MIMIC DIF Testing When the Latent Variable Variance Differs Between Groups

by

Ian A. Carroll

Submitted to the Department of Psychology

and the Faculty of

the Graduate School of the University

of Kansas in partial fulfillment of

the requirements for the degree of

Master of Arts

_____

Chairperson: Carol Woods, Ph.D

_____

Wei Wu, Ph.D

_____

William Skorupski, Ph.D

Date Defended: May 14th, 2014

The thesis committee for Ian Carroll

certifies that this is the approved version of the following thesis:

MIMIC DIF Testing When the Latent Variable Variance Differs Between Groups

_____

Chairperson: Carol Woods, Ph.D

Date approved: December 19th, 2014

ABSTRACT

Multiple indicators multiple causes (MIMIC) models (Joreskog & Golberger 1975) can be employed in a psychometric context to test for differential item functioning (DIF) between groups on the measurement of a latent variable (Muthén 1989). MIMIC DIF models can be attributed some favorable properties when compared to alternative DIF testing methods (i.e., Item Response Theory- Likelihood Ratio DIF) such as having generally small sample size requirements while simultaneously maintaining reliably low Type 1 error rates and sufficient DIF detection power (Woods 2009). The mechanism by which MIMIC models test for DIF is to regress a latent variable and its non-anchor indicators onto an exogenous (grouping) variable. This allows the model to account for differences in the mean of the latent variable across groups, while also testing for uniform DIF in individual items. However, the model does not allow heterogeneity in the covariance structure of the latent variables themselves—it is assumed to be equal across groups.

A simulation study was conducted to examine the consequences of violating this assumption for the MIMIC DIF model. In this simulation, the following characteristics were varied: sample size, DIF effect magnitude, heterogeneity in latent variance between groups, magnitude of the group mean difference on the latent variable, and the ratio of focal group size to reference group size. Results suggest that violating the model's equality of latent covariance structure assumption leads to systematically biased parameter estimates on factor loadings and estimates of the latent group mean difference, inflated Type 1 error in DIF detection, and several other undesirable statistical side-effects.

Table of Contents

Introduction

Psychometric testing, in both psychological and educational contexts, aims to accurately measure unobservable attributes of individuals. This is done by measuring and evaluating other observable characteristics and responses to test items that are theoretically indicative of the presence (or lack thereof) of the aforementioned unobservable attribute. Given that the results of these tests can significantly impact the lives of the examinees (e.g., standardized testing to determine acceptance into institutions of higher learning, or psychological evaluations to identify and determine an optimal approach for treatment of a disorder), test "fairness" is both statistically and politically of paramount concern (Cole, Holland, & Wainer, 1993, Ch 2, p.28-29). As such, an entire field of research has developed over approximately the last half century in order to meet this demand.

Differential Item Functioning

Differential Item Functioning (DIF) can be defined as "when people from different manifest groups (e.g., males and females) do not have equal probability of a correct answer, even if they have the same level of ability" (de Ayala, 2009), or "(when there exist) differences in item functioning *after* groups have been matched with respect to the ability or attribute that the item purportedly measures" (Dorans, Holland & Wainer, 1993, p.37). Both of these definitions, despite one being presented in a specific context (a circumstance where a correct answer exists) and one being more generalized, get at the core issue: an item does not behave the same way across groups, indicating that variability in scores on the item in question may be attributable to unmeasured and theoretically irrelevant variables.

There exist an abundance of methods available for researchers to test DIF, in both the Classical Test Theory (CTT) and latent variable paradigms. This project focuses on multiple-

indicators, multiple causes (MIMIC) models for DIF detection, which is a method rooted within the latent variable paradigm. As such, the following explanation of DIF caters specifically to how DIF is operationalized in the latent variable context.

An understanding of reference groups and focal groups is necessary in order to understand how DIF is operationalized. In essence, the reference group is the (typically larger) group to which a focal group (typically smaller) is compared. Designating one group as reference and another group as focal is a process that is either theory-driven or arbitrary, as this designation merely dictates how parameters are interpreted. In terms of dichotomous grouping variables, "0" is usually representative of the reference group, whereas "1" designates the focal group. Consequently, results estimated using this variable are often interpreted as an effect of being a member of the focal, rather than reference, group.

Furthermore, many modern methods (including the implementation of MIMIC DIF in this study) incorporate designated anchor items (items that are thought to be invariant across groups), which provide a common set of invariant items amongst groups so that other items can be investigated for noninvariance.

Statistically, the presence of DIF indicates a significant item-level group difference in responses while controlling for mean differences on a latent variable. This phenomenon can manifest in two ways: in the literature, these item-level group differences are called either uniform or non-uniform DIF.

Uniform DIF exists when this difference is merely a shift in the item intercept [or, in item response theory (IRT) parameterization, a difference in only the *b* parameter], where the DIF effect is then necessarily favoring one group over another over the entire range of the latent variable. This is contrasted by non-uniform DIF, which occurs when a group difference in item

score significantly interacts with the latent variable (necessarily the *a*, but also potentially the *b*, parameters in IRT). Figure 1 depicts uniform and non-uniform DIF in the form of item-characteristic curves (ICC), as seen in IRT analyses.

Each ICC represents the probability of answering a binary response as "1", and how it changes monotonically across the continuum of the latent variable. The left graph demonstrates uniform DIF: one group has an increased probability of answering "1" across the entire latent continuum. In other words, one group is favored uniformly over the other group.

The graph on the right represents non-uniform DIF; in this instance, it is observable that one group can have a comparatively higher *or* lower probability of responding "1", entirely dependent on the value of the latent variable. One group is not uniformly "favored" over another, and this is consequently non-uniform DIF.

This study focuses solely on uniform DIF with MIMIC models. Newer methods for addressing non-uniform DIF with MIMIC models will be addressed in the discussion.

MIMIC DIF models

The seminal publication on MIMIC models was authored by Joreskog and Goldberger in 1975. The principal idea behind a MIMIC model is that a Confirmatory Factor Analysis (CFA) model can include observed exogenous variables that are predictive of any latent variables, such that variability in scores on the latent variables is not solely explained through disturbance terms and latent variable covariances. This is achieved by regressing the latent variables onto the aforementioned observed exogenous variables, resulting in a model where variability in the latent variable can be attributed to exogenous variables. By incorporating information from exogenous variables into the traditional CFA model, one is able to construct models more reflective of the underlying relationships between observed and unobserved variables, whereas

otherwise any variability in latent scores caused by exogenous variables might be erroneously attributed to other sources. MIMIC models might therefore be understood as CFA models that attempt to account for population heterogeneity in latent constructs by adding a regression structural component to the model, which is why MIMIC models are alternatively known as "CFA with covariates" (Brown 2006).

However, Joreskog and Goldberger (1975) did not expand the method beyond continuous and normally distributed manifest variables, and consequently methods of analyzing MIMIC models with dichotomous and ordinal variables were developed later (Muthén 1984). These methods were developed further in Muthén (1989), which presented a method through which testing for measurement invariance (and, therefore, DIF) using MIMIC models became possible.

MIMIC models are useful in this regard because incorporating grouping variables is simple: it requires merely the addition of a regression pathway where group membership predicts scores on the latent variable. This allows the researcher to estimate potential group mean differences on the latent variable without having to estimate an additional latent variable covariance matrix (in the case of multiple group confirmatory factor analysis), which might not be possible due to limited sample size.

MIMIC DIF testing is conducted by regressing potential DIF items and (simultaneously) the latent variable onto an exogenous variable. This exogenous variable can be either continuous or categorical in nature (the DIF effect is merely a regression pathway), though often in psychological or educational research it is some sort of dichotomous grouping variable. Unlike methods that require an entirely separate covariance structure to be estimated for each individual group, MIMIC DIF requires only the addition of a regression component of the model. The

contribution to model fit attributable to this additional regression pathway can be evaluated either through a nested-model chi-square deviance test (known alternatively as a likelihood ratio test) or a Wald test for parameter significance.

MIMIC DIF model formulation

This section will explain in more explicit terms how a MIMIC DIF model is specified in the factor-analytic framework.

The model below features *p* indicators, *m* factors, and *q* exogenous variables *x*.

First, we have the measurement model,

$$y^* = \mu_y + \Lambda\eta + \varepsilon$$

where y* is a vector with *p* elements, Λ is a *p* x *m* matrix of factor loadings, η is a vector of *m* factors, and ε is a *p*-dimensional vector of error terms.

In accordance with notation set forth by Muthén (1989), the manifest variable in the measurement model has a superscripted asterisk. This indicates that y* is a latent, standard normally-distributed variable that theoretically underlies a categorical observed variable. The model is therefore fitted to a matrix of tetrachoric or polychoric correlations, where bivariate normality is assumed to exist amongst the y* variables. The common factor model

$$\Sigma_{yy} = \Lambda\Phi\Lambda^T + D_\psi$$

is then estimable using least squares estimation.

An additional structural regression component is added with MIMIC models, as seen below.

$$\eta = \alpha + \beta\eta + \Gamma x + \zeta$$

where α is an *m*-dimensional vector of factor intercepts, β and Γ represent *m* x *m* and *m* x *q* matrices of regression coefficients (respectively), and ζ is an *m*-dimensional vector of error terms for the factors.

A concrete example can be useful, so here is an example of a particular model with one DIF tested item (with one exogenous variable), 5 ordinal items with 5 categories each, and one factor. Below are selected matrices from the model estimation process.

$$\Lambda = \begin{pmatrix} 0 & 1 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{41} & 0 \\ \lambda_{51} & 0 \end{pmatrix}, \Phi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\beta = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix}, \Gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

The potential DIF item is treated as a latent variable, and its estimated loading onto the factor is in the form of the regression coefficient β_{21}. It also bears mentioning that anchor items were not designated in Λ above, though they necessarily must be specified in order to set a common scale for group comparison. The estimated group mean difference on the latent construct is γ_1, and the estimated DIF effect is γ_2. The model additionally estimates p*(*C*-1) thresholds (with *C* being the number of categories for each item), which in this case totals to 20.

Advantages and disadvantages of MIMIC DIF compared to other methods

Requiring fewer additional parameters to test group differences is an advantageous characteristic of MIMIC DIF models, in the sense that lower sample sizes are viable, whereas the larger number of required parameters in other procedures (i.e. multiple-group CFA models)

necessitate larger sample sizes to obtain reliable results (Woods 2009). Additionally, "because a single input matrix is used, the advantages of MIMIC models over multiple-groups CFA include their greater parsimony (MIMIC entails fewer freely estimated parameters), their relatively greater ease of implementation when several groups are involved (i.e., depending on the complexity of the measurement model, multiple-groups CFA may be cumbersome when the number of groups exceeds two), and their less restrictive sample size requirements (i.e., multiple-groups CFA requires a sufficiently large sample size for each group)" (Brown 2006).

Another advantageous characteristic of MIMIC DIF models is that testing for measurement and item invariance can be (using Wald tests instead of nested-model deviance tests) a one-step process, rather than the multi-step process required by other latent-variable invariance testing methods (i.e. invariance testing in multiple-group CFA). The comparative simplicity in this regard is all the more apparent when categorical data are introduced, where correct implementation of multi-group CFA invariance testing requires additional knowledge (proper specification of item threshold constraints, etc.) that might make the process more difficult (and consequently perhaps less appealing) to some researchers.

MIMIC DIF models can also flexibly be realized in an IRT framework as well as the traditional factor-analytic framework, where constraints on *a* and *b* parameters lead to nested-model deviance testing to determine if a certain parameter is invariant across groups. This further widens the range of potential research settings in which MIMIC DIF can be employed. Despite the fact that MIMIC DIF models can be run within the IRT parameterization, research providing methods to equate MIMIC DIF model parameters calculated in CFA to IRT parameters (MacIntosh & Hashim, 2003**)** has also been explored.

One disadvantage of this approach, however, is that "MIMIC models...examine just two potential sources of invariance (indicator intercepts, factor means)" (Brown 2006). This means that testing for invariance in factor loadings and factor (co)variances is not possible when using MIMIC DIF. This is the result of an assumption made when only a single latent covariance matrix is estimated in the model: the latent covariance matrix is treated as equal for all groups. As such, even in the unidimensional case of a single factor, demonstrated in Figure 2, equality of latent variances theoretically must exist between substantively relevant groups for the estimation procedure to produce unbiased results. The primary aim of this study is to assess the performance of the MIMIC DIF method when this assumption is violated.

Method

The six independent variables in this simulation were sample size, group mean difference on the latent variable, group latent variable variance difference, magnitude of DIF effect, reference to focal group size ratio, and test length.

Sample Size

MacCallum et al. (1999) argue that "common rules of thumb regarding sample size in factor analysis are not valid or useful", and consequently sample size values in this simulation were selected based on two criteria: (1) every condition has adequate sample size to correctly estimate its corresponding model, and (2) there exists meaningful variability in sample size values such that any effect it contributes is readily apparent in the results. A pilot study was conducted in order to determine sample sizes that fulfill both of the aforementioned criteria, with the result of "small", "medium", and "large" samples corresponding to a combined number of subjects for the reference and focal groups equaling 200, 500, and 1000.

Group Size Ratio

Group size ratio between reference and focal groups also varied, with values of 1:1 and 7:3. These values were selected to be representative of potential group size ratios a researcher might encounter when examining substantive studies' dichotomous grouping variables, with 1:1 being more in line with many analyses of gender groups, and 7:3 corresponding better (in some circumstances) to groups defined by ethnicity.

Group Mean Difference

Group Mean Difference (GMD) varied from -0.5 to 1.5, at intervals of .5 (where positive values indicate a higher latent mean for the focal group). This set of values not only allows evaluation of circumstances where the focal group latent mean is above or below the reference group's, but also provides conditions where the magnitude of the latent mean difference ranges from nonexistent to large (in the sense that the latent ability densities clearly exhibit limited overlapping area).

Magnitude of DIF Effect

DIF (here discussed as the magnitude of the regression pathway to a potentially noninvariant item) had values of 0, .1, .25, and .5. Uniform DIF in the CFA framework is conceptualized as a difference in intercept between groups, controlling for potential GMD. Given the distribution from which the indicators are constructed, the DIF condition values featured here increasingly large discrepancies in item intercept in favor of the focal group. Preliminary simulations suggested that .1, .25, and .5 were representative of small, medium, and large DIF effects in this context, based upon probability of observing a significant estimated regression coefficient. Specifically, .5 (and, to an understandably lesser extent, .25) almost

invariably manifested as significant when data were generated, regardless of sample size, group mean difference, and even (in)equality of latent variances. On the contrary, .1 was a sufficiently small effect that only under the best available circumstances (larger sample size, equal latent variances) were able to reliably label the effect as significant.

Latent Variable Variance of Focal Group

The latent variable variance for the focal group (FG LV) varied from .5 to 2, in intervals of .5. This produced conditions where the focal variance could be either smaller or larger than the reference group latent variance, as well as conditions where the variance difference between the two latent distributions is rather pronounced (with the reference group latent variable distribution set at ~N(0,1)).

Number of Test Items

Scale length was also varied. Indicator sets of 5, 10, and 15 variables were employed in order to determine if scale length in any way impacted MIMIC DIF testing performance when unequal latent variances are present. These values were selected to represent both small and medium-sized scales. There undoubtedly exist many scales that extend well beyond 15 items; the inclusion of this condition is not to necessarily perfectly match and include what would be a "large" scale length in psychological or education research (especially when "large" varies between disciplines and specializations within disciplines), but merely to provide sufficient range in number of indicators such that an effect attributable to scale length can be detected if it does in fact exist.

Model

Though the number of indicators varied (5,10 or 15), the estimated CFA model always

featured one latent variable, with 40% of the available items as anchors and 20% of the available items as DIF items. All anchor sets in this simulation were correctly specified (i.e., DIF-free). The latent variable was scaled at ~N(0,1), allowing free estimation of non-anchor loadings. A set of values for factor loadings selected prior to the simulation, with values ranging from .5 to .8 (interpretable because the models were fitted to polychoric correlation matrices). This selection process aimed to create a set of manifest variables that is comprised of medium-to-high strength indicators for the latent variable, while still maintaining a certain degree of variability amongst the values of the factor loadings themselves.

Estimation procedure

Weighted least squares (WLS) was the estimator of choice in this simulation, given the ordinal indicators conceptualized in the factor-analytic framework. More precisely, this is the WLSMV estimator available in Mplus, which is a form of WLS that uses a diagonal weight matrix instead of a full weight matrix in the fit function for parameter estimates. Standard errors and chi-square-based statistics are then calculated using a mean- and variance- adjusted full weight matrix (Muthén & Muthén, 1998-2011). The model is fit to a matrix of tetrachoric and/or polychoric correlations.

Wald tests versus Likelihood Ratio Tests

Though it is has been argued that the likelihood ratio test (LRT) method for selecting significant DIF pathways is preferable (Patiwan, p.47-48), there are certain practical considerations that make Wald testing more reasonable for the models featured in this study.

Wald tests are all calculated within one model, and are therefore obtained simultaneously. In LRTs, an additional model must be fitted for each parameter being evaluated for significance.

LRTs would also theoretically provide no functional benefit over using Wald tests for the conditions featured in this study; Wald tests traditionally underperform with very small sample sizes and sparse data, and neither of those circumstances is purposefully simulated here. Consequently, Wald tests were used because they are easily implemented, more computationally efficient than LRTs, and, most importantly, LRTs would provide no functional benefit over Wald tests given the data simulated in this study.

Outcomes

The primary goal of purposefully violating the equality of latent variances assumption inherent to MIMIC DIF models in this case was determining the degree of bias introduced by this violation. As such, measures of raw bias (in accordance with measuring any systematic, directional effect) as well as mean square error (MSE) are the targeted outcome variables of interest.

Software

Mplus

Data were generated and analyzed using Mplus (Muthén & Muthén, 1998-2011), with each condition being assessed across 500 replications. Categorical variables in Mplus are generated by specifying K-1 thresholds (where K is the number of categories desired in the variable), where each threshold represents a cutoff on a standard normal distribution. This in line with how polychoric correlations are calculated (and conceptualized): for each categorical variable there exists an underlying normally-distributed latent variable, and as such each categorical variable is simply an imprecise realization of this latent variable. The items generated in these analyses are 5-category ordinal items.

R v. 3.0.2

R was used to automate the Mplus Monte Carlo procedure, aggregate results with the MplusAutomation package (Hallquist & Wiley, 2013), and produce relevant tables and other visuals related to key results.

Results

A general summary (prior to more thorough elaboration on the more nuanced aspects of the results) of these results is that MIMIC DIF testing exhibited increasingly severe degrees of bias in estimation with regard to factor loadings (FLs), group mean differences (GMDs), and DIF effects as the difference in group latent variable variances became more pronounced. These effects were often systematic in their direction and visuals have been chosen to highlight those circumstances. However, given the sheer number of potential visuals (with 6 independent variables and 1440 conditions in total), the graphs and tables in this section were selected because they are representative of a pattern seen consistently throughout all conditions to which the pattern pertains.

Factor Loadings

Figure 3 displays how violating the equality of variances assumption systematically biases estimated factor loadings. Further exploration of these error distributions for unequal latent variance conditions reveals an interaction in correct estimation of factor loadings between latent variance and group membership ratio between reference and focal groups. As seen in Figure 4, conditions where the reference to focal group size ratio was unequal (in favor of the reference group) were less biased in estimation of factor loadings than conditions where the two compared groups were of equal size. This result demonstrates that the constraints imposed on

parameters for the sake of model identification can somewhat arbitrarily affect the results. For example, conditions where 70% of the sample consisted of subjects belonging to the reference group are in essence lesser violations of the equality of latent variances assumption, merely because more subjects have data from where the scale was set (meaning that compared to 1:1 group size ratio conditions, there exist a smaller number of subjects whose data do not adhere to the assumed latent covariance structure).

Estimation of Group Mean Difference

Figures 5.1-5.4 display the latent variance of the focal group plotted against bias in the estimated group mean difference on the latent variable. The reference group latent variable distribution was always set at ~N(0,1), and accurate estimation of group mean difference is clearly demonstrated when the focal group's ability variance is also equal to 1. However, departures from this equality yield systematically inaccurate results, with the general trend being that smaller focal group ability variance leads to bias in the direction of the GMD, whereas larger focal group ability variance leads to bias opposite the direction of the GMD. When the true value of the GMD was zero, even estimates in violation conditions were not clearly biased in any direction; however, the conditions with equal latent variances between the two groups still demonstrate the most precise estimates.

Much as it did in the estimation of factor loadings, group size ratio also has a clear effect upon bias in estimation of the group latent mean difference. Figure 6 features a representative selection of GMD and FG LV combinations, demonstrating the existence of the reference:focal effect in unequal variance conditions, previously seen in Figure 4. Once again, the condition with equal latent variances is the only condition where systematically biased estimation does not

occur.

Estimating the Magnitude of DIF

Figures 7.1-7.4 depict, much like Figures 5.1-5.4, systematically biased parameter estimates with further departures from equal latent variable variance, though in this case the parameters of interest are estimated DIF effects. In conditions where no DIF effects were simulated, unequal latent variances produce seemingly unbiased but clearly imprecise estimates of the DIF effect, whereas equal latent variance conditions were consistently accurate. As DIF effects are introduced and increase, the systematic bias observed in estimation of factor loadings and group mean difference becomes increasingly apparent. As demonstrated previously in estimates of factor loadings and group mean difference, the group size ratio between the reference and focal groups also plays a role in amount of bias observed in DIF estimates, as seen in Figure 7.5.

Larger sample sizes in unequal variance conditions are not less biased than lower sample size conditions, though by sole virtue of smaller standard errors on the DIF parameter, DIF is more frequently detected. This is demonstrated by comparing tables 8.1 versus 8.2, 8.3 versus 8.4, and 8.5 versus 8.6. Though this might initially seem practically advantageous, items flagged for DIF are not necessarily removed from the item pool. There are circumstances where researchers might instead attempt to account for the DIF effect, which is a biased estimate in the case of unequal latent variances.

GMD LV DIF significance interaction

Figures 8.1 to 8.4 demonstrate an interaction regarding probability of obtaining a significant DIF effect between latent variable variance and latent group mean difference. These

figures differ from previous figures in that the x-axis is now representative of the latent group mean difference. As such, cross-sections of results are obtained by selecting specific values of the focal group latent variance (as well as sample size and R:F, in an attempt to further reveal the signal underneath the noise). The decision to change the x-axis from previous figures was made in order to demonstrate trends between significant DIF estimates, group latent mean difference, and magnitude of the DIF effect, all under the umbrella of a single FG LV value. This allows for direct visual comparison of patterns between different FG LV values, and is consequently in line with this project's goal of evaluating aspects of MIMIC model performance when unequal group latent variances exist.

Results from Figures 8.1 and 8.2 can be explained thusly: comparatively smaller focal group latent variance results in overall less latent variance than the model is specified to have. This, coupled with the fact that all DIF conditions favor the focal group, suggests that group mean differences progressively moving in the opposite direction of the DIF effect lead to fewer than otherwise expected simulated respondents endorsing higher values on the DIF item (in this case, "4" or "5"). Consequently, group mean differences progressing in the opposite direction of the DIF effect decrease the probability of obtaining a significant DIF estimate.

By the same token, group mean differences progressing in the same direction as the DIF effect increase the probability of obtaining a significant DIF estimate, given smaller focal group latent variance, because fewer than otherwise expected simulated respondents endorse lower values on the DIF item (in this case, "1" or "2").

In the case of larger focal group latent variance, as demonstrated in Figures 8.3 and 8.4, the opposite direction of the same relationship is observed. Once again, this can be understood

by viewing the result in terms of actual item category endorsement probability versus model-expected item category endorsement probability.

Figures 8.5 and 8.6 have been added (demonstrating that in the case of equal latent variances, this interaction does not occur). These added tables represent the ideal scenario: no incidental multivariate interaction results in inflated or deflated probability of detecting a significant DIF effect.

Number of Indicators

The number of indicators does not appear to affect the outcomes of interest in any tangible way (seen in figures 9.1 and 9.2, where 5 and 15 item results for a certain representative set of conditions are displayed).

Though the number of factors:number of indicators ratio increased, the percentage of anchor and DIF tested items remained proportionally identical across the models. As such, it appears to be the case that increasing the number of items per factor had no impact on parameter estimation, given that the intrinsic properties of the item sets remained identical across conditions.

Sample size

Sample size, as demonstrated previously through tables 8.1-8.6, influences the probability of detecting a significant DIF effect. As would be expected in any circumstance where a model is fundamentally misspecified, no improvement with regard to actual accuracy of estimates is observed, a result demonstrated in Figures 10.1-10.6. Even the dispersion of bias here is not improved by increasing sample size.

Discussion

Naturally, a model that is fundamentally misspecified is likely not to return accurate parameter estimates. The issues this study attempts to address are specifically which parameters in a MIMIC DIF model are influenced by unequal latent variances, and additionally the degree to which estimates and inference regarding these parameters can be influenced.

The results of this simulation suggest that MIMIC DIF models fail to properly estimate factor loadings, group latent mean differences, DIF effects, as well as improperly balance Type-1 and Type 2 error as a result of violating the equality of latent (co)variances assumption inherent to the model.

Needless to say, a variety of different parameter estimates can be influenced greatly by unequal latent variances between groups, leading to entirely different point estimates than the true values of those parameters. That being said, when the assumptions inherent to the model are met, MIMIC DIF performs admirably, with unbiased estimates for all aspects of invariance a MIMIC DIF model can evaluate (item intercept, group mean difference) as well as unbiased estimates for factor loadings. Concordantly, Type 1 and Type 2 error rates are well controlled and perform reliably even at lower sample sizes.

Specifically pertaining to inference, larger magnitude DIF effects, as well as larger sample sizes, increase the probability of detecting a DIF item despite having unequal latent variances. This finding has a practical significance in the sense that any egregiously non-invariant items are not likely to avoid being flagged when testing for DIF with a MIMIC model. Gelin (2005) goes into great depth regarding whether an item should be removed, retained, or revised for significant DIF coefficients obtained in MIMIC DIF models. Understandably, there does not appear to be any universal answer: this is expected even in the case that all of the

assumptions of the model are met. The conclusion that "DIF is not a replacement for item reviews" (Gelin 2005) is of particular help in the scenario explored in this study, in which the researcher has to determine if a MIMIC DIF model is appropriate based on distributions of variables that are inherently unobservable.

As it is with essentially any statistical procedure, the question "should I use it" is necessarily met with "it depends." In circumstances where a researcher has a large sample size, this author sees no reason whatsoever to use methods that, by making more and/or less tenable assumptions, increase the probability of obtaining an incorrect result. As such, if sample size permits use of Multiple Group CFA or IRT-LR-DIF, which do not require as many assumptions on the part of the researcher and allow testing for more types of measurement noninvariance, it is likely optimal to use either of the aforementioned methods rather than MIMIC DIF. However, as long as there exist low sample size studies where multiple group comparisons are desired, MIMIC DIF testing will remain a viable option (and will perform well when its intrinsic assumptions are met).

There are certain limitations to this study that are worth noting. Only single-factor models were analyzed, and across models key determinants of measurement performance were held equal (a pure anchor set containing 40% of available items, only 20% DIF items, medium/strong factor loadings for all manifest variables). Addressing any one of these individual model characteristics is a potential direction for future research, as they all logically could impact proper estimation of the MIMIC model (without or without equal latent variances). Furthermore, because each model only featured a single factor, circumstances where there exist equal latent variances but unequal latent covariances could not be investigated. Additionally, any of these aforementioned issues could be explored in the IRT framework as well.

Another potential direction for future research is incorporating non-uniform DIF into the MIMIC DIF method, with Woods and Grimm (2011) utilizing an added interaction term between the grouping variable and latent variable, allowing MIMIC DIF analyses to go beyond conceptualizing DIF as a solely unidirectional phenomenon (and, in turn, attempting to detect the presence of both simultaneously). Unfortunately, this is complicated by the fact that there are few, if any, implementations of this method in easily accessible software. In the case of Woods and Grimm (2011), the Mplus method for specifying the interaction component assumes normally distributed variables: this is problematic given that the exogenous variable is typically categorical in these analyses. This issue could be addressed by eschewing the frequentist framework entirely in favor of a Bayesian approach, which might avoid the distributional rigidity necessarily imposed to allow traditional methods to estimate properly.

References

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research.* The Guilford Press, New York, NY.

Finch, H. (2005).The MIMIC Model as a Method for Detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement. 29*: 278. doi:10.1177/0146621605275728

Gelin, M. N. (2005). Type 1 Error Rates of the DIF MIMIC Approach Using Joreskog's Covariance Matrix with ML and WLS Estimation. Unpublished Doctoral Thesis.

Hallquist, M. & Wiley, J. (2013). MplusAutomation: Automating Mplus Model Estimation and Interpretation. R package version 0.6-2. http://CRAN.R-project.org/package=MplusAutomation

Holland, P. W., Wainer, H., & Educational Testing Service (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 351.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample Size in Factor Analysis. *Psychological Methods*. doi:10.1037/1082-989X.4.1.84

MacIntosh & Hashim (2003). The MIMIC Model as a Method for Detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement*, *29*: 278-295.

Mantel, N., & Haenszel, W.M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22,* 719-748.

Muthén, B. (1984). A general structural equation model with dichotomous ordered categorical and continuous latent variable dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49,* 115–132.

Muthén, B. O. (1989). Latent Variable Modeling in Heterogeneous Populations. *Psychometrika, 54*(4), 557-585.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford: Clarendon Press.

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

Woods, C.M., Grimm, K.J. (2011). Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models. *Applied Psychological Measurement*, 35: 339. doi:10.1177/0146621611405984

Tables and Figures

Table 1

Variable Name	Values				
Sample Size	200	500	1000		
Latent Group Mean Difference	-.5	0	.5	1	1.5
Magnitude of DIF effect	0	.1	.25	.5	
Focal Group Latent Variable Variance	.5	1	1.5	2	
Group Size Ratio	1:1	7:3			
Number of Items	5	10	15		
DIF=.1	-1.7507	-0.7388	0.5828	1.3408	
DIF=.25	-1.9599	-0.8416	0.4538	1.0364	
DIF=.5	-2.612	-1.0364	0.2533	0.6745	

Figure 1

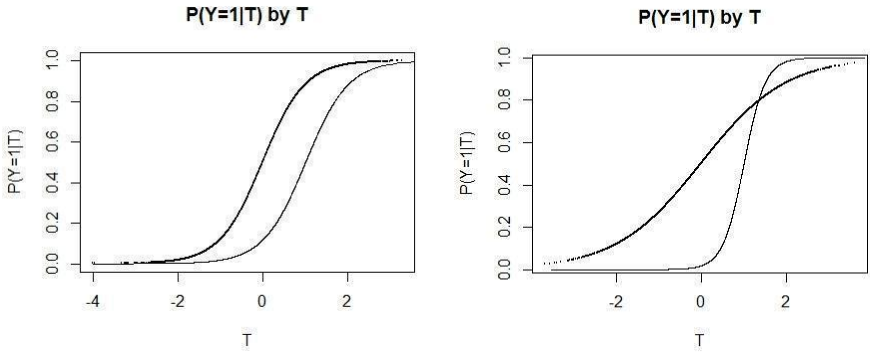

P(Y=1|T) by T
P(Y=1|T)
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
T
P(Y=1|T) by T
P(Y=1|T)
0.0
0.2
0.4
0.6
0.8
1.0
-2
0
2
T

Figure 2

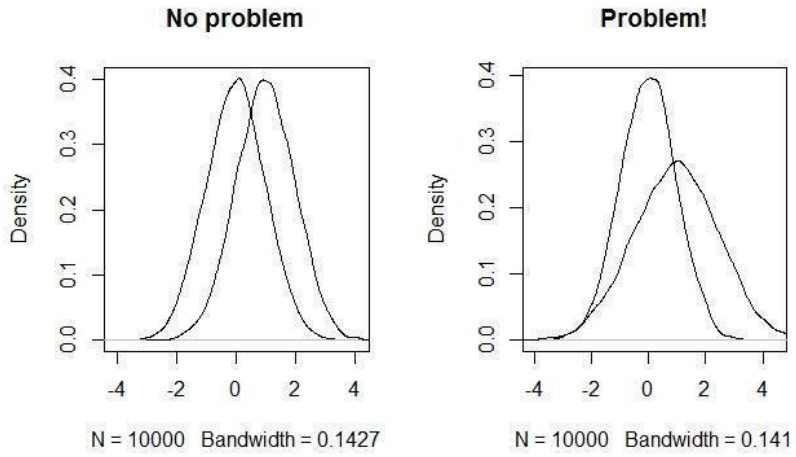
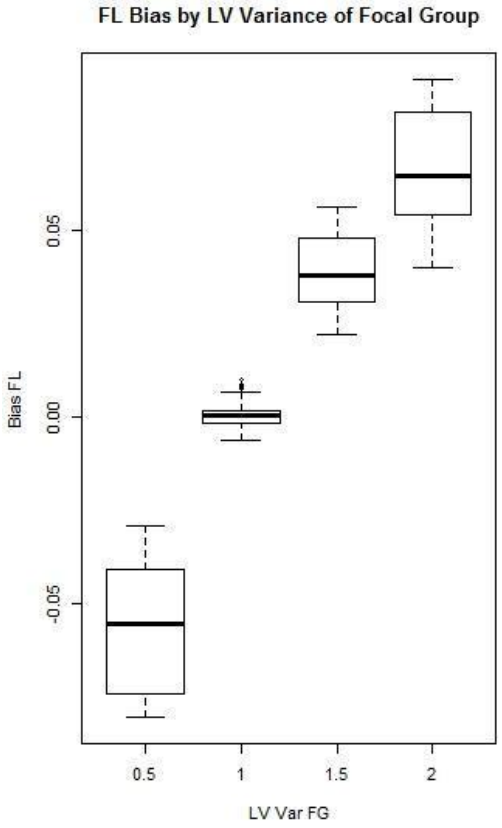

No problem
Problem!
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
N = 10000 Bandwidth = 0.1427
N = 10000 Bandwidth = 0.141

Figure 3



	0.5	1	1.5	2
min	-0.0806	-0.0064	0.0219	0.04
25%	-0.07435	-0.0019	0.0306	0.05385
median	-0.0555	2e-04	0.0379	0.06415
75%	-0.04075	0.00165	0.0476	0.0813
max	-0.0294	0.0065	0.0561	0.0902
N	120	120	120	120
MSE	0.00356	1e-05	0.00159	0.00465
Avg Bias	-0.05726	5e-05	0.03874	0.06664

Figure 4

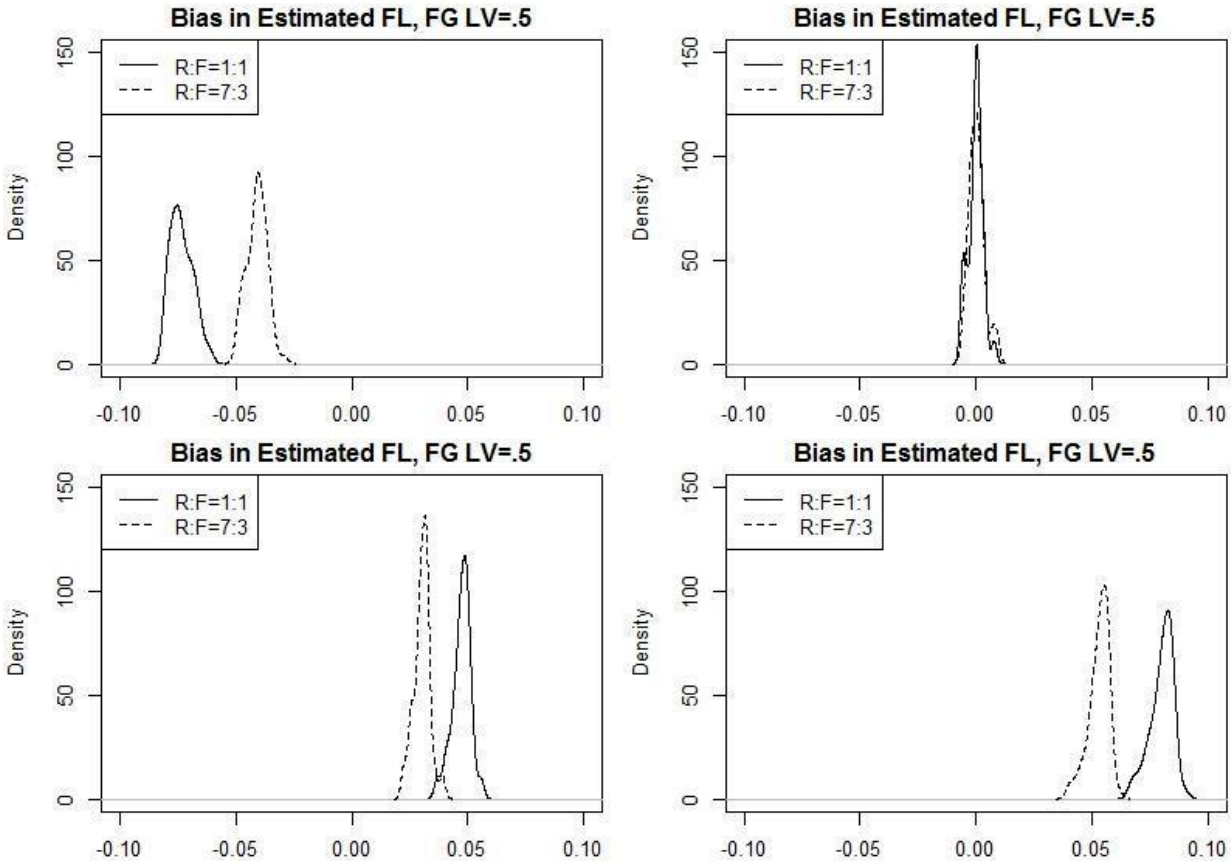
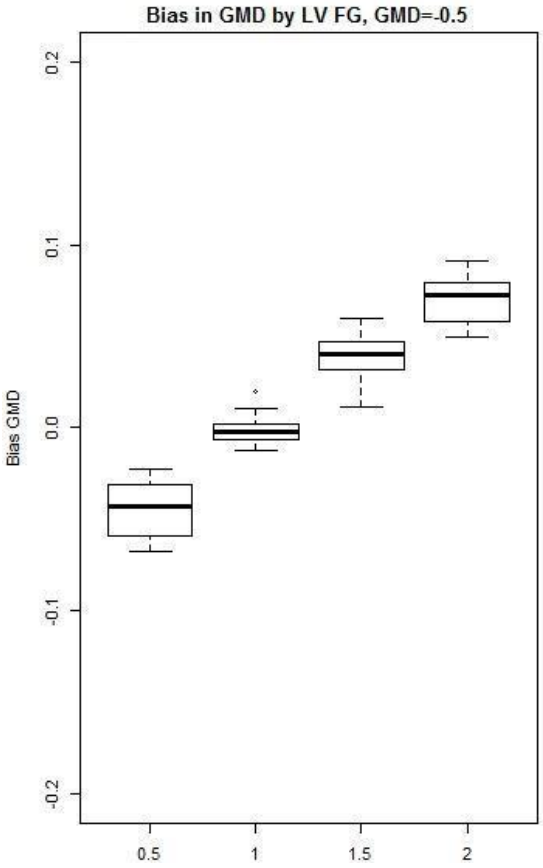
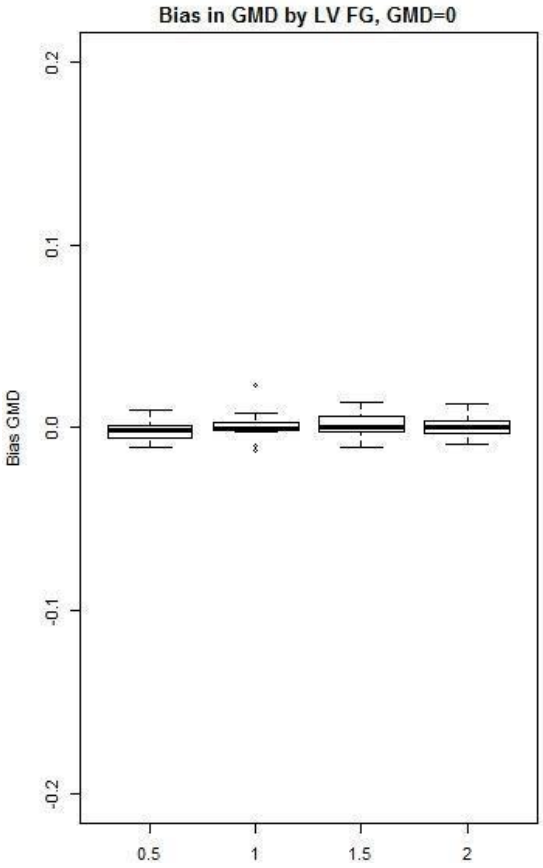

Bias in Estimated FL, FG LV=.5
R:F=1:1
R:F=7:3
Density
0
50
100
150
-0.10
-0.05
0.00
0.05
0.10
Bias in Estimated FL, FG LV=.5
R:F=1:1
R:F=7:3
Density
0
50
100
150
-0.10
-0.05
0.00
0.05
0.10
Bias in Estimated FL, FG LV=.5
R:F=1:1
R:F=7:3
Density
0
50
100
150
-0.10
-0.05
0.00
0.05
0.10
Bias in Estimated FL, FG LV=.5
R:F=1:1
R:F=7:3
Density
0
50
100
150
-0.10
-0.05
0.00
0.05
0.10

Figure 5.1



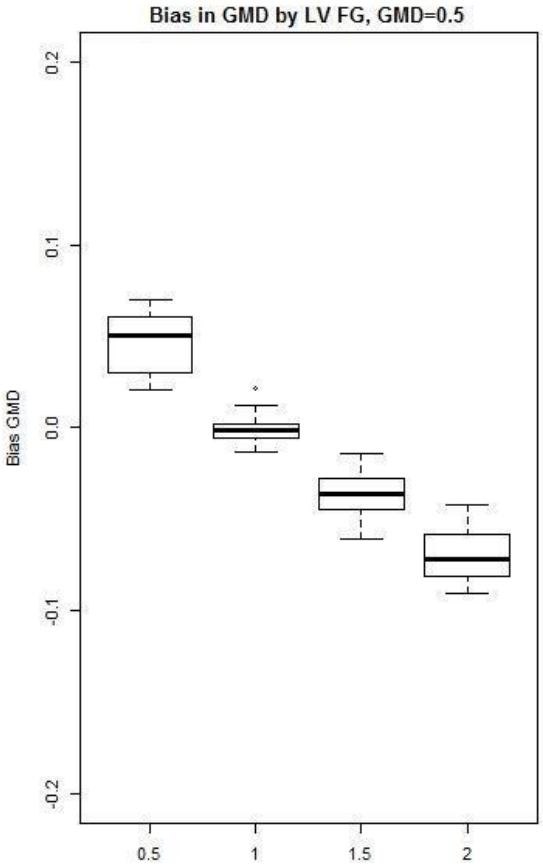
	0.5	1	1.5	2
min	-0.0677	-0.0126	0.0116	0.0494
25%	-0.0589	-0.0065	0.03145	0.0579
median	-0.04305	-0.0018	0.04045	0.07285
75%	-0.0307	0.00215	0.04675	0.07915
max	-0.0224	0.0109	0.0601	0.0912
N	24	24	24	24
MSE	0.0022	5e-05	0.0016	0.0051
Avg Bias	-0.04435	-0.00137	0.03828	0.07026

Figure 5.2



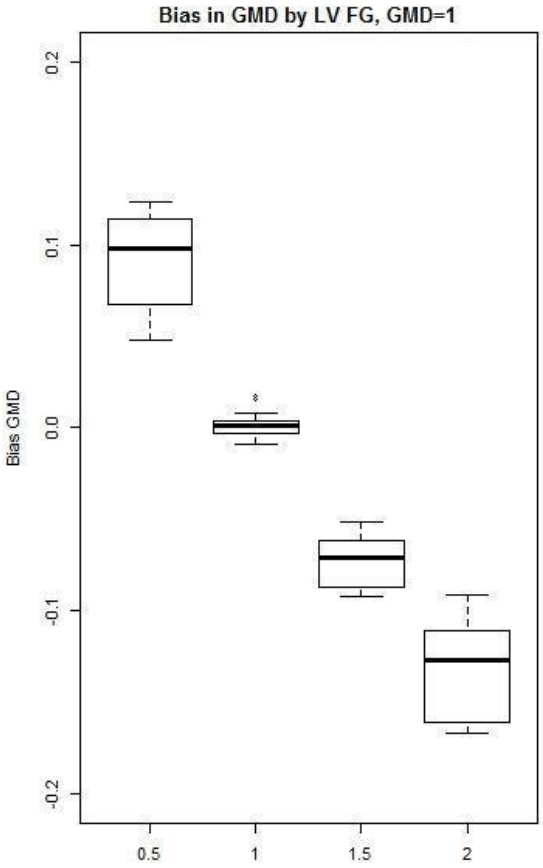
	0.5	1	1.5	2
min	-0.0105	-0.0022	-0.011	-0.0085
25%	-0.0053	-0.00145	-0.00195	-0.0026
median	-0.00115	-5e-05	0.00065	0.00075
75%	0.0016	0.00255	0.0066	0.0038
max	0.0094	0.0084	0.0141	0.0131
N	24	24	24	24
MSE	2e-05	4e-05	4e-05	2e-05
Avg Bias	-0.00143	0.00097	0.00141	0.00081

Figure 5.3



	0.5	1	1.5	2
min	0.0207	-0.0134	-0.0608	-0.0901
25%	0.0299	-0.0057	-0.04415	-0.0808
median	0.05015	-0.0012	-0.0363	-0.0713
75%	0.06045	0.0021	-0.0278	-0.0584
max	0.0701	0.0119	-0.0142	-0.0423
N	24	24	24	24
MSE	0.00233	6e-05	0.00143	0.00497
Avg Bias	0.04554	-0.00057	-0.03616	-0.06933

Figure 5.4



	0.5	1	1.5	2
min	0.048	-0.0085	-0.092	-0.1666
25%	0.06745	-0.0027	-0.08715	-0.16095
median	0.09815	0.0011	-0.07075	-0.12695
75%	0.11445	0.0036	-0.0615	-0.1107
max	0.1235	0.0084	-0.0511	-0.091
N	24	24	24	24
MSE	0.00906	4e-05	0.00561	0.01854
Avg Bias	0.09192	0.00137	-0.0735	-0.13367

Figure 6

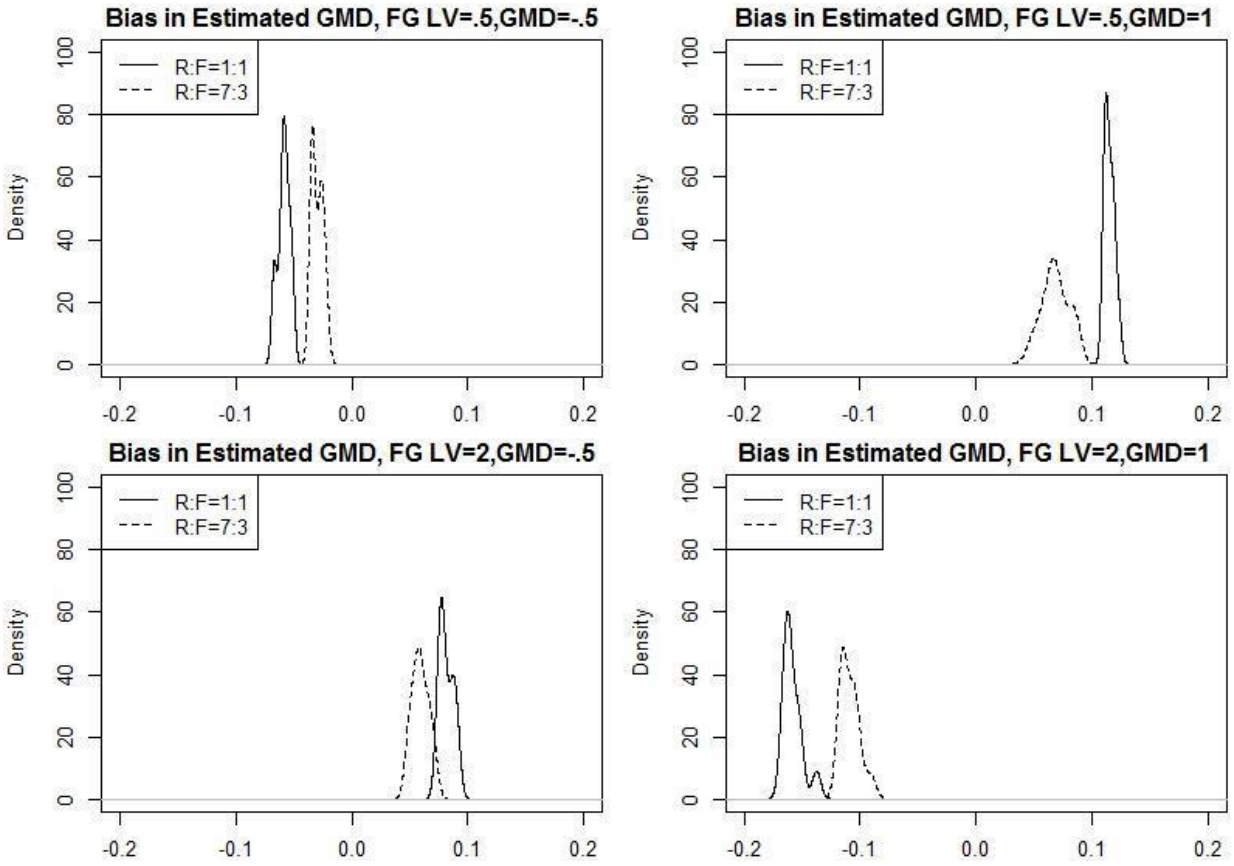
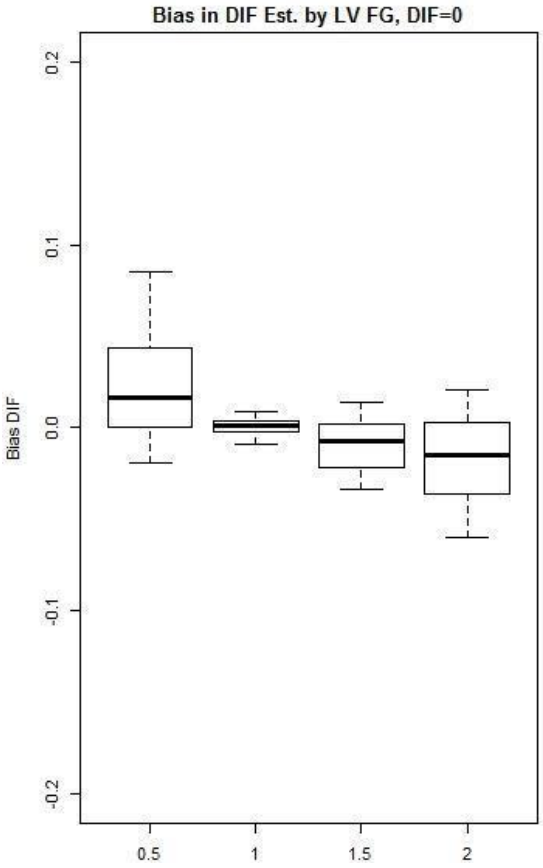
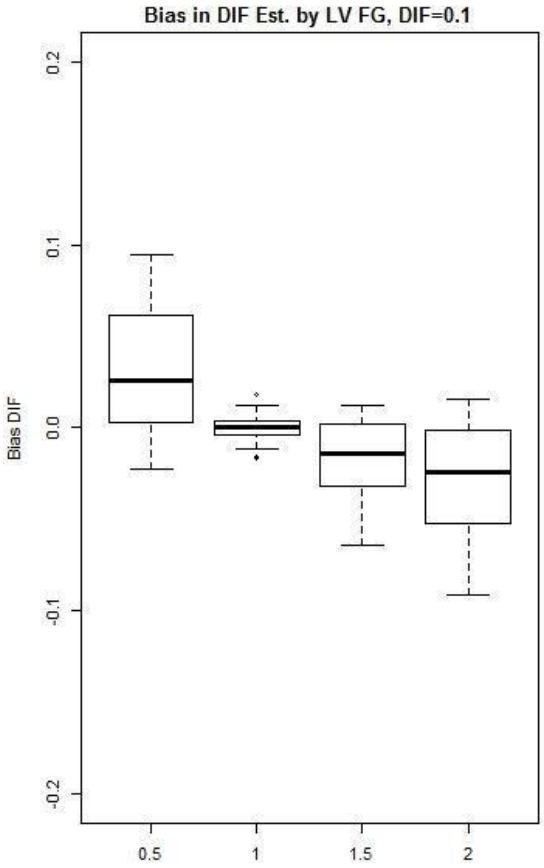

Bias in Estimated GMD, FG LV=.5,GMD=-.5
Bias in Estimated GMD, FG LV=.5,GMD=1
Bias in Estimated GMD, FG LV=2,GMD=-.5
Bias in Estimated GMD, FG LV=2,GMD=1
R:F=1:1
R:F=7:3
Density
0
20
40
60
80
100
-0.2
-0.1
0.0
0.1
0.2

Figure 7.1



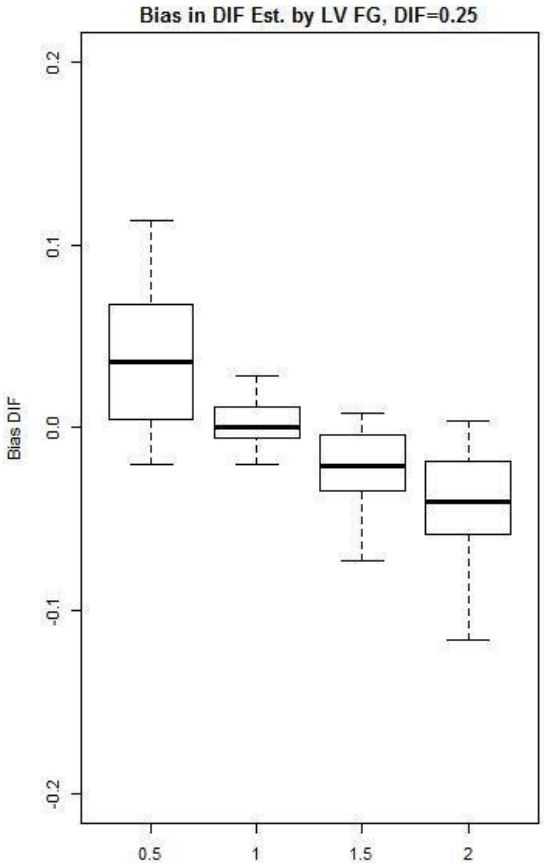
	0.5	1	1.5	2
min	-0.0195	-0.009	-0.0332	-0.0601
25%	4e-04	-0.0024	-0.0217	-0.0362
median	0.01655	0.0011	-0.0076	-0.01465
75%	0.0437	0.0036	0.0024	0.0029
max	0.0854	0.0089	0.0144	0.0204
N	30	30	30	30
MSE	0.00128	2e-05	0.00026	0.00078
Avg Bias	0.02106	0.00085	-0.00922	-0.01553

Figure 7.2



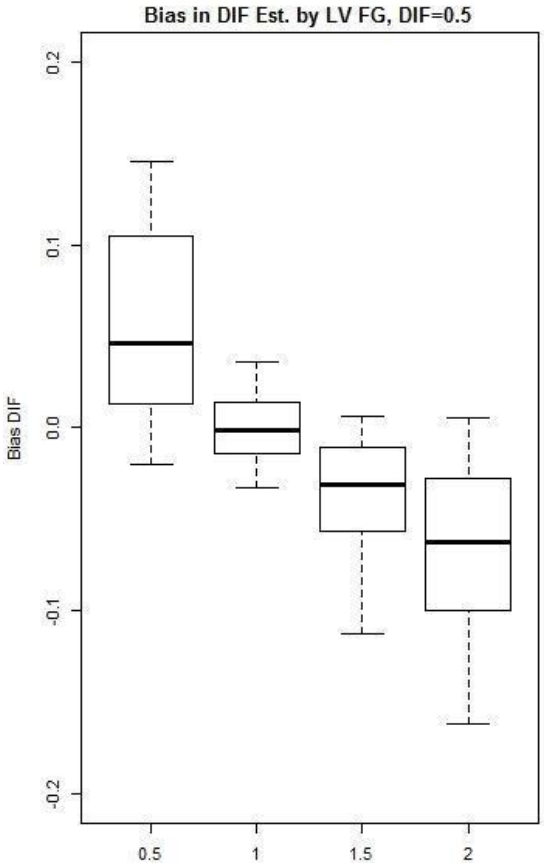
	0.5	1	1.5	2
min	-0.02268	-0.0118	-0.0639	-0.0909
25%	0.00312	-0.004	-0.032	-0.0525
median	0.02605	0.00026	-0.0144	-0.02445
75%	0.0619	0.0036	0.00222	-0.00098
max	0.0946	0.0124	0.01262	0.01543
N	30	30	30	30
MSE	0.0021	6e-05	0.00074	0.00169
Avg Bias	0.02993	0.00019	-0.01613	-0.02667

Figure 7.3



	0.5	1	1.5	2
min	-0.01978	-0.0202	-0.0728	-0.1162
25%	0.00448	-0.0053	-0.03423	-0.05842
median	0.03563	0.00082	-0.02041	-0.04031
75%	0.0676	0.0115	-0.00407	-0.01802
max	0.1131	0.0282	0.00822	0.00403
N	30	30	30	30
MSE	0.00363	0.00013	0.00117	0.00302
Avg Bias	0.0405	0.00261	-0.02456	-0.04303

Figure 7.4



	0.5	1	1.5	2
min	-0.0197	-0.0323	-0.1127	-0.1619
25%	0.0128	-0.0137	-0.0568	-0.0996
median	0.0464	-0.00085	-0.0309	-0.06275
75%	0.1049	0.0137	-0.011	-0.0274
max	0.1456	0.0362	0.0067	0.0053
N	30	30	30	30
MSE	0.00545	0.00035	0.00251	0.0064
Avg Bias	0.05289	9e-05	-0.03745	-0.06675

Figure 7.5

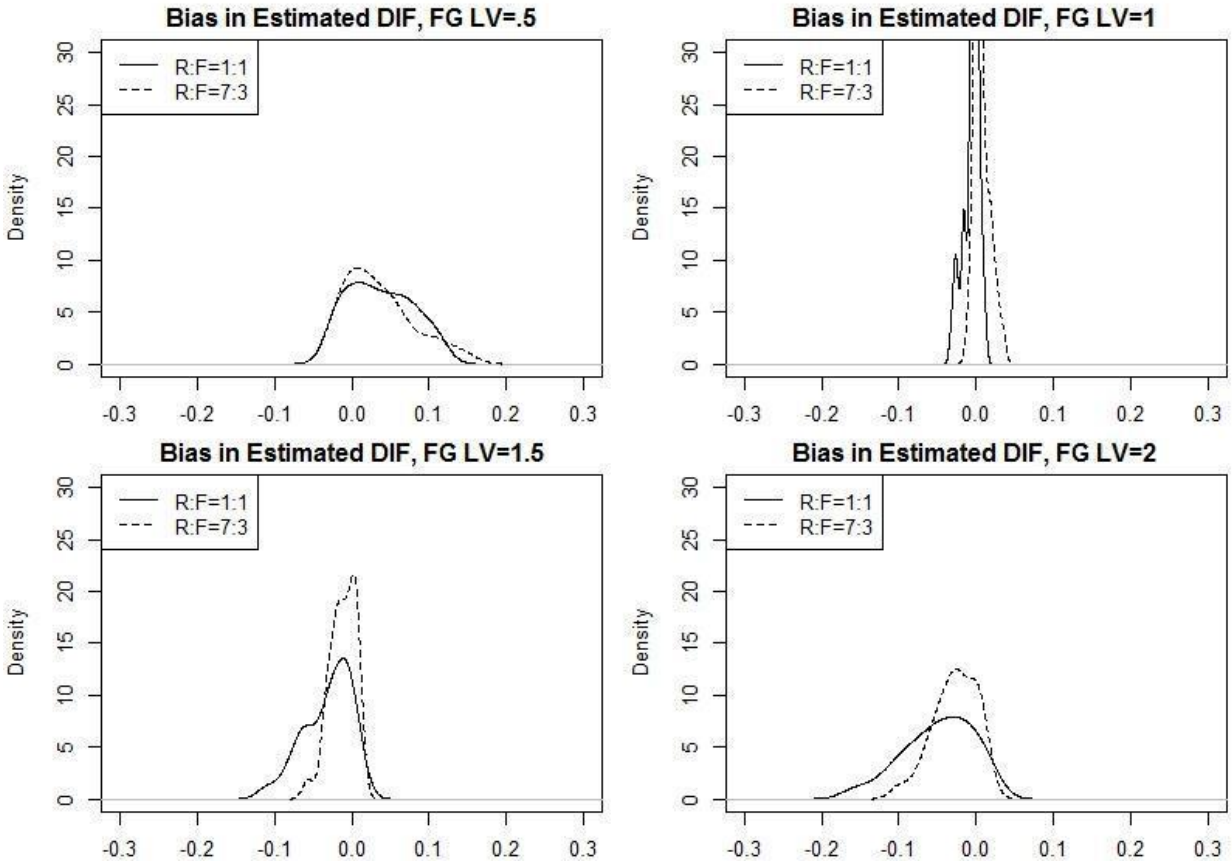

Bias in Estimated DIF, FG LV=.5
Bias in Estimated DIF, FG LV=1
Bias in Estimated DIF, FG LV=1.5
Bias in Estimated DIF, FG LV=2
R:F=1:1
R:F=7:3
Density
0
5
10
15
20
25
30
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3

Figure 8.1

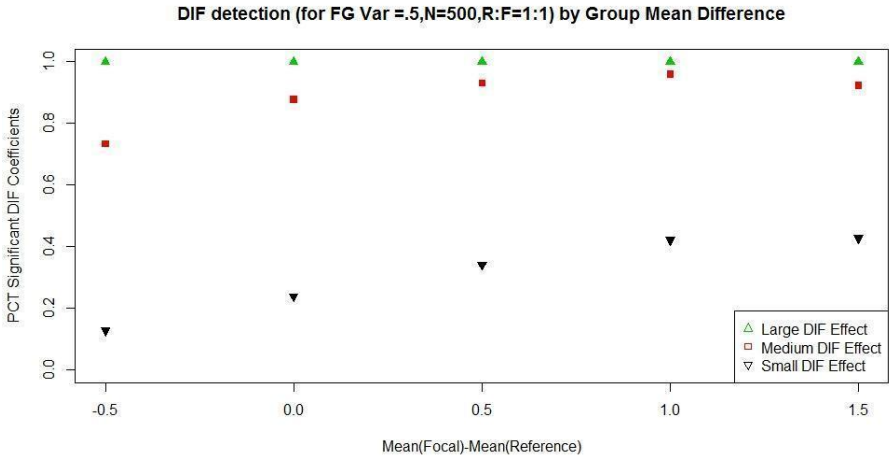

DIF detection (for FG Var =.5,N=500,R:F=1:1) by Group Mean Difference
PCT Significant DIF Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
Mean(Focal)-Mean(Reference)
Large DIF Effect
Medium DIF Effect
Small DIF Effect

Figure 8.2

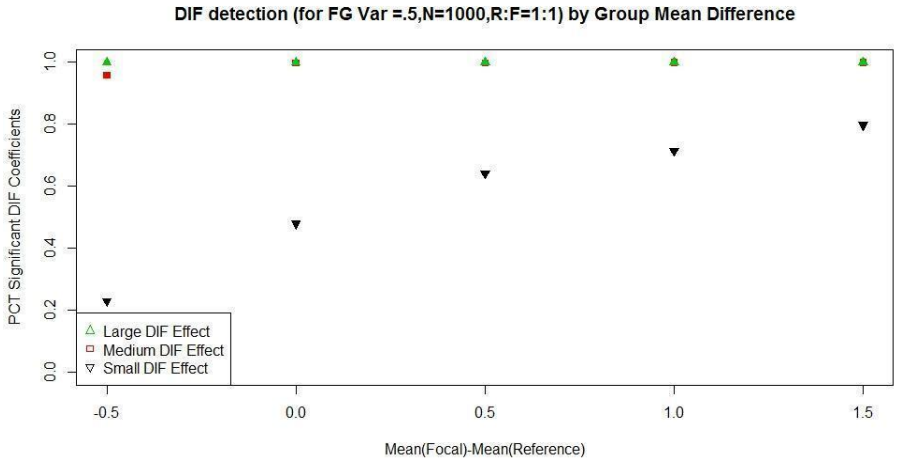

DIF detection (for FG Var =.5,N=1000,R:F=1:1) by Group Mean Difference
PCT Significant DIF Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
Large DIF Effect
Medium DIF Effect
Small DIF Effect
-0.5
0.0
0.5
1.0
1.5
Mean(Focal)-Mean(Reference)

Figure 8.3

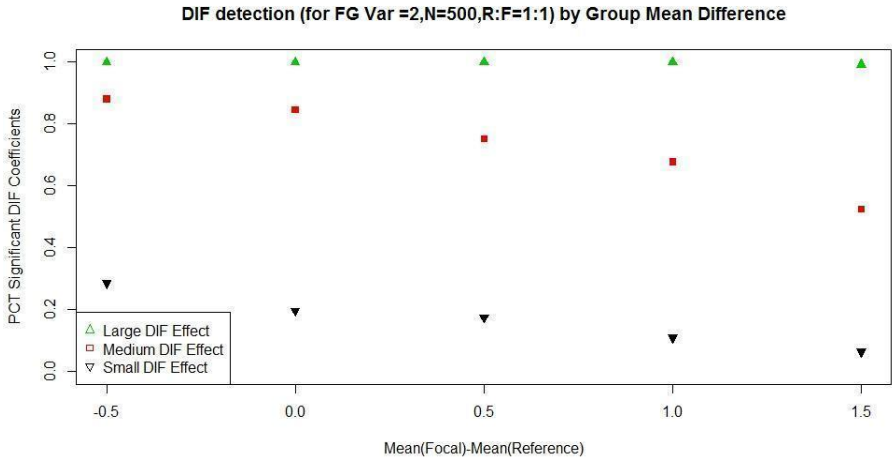

DIF detection (for FG Var =2,N=500,R:F=1:1) by Group Mean Difference
PCT Significant DIF Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
Large DIF Effect
Medium DIF Effect
Small DIF Effect
-0.5
0.0
0.5
1.0
1.5
Mean(Focal)-Mean(Reference)

Figure 8.4

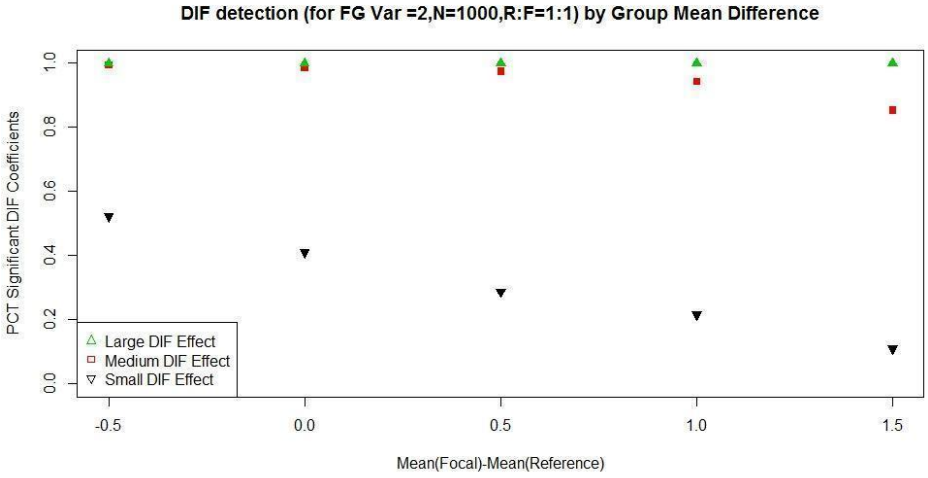

DIF detection (for FG Var =2,N=1000,R:F=1:1) by Group Mean Difference
PCT Significant DIF Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
Large DIF Effect
Medium DIF Effect
Small DIF Effect
-0.5
0.0
0.5
1.0
1.5
Mean(Focal)-Mean(Reference)

Figure 8.5

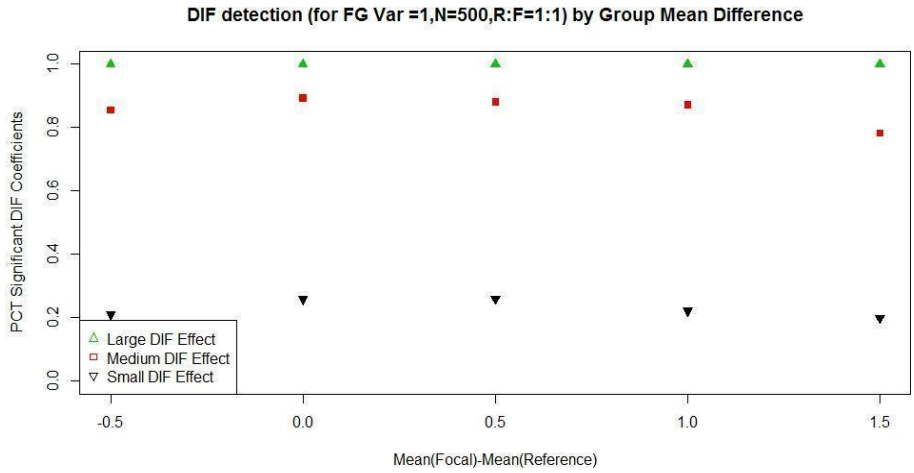

DIF detection (for FG Var =1,N=500,R:F=1:1) by Group Mean Difference
PCT Significant DIF Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
Large DIF Effect
Medium DIF Effect
Small DIF Effect
-0.5
0.0
0.5
1.0
1.5
Mean(Focal)-Mean(Reference)

Figure 8.6

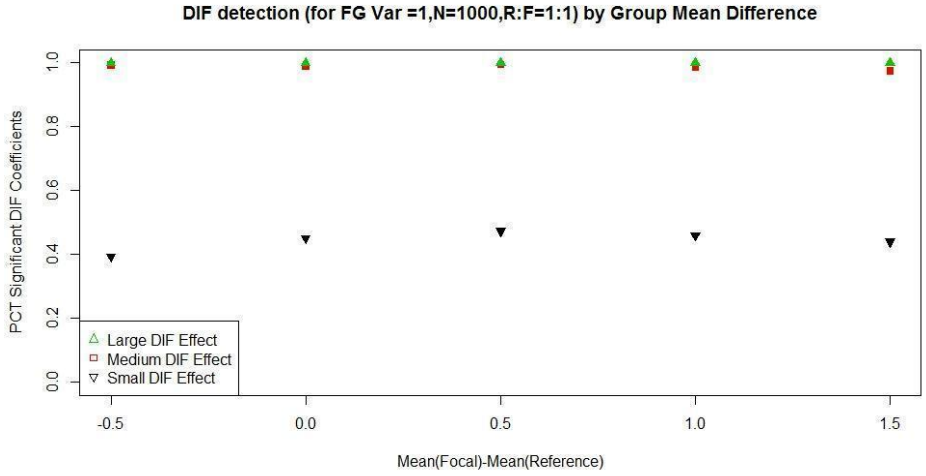

DIF detection (for FG Var =1,N=1000,R:F=1:1) by Group Mean Difference
PCT Significant DIF Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
Large DIF Effect
Medium DIF Effect
Small DIF Effect
-0.5
0.0
0.5
1.0
1.5
Mean(Focal)-Mean(Reference)

Figure 9.1, 5 Items

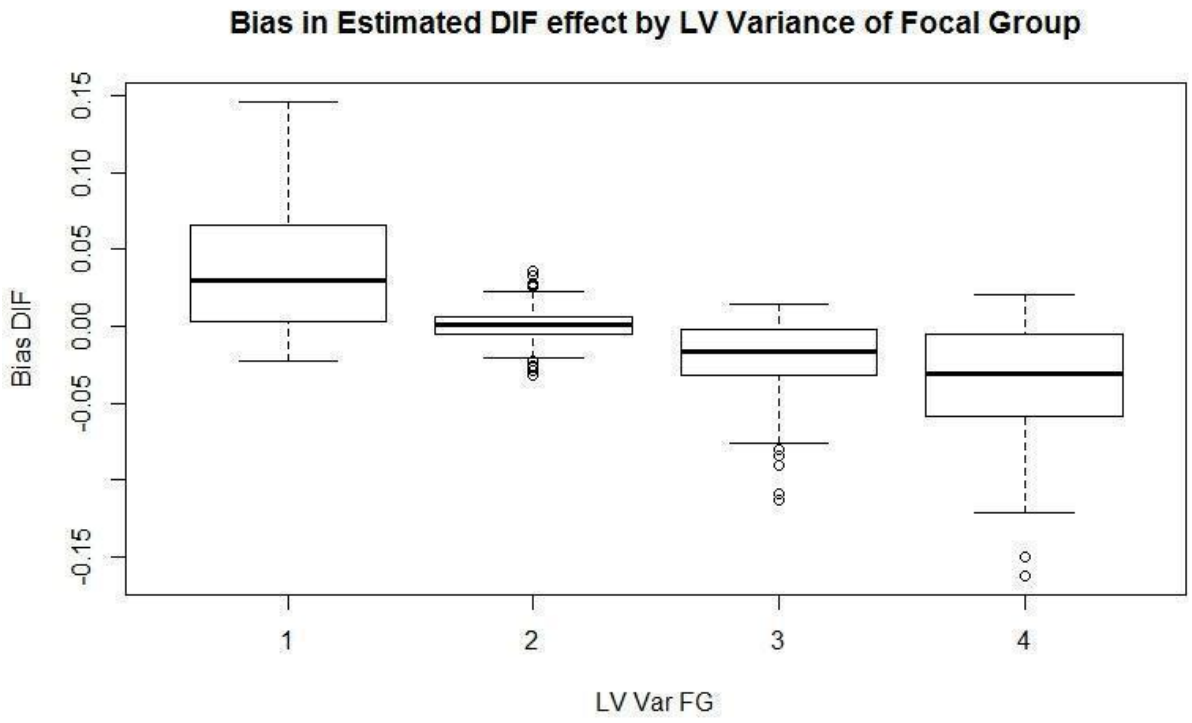


Figure 9.2, 15 Items

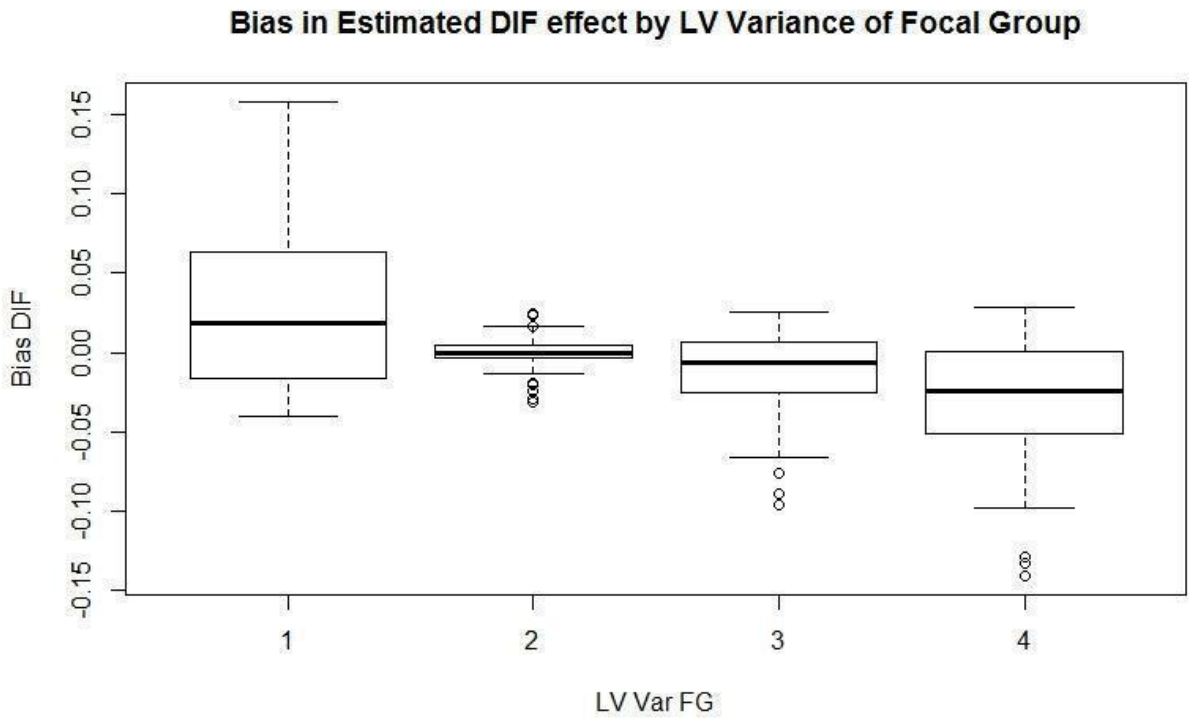
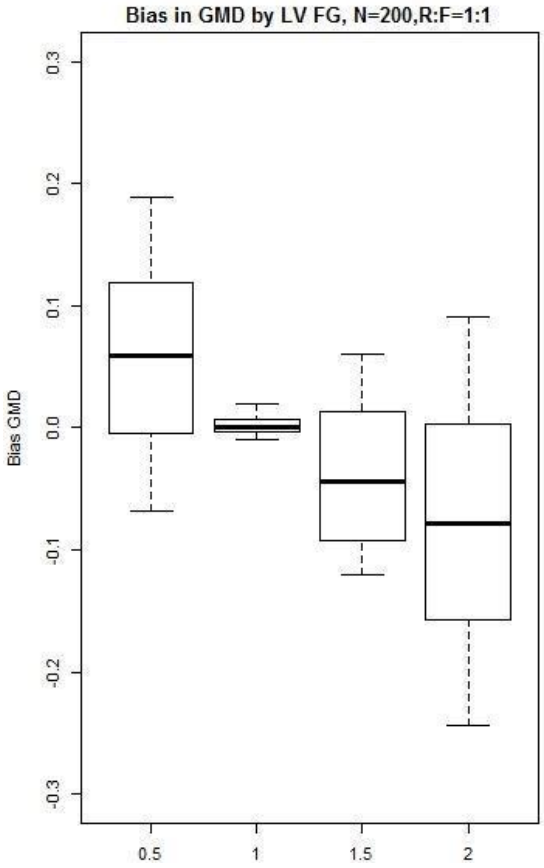
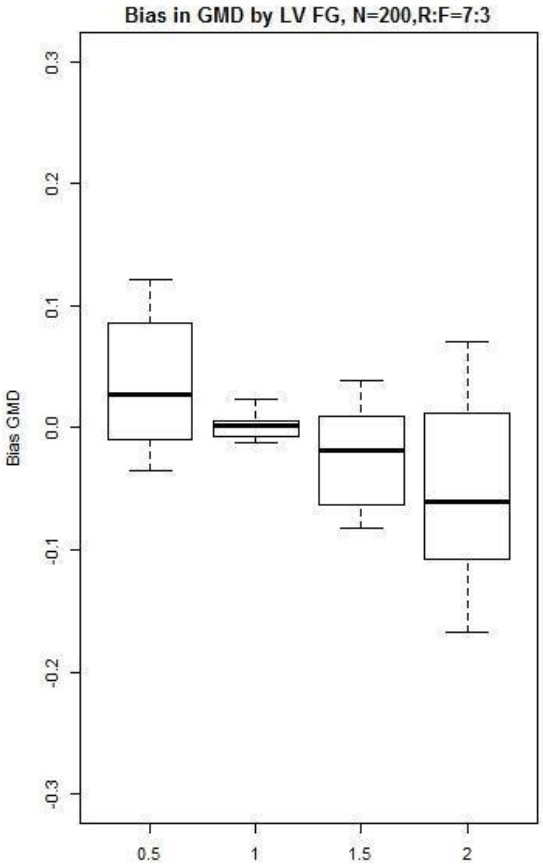


Figure 10.1



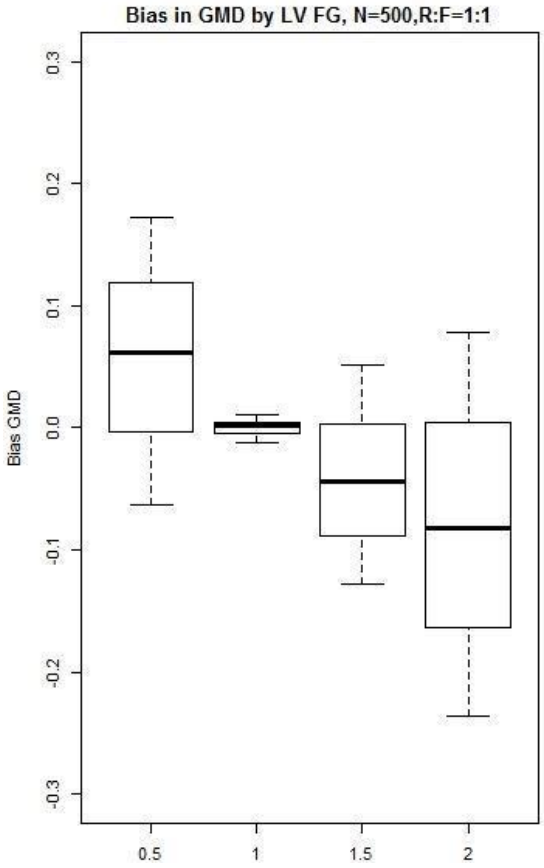
	0.5	1	1.5	2
min	-0.0677	-0.0094	-0.12	-0.2439
25%	-0.00465	-0.0036	-0.092	-0.15755
median	0.05945	0.00105	-0.04435	-0.0778
75%	0.1193	0.0064	0.0129	0.0031
max	0.1884	0.02	0.0601	0.0912
N	20	20	20	20
MSE	0.01051	6e-05	0.00519	0.01814
Avg Bias	0.05836	0.00188	-0.0381	-0.07528

Figure 10.2



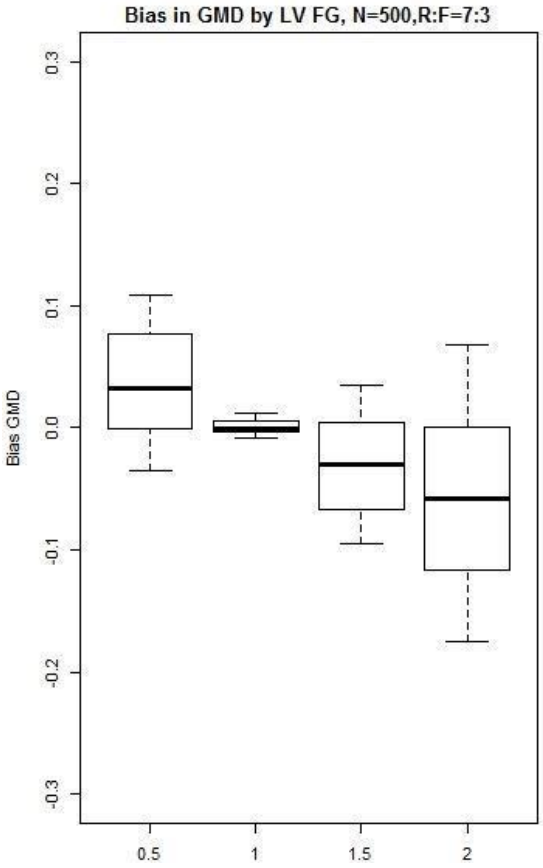
	0.5	1	1.5	2
min	-0.0345	-0.0126	-0.0824	-0.1668
25%	-0.00965	-0.0066	-0.06335	-0.10695
median	0.0273	0.00185	-0.01885	-0.0603
75%	0.0857	0.0058	0.0095	0.012
max	0.1215	0.0231	0.0388	0.0711
N	20	20	20	20
MSE	0.00395	1e-04	0.00214	0.00894
Avg Bias	0.03541	0.002	-0.0259	-0.05216

Figure 10.3



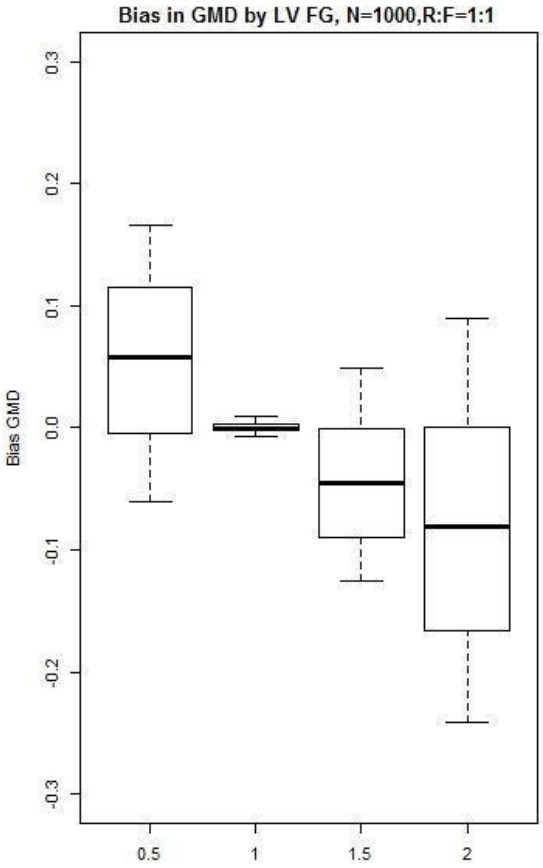
	0.5	1	1.5	2
min	-0.063	-0.0122	-0.1277	-0.2365
25%	-0.0029	-0.00485	-0.08895	-0.1641
median	0.06155	0.00135	-0.0435	-0.08205
75%	0.11955	0.00385	0.0036	0.0042
max	0.1723	0.0109	0.0521	0.078
N	20	20	20	20
MSE	0.00938	4e-05	0.0055	0.01824
Avg Bias	0.05711	-0.00017	-0.04212	-0.07895

Figure 10.4



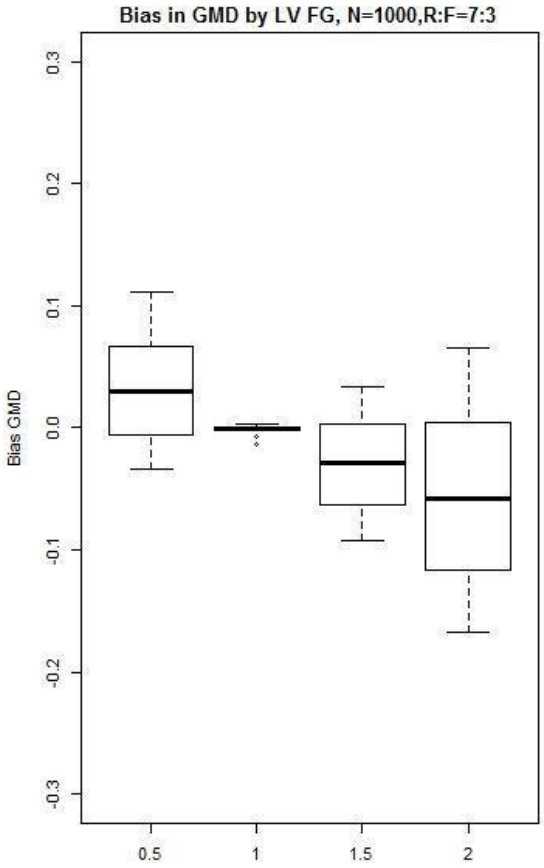
	0.5	1	1.5	2
min	-0.0352	-0.0082	-0.0943	-0.1755
25%	-6e-04	-0.0032	-0.06625	-0.1162
median	0.0321	-0.0011	-0.0297	-0.05795
75%	0.0773	0.00635	0.0042	0.00105
max	0.1082	0.0122	0.0347	0.0683
N	20	20	20	20
MSE	0.00373	4e-05	0.00271	0.00933
Avg Bias	0.03563	0.00086	-0.03052	-0.0556

Figure 10.5



	0.5	1	1.5	2
min	-0.06	-0.0069	-0.1248	-0.2406
25%	-0.0041	-0.0023	-0.0894	-0.1655
median	0.05745	-6e-04	-0.04555	-0.0808
75%	0.11455	0.0028	-5e-04	0.00025
max	0.1663	0.009	0.0484	0.0892
N	20	20	20	20
MSE	0.009	1e-05	0.00542	0.01876
Avg Bias	0.05426	8e-05	-0.04234	-0.07911

Figure 10.6



	0.5	1	1.5	2
min	-0.034	-0.0021	-0.0924	-0.1669
25%	-0.0053	-0.0019	-0.0627	-0.1158
median	0.0305	-8e-04	-0.0289	-0.0584
75%	0.06735	0.0012	0.0037	0.0044
max	0.1114	0.0036	0.0337	0.0649
N	20	20	20	20
MSE	0.00332	1e-05	0.00256	0.0094
Avg Bias	0.03328	-0.00099	-0.02817	-0.05492