The MIMIC Model as a Method for Detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio

Holmes Finch, Department of Educational Psychology, Ball State University

This study compares the ability of the multiple indicators, multiple causes (MIMIC) confirmatory factor analysis model to correctly identify cases of differential item functioning (DIF) with more established methods. Although the MIMIC model might have application in identifying DIF for multiple grouping variables, there has been little examination of how well the technique works in terms of correct and incorrect identification of DIF. A Monte Carlo methodology is used in this study, with manipulation of the number of items, number of examinees, differences between the mean abilities of the reference and focal groups, level of DIF contamination of the anchor items, and amount of DIF in the target item. Results indicate that the MIMIC model is effective for DIF identification for 50 items or when the two-parameter logistic model underlies the data but has a very high rate of incorrect DIF identification for 20 items with three-parameter logistic data. *Index terms: differential item functioning, MIMIC model, likelihood ratio test, Mantel-Haenszel statistic, SIBTEST, item response theory.*

Recently, several authors have discussed the link between item response theory (IRT) models and confirmatory factor analysis (CFA) for dichotomous variables (Fleishman, Spector & Altman, 2002; Glockner-Rist & Hoitjink, 2003; MacIntosh & Hashim, 2003). Both models can be used to link responses on a test item with a latent examinee ability. Both of these modeling approaches yield parameter estimates describing the nature of this relationship and the items and examinees themselves, including the item difficulty and discrimination, and the underlying ability of each examinee. It has been shown that the parameter estimates obtained using the multiple indicator, multiple causes (MIMIC) CFA model can be easily converted to the parameter estimates common to the IRT model (MacIntosh & Hashim, 2003; Muthen, Kao, & Burstein, 1991).

The three-parameter logistic (3PL) item response function (IRF) takes the following form:

$$P(U_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}, \quad (1)$$

where $\theta =$ examinee ability, $c_i =$ pseudo-guessing parameter for item i, $a_i =$ discrimination parameter for item i, and $b_i =$ difficulty parameter for item i.

This IRF can be estimated for each item on an exam, yielding the relationship between theta and the probability of producing a correct response. When pseudo-guessing is not present, the data come from a two-parameter logistic (2PL) IRF, which can be expressed as

$$P(U_i = 1|\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}.$$

Applied Psychological Measurement, Vol. 29 No. 4, July 2005, 278–295
DOI: 10.1177/0146621605275728

In turn, the CFA model takes the following form:

$$y_i^* = \lambda_i \eta + \varepsilon_i, \tag{2}$$

where y_i^* = latent response variable i (when $y_i^* > \tau_i$, an observed variable, $y_i = 1$; τ_i, the threshold parameter, is related to item difficulty), η = latent trait, λ_i = factor loading for variable i, and ε_i = random error.

This factor-analytic model for dichotomous item response data is equivalent to the normal ogive model described by McDonald (1967) and Lord and Novick (1968). Muthen et al. (1991) extended this work for the MIMIC model, demonstrating that the discrimination parameter, a, in the IRT formulation can be obtained using the value of λ from the MIMIC CFA model:

$$a_i = \frac{\lambda_i}{\sqrt{(1-\lambda_i^2)}\sqrt{\sigma_{\eta\eta}}},$$

where $\sigma_{\eta\eta}$ = variance of the latent trait.

In turn, in the MIMIC framework, the IRT difficulty parameter, b, involves the use of both λ and τ:

$$b_i = [(\tau_i - \beta_i z_k)\lambda_i^{-1} - \mu_\eta]\sigma_{\eta\eta}^{-1/2},$$

where z_k = group indicator, where 1 indicates membership in focal group and 0 indicates reference; β_i = measure of relationship between group and item response (significant value of β_i indicates presence of differential item functioning [DIF]); and μ_η = mean of the latent trait.

For a complete explication of these relationships, see Muthen et al. (1991) or MacIntosh and Hashim (2003).

Given the correspondence between these two classes of models, it seems reasonable to consider application of one to help answer questions raised by the other. For example, an important issue in IRT is the concept of uniform DIF, which occurs when the probability of responding correctly to an item is uniformly higher for one of two groups, generally referred to as the reference and focal groups, across all levels of ability. For example, if the latent trait is taken to be a measure of some cognitive ability, then DIF occurs when people in the focal group who are at the same level of the ability as members of the reference group have lower probabilities of answering the item correctly. DIF has been the focus of a great deal of investigation, both in terms of its causes and methods for identifying it (for extensive methodological reviews, see Camilli & Shepard, 1994; Millsap & Everson, 1993).

Two broad categories of DIF have been identified. The first, uniform DIF, is described above. Nonuniform DIF, on the other hand, refers to the case where an item discriminates differently for the groups in question. In this context, the shapes of the IRFs differ between the two groups, and often the graphs of the IRFs will cross. Ackerman (1992) points out that in general, DIF can be thought of as being caused when an item measures more than one latent construct. The item will have a primary ability that it is designed to measure, as well as one or more "nuisance" dimensions. DIF occurs when the groups have different ability distributions on this "nuisance" dimension. The focus of this article is on the assessment of uniform DIF.

A large number of methods have been developed and adapted for identifying DIF. Some of these approaches, such as the Mantel-Haenszel (MH) chi-square, logistic regression, and log-linear analysis, are based on statistical models developed for categorical data. Other techniques in use are based on differences in specific IRFs between the groups of interest (Millsap & Everson, 1993). These methods assume a specific form of the function relating the probability of a correct

response to an item and ability. Another method that has been proposed, SIBTEST, standardizes the two groups of interest to have a common distribution of the latent trait and then estimates the expected difference in scores between the groups (Shealy & Stout, 1993).

The methods described above have been studied fairly extensively in terms of their ability to correctly identify DIF when it is present for particular items (Jiang & Stout, 1998; Navas-Ara & Gomez-Benito, 2002; Rogers & Swaminathan, 1993; Roussos & Stout, 1996; Zwick, 1990). In terms of identifying the presence of uniform DIF, researchers have found that both the MH and SIBTEST approaches are very effective tools (Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993; Roussos & Stout, 1996). (A more complete summary of the technical details underlying each can be found in Roussos & Stout, 1996). Generally, they have similar rates of correct detection (power) and incorrect detection (Type I error), with better performance associated with longer tests, a larger number of examinees, and less contamination of the anchor items used to estimate the latent trait. See Camilli and Shepard (1994) for a more thorough review of the performance of these and other established methods of DIF detection.

Another approach to DIF detection that has shown promise is the IRT likelihood ratio (LR) test. This method, described by Thissen, Steinberg, and Gerrard (1986) and further expanded on later (Thissen, Steinberg, & Wainer, 1988), allows for the comparison of model fit, assuming equality of the parameter estimates for the item in question across the reference and focal groups (compact model), with the model fit when this constraint is relaxed and differences for the item parameters across the groups are allowed (augmented model). (Note that for both models, the parameter estimates for a set of anchor items are constrained to be equal for both groups.) This approach begins by comparing the two groups on all item parameters simultaneously. The test statistic takes the following form:

$$LR = -2\ln L_c - (-2\ln L_A), \tag{3}$$

where $L_c =$ log likelihood of the compact model, and $L_A =$ log likelihood of the augmented model.

For a 3PL model, the resulting statistic is distributed as a chi-square with 3 degrees of freedom. If this value is statistically significant, subsequent tests are computed to compare the fit of models to the two groups, holding all of the item parameters equal except for one. For example, to test for the presence of uniform DIF when an overall significant result for the item in question has first been obtained, a subsequent LR statistic is calculated, with the difficulty parameter allowed to vary in the augmented model and, as before, the compact model holding all three parameters to be equal:

$$LR_b = -2\ln L_c - (-2\ln L_{Ab}), \tag{4}$$

where $L_c =$ log likelihood of the compact model, and $L_{Ab} =$ log likelihood of the augmented model, with the difficulty parameter allowed to vary between groups.

The difference between the $-2\ln L$ values for the two models in this case is distributed as a chi-square statistic with 1 degree of freedom and tests for a significant difference only between the difficulty parameters of the two groups, and thus for the presence of uniform DIF. Thissen (2001) has formalized this methodology in a software program known as IRTLRDIF.

Compared to other methods, there has been relatively little research done assessing the effectiveness of this approach to DIF detection (Thissen, 2001). One reason for this dearth of investigation appears to be the heretofore time-consuming nature of the calculations necessary to compute the test statistics and the relatively recent development of the software to do this easily. One such study did find that the Type I error rate for the LR statistic was somewhat above the nominal 0.05 for 3PL models for samples of both 250 and 1,000 examinees for a 50-item test

(Cohen, Kim, & Wollack, 1996). Although this study did not examine the power of the IRT LR approach, Wanichtanom (2001) did and found that, averaged across group differences in the b_i parameters (set at 0.2, 0.5, and 0.8), it had a 0.97 detection rate for uniform DIF. This study simulated 1,000 respondents in each of the focal and reference groups, with a total of 50 items.

Another approach to DIF assessment that has been discussed both in terms of uniform and nonuniform DIF involves the application of CFA models to item response data (Camilli & Shepard, 1994; Fleishman et al., 2002; Muthen & Lehman, 1985; Navas-Ara & Gomez-Benito, 2002; Oort, 1998). Camilli and Shepard (1994, p. 36) suggested that CFA has potential in DIF detection because it allows for the comparison of group differences on a secondary factor. A number of approaches for using CFA in this regard have been developed, including the use of modification indices, multigroup CFA, and MIMIC models (MacIntosh & Hashim, 2003; Muthen, 1988; Muthen et al., 1991; Oort, 1996, 1998). Wanichtanom (2001) extended work done initially by Oort (1998), using the model modification indices (MIs) that can be obtained from a CFA to identify items that are associated with a group variable, thus indicating the presence of DIF. Muthen and Lehman (1985) and Glockner-Rist and Hoijtnik (2003) both discussed the use of multiple-group CFA as a means for assessing the presence of uniform and nonuniform DIF. In the latter study, the authors describe potential problems with using this approach, particularly the need for large sample sizes, especially if more than two groups are to be examined. In addition, the multiple-groups approach does not allow for the direct modeling of the relationship of group membership and the latent trait.

MIMIC models have been used in practice for checking the presence of uniform DIF for specific items (see, e.g., Gallo, Anthony, & Muthen, 1994), and as discussed earlier, researchers have demonstrated how model parameter estimates from this approach can be converted to IRT-style item difficulty estimates for determining the direction and magnitude of DIF. The MIMIC model in the DIF context takes the following form:

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i, \tag{5}$$

where y_i^*, λ_i, η, and ε_i are as defined in equation (2); $z_k =$ a dummy variable indicating group membership; and $\beta_i =$ slope relating the group variable with the response.

The basic method underlying the assessment of DIF with MIMIC models involves the estimation of both direct and indirect effects for a grouping variable. The indirect effect regresses the latent trait onto the group variable and indicates whether the mean of this latent variable is different across the groups, thus accounting for group differences on the latent trait. The direct effect regresses the item response onto the group variable and indicates whether there is a difference in the response probabilities across the groups. This relationship is the actual assessment of uniform DIF, after controlling for differences in the mean of the latent trait for the groups.

Fleishman et al. (2002) discussed the possibility of using latent variable approaches such as the MIMIC model to help control for the effect of DIF in understanding the performance of an entire test or instrument. In addition, with this approach, it is possible to check for the presence of DIF for more than two groups, and multiple background variables (including both categorical and continuous) can be included in the analysis (Glockner-Rist & Hoitjink, 2003; Muthen, 1988). Muthen (1988) also noted that the MIMIC model allows not only for the assessment of DIF but also a more complete examination of the relationship between background variables and the latent trait. All of these advantages make the MIMIC model worth investigating as a tool for psychometricians interested in the issue of DIF.

There has not been a great deal of Monte Carlo simulation research done assessing the effectiveness of the MIMIC model in detecting DIF under a variety of conditions. Oort (1998) found

that a CFA approach for modeling dichotomous items worked reasonably well at detecting DIF but that the rate of false positives ranged between 0.15 and 0.20. Wanichtanom (2001), using 25 replications in a simulation study, reported a detection rate of 0.98 for uniform DIF when the difference between the IRFs was 0.5, considered moderate DIF (Oort, 1998), which is comparable to his result for the IRT LR method reported above. Note that in both Wanichtanom and Oort, the latent trait was simulated from a standard normal ($\mu = 0$, $\sigma = 1$) distribution. This outcome for the MIMIC model was consistent regardless of the values of the discrimination and difficulty parameters used to generate the data. A third study reported a power value of 0.71 and a Type I error rate of 0.36 for a simulated exam containing 25 items, 10 of which exhibited DIF, with 1,000 examinees and anchor item contamination (Navas-Ara & Gomez-Benito, 2002). The level of DIF was set at a difference of 0.75 between the reference and focal group difficulty indices, and three replications were simulated.

A review of the literature found a number of recent papers that report using the MIMIC model to assess DIF for actual data sets (see, e.g., Fleishman et al., 2002; Glockner-Rist & Hoitjink, 2003; Immekus, Maller, & French, 2003; Levine et al., 2003). Given this evidence of increasing interest in this approach to testing for DIF, as well as the relative paucity of extensive simulation research assessing its effectiveness, it appears some effort should be made to determine when the MIMIC model works well in DIF assessment and when it does not. It is, therefore, the intent of this article to use Monte Carlo methods to study the performance of the MIMIC model in identifying DIF under a variety of conditions and to compare this performance with other well-known methods of DIF detection.

Often, a part of the DIF detection process includes what is known as *item purification*. This term refers to the process of identifying items that display DIF and removing them from the set of items used to estimate the underlying latent trait for each individual, often referred to as the matching or anchor items. This recalculated estimate is then used as the matching criteria for examinees in testing subsequent items for DIF. It has been shown that when test purification is not done, the ability of statistical procedures to detect DIF may be compromised (Kim & Cohen, 1992). Methods of purification for the Mantel-Haenszel test, SIBTEST, and the IRT LR method involve an initial assessment of DIF and then a reanalysis of items using only those not identified by the initial DIF results in determining the matching criteria. In the context of the MIMIC model, purification occurs in a stepwise fashion. In the initial step, items are placed in descending order by level of DIF, as determined in a preliminary analysis. The estimation of the latent trait is then done, excluding the item with the largest level of DIF, as identified in this initial step, and DIF indices are recalculated for all of the remaining items. This technique is repeated in subsequent steps until none of the remaining items are identified as exhibiting DIF, at which point the analysis is completed (Navas-Ara & Gomez-Benito, 2002). For the purposes of this study, test purification was not conducted because the goal of the research is to assess the performance of these methods in a worst-case scenario. This would be analogous to the initial step in a DIF analysis, when nothing is known about the presence of DIF among the items.

Method

The current study seeks to expand on the work described above, particularly in terms of comparing the ability of the MIMIC approach to detect DIF (power) and the rate of incorrect DIF identification (Type I error) with more established methods, including the MH statistic, SIBTEST, and IRT LR. Previous simulation studies of the MIMIC model for DIF detection have

been fairly limited in terms of the parameters manipulated, leaving questions as to its performance under a variety of conditions. Therefore, this study varies several factors that have not been examined in the MIMIC context previously. The data are simulated using IRT-Lab (Penfield, 2003) for 2PL and 3PL models, with item parameters based on the SAT verbal items used by Donoghue and Allen (1993). The base item parameters used in this study appear in Table 1. The values of the latent traits for respondents are generated using a standard normal (0,1) distribution. To assess the performance of each method of DIF detection when the means of the latent trait differ between the reference and the focal groups, two conditions are simulated. In the first, the latent traits are both generated with means of 0, whereas in the second condition, the reference group has a mean of 1 and the focal group has a mean of 0. Two focal group sizes are used, 100 and 500, and the reference group has 500 respondents in all conditions. Two exam lengths are used: 20 and 50 items. The amount of DIF contamination present in the set of anchor items is also manipulated, with one condition having no contamination and the other 15% contamination. Navas-Ara and Gomez-Benito (2002) argue that this is a crucial issue in DIF detection because if items exhibiting DIF are present in the anchor set, the estimate of theta will be biased. Thissen et al. (1988) recommend that practitioners using the IRT LR statistic purify (remove contaminated anchor items) the anchor set with the MH test prior to the detection of DIF for an item. Therefore, this study addresses the impact of item contamination on all four methods in the case in which screening of items for DIF has not been done. This represents the case in which an initial examination of the data must be done to identify potential DIF items. It allows for the determination of DIF detection performance in a worst-case scenario. All of the study conditions are crossed with one another, and for each of the combinations, 500 replications are generated. The amount of DIF (separation between the focal and reference groups' IRFs) present in the target item is either 0 (to assess the Type I error rate) or 0.6 for power. This latter value is comparable to those used in previous research, including 0.4 (Wang & Yeh, 2003); 0.75 (Navas-Ara & Gomez-Benito, 2002); 0.2, 0.5, and 0.8 (Wanichtanom, 2001); 0.2, 0.4, 0.6, and 0.8 (Rogers & Swaminathan, 1993); and 0.48 and 0.64 (Swaminathan & Rogers, 1990). All cases of DIF favor the reference group.

Table 1

Item Parameter Values Used in Generation of Simulated Data

Item	*a*	*b*	*c*
1	1.1	−0.7	0.20
2	0.7	−0.6	0.20
3	0.9	−0.4	0.20
4	1.4	0.1	0.20
5	0.9	0.9	0.16
6	1.2	0.7	0.12
7	0.9	0.3	0.20
8	0.4	0.8	0.20
9	1.6	1.1	0.06
10	2.0	1.1	0.05
11	0.9	−1.5	0.20
12	1.4	−0.4	0.20
13	1.6	−0.1	0.16

(continued)

Table 1
(continued)

Item	*a*	*b*	*c*
14	1.2	0.5	0.20
15	1.2	1.4	0.11
16	1.8	1.4	0.12
17	2.0	1.6	0.16
18	1.0	1.6	0.13
19	1.5	1.7	0.09
20	1.2	1.6	0.09
21	0.7	−0.5	0.20
22	1.2	−0.3	0.20
23	0.9	0.2	0.20
24	0.7	−0.4	0.20
25	0.6	0.2	0.20
26	1.0	0.7	0.15
27	0.6	1.2	0.12
28	1.6	1.1	0.12
29	1.1	2.0	0.06
30	1.1	2.4	0.09
31	2.0	1.4	0.11
32	1.7	1.3	0.17
33	0.9	1.0	0.15
34	0.5	0.4	0.20
35	0.5	−0.6	0.20
36	0.9	1.6	0.11
37	1.3	0.4	0.18
38	1.3	1.4	0.06
39	1.1	1.2	0.05
40	1.2	1.1	0.05
41	0.9	0.8	0.20
42	0.5	0.5	0.20
43	1.3	0.2	0.20
44	1.6	0.9	0.20
45	0.5	−0.8	0.20
46	0.7	0.5	0.20
47	0.4	−0.4	0.20
48	0.8	−0.7	0.20
49	1.0	1.1	0.13
50	0.6	0.2	0.20

The MIMIC model is estimated using the software Mplus (Muthen & Muthen, 2003), which uses robust weighted least squares to obtain parameter estimates (Muthen & Satorra, 1995). To use the MIMIC method for assessing DIF, a model relating a single factor to each item is estimated, along with a direct effect from the group variable to both the factor and the item being assessed. DIF is determined to be present when the direct relationship between the group variable and the item in question is statistically significant. The Mplus program used to estimate the MIMIC model appears in the appendix.

Results

Type I Error Rate

The Type I error rates for each DIF identification technique, by all combinations of the factors included in the study, appear in Table 2. The nominal Type I error rate for this study is 0.05.

Descriptively, it appears that overall, the Type I error rate of the MH procedure is the lowest of all four methods studied here, in both the 2PL and 3PL cases. Indeed, when there is no contamination of the anchor items, the MH Type I error rate for 3PL-generated data is always below 0.05, except for 20 items, 500 individuals in the focal group, and a difference in the group abilities of 1, whereas in the 2PL case, it is always below 0.05. The only exceptions to these low error rates in the 3PL case occur when the contamination of the anchor items is at 15%, the focal group size is 500, and there are no differences in the mean abilities of the two groups. This outcome holds for 20 or 50 items. On the other hand, in the 2PL case, the Type I error rate for MH is always well above 0.05 when the anchor items are contaminated.

SIBTEST has a similar pattern of performance to the MH, although overall, the Type I error rate is a bit higher. It tends to perform less well when there is contamination of the anchor items and the focal group is large. There is one seemingly anomalous result where the Type I error rate exceeds 0.30, which is associated with no anchor item contamination, 20 items, 500 in the focal group, and a difference in the groups' abilities. In the 2PL case with contaminated anchor items, 500 respondents in the focal group, and no difference in the mean abilities between the groups,

Table 2

Type I Error Rates by Percentage of Anchor Items With DIF, Number of Items, Focal Group Size(n), and Difference in Group Mean Thetas (Δ)

DIF%	Number of Items	N	Δ	SIBTEST	MH	MIMIC	IRT LR
0	20	100	0	.058/.048	.042/.024	.054/.234	.054/.038
			1	.068/.096	.038/.046	.050/.380	.048/.038
		500	0	.038/.038	.034/.030	.038/.124	.034/.036
			1	.082/.328	.042/.076	.064/.192	.058/.046
	50	100	0	.054/.064	.032/.028	.045/.038	.054/.050
			1	.058/.076	.048/.048	.055/.094	.052/.062
		500	0	.058/.038	.026/.048	.075/.042	.070/.067
			1	.046/.106	.050/.044	.062/.045	.050/.032
15	20	100	0	.176/.084	.188/.090	.142/.312	.216/.226
			1	.112/.066	.106/.026	.082/.408	.180/.192
		500	0	.362/.164	.446/.298	.224/.440	.608/.462
			1	.116/.138	.218/.068	.100/.234	.244/.294
	50	100	0	.174/.120	.138/.070	.122/.062	.202/.236
			1	.128/.104	.082/.052	.083/.054	.164/.224
		500	0	.404/.258	.376/.278	.134/.094	.448/.472
			1	.216/.084	.242/.090	.075/.073	.366/.266

Note. The first number in each cell is the observed Type I error rate for the two-parameter logistic model, and the second number is the observed Type I error rate for the three-parameter logistic model. DIF% refers to the level of differential item functioning (DIF) contamination in the anchor items. MH = Mantel-Haenszel test; MIMIC = multiple indicators, multiple causes model; IRT LR = item response theory likelihood ratio test.

SIBTEST and MH both have much higher Type I error rates than for other combinations of the conditions. To verify that this anomaly was accurate and not the result of either a problem with the seed used to run SIBTEST or the simulation of the data itself, further analyses were conducted. First, the seed given to the SIBTEST program was changed and the analysis rerun. The results were very similar to that reported in Table 2. In addition, a new set of simulated data was generated, using the same item parameters and conditions as that which created the anomalous result above, and again, the results were very similar to those reported in Table 2. It would appear, then, that this unusual outcome is indeed real.

The Type I error rate for the MIMIC model follows a very different pattern than either SIBTEST or the MH. In the 3PL case, when there are only 20 items, the rate is uniformly much higher than the nominal cutoff of 0.05, never below 0.10, and most often well above 0.20. This inflated error rate appears to be exacerbated by the presence of contamination in the anchor items. On the other hand, when there are 50 items, the error rate is always below 0.10 and generally below 0.05 when there is no anchor item contamination. In contrast to these results, when pseudo-guessing is not present, the MIMIC approach to DIF detection has Type I error rates that are much closer to the nominal 0.05, even when there are 20 items. Indeed, while it has a higher error rate than the other methods examined here for the 3PL and 20-items conditions, it has comparable rates for 20 items and the 2PL model.

The IRT LR technique performs comparably to the MH approach when there is no contamination of the anchor items, particularly for shorter exams, with only two cases of Type I error above 0.05. However, when DIF is present in the anchors, the false detection rate increases markedly, to levels much higher than those of either the MH statistic or SIBTEST. In fact, when contamination is present, the IRT LR statistic has a Type I error rate below 0.20 for only one combination of conditions. In general, the Type I error rate performance of the IRT LR method does not appear to be greatly influenced by the type of model (2PL or 3PL) underlying the data.

To better understand which factors or combinations of factors influence the Type I error rates for each of the four procedures studied here, a variance components analysis is carried out using SPSS, Version 12. Variance components analysis is a statistical methodology, related to analysis of variance (ANOVA), that allows for the identification of the amount of variation in the response variable that is due to one or more independent variables (see, e.g., Glass & Hopkins, 1996). In this context, the response of interest is the Type I error rate for each method of DIF detection, whereas the independent variables include the size of the focal group, the model used to generate the data, the percentage of DIF among the items, the number of items, and the difference between the two groups' thetas. The actual numbers contained in this table reflect the amount of variation in the Type I error rate accounted for by each factor. The results of this analysis appear in Table 3.

The SIBTEST and the MH procedures exhibit somewhat similar results for the variance components analysis. In both cases, the three-way interaction of the focal group size by the percentage of contamination in the anchor items by the difference in abilities between the focal and reference groups is the highest level interaction that contributed the most to variance in the Type I error rate. Table 4 contains the Type I error rates for both SIBTEST and MH by the level of contamination, size of the focal group, and the difference between abilities for the two groups. These results are the same as those presented in Table 2, collapsed across the number of items, which did not appear to be a very important factor in determining the Type I error rate for SIBTEST and MH.

For both DIF identification techniques, when there is no contamination of the anchor items, the Type I error rate increases as the mean abilities of the two groups go from being equal to

Table 3

Results of Variance Components Analysis for Each Four Methods of DIF Detection, Type I Error Rate

Factor	SIBTEST	MH	MIMIC	IRT LR
Focal *N* (*N*)	0.004	0.001	0	0
Items (*I*)	0	0	0.001	0
DIF% (*D*)	0	0.003	0	0.028
Theta (*T*)	0	0	0	0
Model (*M*)	0	0	0	0
*N*I*	0	0	0	0
*N*D*	0	0.004	0.001	0.006
*N*T*	0	0	0.001	0.001
*N*M*	0	0	0	0
*I*D*	0.001	0	0.001	0
*I*T*	0.001	0	0	0
*I*M*	0	0	0.011	0
*D*T*	0.003	0.003	0.002	0.001
*D*M*	0.003	0.003	0	0
*T*M*	0	0	0	0
*N*I*D*	0.001	0	0	0
*N*I*T*	0	0	0	0
*N*I*M*	0	0	0.001	0
*N*D*T*	0.006	0.003	0	0.005
*N*D*M*	0.002	0.001	0.001	0.001
*N*T*M*	0.001	0	0.001	0
*I*D*T*	0.001	0.001	0	0
*I*D*M*	0	0	0	0
*I*T*M*	0.001	0	0	0
*D*T*M*	0	0	0	0
*N*I*D*T*	0	0	0.003	0
*N*I*D*M*	0	0	0	0
*N*I*T*M*	0.002	0	0.001	0
*N*D*T*M*	0	0	0	0
*I*D*T*M*	0	0	0	0.001
*N*I*D*T*M*	0	0	0.001	0.003

Note. DIF = differential item functioning; MH = Mantel-Haenszel test; MIMIC = multiple indicators, multiple causes model; IRT LR = item response theory likelihood ratio test.

differing by 1. This effect is more dramatic when the focal group has 500 respondents versus when it has 100, especially for SIBTEST. On the other hand, when there is 15% anchor item contamination, the pattern for both MH and SIBTEST is reversed, with a lower Type I error rate associated with a difference between the means of the group abilities. These results hold for both the 2PL and 3PL cases. In addition, when contamination of the anchor items is present, the error rate is higher in the 2PL case for both SIBTEST and the MH, which corresponds to the contribution in the variance components analysis of the DIF contamination by model interaction.

The variance components analysis results displayed in Table 3 demonstrate that, by far, the most important factor in accounting for variance of the Type I error rate for the MIMIC model

Table 4

Type I Error Rate for SIBTEST and MH by Percentage Contamination, Focal Group Size, and Difference in Abilities, 2PL/3PL

DIF%	Focal *N*	ThetaΔ	SIBTEST	MH
0	100	0	.056/.056	.037/.026
		1	.063/.086	.043/.047
	500	0	.048/.038	.030/.039
		1	.064/.217	.046/.060
15	100	0	.175/.102	.163/.080
		1	.120/.085	.094/.039
	500	0	.383/.211	.411/.288
		1	.163/.111	.230/.079

Note. 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; DIF = differential item functioning; MH = Mantel-Haenszel test.

method of DIF detection is the interaction between the number of items and the model used to generate the data. Indeed, this term accounts for nearly four times the amount of variation accounted for by the next most important term. Table 5 includes the error rates for all four methods at both 20 and 50 items. In the 3PL case, there is a marked decline in the MIMIC Type I error rate from 20 to 50 items, although it remains above the nominal value of .05. It should be noted that this rate is lower for 50 items than that of the other four methods studied here. Even when no contamination of the anchor items is present, which is the best possible scenario for the other three methods, the performance of the MIMIC technique is very comparable to the other methods in terms of Type I error. In the 2PL case, however, the MIMIC approach to DIF detection had comparable or better error rates than the other approaches, regardless of the number of items. For example, when the anchor items are contaminated and no pseudo-guessing is

Table 5

Type I Error Rate by Number of Items, 2PL/3PL

	Overall			
Number of Items	SIBTEST	MH	MIMIC	IRT LR
20	.13/.12	.14/.08	.09/.29	.18/.17
50	.14/.11	.12/.08	.08/.06	.18/.18
		0% Contamination		
20	.06/.13	.04/.04	.05/.23	.05/.04
50	.05/.07	.04/.04	.06/.05	.06/.05
		15% Contamination		
20	.19/.11	.24/.12	.14/.35	.31/.29
50	.23/.14	.21/.12	.10/.07	.29/.30

Note. 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; MH = Mantel-Haenszel test; MIMIC = multiple indicators, multiple causes model; IRT LR = item response theory likelihood ratio test.

present, MIMIC has a lower Type I error rate than do the other approaches, but when there is no contamination, the MIMIC error rate is virtually identical to that of the other approaches for both 20 and 50 items.

The variance components analysis found that the Type I error rate of the IRT LR procedure is also strongly influenced by a single factor—in this case, the level of contamination in the anchor items. With no contamination, the IRT LR Type I error rate is 0.049, whereas when the anchor items are contaminated, the error rate is 0.209. In short, when there is no contamination of the anchor items, the IRT LR approach to DIF detection maintains the nominal Type I error rate of 0.05. However, when the anchors are contaminated at 15%, the rate of misidentification jumps to nearly 0.30, which represents more than a sixfold increase in the error rate. The underlying model appears to have little to no impact on the Type I error rate of this method.

Power

Because of the large number of conditions in which power cannot be reported due to inflated Type I error rates, variance components analysis is not used with the power results. Descriptive statistics are used, however, to further investigate which factors appear to affect power. The power (correct identification of DIF) for each procedure by contamination, number of respondents, number of items, ability differences, and underlying model appears in Table 6. Note that in cases where the Type I error rate is above 0.10, no power results are reported. In such instances, the standard definition of power at the nominal level of alpha is not meaningful.

Table 6

Power by Percentage DIF Contamination, Number of Items, Focal Group Size (*N*), and Difference in Group Mean Thetas (Δ), 2PL/3PL

Number of Items	*N*	Δ	DIF	SIBTEST	MH	MIMIC	IRT LR
20	100	0	0	.984/.884	.982/.924	1/***	.994/.950
			15	***/.822	***/.842	***/***	***/***
		1	0	.958/.928	.980/.930	.984/***	.990/.844
			15	***/.768	***/.714	.940/***	***/***
	500	0	0	1/1	1/***	1/***	1/1
			15	***/***	***/1	***/***	***/***
		1	0	1/***	1/.998	1/***	1/.986
			15	***/***	***/1	1/***	***/***
50	100	0	0	.972/***	.996/.932	.996/.958	.996/.956
			15	***/***	***/.764	***/.882	***/***
		1	0	.864/.900	.942/.942	.988/.984	.994/.876
			15	.799/.724	***/.738	.929/.924	***/***
	500	0	0	1/1	1/1	1/1	.998/.998
			15	***/***	***/.996	***/.890	***/***
		1	0	1/***	1/***	1/1	.998/.994
			15	***/.996	***/.994	.836/.770	***/***

Note. 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; DIF = differential item functioning; MH = Mantel-Haenszel test; MIMIC = multiple indicators, multiple causes model; IRT LR = item response theory likelihood ratio test.

Both SIBTEST and the MH statistic generally have greater power when the focal group is 500 as opposed to 100, although in a number of such cases, the former method also has an inflated Type I error rate. Furthermore, for the MH statistic in the 3PL case, contamination of the anchor items appears to lead to a diminution in power when the focal group is small but has little impact for the larger sample size. For 2PL data, the Type I error rate of MH when there is contamination is uniformly above 0.10 except for 50 items, 100 focal group respondents, and a difference between the two groups' mean abilities, so that power is not reported in other cases. Neither the number of items nor differences in mean group abilities appears to have an impact on the power of SIBTEST or the MH statistic, for those power results that can be reported. In general, the level of power for both approaches is very comparable across all factors included in this study.

Power for the MIMIC model with 3PL data is only reported in the 50-item case because for all conditions in which the number of items is 20, the Type I error rate is greatly inflated. When the exam contains 50 items, the power of the MIMIC model in detecting DIF appears to be largely immune from the influences of any of the factors included in this study. Perhaps the most marked change occurs when the level of anchor item contamination increases from 0% to 15% with 50 items and 500 respondents. Otherwise, it appears that MIMIC's power is largely impervious to the manipulated factors. With respect to 2PL data, the power of MIMIC appears to be similar to that of the other approaches, both for 20 and 50 items. In general, the ability of the MIMIC approach to correctly identify DIF is at least as high as that of both SIBTEST and the MH statistic and, in some cases, higher, when no pseudo-guessing is present.

The power for the IRT LR statistic is presented only for no item contamination in both the 2PL and 3PL conditions because in all cases in which the level of contamination is 15%, the Type I error rate is inflated above 0.10, making the power results meaningless. In situations where power is meaningful, it appears that IRT LR is comparable to the other methods studied here, except when there are 100 respondents in the focal group and the two groups differ in terms of theta, when its power is somewhat lower. Unlike the other three methods, the power of IRT LR does not seem to be greatly influenced by the underlying model.

In summary, it appears that for both SIBTEST and the MH, the size of the focal group is the most important factor to influence power. In addition, it appears that both procedures have somewhat higher power in the 2PL rather than 3PL situation. In contrast, the MIMIC model does not experience such a severe reduction in power when the anchor items are contaminated. Indeed, this approach has comparable or higher power than any of the others except for the large focal group size with contaminated anchor items and a difference in the means of the two groups. As with SIBTEST and MH, power is higher when pseudo-guessing is not present in the data. Finally, the power of the IRT LR statistic is very similar to that of the other methods studied here, when there is no anchor item contamination. If there is DIF present in the anchor items, the inflated Type I error rate of this approach makes interpretation of power impossible.

Conclusions

Type I Error Rate

The focus of this article is on the performance of the MIMIC model in detecting DIF. To that end, both the Type I error rate and power for this approach have been compared to those of three established methods: SIBTEST, the MH statistic, and the IRT LR statistic. The results of this study appear to support the utility of the MIMIC model in some circumstances and not in others.

Specifically, when an exam is relatively short (20 items) and the model underlying the data is 3PL, the MIMIC approach will falsely indicate the presence of DIF at a rate well above the nominal Type I error rate of 0.05. Indeed, none of the other procedures examined here had Type I error rates nearly as high as those of MIMIC for 20 items. On the other hand, for longer exams or when pseudo-guessing is not present, the MIMIC model is very comparable to the more established methods and indeed outperforms them under several conditions. For example, for the longer tests, MIMIC does not have the inflation of the Type I error rate associated with the contamination of anchor items that is evident for the other three techniques, particularly the IRT LR and, to a lesser extent, SIBTEST. In short, it is possible to conclude that when the exam contains 50 items and/or in the 2PL case, the MIMIC model is very competitive with the other approaches in terms of Type I error rate, regardless of the focal group size, differences in mean group abilities, and level of contamination of the items.

These Type I error results for the MIMIC approach are similar to some that have already been reported. Navas-Ara and Gomez-Benito (2002) found that with 25 items and 40% of the anchor items contaminated with DIF, the Type I error rate was 0.36, which is similar to what is reported in this article with 20 items. However, whereas Navas-Ara and Gomez-Benito attributed this result to anchor item contamination, the results of this study support the conclusion that the inflated error rate might be due more to the number of items. Other researchers who have studied the performance of the MIMIC model in DIF analysis have focused their attention on the rate of correct detection rather than the Type I error rate. Wanichtanom (2001) found the power of the MIMIC model to be very high, with values approaching 1.0. On the other hand, Navas-Ara and Gomez-Benito report much lower power (0.71) when the item contamination is 40%, which is much higher than that used in the current study.

With respect to the MH and SIBTEST statistics, incorrect identification of DIF is generally lower than that of the MIMIC approach for shorter exams and 3PL data, and typically, the Type I error rate is a bit higher for SIBTEST than the MH statistic. The error rate for both is most influenced by the interaction of the focal group size, level of contamination, and difference between the mean group abilities. In the case of the MH, the largest error rate is seen with a larger focal group, high item contamination, and no difference between the two group means. The error rate drops dramatically when the two groups differ on average ability. The Type I error rate for SIBTEST exhibits a similar pattern to the MH, with one exception. Although there is a lower error rate when the group mean thetas differ with high item contamination and a large focal group, the rate is higher when there is no contamination, the focal group is large, and the group means differ. This result was not seen for the MH statistic. Otherwise, the two statistics perform similarly with respect to the Type I error rate, with SIBTEST having a somewhat higher rate.

The Type I error rate for the IRT LR statistic is by far most influenced by the level of contamination in the anchor items. When there is no contamination, the Type I error rate is above 0.05 in only two of the conditions used in this study. On the other hand, when there is 15% contamination of the anchor items, the error rate is well above 0.05 in every condition, regardless of the size of the sample, the number of items, or the level of difference between mean group abilities. This would appear to support the suggestion by Thissen et al. (1988) that prior to using the LR to identify the presence of DIF for a particular item, an exam should be screened and contaminated items removed from the anchor set.

Power

Because the Type I error rates for the methods studied here are not all below 0.10, interpretations of power cannot be made for all conditions examined in this study. With this caveat in

mind, the following conclusions can be made regarding the relative power of these four approaches to DIF detection. First, the MIMIC model is as powerful at detecting DIF as the other three approaches across conditions used in this study when the test contains 50 items or when the data are generated with a 2PL model. Indeed, in some instances, the MIMIC approach actually has higher power. For example, SIBTEST and the MH statistic both have a notable decline in power when the focal group has 100 respondents and the level of contamination is 15%. In contrast, although contamination does lower the power of the MIMIC model somewhat, this effect is less severe than for the other approaches with one exception: when the focal group is large and there are differences in the mean abilities of the groups. This result is especially interesting because the other two methods have very high power in this case. As discussed previously, the power of IRT LR is not meaningful when the anchor items are contaminated because its Type I error rate is always inflated.

The ability of SIBTEST and the MH statistic to correctly identify instances of DIF is higher for a larger focal group. This result is more interpretable for MH than SIBTEST because the latter tends to suffer from inflated Type I error when the focal group has 500 respondents, making power unreportable in some cases. In addition, when the focal group contains 100 examinees, there is the aforementioned diminution of power for both statistics when the anchor items have 15% DIF contamination, with no such effect being evident for the 500-examinee condition. In all cases where no anchor item contamination is present, the IRT LR method exhibited comparable power to the other methods examined here. Finally, power of all four procedures is higher in the 2PL versus the 3PL case, although in some instances, those differences are very small.

Summary

The results discussed above suggest that the MIMIC model approach to DIF detection has an inflated Type I error rate for shorter exams when pseudo-guessing is present, but for longer exams or when there is no pseudo-guessing, it is a viable option to more traditional techniques. This method has the smallest overall Type I error rate and comparable or higher power for the 50-item condition. Furthermore, whereas the performances of the other three techniques are adversely affected by contamination of the anchor items, the MIMIC model demonstrates only a small increase in the Type I error rate for greater contamination and a small decrease in power compared to the others. Thus, in cases where it is feared that a large number of the items may exhibit DIF and screening of items needs to be done, the MIMIC model might be preferable to the others because it does not seem to be so easily influenced by the contamination, at least up to 15% of the anchor items. In addition, the power of the other methods examined here appears to be somewhat more influenced by the size of the focal group than does that of the MIMIC model, especially when there is contamination of the anchor items. At the same time, it must be noted that for the shorter exam condition with a 3PL model, the MIMIC approach to DIF detection does not compare favorably to the others because the Type I error rate is so high.

In terms of the other methods of DIF detection studied here, contamination of the anchor items appears to be an important issue, particularly for the IRT LR statistic. Given an uncontaminated anchor set, the MH and IRT LR seem to be optimal for finding uniform DIF, especially for shorter exams. SIBTEST also performs very well, with just slightly higher Type I error rates than the other two and comparable power.

In short, then, it appears that a major recommendation to come out of this study is that when an exam is of 50 items in length, regardless of the presence or absence of pseudo-guessing, the

MIMIC model should be considered for DIF detection. In addition, if it is known that the data come from a 2PL framework, the MIMIC model is viable for DIF detection even with as few as 20 items. It might be especially useful in cases where multiple exogenous variables or multiple groups need to be considered, as described by Fleishman et al. (2002). At the same time, it should be recognized that further research is required to investigate the performance of the MIMIC approach when multiple exogenous variables are being considered. Although examples exist in the literature demonstrating the use of the MIMIC model in this context with actual data (Fleishman et al., 2002; Glockner-Rist & Hoitjink, 2003), there do not appear to be any simulation studies examining its performance under controlled conditions. Further research would also be useful in identifying the test length for which the MIMIC model performs comparably to the other methods in the 3PL case. Although it is clear that this is true for 50 items and not for 20, further investigation with test lengths between these values would help practitioners in selecting an appropriate method to use.

Appendix

```
    Title:  Test of differential item functioning
     Data:  File is groups.txt; format is i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x,
            i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x,
            i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x,
            i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x,
            i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x,
            i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1, 1x, i1
 Variable:  Names are group i1-i50; categorical are i1-i50; usevariables are group i1-i50
 Analysis:  estimator = wls
    Model:  f1 by i1-i50; f1 on group; i1 on group
 Savedata:  File (results) is results.out
```

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

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

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

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics*, *18*, 131-154.

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*(5), S275-S283.

Gallo, J. J., Anthony, J. C., & Muthen, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences*, *49*, P251-P264.

Glass, G. V, & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Needham Heights, MA: Simon & Schuster.

Glockner-Rist, A., & Hoitjink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*(4), 544-565.

Immekus, J. C., Maller, S. J., & French, B. F. (2003, April). *TIMSS 1999 factor invariance*

across U. S. samples of males and females. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, *23*(4), 291-322.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 551-566.

Levine, D. W., Bowen, D. J., Kaplan, R. M., Kripke, D. F., Naughton, M. J., & Shumaker, S. A. (2003). Factor structure and measurement invariance of the Women's Health Initiative Insomnia Rating Scale. *Psychological Assessment*, *15*(2), 123-136.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372-379.

McDonald, R. P. (1967). *Nonlinear factor analysis*. Psychometric Monographs, No. 15.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

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

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

Muthen, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*, 133-142.

Muthen, B. O., & Muthen, L. K. (2003). Mplus software, version 2 [Computer software]. Los Angeles: Author.

Muthen, B. O., & Satorra, A. (1995). Technical aspects of Muthen's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, *60*(4), 489-503.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328.

Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment*, *18*(1), 9-15.

Oort, F. J. (1996). *Using restricted factor analysis in test construction*. Unpublished doctoral dissertation, University of Amsterdam, Amsterdam, the Netherlands.

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

Penfield, R. D. (2003). IRT-Lab: Software for research and pedagogy in item response theory. *Applied Psychological Measurement*, *27*(4), 301-302.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*(2), 215-230.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Chapel Hill: 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., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer &

H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

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

Wanichtanom, R. (2001). *Methods of detecting differential item functioning: A comparison of item response theory and confirmatory factor analysis*. Unpublished doctoral dissertation.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, *15*, 185-198.

Acknowledgment

The author would like to acknowledge the efforts of Dr. Brian Habing, who read the manuscript during its development and provided valuable insights and suggestions.

Author's Address

Address correspondence to Holmes Finch, Department of Educational Psychology, Ball State University, Muncie, IN 47304; phone: (765) 285-3668; e-mail: whfinch@bsu.edu.