한국심리학회지 : 일반
Korean Journal of Psychology: General
2011, Vol. 30, No. 4, 1083-1110

Empirical Comparisons of Analytic Strategies for MIMIC DIF Analysis: A Potential Solution for Biased Anchor Set

Jaehoon Lee

University of Kansas

The purpose of this Monte Carlo study was to evaluate the performance of the multiple indicators and multiple causes (MIMIC) confirmatory factor analysis (CFA) for detecting differential item functioning (DIF). Specifically, this study compared different application strategies including two conventional testing approaches (forward-inclusion, backward-elimination) and five test statistic values (uncorrected or Bonferroni-corrected LR, △CFI of 0.01 or 0.002, △SRMR of 0.005) across conditions of different item type, test length, sample size, impact, and DIF type and DIF size in a target item and an anchor set. In addition, the author proposed an alternative testing approach (effects-coded backward-elimination) as a potential solution for arbitrary choice of a DIF-free anchor set. Simulation results indicated that when an anchor set was truly biased, only the proposed approach performed adequately under several conditions. False positive rates were controlled at the nominal alpha level (with Bonferroni-corrected LR) or slightly inflated (with uncorrected LR) as the DIF contamination rate in a scale decreased.

Key words : MIMIC, DIF, biased anchor set, testing approach.

† 교신저자 : Jaehoon Lee, Center for Research Methods and Data Analysis, University of Kansas, 1415 Jayhawk Blvd. Room 470, Lawrence, KS 66045-7556, USA.
E-mail : jaehoon@ku.edu

It has been a common practice for applied psychologists to recognize measurement equivalence (ME) for fair use of a test[1] (see AREA, APA, NCME, 1999). Moreover, many researchers have emphasized evaluating ME as a prerequisite for meaningful group comparison (e.g., Drasgow, 1984; Little, 1997; Raju, Laffitte, & Bryne, 2002; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). To the extent that a scale does not hold ME, any interpretations of group differences, as indicated by differences in scores at the item level, at the scale level, or at both, are necessarily open to question-observed score differences may represent true differences in underlying (latent) trait across groups, measurement artifacts related to the instrument, or both (Byrne & Stewart, 2006; Drasgow, 1987; Lee, Little, & Preacher, 2010; Stark, Chernyshenko, & Drasgow, 2004). Thus, such practice of assessing ME requires a methodology that can distinguish measurement artifacts, or lack of ME, from true differences in the relevant construct (Stark, Chernyshenko, & Drasgow, 2006).

Among various techniques currently available for assessing ME, those most commonly used are based on either item response theory (IRT) or confirmatory factor analysis (CFA), a special case of structural equation modeling (SEM) (Teresi, 2006). Although the concept of *item-level* measurement (non)equivalence (i.e., differential item functioning; DIF) was originated in the IRT literature (Camilli & Shepard, 1994), recent demonstrations regarding the link between IRT and CFA allow researchers to address the problems of DIF within the CFA framework (see Kamata & Bauer, 2008; MacIntosh & Hashim, 2003; Muthén, Kao, & Burstein, 1991). The CFA-based DIF analysis employs either mean and covariance structure model (MACS; Sörbom, 1974) or multiple causes multiple indicators model (MIMIC; Jöreskog & Goldberger, 1975). Using the MIMIC technique, for example, researchers have successfully detected DIF in questionnaires of mental health (e.g., Jones, 2006; Woods, Oltmanns, & Turkheimer, 2009). Moreover, a few simulation studies have supported the utility of this method-for instance, it provides reasonable control for Type I error and adequate power, unless smaller group size is less than 100 (Woods, 2009a); three-parameter logistic IRT model underlies the responses on a short scale (Finch, 2005); or the number of DIF items are too large relative to the number of items in a scale (Finch, 2005; Navas-Ara & Gómez-Benito, 2002). The present study also focuses on evaluating the accuracy of the MIMIC DIF analysis under various conditions, especially with regard to its implementation strategy.

Although previous simulation studies have provided some valuable insights and practical implications, there are some analytic issues that may arise when researchers conduct the MIMIC DIF analysis (see Lee, 2009). Also, it is

[1] The terms test and scale are used synonymously in this article.

premature to strongly advocate a particular procedure because there have been little or no direct comparisons of different application strategies[2]. Accordingly, this article presents a simulation study that compares two common testing approaches and an alternative approach in terms of efficiency. The latter approach has been designed as a potential solution for arbitrarily choosing an unbiased anchor set. This study also examined the use of different test statistics, along with other factors known to impact the efficiency of the MIMIC DIF analysis. The organization of this article is as follows. The next sections demonstrate the MIMIC models and discuss some methodological issues in the context of DIF analysis. A Monte Carlo study and simulation results are presented in the following sections. In the final section, the author discusses study findings and implications as well as limitations and directions for future research. The present study would contribute to the literature by cautioning researchers and practitioners against the use of an innocuously chosen analytic strategy when conducting the MIMIC DIF analysis.

MIMIC Terminology and Specification

To illustrate the MIMIC models, this section starts with the MACS model. In case of a single latent trait, the MACS model can be written as

$$y_i^* = \tau_i + \lambda_i \eta + \epsilon_i, \quad (1)$$

where y_i^* is the latent item responses ($i = 1, \cdots, p$) (when $y_i^* > \kappa_i$, an observed item response $y_i = 1$; κ_i is the item threshold), τ_i is the item intercepts, λ_i is the item loadings, η is the latent trait, and ϵ_i is the unique factor scores that are assumed to have a normal distribution. The MIMIC model is a simple extension of the MACS model-it incorporates the impacts of (observed) covariates on the trait (Jöreskog & Goldberger, 1975; standard MIMIC model, hereafter). Muthén and colleagues further extended the standard MIMIC model such that the covariates also influence the responses (Gallo, Anthony, & Muthén, 1994; Muthén, 1988; MIMIC-DIF model, hereafter). The MIMIC-DIF model can be written as

$$y_i^* = \tau_i + \lambda_i \eta + \sum_{i=1}^{q} \beta_{ij} x_j + \epsilon_i, \quad (2)$$

where x_j is the covariates ($j = 1, \cdots, q$) and β_{ij} is the regression coefficients that correspond to the impacts of the covariates on the responses. The trait score η can be obtained by

$$\eta = \alpha + \Gamma x + \zeta, \quad (3)$$

[2] In fact, different analytic strategies including testing approach, test statistic, and scaling method have been empirically compared in case of the MACS DIF analysis (e.g., Start et al., 2006; Lee, 2009).

where x is a $q \times 1$ vector of the covariates, Γ is an $1 \times q$ vector of the regression coefficients γ_j that correspond to the impacts of the covariates on the trait, and ζ is a scalar of the disturbance that is assumed to have a multivariate normal distribution with mean of 0 and variance ψ. The two error terms ϵ_i and ζ are assumed to be independent of each other and η.

The regression coefficients γ are termed *indirect effects* as they represent the impacts on the responses through the trait ($\lambda_i \times \gamma_j$). Given a grouping covariate, the indirect effects account for group differences in trait mean. The regression coefficients β are termed *direct effects* as they represent the influences on the responses, unmediated by the trait (Dorans & Holland, 1993; Jones, 2006). The direct effects capture group differences in responses after controlling for the differences in trait mean across groups, which is the definition of DIF (Fleishman, 2005). Thus, an item is considered as having DIF when a corresponding direct effect is statistically significant (Jones, 2006).

As observed in Equation 2, the MIMIC models presume identical trait variances and, more importantly, equal loadings across groups. Consequently, an apparent limitation of the MIMIC technique is that there is no test for non-uniform DIF. However, Woods and Grimm (2011) recently demonstrated the use of MIMIC models for testing both uniform and non-uniform DIF with categorical covariates. They expanded the MIMIC-DIF model by incorporating latent interactions between a latent trait and categorical covariates to identify non-uniform DIF items in a scale. Because the MIMIC method for detecting non-uniform DIF is beyond the scope of the present study, the Woods and Grimm 's model is not discussed in this article.

Some Analytic Issues

This section discusses some methodological issues that may arise when researchers conduct the MIMIC DIF analysis. The related testing approaches and test statistics are examined in the current simulation study.

Scaling

In any CFA model, the scale for a latent construct (trait) needs to be identified to obtain a unique solution for every parameter (Bollen, 1989). In a simple MACS model, given three or more items[3], scaling is often achieved by fixing one of the loadings and a corresponding intercept (e.g., to 1 and 0, respectively) (*marker-variable* method); fixing the latent variance and mean (e.g., to 1 and 0, respectively) (*fixed-factor* method); or constraining

[3] Fewer than three items per trait would result in an under-identification problem, increasing likelihood of obtaining an infeasible solution (Bollen, 1989). Thus, discussions focus on the cases of three or more items for each trait.

the loadings and intercepts to average particular values (e.g., to 1 and 0, respectively) (*effects-coded* method) (Little, Slegers, & Card, 2006). The same scaling methods are used to set the scale of a trait defined in the MIMIC model (scaling part i). Since the MIMIC models involve estimating indirect and direct effects, scaling also should take into account these additional parameters (scaling part ii). This latter part of scaling can be done by further fixing the direct effects for a set of items (at least one item; anchor) to 0 (e.g., Finch, 2005); or imposing a constraint on the direct effects. Although any combinations of the scaling methods between the scaling parts i and ii are simple reparameterizations of one another and, as a result, they provide identical model fit -i.e., by no means change the DIF test results, the scaling part ii is closely related to the choice of an anchor set and testing approach.

Biased Anchor Set

Fixing the direct effects for a set of anchor items is essentially the same as assuming that the anchor set is truly free from DIF. However, if an anchor set is contaminated by DIF, the direct effects for other (non-anchor) items may be erroneously estimated (Cheung & Rensvold, 1999; Millsap, 2005). Indeed, both Finch (2005) and Navas-Ara and Gómez-Benito (2002) showed that the accuracy of the MIMIC DIF analysis is adversely influenced by the presence of DIF in an anchor set. To rule out this possibility, various empirical solutions have been proposed in the literature (e.g., Fleishman, Spector, & Altman, 2002; Christensen, MacKinnon, Korten, & Jorm, 2001; Mackinnon, Jorm, Christensen, Korten, Jacomb, & Rodgers, 1999, Woods, 2009b). Although such solutions could provide an unbiased anchor set much of the time, they increase necessarily the number of nested-model comparisons, which inflates Type I error and more severely if no item appears to be DIF-free; or involve the risk of capitalizing on chance by the use of data-driven modification index (MI).

Testing Approach

Two testing approaches are often used for the MIMIC DIF analysis. First, *forward-inclusion* approach tests DIF one item at a time, assuming that all other items in the scale are DIF-free anchor items (e.g., Christensen et al., 1999; Finch, 2005; Muthén & Asparouhov, 2002). This approach starts with a baseline model where no direct effects have been specified (i.e., standard MIMIC model). Once the baseline model is fitted to the data, model fit is compared against each of p nested models (where p = number of items in the scale), where a direct effect is added or freely estimated for only one item at a time (see Figure 1A). An item is considered as having DIF if the inclusion of a corresponding direct effect improves model fit significantly.

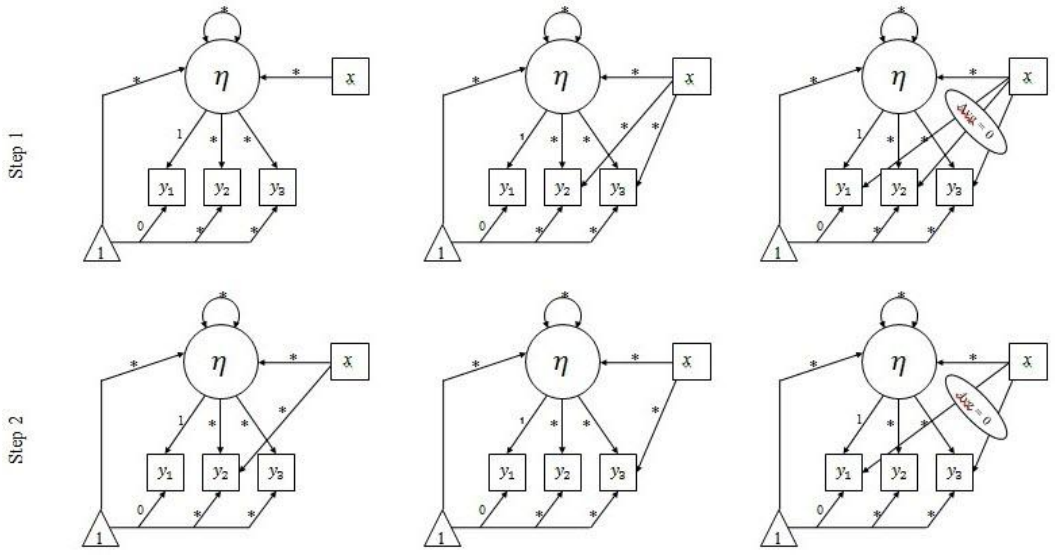
Second, *backward-elimination* approach tests DIF

one item at a time, assuming that, unlike the forward-inclusion approach, all other items are not necessarily free from DIF (e.g., Fleishman et al., 2002; Woods, 2009a; Woods, Oltmanns, & Turkheimer, 2008). Accordingly, the baseline model includes all possible direct effects, except for an anchor set (at least one item) needed for scale setting. The fit of this baseline model is compared against each of the nested models, where the direct effect for an item being tested (target item) is eliminated or fixed to 0 (see Figure 1B). Uniform DIF is indicative if eliminating a direct effect worsens model fit significantly.

By employing the effects-coding schema, a variant form of the backward-elimination approach can be constructed. This alternative approach, named *effects-coded backward-elimination*, constrains all possible direct effects to average 0 in the baseline model, assuming that all items in the scale are not necessarily DIF-free. It can be written as

$$\sum_{i=1}^{p}\beta_{ij}=0 \text{ for each of the covariates } x_j, \tag{7}$$

where $i = 1$ to p refers to summation across the set of p unique items for a given latent trait. The baseline model fit is compared against each of the nested models, where the direct effect for a target item is fixed to 0 but the



A. Forward-Inclusion B. Backward-Elimination C. Effects-Coded Backward-Elimination

Figure 1. MIMIC Baseline and Constrained/Augmented Models

Note. For simplicity, these models include only three items and a single covariate and the unique factor variances are omitted in this figure. The marker-variable method is chosen for the scaling part i.

constraint still holds for other items in the scale (see Figure 1C).

These three different testing approaches would likely to yield different results of DIF analysis because this post-hoc analysis relies on the examination of individual parameters. More important, from statistical standpoint, the forward-inclusion approach is not theoretically suitable. In the likely cases where a scale includes one or more DIF items, the baseline model for the forward-inclusion approach (standard MIMIC model) may not fit adequately because this model assumes no DIF in the scale (Stark et al., 2006; see Maydeu-Olivares & Cai, 2006). In the present Monte Carlo study, therefore, the backward-elimination and effects-coded backward-elimination approaches are expected to outperform the forward-inclusion approach under ideal conditions (i.e., unbiased anchor set). However, when the anchor set is contaminated by DIF (e.g., a single DIF anchor item or one DIF item in the anchor set, depending on the testing approach used), the effects-coded backward-elimination approach is anticipated to outperform the backward-elimination approach. This is because the DIF contamination rate in the anchor set is always lower in the former approach than in the latter approach, given the same scale.

Test Statistics

As described earlier, statistical significance of a direct effect, or equivalently uniform DIF, is determined by assessing change in model fit between two nested models[4]. Log-likelihood or chi-square goodness-of-fit difference (likelihood ratio [LR] statistic) is the most frequently used test statistic. Although LR statistic has a problem of inherent dependency on sample size (Brannick, 1995), Stark et al. (2006) showed that the use of Bonferroni-corrected LR test substantially decreased Type I error of the (MACS) DIF analysis in some cases (e.g., large sample, large DIF). Lee (2009) also found that Bonferroni correction can almost eliminate Type I error in cases of comparable large groups, regardless of item type, DIF type, and test length.

Other recent studies have provided empirical sampling distributions for some key fit measures in regard to DIF analysis at the scale level. For example, a change in comparative fit index (CFI) by 0.01 (Cheung & Rensvold, 2002) or 0.002 (Meade, Johnson, & Braddy, 2008) and a 0.005 change in standardized root mean square residual (SRMR) (Chen, 2007) have been suggested as optimal criteria[5]. However, there is no such

[4] Some previous studies have used the Wald test results from fitting only a single MIMIC-DIF model (i.e., no model fit comparison). However, Wald statistic is not a stable measure for statistical significance (Brown, 2006). For example, when different but statistically equivalent scaling methods are used, they can provide different standard errors and consequently different Walt statistics (see González & Griffin, 2001).

[5] Information-theoretic fit measures (e.g., Akaike information criterion, Bayesian information criterion) are also suitable for evaluating scale-level DIF, but

standard proven useful in case of testing DIF at the item level. Thus, the current simulation study examines the use of these scale-level global fit criteria. It is expected that the suggested criterion values, △CFI of 0.01 and 0.002 and △SRMR of 0.005, are too stringent and thus will provide biased, unreasonable conclusions about DIF.

As is evident from the discussions above, there are at least two important differences in implementing the MIMIC DIF analysis: testing approach (forward-inclusion, backward-elimination, effects-coded backward-elimination) and test statistics (LR, Bonferroni-corrected LR, △CFI, △SRMR). Accordingly, the primary goal of this investigation was to compare the performance of the MIMIC (uniform) DIF analysis for different analytic strategies via a Monte Carlo simulation described below.

Method

Condition Factors

Condition factors included those that have been commonly examined in the DIF literature-item type, test length, sample size, impact (true difference in latent trait mean), DIF type and DIF size in a target item, and DIF type and DIF size in an anchor set. Considering usual practice in DIF studies, a dummy-coded group variable (i.e., focal versus reference) was used as a single covariate.

Item Type. Item responses were either dichotomous or ordinal (5-point Likert scale). They were conceptualized as observed categorical responses y, wherein underlying responses y^* are completely latent and continuous (Mellenbergh, 1994). As a normally distributed latent response exceeded certain threshold value(s), the observed response took higher score(s). In other words, examinees who chose a particular category had more of characteristic of the trait than others who chose a lower category.

Test Length. The test consisted of six or 12 items. The second item (Item 2) always served as the target item. Other items in the test served as an anchor set (i.e., all the remaining items under the forward-inclusion or effects-coded backward-elimination approach; the first item [a single anchor item; Item 1] under the backward-elimination approach). When DIF was simulated, it appeared only on Item 1, only on Item 2, or both. Consequently, the rate of DIF contamination ranged from 0 to 33%.

Sample Size. Three combinations of sample sizes were constructed; focal group N_f = (a) 100, (b) 250, or (c) 500 and reference group N_r = (a) 900, (b) 750, or (c) 500, respectively. Total sample size was always

their effectiveness has not been supported in the literature beyond the fit measures discussed here.

$N_f + N_r = 1,000$, so as to not confound differences in sample size with total sample size.

Impact. When there was no impact, the latent trait followed a standard normal distribution ($\eta \sim N[0, 1]$) in each group. When a moderate impact was present, the trait means differed by 1 standard deviation so that the focal group had a smaller trait mean ($\eta \sim N[-1, 1]$) compared to the reference group.

DIF Type. DIF can be either uniform or non-uniform depending on the item parameter that differs across groups. Uniform DIF is present when the item intercept(s) differs across groups; non-uniform DIF exists when the item loading differs across groups, regardless of the invariance of the item intercept(s). In the present study, uniform DIF was created by varying an item's threshold(s) between two groups-the threshold(s) for the focal group was raised by 0.8 (large DIF), making uniform DIF items more difficult (less attractive) for this group compared to the reference group. For ordinal responses, all four thresholds were shifted by the same amount, which is analogous to varying all the location parameters in the graded response model (GRM; Samejima, 1969). Non-uniform DIF was created by varying an item's loading between two groups-the loading for the focal group was reduced by 0.4 (large DIF) so that non-uniform DIF items were less discriminative for this group. It should be noted that non-uniform DIF also could be simulated by varying "both" an item's loading and threshold(s) but only the loading parameter was manipulated so as to isolate the effects of varying the threshold(s) and/or loading from each other, the two primary sources of DIF.

Data Generation

Both dichotomous and ordinal responses were generated as follows. First, population parameter values were specified such that the same factor structure underlay each of two groups. To isolate the effects of varying item difficulty and/or latent trait distribution from each other[6], single common factor model was applied for data generation. This model can be written as

$$y_i^* = \lambda_i \eta + \beta_i \epsilon_i, \tag{8}$$

where β_i is the loadings on the unique factor scores ϵ_i that are assumed to be normally distributed. The unique factor loadings were given by $\sqrt{1-\lambda_i^2}$, thereby yielding the item variances of unity. The item loadings λ_i were equal between two groups, except for the target and anchor items. The population parameter

[6] An intercept difference does not necessarily indicate an item mean difference because the item mean is dependent on both the loading and the latent mean (Stark et al., 2006; see Equation 3). In other words, lowering the intercept of an item increases difficulty of this item when and only when the loading and the latent mean are equal across groups.

values used for data generation are shown in Table 1.

For each group, the trait scores and unique factor scores sampled from a normal distribution and the loadings a priori defined were substituted into Equation 8 to create continuous responses. Once continuous responses were generated, they were transformed into discrete responses under two or five categories. For dichotomous responses, a threshold parameter value was chosen according to 50% of the area under the normal curve-if continuous responses were greater than the threshold $\kappa_i = 0$, they were scored as 1; otherwise, they were scored as 0. For ordinal responses, four threshold values were chosen with an equal interval according to approximately 3.6%, 23.8%, 45.1%, 23.8%, and 3.6% of the area under the normal curve.

Table 1. Population Item Parameters

		Dichotomous item			Ordinal item					
Group	Item	λ	κ	β	λ	κ_1	κ_2	κ_3	κ_4	β
Reference	1	0.90	0.00	0.19	0.90	-1.80	-0.60	0.60	1.80	0.19
	2	0.80	0.00	0.36	0.80	-1.80	-0.60	0.60	1.80	0.36
	3	0.70	0.00	0.51	0.70	-1.80	-0.60	0.60	1.80	0.51
	4	0.60	0.00	0.64	0.60	-1.80	-0.60	0.60	1.80	0.64
	5	0.50	0.00	0.75	0.50	-1.80	-0.60	0.60	1.80	0.75
	6	0.40	0.00	0.84	0.40	-1.80	-0.60	0.60	1.80	0.84
	7	0.85	0.00	0.28	0.85	-1.80	-0.60	0.60	1.80	0.28
	8	0.75	0.00	0.44	0.75	-1.80	-0.60	0.60	1.80	0.44
	9	0.65	0.00	0.58	0.65	-1.80	-0.60	0.60	1.80	0.58
	10	0.55	0.00	0.70	0.55	-1.80	-0.60	0.60	1.80	0.70
	11	0.45	0.00	0.80	0.45	-1.80	-0.60	0.60	1.80	0.80
	12	0.35	0.00	0.88	0.35	-1.80	-0.60	0.60	1.80	0.88
Focal										
Non-uniform DIF	1	0.50	0.00	0.75	0.50	-1.80	-0.60	0.60	1.80	0.75
	2	0.40	0.00	0.84	0.40	-1.80	-0.60	0.60	1.80	0.84
Uniform DIF	1	0.90	0.80	0.19	0.90	-1.00	0.02	1.40	2.60	0.19
	2	0.80	0.80	0.36	0.80	-1.00	0.02	1.40	2.60	0.36

Note. Items 1 and 2 were used as an anchor item and a target item, respectively. The parameter values for other items are not shown in the focal group because they were equal to those in the reference group.

Ordinal responses were assigned as such $y_i = 1$ if $y_i^* \leq -1.8$; $y_i = 2$ if $-1.8 < y_i^* \leq -0.6$; $y_i = 3$ if $-0.6 < y_i^* \leq 0.6$; $y_i = 4$ if $0.6 < y_i^* \leq 1.8$; and $y_i = 5$ if $y_i^* > 1.8$.

Analysis

The MIMIC DIF analysis was conducted using M*plus* 6.0 (Muthén & Muthén, 1998 - 2010). In each of 500 replications, a baseline model and a constrained or augmented model, constructed as described earlier, were successively fitted to the generated data. The null hypothesis of no DIF was tested by using each of three testing approaches (forward-inclusion, backward-elimination, effects-coded backward-elimination). The M*plus* syntax examples used to specify the MIMIC models appear in the appendix.

Six test statistics including uncorrected LR and corrected LR (Bonferroni-corrected $p = .05/n$, where *n* is the number of possible DIF tests within the scale; Stark et al., 2006), △CFI of 0.01 and 0.002, and △SRMR of 0.005 were calculated within each nested-model comparison. Outcome variables of interest were false positive (FP) rate and true positive (TP) rate. The FP rate was computed as the proportion of times that the unbiased target item was erroneously identified as having DIF (i.e., reject true null hypothesis) out of 500 replications of each condition. Similarly, the TP rate was calculated as the proportion of times that DIF was correctly detected in the biased target item (i.e., reject false null hypothesis).

Using SAS 9.2 (SAS Institute, 2002 - 2008), variance components analysis was also conducted to assess the relative influences of the condition factors and application strategies on the study outcomes. All the effects, except for an intercept, were treated as random via minimum variance quadratic unbiased estimation (MIVQUE).

Results

When impact was present such that the latent trait means differed by 1 standard deviation between the focal and reference groups, the false positive (FP) rates based on the LR statistic (either uncorrected or corrected) were severely inflated and the true positive (TP) rates were artificially high in nearly all simulated conditions, regardless of testing approach used in combination (forward-inclusion, backward-elimination, or effects-coded backward-elimination). For the △CFI or △SRMR, the TP rates were around 0 in nearly all conditions. Therefore, following discussions about simulation results are focused on the cases where the trait means were equal between two groups (i.e., no impact). As noted previously, the MIMIC method is not suitable for detecting non-uniform DIF because the models assume equal loadings across groups. Supporting this limiting assumption, the TP rates for detecting non-uniform DIF were either very low or spuriously raised in all study

conditions. Thus, discussions are further limited to the cases where, if present, only uniform DIF appeared in the target item. More complete results will be available to interested readers by request.

False Positive Rate

Table 2 presents the FP rates, by all combinations of the conditions factors, separately for different testing approaches and test statistics. In the favorable cases of unbiased anchor set, each of the three testing approaches provided reasonable control for the FP rate. For the backward-elimination approach, the FP rates were equal to 0 in all conditions, regardless of test statistic used together. The effects-coded backward-elimination approach produced the rates below the nominal alpha value (.006 – .050) with uncorrected LR; and less than .006 with other test statistics. Similar results were observed for the forward-inclusion approach, except for a few conditions-slightly elevated FP rates for the binary responses from the groups of greater than 100 examinees (.086 – .124).

The presence of large uniform DIF in the anchor set (i.e., a single DIF anchor item or one DIF item in the anchor set) severely inflated the FP rates, especially when the backward-elimination approach was utilized, as expected. For example, the rates easily approached 1 in almost all conditions when this approach was used with (either uncorrected or corrected) LR or △CFI of 0.002. In contrast, even when the anchor set was biased by uniform DIF, the effects-coded backward-elimination approach still maintained reasonable control for the FP rate in some cases. When the scale consisted of 12 items, this approach provided the rates less than .014 with corrected LR, △CFI, or △SRMR; and those ranged from .002 to .206 (median < .10) with uncorrected LR. For the forward-inclusion approach, the FP rates were always below the nominal alpha value except for a few conditions (e.g., .056 – .098 for uncorrected LR), regardless of test statistic. Nevertheless, this overall reduction in the FP rate had little practical implications when considering very low TP rates in general (see True Positive Rate section below).

Generally, a biased, non-uniform DIF anchor set did not inflate the FP rates of the MIMIC (uniform) DIF analysis. Unless uncorrected LR was utilized in combination, each of the three testing approaches controlled the FP rates at the nominal alpha level. Even with the uncorrected LR, the backward-elimination approach provided the rates less than .032; the effects-coded backward-elimination less than .098; and the forward-inclusion approach less than .134. This finding was not surprising because the performance of testing uniform DIF may be influenced negligibly by having different type of bias in the anchor set (i.e., non-uniform DIF) and less influenced by having the same type of bias in this set (i.e., uniform DIF) (see Lee, 2009).

Table 2. False Positive Rates

Item type	Test length	Sample size	DIF in anchor set	Forward-inclusion				
				Uncor.LR	Cor. LR	△CFI 0.01	△CFI 0.002	△SRMR 0.005
Binary	6 items	100/900	No DIF	0.008	0.002	0.000	0.000	0.000
			Non-uniform	0.010	0.002	0.000	0.000	0.000
			Uniform	0.018	0.002	0.000	0.000	0.000
		250/750	No DIF	0.100	0.016	0.000	0.000	0.000
			Non-uniform	0.134	0.018	0.000	0.000	0.000
			Uniform	0.050	0.006	0.000	0.000	0.000
		500/500	No DIF	0.086	0.014	0.000	0.000	0.000
			Non-uniform	0.106	0.016	0.000	0.000	0.000
			Uniform	0.036	0.002	0.000	0.000	0.000
	12 items	100/900	No DIF	0.008	0.002	0.000	0.000	0.000
			Non-uniform	0.008	0.002	0.000	0.000	0.001
			Uniform	0.010	0.002	0.000	0.000	0.000
		250/750	No DIF	0.124	0.006	0.000	0.000	0.000
			Non-uniform	0.134	0.010	0.000	0.000	0.000
			Uniform	0.098	0.006	0.000	0.000	0.000
		500/500	No DIF	0.110	0.004	0.000	0.000	0.000
			Non-uniform	0.116	0.004	0.000	0.000	0.000
			Uniform	0.070	0.000	0.000	0.000	0.000
Ordinal	6 items	100/900	No DIF	0.018	0.002	0.000	0.000	0.000
			Non-uniform	0.024	0.004	0.000	0.000	0.000
			Uniform	0.056	0.004	0.000	0.000	0.000
		250/750	No DIF	0.052	0.002	0.000	0.000	0.000
			Non-uniform	0.058	0.004	0.000	0.000	0.000
			Uniform	0.022	0.000	0.000	0.000	0.000
		500/500	No DIF	0.060	0.000	0.000	0.000	0.000
			Non-uniform	0.062	0.000	0.000	0.000	0.000
			Uniform	0.002	0.002	0.000	0.000	0.000
	12 items	100/900	No DIF	0.006	0.000	0.000	0.000	0.000
			Non-uniform	0.006	0.000	0.000	0.000	0.000
			Uniform	0.024	0.002	0.000	0.000	0.000
		250/750	No DIF	0.048	0.002	0.000	0.000	0.000
			Non-uniform	0.058	0.002	0.000	0.000	0.000
			Uniform	0.024	0.000	0.000	0.000	0.000
		500/500	No DIF	0.048	0.000	0.000	0.000	0.000
			Non-uniform	0.044	0.000	0.000	0.000	0.000
			Uniform	0.020	0.000	0.000	0.000	0.000

Table 2. False Positive Rates (Continue)

Backward-elimination					Effects-coded backward-elimination				
Uncor.LR	Cor. LR	△CFI 0.01	△CFI 0.002	△SRMR 0.005	Uncor.LR	Cor. LR	△CFI 0.01	△CFI 0.002	△SRMR 0.005
0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000
0.010	0.000	0.000	0.000	0.000	0.022	0.002	0.000	0.000	0.000
1.000	1.000	0.004	1.000	0.102	0.392	0.092	0.000	0.002	0.000
0.000	0.000	0.000	0.000	0.000	0.049	0.002	0.000	0.000	0.000
0.032	0.012	0.000	0.000	0.000	0.122	0.008	0.000	0.000	0.000
1.000	1.000	0.700	1.000	0.976	0.214	0.014	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.037	0.000	0.000	0.000	0.000
0.032	0.008	0.000	0.000	0.000	0.098	0.002	0.000	0.000	0.000
1.000	1.000	1.000	1.000	1.000	0.400	0.074	0.000	0.002	0.000
0.000	0.000	0.000	0.000	0.000	0.006	0.002	0.000	0.000	0.000
0.006	0.000	0.000	0.000	0.000	0.011	0.003	0.000	0.000	0.000
1.000	1.000	0.000	0.288	0.000	0.044	0.002	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.050	0.002	0.000	0.000	0.000
0.024	0.002	0.000	0.000	0.000	0.076	0.004	0.000	0.000	0.000
1.000	1.000	0.000	0.994	0.000	0.002	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.046	0.000	0.000	0.000	0.000
0.028	0.000	0.000	0.000	0.000	0.056	0.000	0.000	0.000	0.000
1.000	1.000	0.000	1.000	0.008	0.006	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.000
0.024	0.004	0.000	0.000	0.000	0.040	0.002	0.000	0.000	0.000
1.000	1.000	0.302	1.000	0.974	0.828	0.438	0.000	0.020	0.000
0.000	0.000	0.000	0.000	0.000	0.026	0.002	0.000	0.000	0.000
0.030	0.004	0.000	0.000	0.000	0.072	0.002	0.000	0.000	0.000
1.000	1.000	1.000	1.000	1.000	0.968	0.748	0.000	0.104	N.A.
0.000	0.000	0.000	0.000	0.000	0.024	0.000	0.000	0.000	0.000
0.018	0.000	0.000	0.000	0.000	0.038	0.000	0.000	0.000	0.000
1.000	1.000	1.000	1.000	1.000	0.998	0.926	0.000	0.288	N.A.
0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000
0.018	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000
1.000	1.000	0.000	0.992	0.000	0.174	0.014	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.022	0.000	0.000	0.000	0.000
0.024	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000
1.000	1.000	0.000	1.000	0.118	0.136	0.010	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000
0.018	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000
1.000	1.000	0.022	1.000	0.996	0.206	0.008	0.000	0.000	0.000

The results of variance components analysis were similar across the five different test statistic values, which were independently used to derive the FP rates. That is, the most influential factor was commonly the two-way interaction, testing approach × DIF in an anchor set (no, uniform, non-uniform), for uncorrected LR (MIVQUE = 0.077), corrected LR (0.090), and △CFI of 0.002 (0.095). This two way interaction × testing length (three-way interaction) contributed the most to the FP variance when the △CFI of 0.01 (0.020) or △SRMR of 0.005 (0.027) was used as a test statistic.

True Positive Rate

On average, the TP rate for testing (large) uniform DIF was highest when DIF was tested by the use of uncorrected LR (.692), followed by corrected LR (.527), △CFI of 0.002 (.246), △SRMR of 0.005 (.078), and △CFI of 0.01 (.040). As we expect, the latter three criteria suggested as optimal for testing DIF at the scale level were somewhat stringent for testing DIF at the item level. However, in general, the use of Bonferroni correction did not reduce the TP rates substantially.

By using uncorrected LR, average TP rates were almost equal between dichotomous responses (.693) and ordinal responses (.692). However, the average rates were higher when an anchor set had no DIF (.877) rather than uniform DIF (.322); and when group sizes were comparably large ($N_f = N_r = 500$; .816) rather than largely different ($N_f = 100, N_r = 900$; .518). Also, the backward-elimination approach (.701) provided the highest average TP rate, followed by the effects-coded backward-elimination approach (.691) and forward-inclusion approach (.686).

The TP rates, by all combinations of the conditions factors, are shown in Table 3. When an anchor set was free from DIF, the TP rates for the backward-elimination approach were equal to or near 1 in all conditions when used with (either uncorrected or corrected) LR. The effects-coded backward-elimination approach provided acceptable TP rates with corrected LR unless group sizes differed largely (.724 – 1); or with uncorrected LR unless group sizes differed largely under 12-item scale conditions (.836 – 1). With uncorrected LR, the forward-inclusion approach also provided the TP rates greater than .80 unless groups differed greatly in size (.834 – .946). However, given the inflated FP rates observed earlier for this approach, this finding needs to be interpreted with caution. Regardless of testing approach, in general, the TP rates were very low when △CFI or △SRMR was used as a test statistic. Some exceptions occurred when the backward-elimination approach was chosen-for example, when △CFI of 0.002 was used with ordinal responses (.0804 – 1).

The presence of uniform DIF in an anchor set severely degraded the TP rates of the MIMIC (uniform) DIF analysis. Regardless of

Table 3. True Positive Rates

Item type	Test length	Sample size	DIF in anchor set	Forward-inclusion				
				Uncor.LR	Cor. LR	△CFI 0.01	△CFI 0.002	△SRMR 0.005
Binary	6 items	100/900	No DIF	0.454	0.154	0.000	0.024	0.000
			Non-uniform	0.450	0.154	0.000	0.032	0.000
			Uniform	0.276	0.056	0.000	0.000	0.000
		250/750	No DIF	0.834	0.516	0.000	0.269	0.000
			Non-uniform	0.844	0.538	0.000	0.424	0.000
			Uniform	0.568	0.284	0.000	0.003	0.000
		500/500	No DIF	0.946	0.808	0.000	0.716	0.000
			Non-uniform	0.956	0.824	0.000	0.836	0.000
			Uniform	0.693	0.461	0.000	0.011	0.000
	12 items	100/900	No DIF	0.484	0.098	0.000	0.000	0.000
			Non-uniform	0.486	0.100	0.000	0.000	0.000
			Uniform	0.408	0.064	0.000	0.000	0.000
		250/750	No DIF	0.842	0.468	0.000	0.000	0.000
			Non-uniform	0.858	0.474	0.000	0.000	0.000
			Uniform	0.736	0.360	0.000	0.000	0.000
		500/500	No DIF	0.914	0.746	0.000	0.000	0.000
			Non-uniform	0.922	0.754	0.000	0.000	0.000
			Uniform	0.856	0.596	0.000	0.000	0.000
Ordinal	6 items	100/900	No DIF	0.384	0.164	0.000	0.102	0.000
			Non-uniform	0.388	0.176	0.000	0.104	0.000
			Uniform	0.190	0.050	0.000	0.000	0.000
		250/750	No DIF	0.890	0.642	0.000	0.792	0.000
			Non-uniform	0.906	0.706	0.000	0.874	0.000
			Uniform	0.514	0.270	0.000	0.000	0.000
		500/500	No DIF	0.968	0.844	0.000	0.980	0.000
			Non-uniform	0.978	0.920	0.000	0.994	0.000
			Uniform	0.712	0.416	0.000	0.000	0.000
	12 items	100/900	No DIF	0.350	0.092	0.000	0.000	0.000
			Non-uniform	0.352	0.092	0.000	0.000	0.000
			Uniform	0.260	0.058	0.000	0.000	0.000
		250/750	No DIF	0.856	0.524	0.000	0.000	0.000
			Non-uniform	0.868	0.546	0.000	0.000	0.000
			Uniform	0.722	0.348	0.000	0.000	0.000
		500/500	No DIF	0.968	0.740	0.000	0.000	0.000
			Non-uniform	0.976	0.804	0.000	0.000	0.000
			Uniform	0.880	0.548	0.000	0.000	0.000

Table 3. True Positive Rates (Continue)

Backward-elimination					Effects-coded backward-elimination				
Uncor.LR	Cor. LR	△CFI 0.01	△CFI 0.002	△SRMR 0.005	Uncor.LR	Cor. LR	△CFI 0.01	△CFI 0.002	△SRMR 0.005
1.000	1.000	0.000	0.982	0.002	0.836	0.412	0.000	0.002	0.000
1.000	0.998	0.000	0.930	0.006	0.826	0.412	0.000	0.002	0.000
0.020	0.000	0.000	0.000	0.000	0.054	0.002	0.000	0.000	0.000
1.000	1.000	0.000	1.000	0.022	0.981	0.817	0.000	0.057	0.000
1.000	1.000	0.036	1.000	0.394	0.990	0.902	0.000	0.069	0.000
0.127	0.006	0.000	0.000	0.000	0.210	0.026	0.000	0.016	0.000
1.000	1.000	0.000	1.000	0.550	1.000	0.978	0.000	0.225	0.000
1.000	1.000	0.158	1.000	0.930	1.000	0.990	0.000	0.255	0.000
0.243	0.038	0.000	0.000	0.000	0.365	0.068	0.000	0.050	0.000
1.000	1.000	0.000	0.010	0.001	0.514	0.084	0.000	0.000	0.000
0.998	0.996	0.000	0.034	0.002	0.514	0.076	0.000	0.000	0.000
0.020	0.000	0.000	0.000	0.000	0.118	0.004	0.000	0.000	0.000
1.000	1.000	0.000	0.054	0.000	0.860	0.460	0.000	0.000	0.000
1.000	1.000	0.000	0.456	0.000	0.890	0.508	0.000	0.000	0.000
0.154	0.002	0.000	0.000	0.000	0.404	0.082	0.000	0.000	0.000
0.998	0.998	0.000	0.378	0.002	0.938	0.766	0.000	0.002	0.000
1.000	1.000	0.000	0.862	0.000	0.956	0.816	0.000	0.002	0.000
0.268	0.012	0.000	0.000	0.000	0.604	0.228	0.000	0.002	0.000
1.000	1.000	0.054	1.000	0.898	0.954	0.724	0.000	0.003	0.000
1.000	1.000	0.096	1.000	0.858	0.938	0.684	0.000	0.003	0.000
0.004	0.000	0.000	0.000	0.000	0.058	0.004	0.000	0.000	0.000
1.000	1.000	0.996	1.000	1.000	1.000	0.998	0.000	0.037	0.000
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.047	0.000
0.052	0.000	0.000	0.000	0.000	0.228	0.026	0.000	0.000	0.000
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.120	0.000
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.192	0.000
0.122	0.008	0.000	0.000	0.000	0.330	0.066	0.000	0.006	0.000
1.000	1.000	0.000	0.846	0.000	0.614	0.192	0.000	0.000	0.000
1.000	1.000	0.000	0.804	0.000	0.586	0.190	0.000	0.000	0.000
0.004	0.000	0.000	0.000	0.000	0.098	0.004	0.000	0.000	0.000
1.000	1.000	0.000	1.000	0.000	0.970	0.810	0.000	0.000	0.000
1.000	1.000	0.000	1.000	0.006	0.978	0.846	0.000	0.000	0.000
0.056	0.000	0.000	0.000	0.000	0.444	0.084	0.000	0.000	0.000
1.000	1.000	0.000	1.000	0.006	1.000	0.962	0.000	0.000	0.000
1.000	1.000	0.004	1.000	0.762	1.000	0.980	0.000	0.000	0.000
0.154	0.000	0.000	0.000	0.000	0.622	0.212	0.000	0.000	0.000

testing approach and test statistics, the TP rates were always less than .80 with only a few exceptions. Generally, a biased, non-uniform DIF anchor set did not decrease the TP rates-the rates were, in fact, similar to those observed when the anchor set was DIF-free.

Based on the variance components analysis results, DIF in the anchor set accounted for the greatest TP variance when uncorrected (MIVQUE = 0.087) or corrected LR (0.103) was used as a test statistic. The testing approach was included in the most influential interactions for the △CFI of 0.01 (testing approach × item type × test length; 0.020) or 0.01 (testing approach × DIF in an anchor set; 0.052) and △SRMR of 0.005 (testing approach × test length; 0.016).

Discussion

Given that the MIMIC DIF analysis has enjoyed increasing attention in the DIF literature, the purpose of the present study was to evaluate the accuracy of this technique by means of Monte Carlo simulation. Specifically, different testing approaches and test statistics were compared in regard to false positive (FP) and true positive (TP) rates, across various conditions of different item type, test length, sample size, impact, and DIF type and DIF size in a target item and an anchor set. This study proposed and empirically tested a new testing approach (effects-coded backward-elimination) as a potential solution for arbitrary choosing a DIF-free anchor set.

It should be noted that the two primary outcome variables, FP and TP statistics, are closely tied to each other -for instance, lowering an alpha level for a test (e.g., through the use of Bonferroni correction or scale-level global fit criteria) generally reduces "both" the FP and TP rates of the test. Besides, in cases where the FP rate is inflated, the standard definition of the TP rate at the nominal alpha level (i.e., power) is not meaningful (Finch, 2005). Thus, any conclusion about performance should not be made solely based on one outcome. Combining the FP and TP outcomes together, the following section discusses the study findings and implications.

Summary of Important Findings and Implications

The current simulation results appear to support the utility of the MIMIC DIF analysis in some circumstances but not in others. As expected, this technique was not suitable for testing non-uniform DIF-in all simulated conditions, the TP rates were either very low or spuriously raised due to inflated Type I error. When impact was present between two groups (i.e., groups truly differ in their levels of latent trait), not only the FP rates were severely inflated but also the TP rates were not acceptable when testing (large) uniform DIF (see González-Romá, Hernández, & Gómez-Benito,

2006[7]), regardless of testing approach and test statistic used in any combination. Thus, unequal trait means should be concerned when using the MIMIC technique, of which models specify common factor parameters and common item parameters across groups. Indeed, Cheung and Rensvold (1999) noted that if trait parameters are constrained across groups when they are not actually equal, biased conclusions of measurement equivalence can occur.

It was also found that, not surprisingly, the global fit criteria known as optimal for testing scale-level DIF (△CFI of 0.01 or 0.002, △ SRMR of 0.005) are fairly strict for the item-level DIF tests, decreasing both the FP and TP rates in most conditions. In contrast, although the use of Bonferroni correction (on the LR statistic) did reduce the FP rates, it maintained acceptable TP rates in some conditions-e.g., either the backward-elimination or effects-coded backward-elimination approach was used along with an unbiased anchor set. Thus, Bonferroni correction would be desired when one conducts the CFA-based DIF analysis (Start et al., 2006)

More important, it appeared that different testing approaches yield different outcomes. When an anchor set was truly unbiased, as expected, both the backward-elimination approach and the effects-coded backward-elimination approach outperformed the forward-inclusion approach in regard to both the FP and TP rates. More interestingly, when an anchor set was biased by large uniform DIF, only the effects-coded backward-elimination approach performed effectively in some conditions. Specifically, the FP rates for this approach were controlled at the nominal alpha level (with corrected LR) or slightly inflated (with uncorrected LR) as the DIF contamination rate decreased (i.e., more [DIF-free] items in the scale). However, the TP rates were not satisfactory in these conditions (medians of .424 and .083 for uncorrected LR and corrected LR, respectively). These findings were supported by the subsequent variance component analyses - testing approach and/or DIF in an anchor set were commonly in the factors that contributed the most to the FP and TP variances.

Taken together, the findings from this Monte Carlo study might suggest a possibility that ameliorates the problems of biased anchor set, which are repeatedly alerted in the DIF literature (see Cheung & Rensvold, 1999; Finch, 2005; Millsap, 2005; Navas-Ara & Gómez-Benito, 2002). That is, even with a biased anchor set (by either uniform or non-uniform DIF), the effects-coded backward-elimination (uniform) DIF test of Bonferroni-corrected LR is expected to eliminate the chances of FP under many conditions, while not substantially reducing the PT rates if used with a relatively large scale and equally large groups (e.g., 500 examinees each) in combination[8]. This proposed analytic

[7] Under the same condition, MACS DIF analysis controls Type I error only when group sizes are equal.

strategy is appealing in theoretical as well as practical standpoints in a number of reasons-(a) any particular direct effect is not necessarily fixed to set the scale and thus no need for a priori designated anchor set; (b) the baseline model provides a proper fit, against which DIF is examined in the subsequent nested models (Maydeu-Olivares & Cai, 2006); and (c) unlike other pervious solutions (e.g., Fleishman et al., 2002; Christensen et al., 2001; Mackinnon et al., 1999, Woods, 2009b), it neither necessarily increases the number of nested-model comparisons and thus maintains Type I error; (d) nor involves the risk of capitalizing on chance. The effects-coded backward-elimination approach is also desirable in the sense that no item is absolutely devoid of DIF. As noted previously, "all" the direct effects in the model are estimated as an optimal balance. Thus, even when a designated anchor set is available, it will provide more accurate conclusions about DIF.

Limitations and Suggestions for Future Research

Although some important findings and implications could be obtained, this study has several limitations and suggestions for further research. First, given that this is a Monte Carlo study, caution should be used in generalizing current results and conclusions beyond the conditions investigated. For example, no missing values were simulated on the responses although conclusions of any DIF analysis likely depend on the missing data mechanism-missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR)-and the amount of missing data. Also, the scales were relatively short, having six or 12 items, and only large DIF was simulated on only one or two items in the scales. Sample sizes were selected so as to represent those often found in psychological assessment. However, smaller sample sizes (e.g., < 100 examinees) will be easily encountered in the study of low-incidence groups. Thus, further investigations, especially with wide-ranging test lengths and additional conditions, are encouraged to continue to evaluate the proposed solution along with the previous solutions for selecting an anchor set. This practice would help practitioners in selecting an appropriate analytic strategy to use.

Second, this study found that the global fit criteria are not optimal for testing DIF at the item level. Compared to IRT, one advantage of using CFA is to have a variety of practical fit measures. Thus, future efforts are needed to empirically examine distributions of those fit measures and find some potential criterion values suitable for item-level DIF test. Any new criteria should be independent of the overall fit of a baseline model; should not be impacted by

[8] Previous simulation studies also have shown that, in general, the performance of the CFA-based DIF analysis improves with the use of Bonferroni correction and/or with lower DIF contamination rates (e.g., Finch, 2005; Meade & Lautenschlager, 2004; Navas-Ara & Gómez-Benito, 2002; Stark et al., 2006).

model complexity; and should not be redundant with other fit measures (Cheung & Rensvold, 2002).

Finally, nested data are often present in the educational research (e.g., students nested within classrooms, further nested with schools) as well as in diverse behavioral and social science settings. Both the theory and the utility regarding multilevel SEM are well demonstrated in the literature (e.g., Raudenbush & Bryk, 2002; Mehta & Neale, 2005; Everson & Millsap, 2004). With respect to the MIMIC DIF analysis, an interesting question will be how the presence of multilevel data impacts the performance and the conclusions of usual, single-level DIF analysis. In fact, Finch and French (2011) showed that Type I error is inflated when single-level (standard) MIMIC model is innocuously fitted, not accounting for the multilevel data structure. More interesting investigation will be to assess the efficiency of the multilevel MIMIC DIF analysis, relative to detecting DIF at the between-cluster level, at the within-cluster level, or at both.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Barrett, P. (2007). Structural equation modeling: Adjusting model fit. *Personality and Individual Differences, 42*, 815-824.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*, 201-213.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321.

Camilli, G., & Shepard, L. A. (1994). *Measurement methods for the social sciences series: Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method, *Journal of Management, 25*, 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Christensen, H., MacKinnon, A. J., Korten, A., & Jorm, A. F. (2001). The "common cause hypothesis" of cognitive aging: Evidence for not only a common factor but also specific associations of age with vision and grip strength in a cross-sectional analysis. *Psychology*

and Aging, 16, 588-599.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale NJ: Lawrence Erlbaum.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin*, *95*, 134 - 135.

Everson, H. T., & Millsap, R. E. (2004). Beyond individual differences: Exploring school effects on SAT scores. *Educational Psychologist, 39*, 157-172.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and the IRT likelihood ratio test. *Applied Psychological Measurement, 29*, 278-295.

Finch, H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling, 18*, 229-252.

Fleishman, J. A. (2005). Using MIMIC models to assess the influence of differential item functioning. Retrieved October, 24 2005, from http://outcomes.cancer.gov/conference/irt/fleishman.pdf

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, 57*, 275-283.

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*, 251-264.

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*, 29-53.

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

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, 10*, 631-639.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136-153.

Lee, J. (2009). *Type I error and power of the MACS CFA for DIF detection: Methodological issues and resolutions.* Unpublished doctoral dissertation, University of Kansas, USA.

Lee, J., Little, T. D., & Preacher, K. J. (2010). Methodological issues in using structural equation models for testing differential item functioning. In E. Davidov, P. Schmidt, and J. Billiet (Eds.), *Cross-cultural data analysis: Methods and applications*. (pp. 57-86). New York, NY: Routledge.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural

data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53-76.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13*, 59-72.

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

Mackinnon, A., Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., & Rodgers, B. (1999). A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual Differences, 27*, 405-416.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*, 55-64.

McDonald, R. P. (1999). *Test theory: Unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Meade, A. W., & Lautenschlager, G. K. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/ invariance. *Structural Equation Modeling*, *11*, 60-72.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in test of measurement invariance. *Journal of Applied Psychology, 93*, 568-592.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259-284.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223-237.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 153-172). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Los Angeles: University of California and Muthén & Muthén.

Muthén, 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*, 1-22.

Muthén, L.K. & Muthén, B.O. (1998 – 2010). *Mplus user's guide*. (6th Ed.). Los Angeles, CA: Muthén & Muthén.

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

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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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.

SAS Institute. (2002 – 2008). *SAS/STAT 9.2 user's guide*. Cary, NC: SAS Institute Inc.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, No. 17.

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. (2004). Examining the effects of differential item/test functioning (DIF/DTF) on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508.

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*, 1292-1306.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care, 44*, 39-49.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-69.

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

Woods, C. M. (2009b). Testing for differential item functioning with measures of partial association. *Applied Psychological Measurement, 33*, 538 – 554.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indictor multiple cause models. Applied Psychological Measurement, 35, 339-361.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-Model DIF Testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment, 31*, 320-330.

1차원고접수 : 2011. 9. 29.
수정원고접수 : 2011. 12. 2.
최종게재결정 : 2011. 12. 5.

한국심리학회지 : 일반
Korean Journal of Psychology: General
2011, Vol. 30, No. 4, 1083-1110

MIMIC DIF 분석 기법의 실증적 비교: 불편정착기준문항의 임의적 선택에 대한 잠재적 해결책

이 재 훈

캔자스대학교

본 연구는 검사 내 존재하는 차별기능(DIF)의 탐지를 위해 사용되는 복수측정변수복수원인모형(MIMIC) 확인적 요인분석의 효율성을 검증한다. 특히 기존의 분석 기법(forward-inclusion, backward-elimination)과 통계치(uncorrected or Bonferroni-corrected LR, DCFI of 0.01 or 0.002, DSRMR of 0.005)에 관한 통계적 검증력을 검사 문항의 종류, 검사의 길이, 표집의 크기, 탐지되는 문항과 정착기준문항 내 차별기능의 종류와 크기를 포함, 다양한 조건에서 조사한다. 또한 차별기능 분석 시 연구자가 불편정착기준문항(들)을 임의적으로 선택해야 하는 것에 대한 대안으로 effects-coded backward-elimination 분석 기법이 제시된다. 몬테카를로 시뮬레이션을 통해, 본 연구는 정착기준문항(들)이 실제로 편향된 경우 제시된effects-coded backward-elimination 기법만이 몇 가지 조건하에서 적합한 통계적 검증력을 가진다는 것을 보여준다. 검사 내 차별기능문항의 비율이 감소할수록, 이 새로운 분석 기법은 위양성률(false positive rate)을 0.05 alpha수준에서 통제하거나(Bonferroni-corrected LR과 함께 사용된 경우) 약간 인상된 수준을 (uncorrected LR과 함께 사용된 경우) 나타내었다.

주요어 : 복수측정변수복수원인모형, 차별기능, 불편정착기준문항, 분석 기법

Appendix

TITLE: Mplus Syntax Example for Forward-Inclusion Approach (Step 1);
DATA: FILE = list.txt; ! *Names data list file*
TYPE = MONTECARLO; ! *Indicates the type of model to estimate*
VARIABLE: NAMES = X Y1-Y6; ! *Defines variable names, covariate X and items Y1-Y6*
MODEL: F BY Y1@1 Y2-Y6; ! *Defines item loadings on latent variable F, fixing the loading of Y1 to 1 (marker-variable method for scaling part i)*
F; ! *Defines latent variable variance*
Y1-Y6; ! *Defines item variances*
[F]; ! *Defines latent variable mean*
[Y1@0 Y2-Y6]; ! *Defines item intercepts, fixing the intercept of Y1 to 0 (marker-variable method for scaling part i)*
F ON X; ! *Regresses covariate on latent variable*
SAVEDATA: RESULTS = forward_step1.fit; ! *Saves model estimates and model fit values*

TITLE: Mplus Syntax Example for Forward-Inclusion Approach (Step 2);
DATA: FILE = list.txt;
TYPE = MONTECARLO;
VARIABLE: NAMES = X Y1-Y6;
MODEL: F BY Y1@1 Y2-Y6;
F;
Y1-Y6;
[F];
[Y1@0 Y2-Y6];
F ON X;
Y2 ON X; ! *Regress covariate on target item*
SAVEDATA: RESULTS = forward_step2.fit;

TITLE: Mplus Syntax Example for Backward-Elimination Approach (Step 1);
DATA: FILE = list.txt;
TYPE = MONTECARLO;
VARIABLE: NAMES = X Y1-Y6;
MODEL: F BY Y1@1 Y2-Y6;
F;

```
Y1-Y6;
[F];
[Y1@0 Y2-Y6];
F ON X;
Y2-Y6 ON X; ! Regress covariate on items except for anchor item
SAVEDATA: RESULTS = backward_step1.fit;

TITLE: Mplus Syntax Example for Backward-Elimination Approach (Step 2);
DATA: FILE = list.txt;
TYPE = MONTECARLO;
VARIABLE: NAMES = X Y1-Y6;
MODEL: F BY Y1@1 Y2-Y6;
F;
Y1-Y6;
[F];
[Y1@0 Y2-Y6];
F ON X;
Y3-Y6 ON X; ! Regress covariate on items except for anchor item and target item
SAVEDATA: RESULTS = backward_step2.fit;

TITLE: Mplus Syntax Example for Effects-Coded Backward-Elimination Approach (Step 1);
DATA: FILE = list.txt;
TYPE = MONTECARLO;
VARIABLE: NAMES = X Y1-Y6;
MODEL: F BY Y1@1 Y2-Y6;
F;
Y1-Y6;
[F];
[Y1@0 Y2-Y6];
F ON X;
Y1 ON X (a); ! Defines regression of covariate on Y1 as "a"
Y2 ON X (b); ! Defines regression of covariate on Y2 as "b"
Y3 ON X (c); ! Defines regression of covariate on Y3 as "c"
Y4 ON X (d); ! Defines regression of covariate on Y4 as "d"
Y5 ON X (e); ! Defines regression of covariate on Y5 as "e"
```

```
Y6 ON X (f); ! Defines regression of covariate on Y6 as "f"
MODEL CONSTRAINT: a = 0 - b - c - d - e - f; ! Constrains regressions of covariate on items to average 0
SAVEDATA: RESULTS = effects_step1.fit;

TITLE: Mplus Syntax Example for Effects-Coded Backward-Elimination Approach (Step 2);
DATA: FILE = list.txt;
TYPE = MONTECARLO;
VARIABLE: NAMES = X Y1-Y6;
MODEL: F BY Y1@1 Y2-Y6;
F;
Y1-Y6;
[F];
[Y1@0 Y2-Y6];
F ON X;
Y1 ON X (a); ! Defines regression of covariate on Y1 as "a"
Y3 ON X (c); ! Defines regression of covariate on Y3 as "c"
Y4 ON X (d); ! Defines regression of covariate on Y4 as "d"
Y5 ON X (e); ! Defines regression of covariate on Y5 as "e"
Y6 ON X (f); ! Defines regression of covariate on Y6 as "f"
MODEL CONSTRAINT: a = 0 - c - d - e - f;
SAVEDATA: RESULTS = effects_step2.fit;
```