grouping variable. Three items with the smallest SS/SE ratios were selected as anchors: 8, 11, and 14. Thus, Items 6, 9, 12, and 13 were tested individually for DIF using a designated anchor.

Table 4 lists the χ^2 test results and the discrimination (a) and threshold (τ_j) parameter estimates from the final model. Items 6, 9, 12, and 13 appear to function significantly ($\alpha = .05$) differently between groups. These items are "6. I am going to act foolish; 9. I will hurt someone; 12. I am going to scream; 13. I am going to babble or talk funny." The estimated mean difference on the latent variable (from the final model) was $\gamma = -.52$ ($SE = .13$), indicating that loss of control was higher for the group coded zero (Whites).

TABLE 4
MIMIC-Model DIF Testing for Items on the Loss of Control Scale
(139 African Americans vs. 757 Whites)

Item	$\chi^2(1)$	*p*	*a (SE)*	τ_1 *(SE)*	τ_2 *(SE)*	τ_3 *(SE)*	τ_4 *(SE)*	β *(SE)*
9	6.41	.011	0.87 (.12)	♦1.42 (.11)	♦2.98 (.16)	♦4.27 (.26)	—	.59 (.23)
6	6.28	.012	1.49 (.15)	♦−2.18 (.14)	♦−0.38 (.10)	♦1.37 (.12)	♦3.55 (.21)	−.58 (.23)
12	5.83	.016	1.42 (.13)	♦−0.43 (.10)	♦1.08 (.11)	♦2.49 (.14)	♦4.40 (.24)	.61 (.26)
13	4.98	.026	1.40 (.14)	♦−1.20 (.11)	♦0.23 (.10)	♦1.70 (.13)	♦3.50 (.20)	.50 (.23)
8	anchor		2.10 (.19)	−0.25 (.12)	2.15 (.16)	4.13 (.25)	6.50 (.45)	n/a
11	anchor		1.61 (.18)	0.73 (.11)	2.33 (.15)	3.64 (.21)	5.29 (.34)	n/a
14	anchor		1.39 (.14)	0.85 (.10)	2.52 (.15)	3.49 (.20)	4.67 (.29)	n/a

Note. $\alpha = .05$; a = estimated discrimination; SE = standard error; τ_j = estimated threshold; ♦ = applies to Whites only; β = regression coefficient showing the group difference in τ_js for this item; n/a = not applicable because the item was an anchor; — = threshold not estimated because respondents did not use all available response categories.

For items with DIF, τ_js given in Table 4 apply only to Whites. The estimates for African Americans are equal to $\tau_j + \beta$, where β is the DIF effect, reflecting the relation 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, African Americans had to possess more loss of control than Whites before endorsing a particular response category for Items 9, 12, and 13, and Whites had to possess more loss of control than African Americans before endorsing any category for Item 6.

CONCLUSIONS AND RECOMMENDATIONS

MIMIC models are useful for DIF testing because they use latent variables to account for measurement error in the responses and offer the flexibility of an SEM context; for example, it is easy to model multiple factors, more than two groups, and categorical or continuous covariates. The present research provides empirical support for an additional virtue: With other study characteristics held constant, the sample size needed for adequate power and reasonably accurate parameter estimation is smaller for MIMIC models than two-group analysis.

With small N_F, tests of uniform DIF with binary or five-category ordinal responses were more accurate with MIMIC models than IRT-LR-DIF. At all values of N_F, Type I error was well below the nominal α level and power was greater for the MIMIC approach than for IRT-LR-DIF. With $N_F = 200$ or 400, MIMIC-model estimates of the mean difference on θ were quite accurate and closer to the true value than those from IRT-LR-DIF, and MIMIC estimates of Birnbaum's 2PL and graded model parameters were as accurate, or more so, than those from IRT-LR-DIF.

Despite the advantages and favorable simulation findings, MIMIC methods are not always preferable to IRT-LR-DIF. With larger N_F, IRT-LR-DIF has performed more accurately than in the present study (Ankenmann et al., 1999; Cohen et al., 1996; Kim & Cohen, 1998; Wang & Yeh, 2003). Also, the present simulation did not fully exploit the capabilities of IRT-LR-DIF because tests for uniform and nonuniform DIF were not carried out. Final models were constructed so that either all or no parameters for each item were permitted to vary between groups. As a result, a_i for items with uniform DIF was estimated separately in each group in the final model, which is a potentially avoidable misspecification. Obviously, free estimation of a_{iF} and a_{iR} permits them to be equal, but it adds to the number of free parameters, and increases in free parameters were associated with reductions in parameter accuracy for IRT-LR-DIF.

An important limitation of MIMIC methods is that they cannot test for nonuniform DIF. With $k > 6$ and $N_F > 25$, IRT-LR-DIF always had greater power to detect nonuniform DIF than MIMIC models, and for all k and N_F, bias was elevated for MIMIC estimates of a_{iF} and ordinal-item b_{ihF} for nonuniform D-F items. In this study, nonuniform DIF was defined as differential functioning of both a_i and b_{ih} because this pattern is frequently observed in published applications of IRT-LR-DIF. However, it is possible for items to have D-F a_i with invariant b_{ih}. MIMIC methods are not designed for such items and are expected to perform poorly in this case.

Some recommendations are offered on the basis of this research. Consider using MIMIC methods when it is substantively justifiable to hypothesize uniform DIF and unlikely that items have D-F a_i with invariant b_{ih}. Items hypothesized to have D-F a_i and b_{ih} (nonuniform DIF) may be included in the analysis as long as F-group item parameters will not be interpreted. Power is likely to be a little lower for nonuniform DIF, but if it is more important to detect as much uniform DIF as possible than to guard against misclassifying some invariant items as D-F, power may be increased by using raw rather than FDR-adjusted p values, $\alpha > .05$, or both.

For scales and data similar to those simulated here, N_F should be at least around 100 (with $N_R \geq 500$) for reasonably powerful, accurate MIMIC results. Because power is greater when item discrimination is larger (Ankenmann et al., 1999), smaller N_F may be acceptable for the highly discriminating items sometimes observed on psychopathology scales (e.g., Rodebaugh et al., 2004). Scale length is unlikely to have much impact on statistical power or item parameter accuracy because differences between $k = 6$, 12, or 24 were slight, but longer scales may produce more accurate estimates of the mean difference.

In future research, it would be useful to evaluate the extent to which the differences in hit rates and false positives between MIMIC models and IRT-LR-DIF translate into practical consequences for score interpretation. Future studies are also needed to determine how well these results and recommendations generalize when there are covariates, multiple factors, or more than two groups.

ACKNOWLEDGMENT

I am grateful to Michael Strube for providing helpful suggestions on a draft of this article.

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.

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.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist-Revised. *Psychological Assessment, 16,* 155–168.

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

Chambless, D. L., Caputo, G. C., Bright, P., & Gallagher, R. (1984). Assessment of 'fear of fear' in agoraphobics: The Body Sensations Questionnaire and the Agoraphobic Cognitions Questionnaire. *Journal of Consulting and Clinical Psychology, 52,* 1090–1097.

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., et al. (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.

Deacon, B. J., Abramowitz, J. S., Woods, C. M., & Tolin, D. F. (2003). The Anxiety Sensitivity Index-Revised: Psychometric properties and factor structure in two nonclinical samples. *Behaviour Research and Therapy, 41,* 1427–1449.

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. *Medical Care, 44,* S134–S142.

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, and Coping, 17,* 49–67.

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

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*, Hillsdale, NJ: Erlbaum.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

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, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8,* 291–312.

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.

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

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.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41,* 55–64.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7,* 361–388.

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

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515.

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., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407–419.

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: 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.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Reider (Eds.), *Equivalence in measurement: Research in management* (pp. 25–50). Greenwich, CT: Information Age Publishing.

*Rodebaugh, T. L., *Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16,* 169–181. *Contributions of the first two authors are equal.

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

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 and Tobacco Research, 9,* 137–145.

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., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1291–1306.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

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., Chen, W.-H., & Bock, R. D. (2003). Multilog (Version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.

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: 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: Erlbaum.

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

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.

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. (2006). Ramsay-curve item response theory to detect and correct for nonnormal latent variables. *Psychological Methods, 11,* 253–270.

Woods, C. M. (in press). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). *Differential item functioning on the Schedule for Nonadaptive and Adaptive Personality.* Manuscript submitted for publication.