The Dissertation Committee for Hwa Young Lee
certifies that this is the approved version of the following dissertation:

Evaluation of Two Types of Differential Item Functioning in Factor Mixture Models with Binary Outcomes

Committee:

S. Natasha Beretvas, Supervisor

Barbara Dodd

Gary Borich

Tiffany Whittaker

Daniel A. Powers

Evaluation of Two Types of Differential Item Functioning in Factor Mixture Models with Binary Outcomes

by

Hwa Young Lee, B.B.A.; M.A.Psy

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2012

Dedicated to my parents, Duk-Keun Lee and Jai-Sun Oh

Acknowledgements

Praise Lord! First and foremost, I am most grateful to Almighty God. He allowed me to do all things possible. “For the joy of the LORD is your strength”—Nehemiah 8:10.

I would like to express how much I appreciate my advisor, Dr. Tasha Beretvas. Without her amazing advising, I would not have been able to finish my dissertation. Whenever I struggled with something that I could not figure out, she always patiently guided me in the right direction. All the while I have worked with her, I have learned so many valuable things from her insight and guidance. It was the greatest luck to have her as my advisor. Dr Beretvas, I really, truly appreciate the mentoring and support you have given to me.

I would also like to thank all other members of my committee—Drs. Barbara Dodd, Gary Borich, Tiffany Whittaker, and Daniel Powers—for their careful review, insightful comments, and valuable suggestions to improve my dissertation.

Most importantly, I would like to thank my wonderful husband Dr. Jae Hak Jung and my lovely daughter Joann Eun-Suh Jung. Thank you so much for your understanding, support, and prayer, offered with all your heart. You have guided me with wonderful advice and have supported me in every way you possibly can. I really appreciated your insightful questions, which I received even from my daughter, at six years old, while I practiced my presentation in front of you. You are amazing! You helped me focus on my dissertation. I could not have finished this journey without your love and support.

I cannot thank my parents enough with mere words. My parents have always prayed for me and supported me. They have always said, “My daughter is the best!” Dear Mom and Dad, I love you so much, and I am really proud that I am your daughter.

I would especially like to thank my pastor David Kim and his wife Michelle Kim. You are amazing! You have shown me how to live like a Christian. I really appreciate that whenever I have asked you to pray for me, you have done so with all your heart. You have comforted me and encouraged me always. If I had not met you in Austin, I would not have lived such a happy life here. To me you are like foster parents even though you are very young. I am so lucky to have met with you in Jesus' name in Austin. I am happy that I am Jesus' disciple with you.

I also wish to thank Mary Chamberlain for her warm, caring support and endless prayers for me and my family. Mary! I cannot find the words to express how much I appreciate you. You are the most wonderful person I have ever met. I am learning how Christians live in the world from the example of your life. And I thank Nancy Davis who, with Mary, prayed for me while I had final orals. I really appreciate their support.

I cherish the support I received from my brothers and sisters in Jesus who are members of my Love Bible study group at The Great Light Presbyterian Church: Eun-Young Lee, Hye-Jeong Kim, Sohee Jo, Ara Cho, Chang-Sik Choi, Dong-Hoo Kim, Larry Gil, and Hyun-Jun Jang. You have always supported me with your food, prayer, and supportive words. And you have also entertained me. I have been really happy to be a member of the group, and I can't wait to meet with you every Friday.

Finally, I would also like to thank my brother Young-Woong Lee and his wife Ji-Na You, along with their children Ju-Un, Ju-Min, and Ju-Eun who live in South Korea. I really love them, and I miss them.

Evaluation of Two Types of Differential Item Functioning in Factor Mixture Models with Binary Outcomes

Hwa Young Lee, Ph.D.

The University of Texas at Austin, 2012

Supervisor: S. Natasha Beretvas

Differential Item Functioning (DIF) occurs when examinees with the same ability have different probabilities of endorsing an item. Conventional DIF detection methods (e.g., the Mantel-Hansel test) can be used to detect DIF only across observed groups, such as gender or ethnicity. However, research has found that DIF is not typically fully explained by an observed variable (e.g., Cohen & Bolt, 2005). True source of DIF may be unobserved, including variables such as personality, response patterns, or unmeasured background variables.

The Factor Mixture Model (FMM) is designed to detect *unobserved* sources of heterogeneity in factor structures, and an FMM with binary outcomes has recently been used for assessing DIF (DeMars & Lau, 2011; Jackman, 2010). However, FMMs with binary outcomes for detecting DIF have not been thoroughly explored to investigate both types of between-class latent DIF (LDIF) and class-specific observed DIF (ODIF).

The present simulation study was designed to investigate whether models correctly specified in terms of LDIF and/or ODIF influence the performance of model fit indices (AIC, BIC, aBIC, and CAIC) and entropy, as compared to models incorrectly specified in terms of either LDIF or ODIF. In addition, the present study examined the recovery of item difficulty parameters and investigated the proportion of replications in which items were correctly or incorrectly identified as displaying DIF, by manipulating DIF effect size and latent class probability. For each simulation condition, two latent classes of 27 item responses were generated to fit a one parameter logistic model with items' difficulties generated to exhibit DIF across the classes and/or the observed groups.

Results showed that FMMs with binary outcomes performed well in terms of fit indices, entropy, DIF detection, and recovery of large DIF effects. When class probabilities were unequal with small DIF effects, performance decreased for fit indices, power, and the recovery of DIF effects compared to equal class probability conditions. Inflated Type I errors were found for invariant DIF items across simulation conditions. When data were generated to fit a model having ODIF but estimated LDIF, specifying LDIF in the model fully captured ODIF effects when DIF effect sizes were large.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In the past half century, as use of high stakes measures including college admission tests, employment tests, and mental health inventories has increased, fairness has become a principal concern in educational and psychological testing. While fairness can be a complex construct to assess, it is fundamentally a commitment "to absence of bias and to equitable treatment of all examinees in the testing process" (AERA, APA, NCME, 1999, p. 74). Fairness clearly requires that examinees' test scores should be comparable regardless of group memberships (for example, gender and ethnicity). If performance on certain test items is easier for members in one group than in another group after controlling for ability then the test could be unfair and associated test-based inferences will be unfair.

When an item is so constructed that it performs differently on the basis of an individual's group membership, the item is considered to exhibit differential item functioning (DIF) (Dorans & Hollad, 1993; Holland &Thayer, 1988; Holland & Wainer, 1993). When the purpose of testing is to compare subgroups, the detection of DIF is particularly critical to meaningful group comparison. Thus, DIF analyses are frequently included in large-scale assessments in education and in social and health sciences (Penfield & Camilli, 2007).

Several commonly used models and associated test statistics have been developed to detect DIF as a function of membership in observed groups (like gender or ethnicity), including the Mantel-Hansel test (Holland & Thayer, 1988), the standardization method (Dorans & Kulick, 1986), the logistic regression model (for example, Swaminathan &

Rogers, 1990), the IRT-based chi-square test (Lord, 1980; Wright & Stone, 1979), the likelihood ratio test (IRT-LRT; Thissen, Steinberg, & Wainer, 1988; Wang & Yeh, 2003), and the multiple indicators multiple causes (MIMIC: Muthén, 1985; 1989) model. Typically, such approaches focus on comparing differences in items' functioning between *observed* groups. For example, researchers use pre-existing groups, such as gender or ethnicity, to investigate whether responses to some items function differently on the basis of the observed group characteristics (after controlling for ability). The approaches are based on the assumption that individuals within an observed group are more likely to be homogeneous than individuals across the observed groups (Samuelsen, 2005). However, numerous studies have suggested that using an observed group (for example, gender) that is frequently considered a source of DIF does not result in fully detecting DIF, because some unobserved or unmeasured factors can lead to DIF (for example, Cohen and Bolt, 2005; De Ayala, Kim, Stapleton, and Dayton, 2002). That is, there may also be a high level of heterogeneity within each observed group. If a researcher fails to consider heterogeneity by making an assumption of homogeneity within each observed group, it can be possible to be lead to erroneous conclusions about DIF (Samuelsen, 2005). In addition, Cohen and Bolt (2005) have cautioned that more traditional approaches do not provide information to explain why DIF occurs, because the focus is not on the dimension causing DIF but simply on the observed examinee characteristic of interest.

Recently, mixture modeling—designed to assess heterogeneity in factor structures across unobserved subpopulations—has been used for identification of DIF. Mixture modeling involves classifying examinees ex post facto into latent subpopulations as a function of examinees' response patterns rather than classifying examinees a priori into their observed groups. The unobserved sub-populations known as latent classes arise among individuals as a result of qualitative differences, for example distinctions in groups' use of different problem solving strategies, different response styles, or different levels of cognitive thinking (Samuelsen, 2005).

Within the family of mixture models, the factor mixture model (FMM) integrates both continuous and categorical latent variables in its framework. Individuals are classified into one of the latent classes, and the within-class factor structure and factor mean differences across latent classes are investigated. Because a latent class variable is unobserved in mixture models, the true number of classes is unknown. Thus, researchers should pre-specify the number of latent classes. Typically, selection of mixture models is decided based on various fit indices, such as the Akaike Information Criteria (AIC; Akaike, 1987), the Bayesian Information Criteria (BIC, Schwartz, 1978), the adjusted BIC (aBIC; Sclove, 1987), and the consistent AIC (CAIC; Bozdogan, 1987). In addition, unknown class membership is estimated based on the probabilities of individuals' most likely latent class assignment, through use of entropy and the highest posterior probability of latent class membership.

Background variables such as gender, ethnicity, or SES as covariate variables can be modeled in FMM, and modeling background variables as covariate effects in FMM helps in the interpretation of latent class membership (Lubke & Muthén, 2005; 2007). There are two kinds of covariate effects that can be specified in mixture models: class-specific and between-class. While the class-specific covariate effect explains variability in latent ability within latent classes, the between-class covariate effect explains between-class variation across classes due to the influence of the covariate on the latent class variable. That is, the class-specific covariate effect reflects a direct effect on a continuous latent factor, and the between-class covariate effect reflects an indirect effect of the covariate on the continuous latent factor (through the latent class mediator). Most studies that have investigated covariate effects in FMM have addressed only between-class covariate effects, and they have supported including even small between-class covariate effects to improve the probabilities of assigning individuals to their true classes (Lubke & Muthén, 2007).

FMMs have been extended to measure binary outcomes, and an FMM with binary outcomes—known as a mixture IRT model—can be used for identifying DIF between *latent* groups. Binary responses (0 or 1) to items that are estimated within a confirmatory factor analytic model can be compared to assess whether measurement invariance holds across latent classes. If individuals' responses to an item differ as a function of latent classes after controlling for latent ability, the item is identified as exhibiting between-class latent DIF. Many studies have found that FMMs with binary outcomes performed well in detecting sources of DIF in comparison to more traditional approaches that consider sources of DIF using pre-specified observed groups. More specifically, use of both FMMs with binary outcomes and more traditional DIF methods identified DIF items well when a source of DIF was observable, but FMMs with binary outcomes outperformed traditional methods in determining a source of DIF if the source was unobservable (Cohen & Bolt, 2005; De Ayala et al., 200; Samuelsen, 2005).

However, unless latent classes are largely separated, it is difficult to interpret the qualitative meaning of latent class memberships identified by the models using FMMs with binary outcomes. In addition, the observed response data alone might make it difficult to estimate parameters precisely, especially for complicated statistical models (Embretson, 2006; Jackman, 2011; Smit, Kelderman, & van der Flier, 1999). So, previous studies have included observed groups in using FMM with binary outcomes, and researchers have found that inclusion of observed grouping variables improved recovery of the composition of the latent classes, the recovery of item parameters (Smit et al., 1999), and detection rates for between-class latent DIF items (Maij-de Meij, Kelderman, & van der Flier, 2011). Most such studies have focused on investigating whether inclusion of observed grouping variables improved the probability of placing members in their correct class, resulting in enhanced recovery of between-class latent DIF.

Even though sources of DIF can be detected by using latent class models, there might be some variability that cannot be explained as a function of latent classes but that can be explained as a function of observed groups. Thus, including observed groups that might have different effects on some items across latent classes makes it possible to detect class-specific observed DIF. For example, Tay, Newman, and Vermunt (2010) investigated whether different item responses by respondents, controlling for latent ability, could be captured as a function of a latent grouping variable on a union citizenship scale (an eight-item test). They found that the model with two latent classes provided the best fit to the data. Additionally, they found that some items had functioned differently based on latent group memberships, but one item had functioned differently based on a latent group membership as well as an observed group membership—gender. These results suggested that including class-specific observed group can make it possible to detect different functioning based on observed group membership within latent classes. As an example of an applied study, Cho, Lee, and Kingston (2012) investigated the effect of testing accommodation on a math assessment for students with disabilities by comparing accommodated versus non-accommodated groups. Unlike findings in numerous studies that accommodation was a source of DIF, they found that latent math ability was an unobserved source of DIF (that is, between-class latent DIF) when the mixture IRT model was used. In addition, they found that accommodation was the source of DIF in only a low math-ability class, not in a high math-ability class (that is, class-specific observed DIF).

Because in real-world situations the unknown but true underlying pattern might be more complicated than the simply hypothesized pattern that contains only unobserved sources or observed sources of DIF, it is important to analyze various sets of simulated data that fit models that have between-class latent DIF (that is, LDIF) and class-specific observed DIF (that is, ODIF). However, there has been no simulation study to investigate both unobserved and observed sources of DIF by including an observed group that has class-specific effects in FMM with binary outcomes. Therefore, it is reasonable to conduct a study to evaluate models by manipulating between-class latent DIF (LDIF) and/or class-specific observed DIF (ODIF) in FMM with binary outcomes, to assess performance of fit indices, success in recovery of item parameters and latent class membership, and detection rates for between-class DIF and/or class-specific observed DIF.

Almost all simulation studies investigating DIF in mixture models have generated correctly specified models to examine how well between-class latent sources of DIF detect or how well item parameters estimate under various simulation conditions including sample size, latent class membership proportion, magnitude of DIF effect, and number of invariant and DIF items. In addition, it is assumed that an observed group is correctly specified when an observed grouping variable is modeled in FMM with binary outcomes. However, it is possible to mis-specify a model by mis-specifying an observed grouping variable's effect. Maij-de Meij et al. (2011) simulated data to examine how a mis-specified observed grouping variable influenced detection of between-class latent DIF. They included an observed grouping variable that was not associated with latent class membership, and then they compared between-class latent DIF detection rates in circumstances when an observed grouping variable was incorrectly included and when an observed grouping variable was excluded in FMM with binary outcomes. They found that even an observed grouping variable that had no effect on latent class membership influenced between-class DIF detection either positively or negatively, depending on latent class proportions. However, it is unknown how an incorrectly specified observed source of DIF influences the performance of FMMs with binary outcomes (for example, parameter recovery, DIF detection rates and correct class membership). That is, it is possible that an observed grouping variable is one of the sources of DIF, but a researcher assumes that the observed grouping variable predicts latent class membership and then helps find between-class latent DIF.

Likewise, there are many possibilities for how observed and unobserved sources of DIF in FMMs with binary outcomes might be mis-specified in real world situations. For example, both between-class latent DIF and class-specific observed DIF may exist, but a researcher may estimate a model that includes only one of these sources of DIF (between-class latent DIF or class-specific observed DIF). On the other hand, there may be only between-class latent DIF, but a researcher may estimate a model that includes both between-class latent and class-specific observed DIF. In addition, there may be only an observed source of DIF (for example, gender), but a researcher may estimate a model that has between-class latent DIF. Lastly, a researcher might estimate a model assuming that there is neither between-class latent DIF nor class-specific observed DIF when, in fact, there truly is both between-class latent DIF and/or class-specific observed DIF. Therefore, it is reasonable to examine how models that are incorrectly specified by including, excluding, or differently specifying between-class latent DIF and class-specific observed DIF impact fit indices, correct class membership, parameter recovery, and DIF detection rates (between-class latent DIF/class-specific observed DIF).

The present simulation study had four goals. First, the study evaluated models correctly or incorrectly specified by including between-class latent DIF and/or class-specific observed DIF, in terms of model fit indices such as AIC, BIC, aBIC and CAIC. Previous studies have focused on determining optimal numbers of latent classes by manipulating various simulation conditions, but the present study focused on how well fit indices perform in correctly specified models as compared to incorrectly specified models when a two-latent class FMM with binary outcomes is specified. Second, to date, no study has investigated parameter recovery of correctly identified DIF items in models that include both between-class latent DIF and class-specific observed DIF. Thus, the present study evaluated the recovery of parameters for between-class latent DIF and/or class-specific observed DIF items when models are correctly specified. Third, the study evaluated whether items are correctly or incorrectly identified as exhibiting DIF when models are correctly and incorrectly specified. While most studies have examined how well inclusion of observed groups helped find between-class latent DIF items by improving the probabilities of individuals belonging to correct classes, the present study included an observed group that has specific effects on some items in each latent class and examined how well correctly specified models detect between-class latent DIF and class-specific observed DIF items, compared to incorrectly specified models. Fourth, the study examined how well correctly specified models, compared to incorrectly specified models, correctly assign individuals into their latent class based on entropy value. To be included as manipulated conditions in the study were class probability (equal vs. unequal), between-class latent DIF effect size (small vs. large), and class-specific observed DIF effect size (small vs. large). To be summarized and compared across conditions were the relative parameter bias and standard error bias of items' difficulties for correctly identified DIF items and of DIF effects as well as the performance of information criteria in terms of correct model identification. In addition, the recovery of correct class membership was summarized and compared across conditions.

Chapter 2: Literature Review

Educational, behavioral, and social science researchers are often interested in detecting differences in one or several dependent measures for two or more groups. For example, researchers might be interested in investigating whether measure(s) of learning strategy skills are different for males versus females or among ethnic groups. To compare groups, researchers might use analysis of variance (ANOVA) or a multivariate analysis of variance (MANOVA). Often, researchers are interested in detecting group differences in an underlying *latent* variable, which is a construct that is difficult or impossible to observe directly. Using a set of items as indicators of the latent construct, researchers can make inferences about mean differences in the latent variable across groups using structural equation modeling.

One of the important assumptions made when testing mean differences is measurement invariance, which means that the same construct is being measured equivalently across different populations. Examinees who have the same level of a latent construct of interest should perform equivalently on each in a set of items regardless of their group membership, defined by factors such as sex, ethnicity, or culture. However, if examinees perform differently on an item—after controlling for the construct measured by the item—as a function of group membership, then the item scores are not measurement invariant. The item is then exhibiting differential item functioning (DIF) as a function of group membership. If DIF is present in one or more test items, then the inference is that the test is measuring groups differently, thereby potentially invalidating tests of mean differences between the groups.

Typically, the group in which examinees are disadvantaged by some test items is referred to as the focal group. The other group, in which examinees are advantaged by some test items, is referred to as the reference group. For example, if a math item is more difficult for females (after controlling for ability) and less difficult for males, then the item has gender DIF.

In order to describe DIF and the types of DIF in more detail, the first section of this chapter contains a description of item response theory (IRT) for dichotomous items. Although IRT models for polytomous items can be used for DIF detection, the present study focuses solely on DIF detection methods for dichotomous items, which are widely used in large-scale assessments.

Dichotomous Item Response Theory

There are three item response theory (IRT) models that are frequently used to model dichotomous items as a function of the ability the items are intended to measure, namely: the one parameter logistic (1PL) model (Rasch, 1960; Wright, 1968), the two parameter logistic (2PL) model (Birnbaum, 1968; Lord, 1952), and the three parameter logistic (3PL) model (Birnbaum, 1968; Lord, 1952).

Figure 1 illustrates two item characteristic curves (ICCs) for items A and B that differ only in item difficulty. For example, the ability level associated with the point of inflection is lower for item A than for item B, so item B is more difficult than item A.

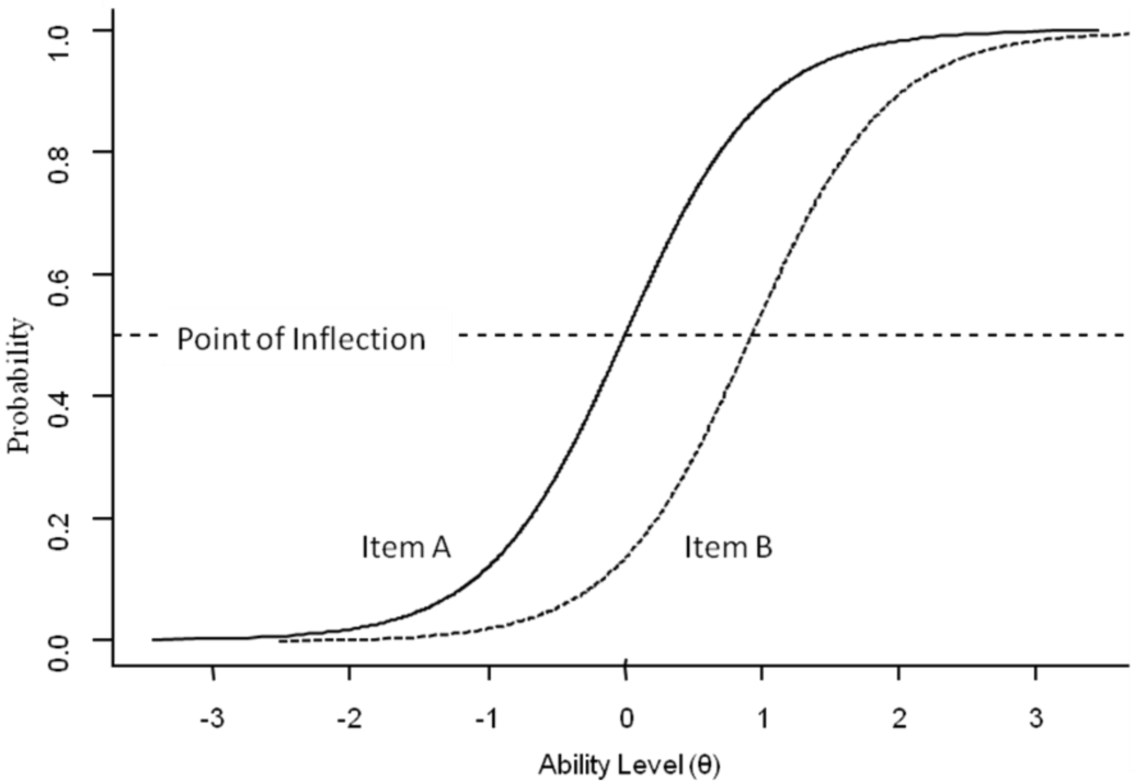

Figure 1. 1PL model ICCs for two items that differ in item difficulty

The 1PL model defines the probability of correctly responding to an item as

$$P_i(\theta) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}, \qquad [1]$$

where θ_j is the ability of examinee *j* and b_i is the difficulty of item *i.* For each item, it is possible to plot an ICC that shows the relationship between the ability scale value (θ) and the probability of responding correctly to the item. The item difficulty *b* represents the point on the θ scale that corresponds to the point of inflection of the ICC, where the slope is at a maximum. Because the 1PL model includes no discrimination parameter [that is, the discrimination parameter (*a*) is constrained to be equal across all items], the ICCs for all items have the same slope but occupy different locations along the ability scale (see Figure 1).

Figure 2 shows two ICCs for items C and D that differ only in their item discriminations. The ICCs occupy the same location along the ability scale but have different slopes. The steeper the slope is at the point of inflection, the higher the item discrimination power is (Hambleton & Swaminathan, 1985) and the better the item can distinguish between examinees who are more proficient from those who are less proficient for a given θ. Thus, in Figure 2, item D better distinguishes the abilities of individuals than item C, because it has a steeper slope at the point of inflection.

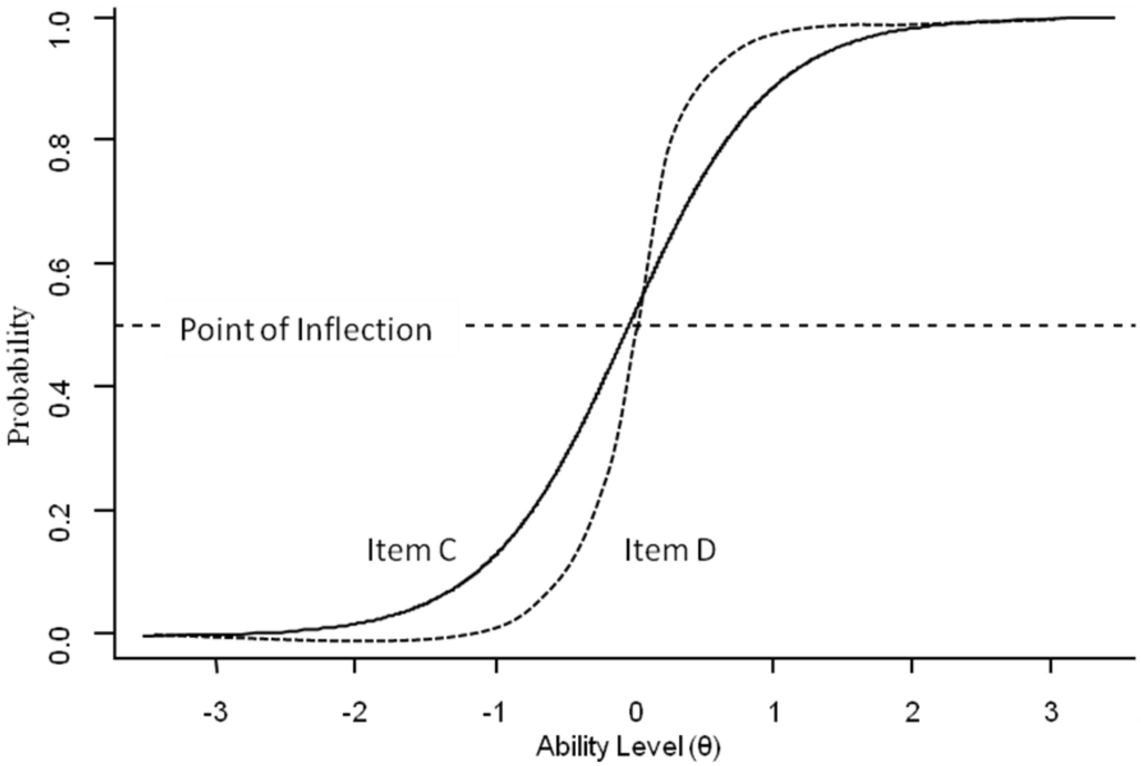

Figure 2. 2PL model ICCs for two items that differ in discrimination power

The 2PL model defines the probability of correctly responding to an item as

$$P_i(\theta) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \qquad [2]$$

where a_i is the discrimination power of item *i*. Unlike the 1PL model, the 2PL model allows each item to have unique discriminations, as indicated by the inclusion of the item-specific discrimination parameter in Equation 2. The item discrimination parameter, *a*, is proportional to the slope of the ICC at the point of inflection.

As shown in Figures 1 and 2, a non-zero guessing parameter is not assumed with the 1PL and 2PL models. Thus, the lower asymptote of each item's ICC is assumed to be zero, so that the probability of a correct response is 50% under the 1PL and 2PL models at the inflection point. If a unique, non-zero guessing parameter is modeled then the model is referred to as the 3PL model. However, a guessing parameter is not of interest in the present study and so only the 1PL and 2PL models are discussed, here.

Types of differential item functioning. If, at each ability level, the probability of endorsing an item is consistently higher or lower for one group than for another group, then the item's differential functioning is referred to as *uniform* DIF (see Figure 3). That is, the item's difficulty parameter differs for each group, but the discrimination parameter can be assumed to be the same across the groups. As can be seen in Figure 3, the probabilities of success on item 1 for group 1 are consistently higher across the ability continuum than the probability of success on that item for group 2

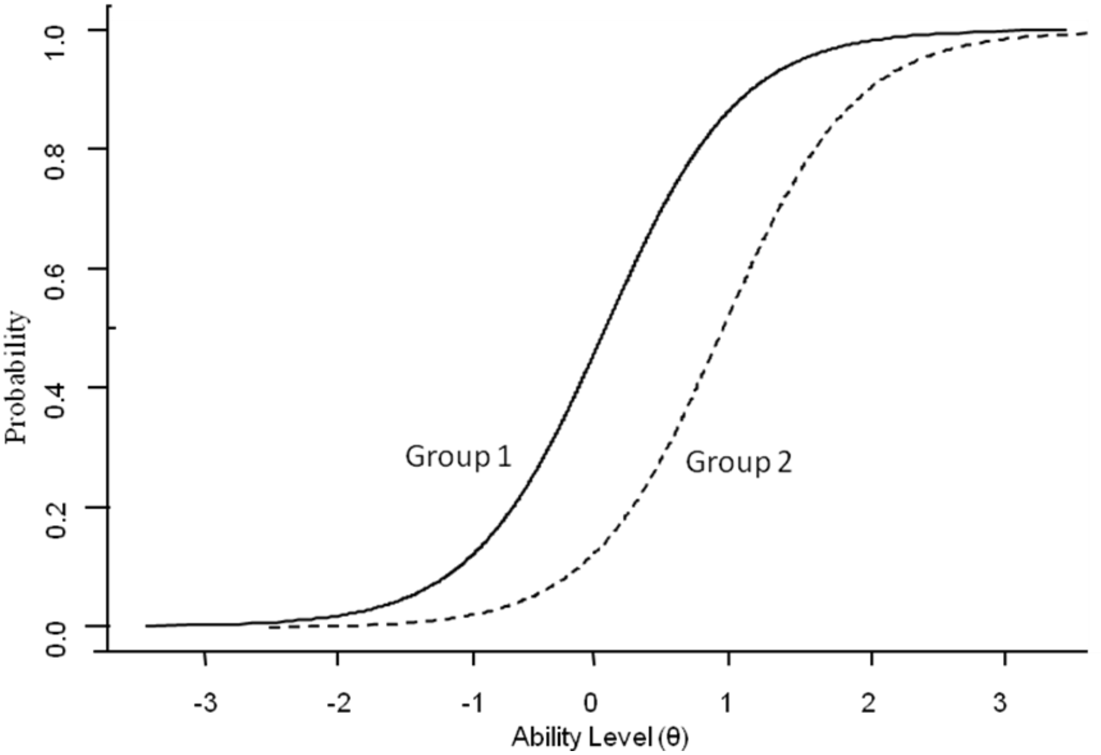

Figure 3. Uniform DIF between two groups

On the other hand, if an item's discrimination parameter differs across groups, then the item would be said to exhibit *non-uniform* DIF (see Figure 4). Under non-uniform DIF differences in the probabilities of success on item 2 for the two groups are not the same at all ability levels. As can be seen in Figure 4, the probabilities of success across ability levels on item 2 for low-ability members of group 1 are higher than the probabilities of success for low-ability members of group 2. However, the probability of success on item 2 for high-ability members of group 1 are lower than the probabilities of success for high-ability members of group 2.

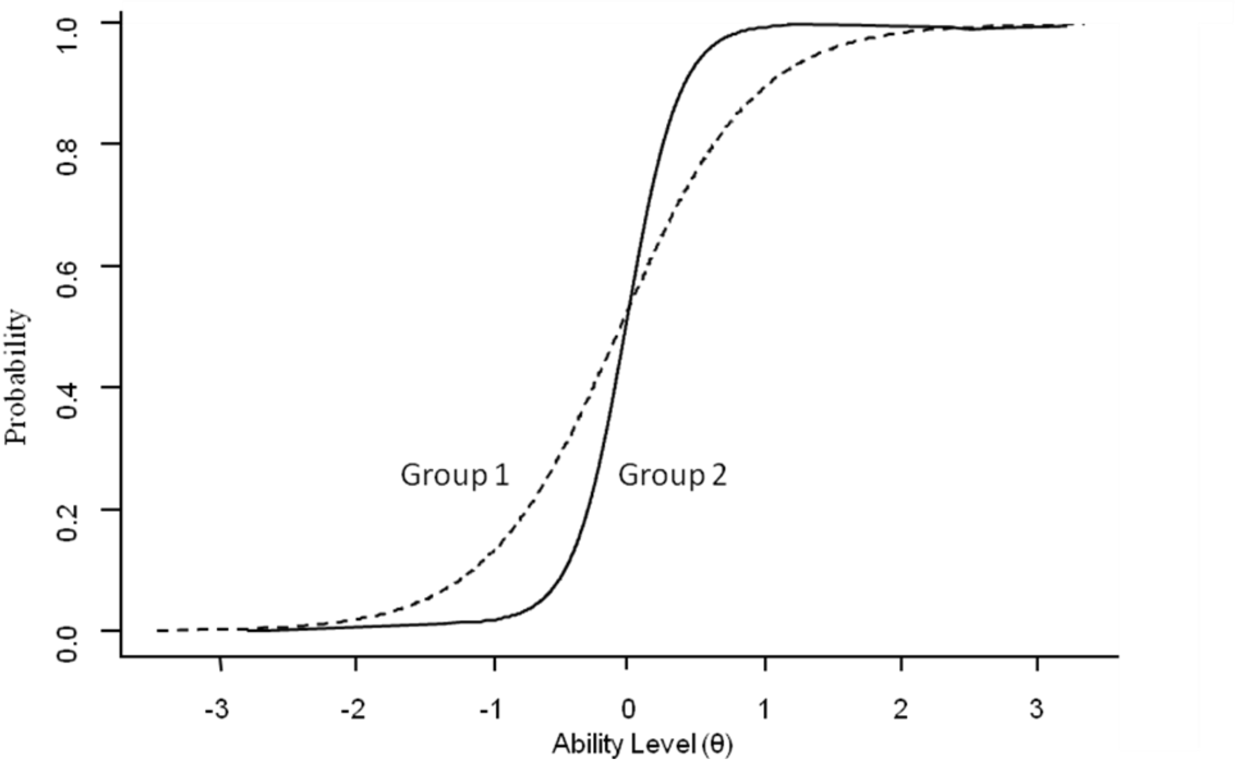

Figure 4. Non-uniform DIF between two groups

Traditional DIF Detection Methods

Many methods have been developed and adapted for identifying DIF. Approaches such as the Mantel-Haenszel (M-H; Holland & Thayer, 1988) test and logistic regression model are based on statistical models developed for categorical data. The M-H test, based on a chi-square distribution, divides the focal and reference groups into ability strata in terms of examinees' overall test scores. The approach is to estimate an odds ratio that is typically denoted by $\alpha_i = \frac{p_{ri} / q_{ri}}{p_{fi} / q_{fi}}$, where *p* represents the proportion of answering an item *i* correctly and *q* is otherwise. The subscripts *r* and *f* represent the reference and focal group, respectively. It tests the overall degree of association.

The logistic regression model for DIF (Swaminathan & Rogers, 1990) uses binary item responses (0 or 1) as outcomes and grouping variables (such as focal and reference groups), total score, and the interaction of the grouping variable and total score as independent variables. In the procedure, the main group effect provides a test for uniform DIF, and the interaction effect of group and total score provides a test for non-uniform DIF.

The M-H test and the logistic regression model use the total test score to take account of an examinee's ability. Because the observed score (that is, total score) contains measurement error, it can be problematic, especially with short scales due to low reliability (Gelin & Zumbo, 2005). Rather, Woods (2011) suggested that latent variable methods, which account for measurement error, are more likely to provide an appropriate approach to detect DIF, such as an IRT-based chi-square test (Lord, 1980; Wright & Stone, 1979), the likelihood ratio test (IRT-LRT; Thissen, Steinberg, & Wainer, 1988; Wang & Yeh, 2003), and the between-item-characteristic-curves test (Raju, 1988, 1990). These approaches test whether item parameters, conditional on ability level, are invariant across reference and focal groups. For example, in an IRT-based chi-square test described by Lord (1980), differences in item difficulty parameters across two groups of participants can be examined using the χ_L^2 statistic (Maij-de Meij, Kelderman, &Van der Flier, 2010):

$$X_L^2 = \frac{(\hat{b}_R - \hat{b}_F)^2}{V(\hat{b}_R - \hat{b}_F)}, \qquad [3]$$

where *V* is the variance of the differences in difficulty parameters of the two groups, $\hat{b}$ is the estimated item difficulty parameter, and the subscripts *R* and *F* refer to the reference group and focal group, respectively. This statistic is asymptotically chi-squared distributed with one degree of freedom. If the χ_L^2 statistic exceeds the critical value for a given level of significance, an item is said to exhibit uniform DIF.

The IRT-LRT procedure compares the likelihood associated with an item's parameters in the two cases of a model with parameter estimates constrained to be equal (that is, the reference group) and a model with parameter estimates allowed to vary (that is, a focal group). A likelihood ratio test can be conducted by computing

$$G^2 = 2\ln\left[\frac{L(A)}{L(C)}\right], \qquad [4]$$

where L(C) represents the likelihood obtained using a model with parameter estimates constrained to be equal across groups, and L(A) represents the likelihood obtained using a model with freely varying parameter estimates across groups. G^2 is distributed approximately as a chi-square variable.

Structural Equation Model (SEM) Framework as DIF Detection Method

In addition to use of the IRT framework approach for DIF identification, the structural equation model (SEM) has been recently extended for DIF identification. Several authors have demonstrated the equivalence between the IRT model's parameterization and the confirmatory factor analysis (CFA) model's parameterization. The equivalence has been derived for scenarios with dichotomous items and, thus, with categorical factor indicators. (See, for example, Baker & Kim, 2004; du Toit, 2003;

Fleishman, Spector, & Altman, 2002; Glockner-Rist & Hoitjink, 2003; MacIntosh & Hashim, 2003; Takane & de Leeuw, 1987). This correspondence means that the CFA model with binary outcomes is the equivalent of the 2PL IRT model (Muthén, Asparouhov, & Rebollo, 2006).

In this sense, the CFA model with binary outcomes can also be used for DIF detection which is more commonly referred to as measurement non-invariance in the SEM framework. In using CFA, the assumption of the equivalence of a factor's measurement across groups should hold before comparing groups' factor mean differences. This means that the assumption of measurement invariance is essential before testing latent mean differences across groups (see, for example, Lubke & Muthén, 2005; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000).

The present study is focused on assessing use of the SEM framework for DIF identification ultimately for dichotomous items. However, to facilitate explanation of the parameterization of the SEM for DIF with binary outcomes, it is easiest to start with a description of SEM with interval-scaled outcomes. Thus, the next section will briefly describe SEM with interval-scaled outcomes followed by a discussion of the use of SEM for dichotomous items' DIF identification.

Multiple Indicators Multiple Causes (MIMIC) Model

The multiple indicators multiple causes (MIMIC; Jöreskog, 1971) model in the SEM family is designed to test latent mean differences among groups under the assumption that all loadings, intercepts, and error variances are equal (that is, strict invariance). The value of the path representing the prediction of the factor mean by a grouping variable can be used to compare groups on the factor mean. In order to compare group mean differences in a MIMIC model, data from different groups are combined into a single sample.

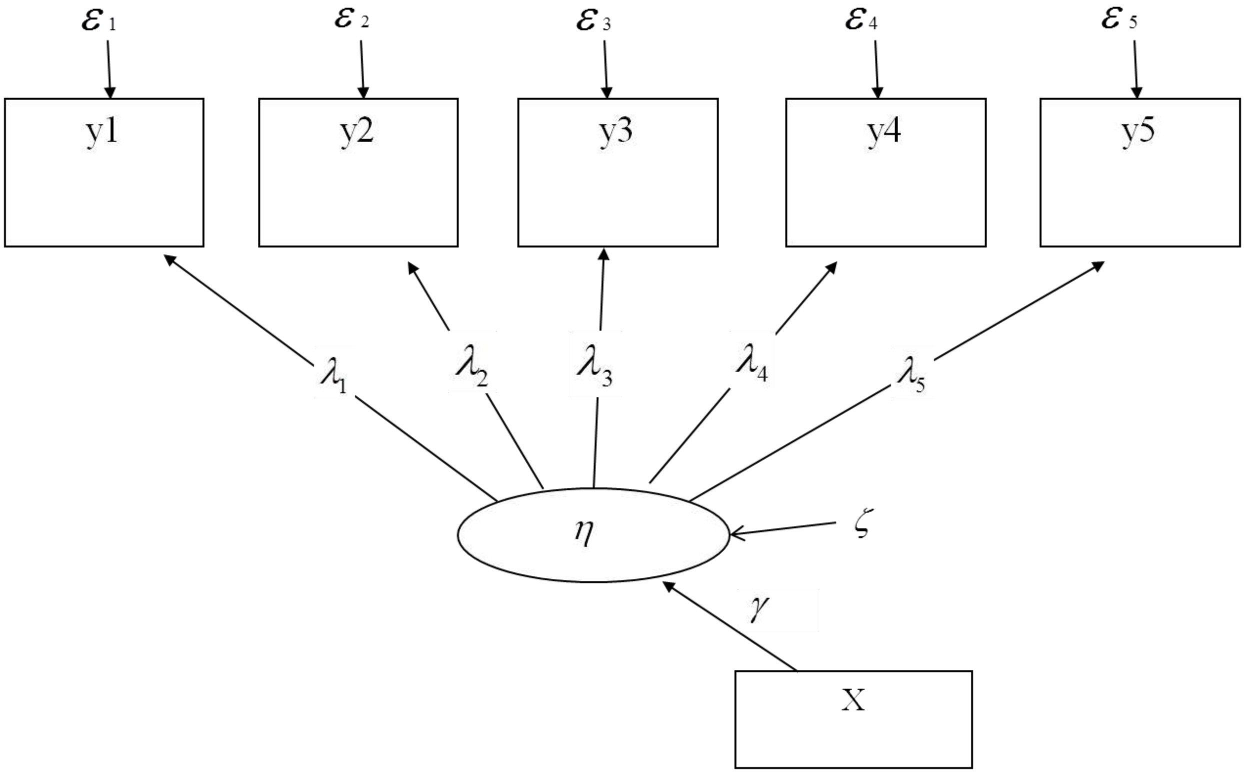

Figure 5. MIMIC model

Let us suppose that a latent construct η (for example, mathematics ability) is indicated by *p* measured variables with a total of *N* observations (Allua, 2007). The MIMIC model is formulated as follows:

$$y_i = \lambda_y \eta_i + \varepsilon_i \qquad [5]$$

where subscript *i* indexes respondents ($i = 1, \ldots, N$), y_i is a $p \times N$ matrix of indicators of the latent construct for individual *i*, λ_y is a $p \times 1$ vector of factor loadings, η_i is a $1 \times N$ vector of factor scores, and ε_i is a $p \times N$ matrix of residuals.

The latent continuous factor η is regressed on the grouping variable *X* (where, for example, *X* might represent gender) consisting of *G* (here, $G = 2$) groups, which can be expressed as

$$\eta_i = \gamma X_i + \zeta_i, \qquad [6]$$

where γ is a $1 \times (G - 1)$ vector of regression coefficients to describe group mean differences in η. X_i is a $(G - 1) \times N$ matrix of grouping variables that are dummy-coded, and ζ_i is a $1 \times N$ vector of disturbances.

Muthén (1985; 1989) popularized an adaptation of the MIMIC model (termed here, the MIMIC DIF model) to investigate measurement non-invariance (that is, DIF). Figure 6 contains a factor model in which the factor, θ (here, math ability), is regressed on *X* (gender). In IRT and DIF language, γ is often referred to as *impact* (Ackerman, 1992; Camilli, 1993). That is, impact refers to potential group differences in factor means on the construct of interest (here, θ) across groups. In contrast to impact, DIF refers to group differences in the probability of getting an item correct, conditional on the construct of interest measured by items. As shown in Figure 6, the five items in the rectangles represent observed variables designed to measure θ, here, math ability. In the model, item 5 is regressed on gender to test for gender-based DIF in responses to that item. Notice that all other factor loadings, variable intercepts and error variances are assumed equal (that is, strictly invariant) across the observed groups in the MIMIC model. Thus, if gender significantly predicts an item response's intercept (item difficulty, path A), controlling for math ability, there is evidence of uniform DIF. In other words, scores on the item cannot be assumed homogeneous across gender. Considerable literature has supported the finding that use of the MIMIC DIF model permits detection of *uniform* DIF (for example, Chen & Anthony, 2003; Christensen et al., 1999; Finch, 2005; Fleishman, Spector, & Altman, 2002; Gelin, 2005; Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Hagtvet & Sipos, 2004; MacIntosh & Hashim, 2003; Mast & Lichtenberg, 2000; B. O. Muthén, Kao, & Burstein, 1991; Oishi, 2006; Schroeder & Moolchan, 2007; Shih & Wang, 2009; Wang & Shih, 2010). Thus, Figure 6 depicts only uniform DIF in the MIMIC-DIF model.

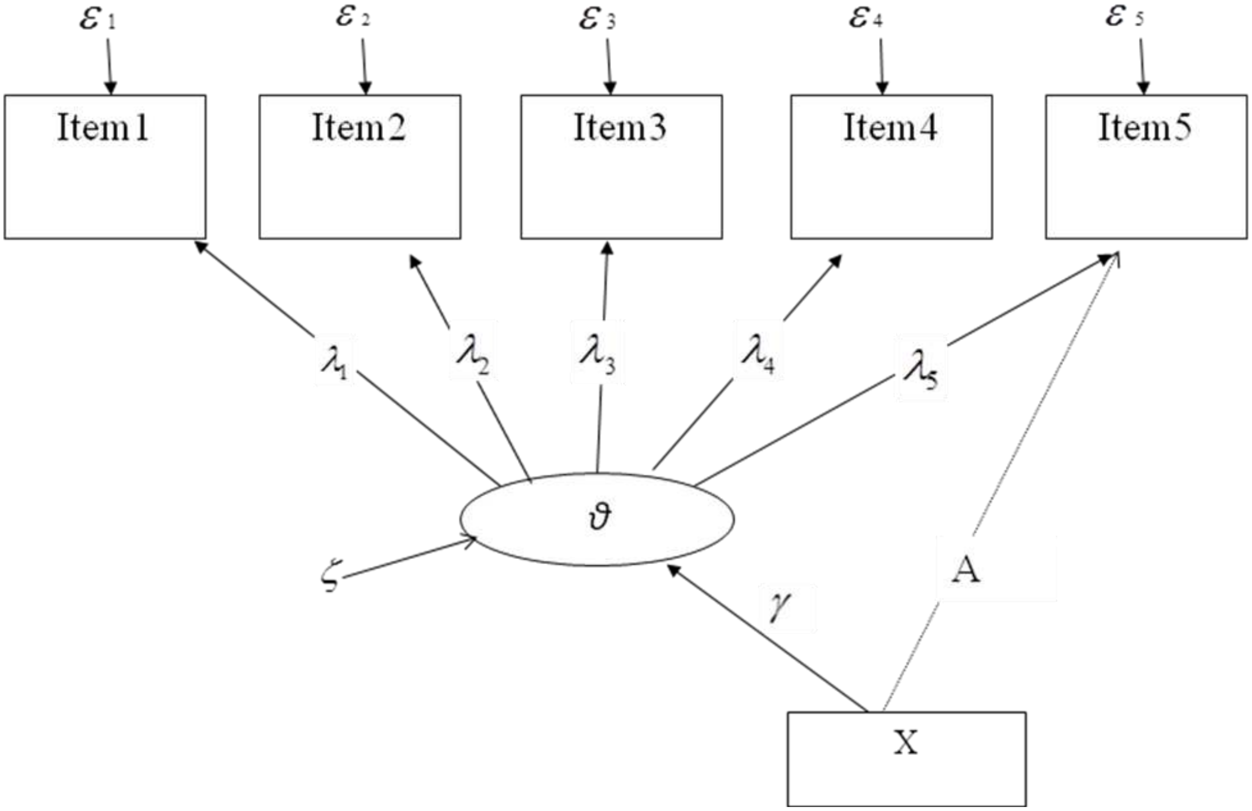

Figure 6. MIMIC DIF model to test observed uniform DIF

As mentioned above, the MIMIC model results in only one model for the combined data from both groups, so it is assumed that the same measurement model holds in both groups. However, the MIMIC model cannot be used for investigating how the factor structure and loadings might differ across observed groups. Therefore, the next section addresses the structured means model (SMM), which provides a more flexible framework for assessing potential measurement heterogeneity between multiple observable groups.

Structured Means Model (SMM)

The structured means model (SMM) permits more flexible modeling and testing of measurement invariance than under the MIMIC model. Any parameters can be estimated uniquely for each of multiple groups provided proper identification. Use of SMM does require the use of a unit-constant pseudo-variable (the “1” in a triangle appearing in Figure 7) to enable comparison of latent variable means across groups.

Figure 7 shows a model in which equality constraints (strict invariance) can be released for any of the factor loadings and variances, residual variances, and intercepts for each grouping variable, *X*. The ellipse that is shaded represents the model being estimated in each group (Hancock, 1997), and the arrow pointing from the shaded box (here, containing *X*) to the ellipse represents the grouping variable. An asterisk * beside a path from the unit-constant pseudo-variable to the latent construct within the ellipse indicates that factor mean α is freely estimated across grouping variable *X*.

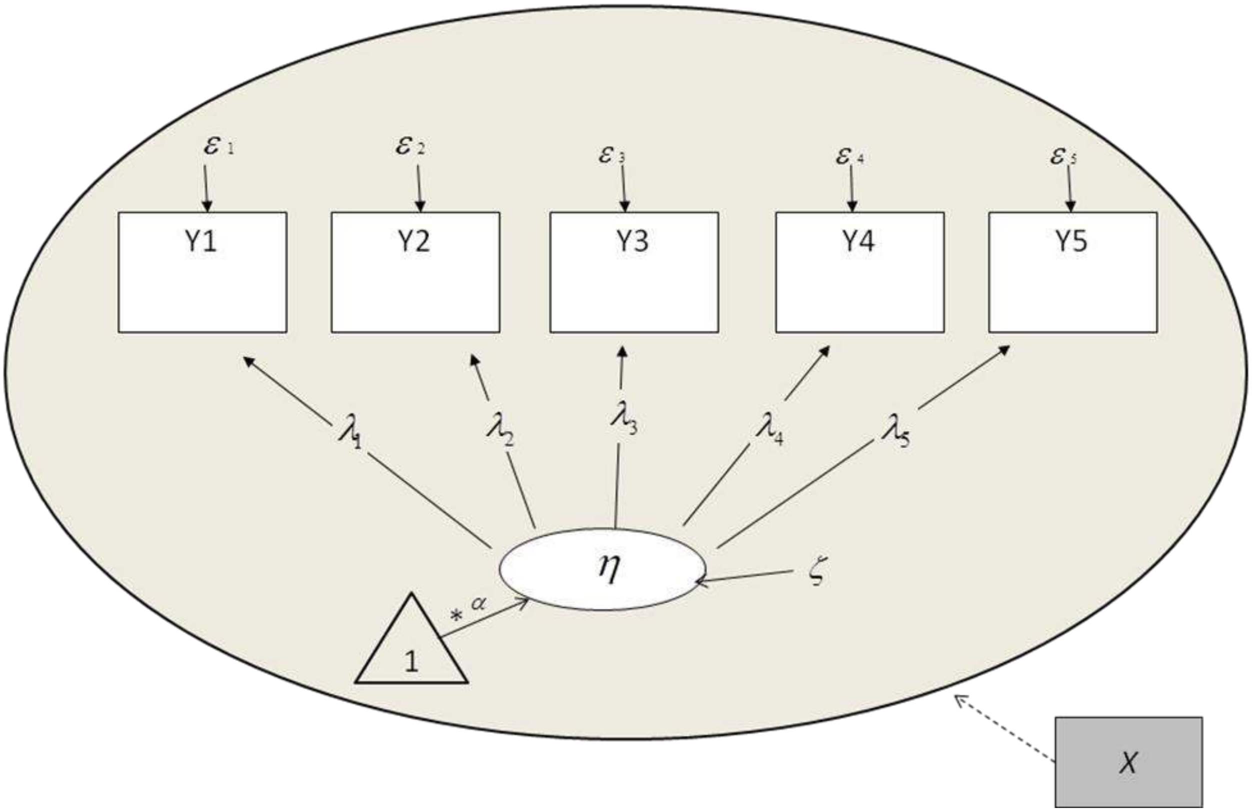

Figure 7. Structured means model (SMM)

SMM is denoted as follows:

$$y_i = v_i + \lambda_y \eta_i + \varepsilon_i, \qquad [7]$$

where

$$\eta_i = \alpha + \zeta_i, \qquad [8]$$

and where v_i is a p×1 vector of intercepts, and α is the mean of the construct (where it is assumed that only a single factor is of interest). Other matrices and vectors are the same as in Equations 5 and 6.

Under the SMM approach, the intercept (ν), factor loading (λ), and residual (ε) can be freely estimated or constrained to be equal in order to test measurement invariance across the grouping variable *X*. For this, three degrees of measurement invariance can be assessed, including: configural invariance, metric invariance, and scalar invariance. Configural factorial invariance assumes that the indicators and their pattern of factor loadings are equivalent across groups (Horn & McArdle, 1992). When configural invariance is supported, metric invariance (involving the assumption that factor loadings are equivalent across groups) can be tested. If metric invariance is supported, then scalar invariance (under which the indicators' factor loadings and intercepts can be assumed invariant) can be tested. Many researchers have asserted that scalar invariance must be supported before comparing groups' latent means (Chueng & Rensvold, 1999; Cohen & Muller, 2006; Vandenberg & Lance, 2000).

If support is not found for metric invariance, then this provides evidence of non-uniform DIF. If an item's intercept cannot be assumed invariant across groups (as is part of the scalar invariance assumption), then evidence has been found for uniform DIF.

Unobserved Heterogeneity

Both the SMM and the MIMIC model can be used to model observed group membership as a source of DIF. However, neither the SMM nor the MIMIC model can be used to identify unobservable sources of DIF. For example, Cohen and Bolt (2005) investigated gender DIF for items on a college-level mathematics placement test, using a likelihood ratio test. While under the likelihood ratio test five items exhibited DIF by gender, under mixture modeling (which will be addressed in the next section) gender was not a significant predictor of latent class membership. More specifically, while about one half of the male students were disadvantaged by items that were identified as favoring males, one half of the female students were advantaged by the same items. And De Ayala, Kim, Stapleton, and Dayton (2002) used five 20-item subtests created from a 50-item college qualification test that contained four “Black slang” items that were biased against White examinees. They found that not all White examinees were disadvantaged by those items; rather, responses to those items for some Black examinees were not different from those of White examinees of the same ability. Thus, researchers suggested that grouping individuals by observed variables such as gender or ethnicity may not fully explain who is being disadvantaged by studied items (Cohen & Bolt, 2005; Dai, 2009). That is, these applied studies have indicated that considering only an observed group(s) as a source of DIF might not provide a full picture of the source of DIF, because DIF might not be fully explained by observed group membership such as gender or ethnicity. Other unobserved sources of DIF may exist. Consistent with results obtained by Cohen and Bolt (2005) and De Ayala et al. (2002), Van Nijlen and Janssen (2008) have also found that observed groups considered to be a main source of DIF such as gender and grade were only partially related to unobserved sources of DIF.

Thus, traditional DIF detection methods that use only observed groups for DIF identification might be inappropriate if DIF occurs as a result of something that is not observable (De Ayala et al., 2002). In order to investigate a source of DIF that might be unobserved, it is necessary to develop a method to model unobserved variables for DIF identification. Therefore, the next section will introduce the factor mixture model (FMM), which seeks an unobserved variable as a source of population heterogeneity.

Factor Mixture Model

A factor mixture model (FMM) is designed to detect population heterogeneity in factor-analytic model parameters between unobserved groups (Lubke &Spices, 2008). FMM features a combination of a latent class analysis (LCA) model and a confirmatory factor analysis (CFA) model (see, for example, Kim, Beretvas, & Sherry, 2010). LCA, introduced by Lazarsfeld and Henry (1968), can be used to classify individuals into categories (or latent classes) based on observed item responses (Nylund, Asparouhov, & Muthén, 2006). Unlike factor analysis, which uses continuous latent variables (that is, factors), the LCA model uses a categorical latent variable, which is called a latent class. In addition, it is assumed that each individual belongs to only one of the latent classes, and that observed variables are mutually independent within a latent class (Goodman, 1979a). On the other hand, the CFA model involves the assumption that associations among observed variables are explained by a latent continuous factor(s), so the CFA model serves to specify a factor structure within a single homogeneous population in FMM. Thus, FMM—which involves both categorical latent variable(s) and one or more continuous latent variables—explains covariances among within-class variables and allows some or all of the factor model's parameters to vary across the classes.

Figure 8 provides a picture of a one-factor mixture model where *c* in the small ellipse represents a categorical latent variable, and the five indicators in the rectangles indicate observed variables measuring a latent factor. Whereas a conventional factor model produces estimates assuming a single population, a FMM produces estimates for each class. The dashed arrow from the latent class variable to the ellipse represents the model being estimated for each latent class. The categorical latent class variable is a predictor variable of the latent continuous factor, η. In this FMM, any of the factor loading, error variance, intercept and factor mean parameters can be modeled as equivalent or varying across classes. Under strict invariance, only the latent mean, α, is modeled as varying across latent classes. This is represented in Figure 8 by the asterisk * beside a path from the constant (one) to the factor η.

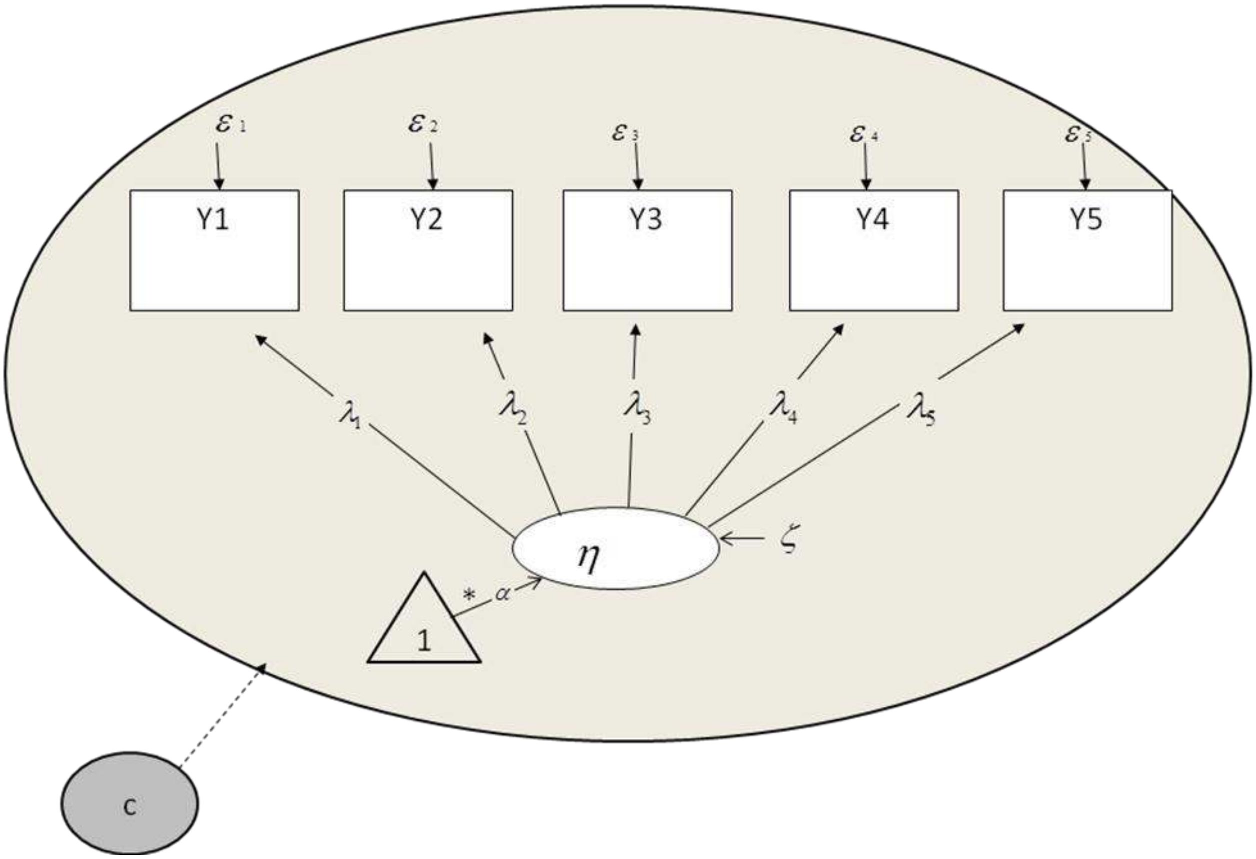

Figure 8. Factor Mixture Model (FMM)

Consider a latent continuous factor η measured by a *p*-dimensional vector of *y,* conditional on a *K*-dimensional vector *c* (representing the latent categorical variable). Here $c_i = c_{i1},...,c_{iK}$ is a multinomial variable where $c_{ik} = 1$ if individual *i* belongs to class *k* and is 0 otherwise.

The regression of observed variables on the latent (continuous) factor can be represented as

$$y_{ik} = v_k + \lambda_{yk}\eta_{ik} + \varepsilon_{ik} \, , \tag{9}$$

where subscripts *i* and *k* index individuals (*i* = 1, …, *N*) and varying parameters across classes (*k* = 1, 2, …, *K*), respectively, y_{ik} is the vector of observed indicator scores of individual *i* in class *k, v_k* is the *p* × 1 vector representing the intercepts of the observed indicators in class *k*, λ_{yk} is a *p* × 1 matrix of factor loadings for the *p* indicators, and ε_{ik} is a 1 × *N* vector of residuals in class *k*.

The regression of the latent continuous factor on the latent categorical class variable *c* is represented as

$$\eta_{ik} = \alpha c_i + \zeta_{ik} \tag{10}$$

where α indicates a (1×*K*) matrix containing intercepts of the factor or a latent mean for each latent class *k*, and ζ_{ik} is a residual vector for individual *i* in class *k*. In order to set the scale, one of the columns of α must be fixed to zero (Sörbom, 1974). The other columns of α contain the mean differences in a latent continuous factor with respect to the arbitrarily chosen reference class with a mean of zero (Lubke & Muthén, 2005).

Estimation of latent class membership. As mentioned, a categorical latent variable is called a latent class, because class membership in mixture models is unobserved, so this parameter, known as the mixing proportion φ, needs to be estimated (Gagné, 2004). Because the two-latent class model is of interest for the present study, formulation of the joint log-likelihood of the mixture model is described for only the case of two latent subpopulations existing in a population. (For a description of a more general formulation, see Gagné, 2004).

The joint log-likelihood of the two-latent class mixture model (Jackman, 2011) is expressed as

$$\prod_{i=1}^{N}[\sum_{k=1}^{2}\ln(\varphi L_{i1}+(1-\varphi)L_{i2})], \qquad [11]$$

where L_{i1} and L_{i2} represent the likelihood of individual *i* being a member of the latent class 1 and the latent class 2, respectively, φ represents the mixing proportion, and *N* is the total number of individuals. Based on Equation 11, individuals obtain a probability of being a member in each of the latent classes, and then they are assigned to the latent class for which they have the highest posterior probability of membership. Probabilities that each respondent should be assigned to their most likely class are averaged. The smaller (closer to zero) that these average posterior probabilities are, the less reliable are the model's estimated classifications by latent class. However, if the average posterior probability is close to one, the indication is that respondents are more likely to be classified into their correct class.

Entropy provides another measure of the classification utility of an FMM. Entropy is based on the average posterior probability of belonging to a class. Entropy values ranging from 0 to 1 indicate that the higher value is, the better is the classification of each respondent into each latent class. For example, if there is a two-latent class mixture model, the probability of belonging in class 1 is 1 minus the probability of belonging to class 2 and conversely. Generally, when entropy values are less than .60, more than 20% of participants are misclassified, so Lubke and Muthén (2007) suggested that an entropy value greater than .80 provides at least 90% correct class assignment.

Model fit criteria for mixture models. In conventional structural equation models, the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), the Standardized Root Mean square Residual (SRMR; Bentler, 1995), and the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980; Steiger, 1990) are typically used for evaluating the fit of a single model in comparison to a baseline model that assumes zero population covariances among the observed variables. A corresponding measure of individual model fit is not available for FMMs (Kim, 2009). Instead, other indices and statistics can be used for comparing the fit of pairs of different mixture models: (a) likelihood-based tests such as the Lo-Mendell-Rubin adjusted likelihood ratio test (aLRT; Lo, Mendell & Rubin, 2001) and a bootstrapped likelihood ratio test (BLRT; McLachlan & Peel, 2000) and (b) information criteria (IC) such as Akaike's Information Criterion (AIC; Akaike, 1987), the consistent AIC (CAIC; Bozdogan, 1987), Bayesian Information Criterion (BIC; Schwartz, 1978), and the sample size-adjusted BIC (aBIC; Sclove, 1987).

Likelihood ratio tests are used to compare nested models (and can be used with latent class models) that differ on the basis of a set of parameter restrictions. In the context of mixture models, a model with *k* classes can be compared with the corresponding model that only has *k*–1 classes. The likelihood ratio test is defined as

$$\mathrm{LR} = -2[\log \mathrm{L}_{\mathrm{restricted}} - \log \mathrm{L}_{\mathrm{unrestricted}}(\hat{\theta}_u)\ ], \qquad [12]$$

where log $\mathrm{L}_{\mathrm{restricted}}$ is the log likelihood associated with the restricted form [here, the (*k*–1)-class model] and log $\mathrm{L}_{\mathrm{unrestricted}}$ is the log likelihood associated with the unrestricted form (here, the *k*–class model). A significant aLRT test statistic indicates that the fit of the *k*-class model is better than that of the (*k*–1)-class model.

Use of the BLRT, as operationalized in M*plus* software, involves repeated generation of bootstrap samples to estimate the sampling distribution of the -2 log likelihood difference between the (*k*–1)-class and the *k*-class models. A statistically significant BLRT indicates that the *k*-class model fits better than does the (*k*–1)-class model (see Nylund, Asparouhov, and Muthén, 2007, for further details about the BLRT).

The AIC, CAIC, BIC, and aBIC are frequently used as indices for comparing the fit of non-nested models in mixture modeling (Nylund et al., 2007). The AIC is defined as

$$\mathrm{AIC} = -2 \log L + 2p, \qquad [13]$$

where *p*, here, is the number of free parameters in the model being estimated and log *L* is the log-likelihood function for the estimated model. The CAIC uses an adjusted sample size that penalizes fit for models with more parameters as a function of sample size.

The CAIC is defined as

$$\text{CAIC} = -2 \log L + p\, (\log (n) + 1). \qquad [14]$$

The BIC is defined as

$$\text{BIC} = -2 \log L + p \log (n). \qquad [15]$$

A sample-size adjusted BIC (aBIC; Sclove, 1987) uses an adjustment of the sample size used in BIC such that

$$\text{aBIC} = -2 \log L + p \log ((n + 2) / 24). \qquad [16]$$

Although several simulation and applied studies have focused on whether there is agreement among model-fit indices concerning the performance of mixture models, there is still no consensus about which model-fit index performs best in terms of consistently identifying the correct model. Nylund et al. (2007) tested the performance of several fit indices such as aLRT, BLRT, AIC, CAIC, BIC, and aBIC when used with various mixture models: the latent class analysis (LCA) model, the factor mixture model (FMM), and the growth mixture model (GMM). According to their assessment of model fit for mixture models, they found that the BLRT performed the best and that the BIC, CAIC, and aBIC performed relatively well in identifying the optimal number of classes for the FMM and GMM. On the other hand, Tofighi and Enders (2008) compared GMM under various simulation conditions using the aLRT, BIC, aBIC, AIC, CAIC, and adjusted sample size CAIC (Bozdogan, 1987). They found that the aBIC and adjusted sample size CAIC indices consistently performed well in supporting better fit of the model with the optimal number of latent classes when fitting models without covariates. In another simulation study, Li and Hser (2010) investigated whether model fit indices used with

GMM correctly identified the optimal number of latent classes when a choice is made between models correctly specified by including a covariate variable and models mis-specified in that they excluded a covariate variable. The BIC, aBIC, LMR, and BLRT were found to support models with the correct number of latent classes. The BLRT performed the worst for mis-specified models.

Factor Mixture Model with a Covariate Variable

Covariates can be included in an FMM to help identify potential predictors of latent class membership (Lubke, 2010; Lubke and Muthén, 2005; 2007). Even though the number of classes is pre-specified before estimating a model, it is unknown which individuals belong to which latent class. Inclusion of demographic variables as covariates, such as gender, SES, education level, or ethnicity as covariates, can help understand the membership of latent classes. For example, researchers may use an FMM to investigate unknown sources of heterogeneity in measuring examinees' math ability and hypothesize that two unobserved groups (latent classes) exist in the data. Then a researcher might hypothesize further that membership in the classes is partly or wholly a function of gender. Here, the gender group might predict the log odds of the probability of belonging to one of the latent classes (here, the *k* class) compared to the probability of belonging to the other latent class (K_{th} class) (that is, between-class covariate effect) (Lubke and Muthén, 2005). Otherwise, the researcher might hypothesize that gender explains some of the variability in the latent continuous factor (here, math ability) and that the gender effect differs by latent classes (that is, class-specific covariate effect). The presentation in the next section describes the potential different FMMs that include between-class or class-specific covariates.

FMM with between-class covariate effects. As mentioned above, an observed variable(s) such as gender can be included to predict latent class membership (called a *between-class covariate*). Figure 9 shows the extension of the FMM to include a between-class covariate affecting latent class (class membership).

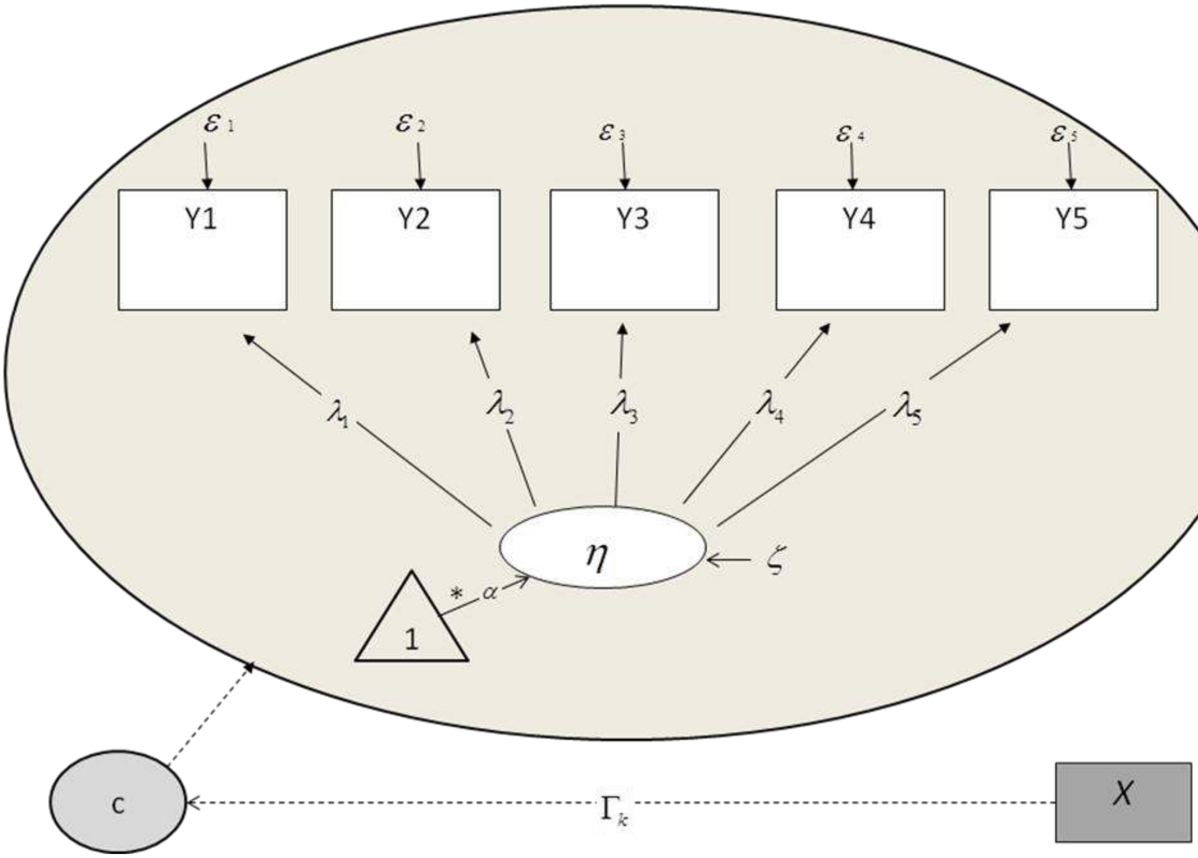

Figure 9. FMM with between-class covariate effects

As shown in Figure 9, the latent class variable (class membership) is regressed on the covariate effect *X* (for example, ethnicity), modeling that class membership is explained by ethnicity. Note that in Figure 9, the classes are distinguished only by their means on the latent variable, η. The indirect effect of the covariate, *X*, on the continuous latent variable through the latent class mediator, *c*, can be denoted as

$$Logit\ (\pi_{ik}) = \ln\left[\frac{P(c_{ik} = 1 \mid X_i)}{P(c_{iK} = 1 \mid X_i)}\right] = \alpha_k + \Gamma_k X_i\,, \qquad [17]$$

where α_k is a (*K*-1)-dimensional parameter vector for a *K*-class model and represents the regression intercept for each class, *k*, and Γ_k represents the regression weights for the covariate for each class. Equation 17 represents a multinomial logistic regression where *X* predicts the log odds of the probability of belonging to class *k* compared to the probability of belonging to the reference class, *K.* Also, the logit can be converted to an odds ratio for the purpose of the interpretation. The odds ratio can be interpreted as a measure of the odds of being a member in latent class *k* relative to latent class *K.*

Most studies investigating covariate effects in FMM have used between-class covariate effects. Lubke and Muthén (2007) conducted a simulation study that focused on whether including an observed group as a between-class covariate variable improved detection of correct latent class membership in FMM. The authors found that inclusion of between-class covariates with even small effects improved the recovery of correct class membership. More specifically, the authors manipulated the degree of class separation using the Mahalanobis Distance (MD) by varying the latent classes' factor mean differences (0.5, 1, 1.5, 2.0) and the covariates' effect sizes (0, .5, 1, and 1.5). The performance of the estimation of the FMM was assessed in terms of optimal (high) posterior class probability and entropy. In scenarios with more class separation and/or stronger covariate effects, individuals were more frequently assigned to their correct class. In addition, Lubke and Muthén indicated that inclusion of even small covariate effects improved correct class assignment, even with small degrees of class separation. Similarly, Muthén (2004) noted that inclusion of covariate variables in GMMs improved the selection of GMMs with the proper number of classes and correct class assignments.

FMM with class-specific covariate effects. It might be hypothesized that covariate effects (for example, ethnicity) explain some of the variability in latent ability (for example, math ability) *within* each class, and the effect can be the same or class-specific across latent classes. In Figure 10, an asterisk * beside a path from the covariate variable (ethnicity) to the latent continuous factor (here, math ability) indicates the coefficient $\gamma_{\eta k}$ is freely estimated within each class (class-specific). For example, the data might fit a two-latent class model in which latent class 1 might function as the high math ability class and latent class 2 as the low math ability class. The researcher might hypothesize that there is no difference between ethnic majority students and ethnic minority students in the high math ability class, while ethnic minority students have even lower math ability than do ethnic majority students in the low math ability class. That is, ethnicity explains some variability in the continuous factor (here, math ability) within classes. Likewise, if a covariate effect does vary across classes, then this is regarded as a moderated direct effect of the covariate on the continuous factor by latent class membership.

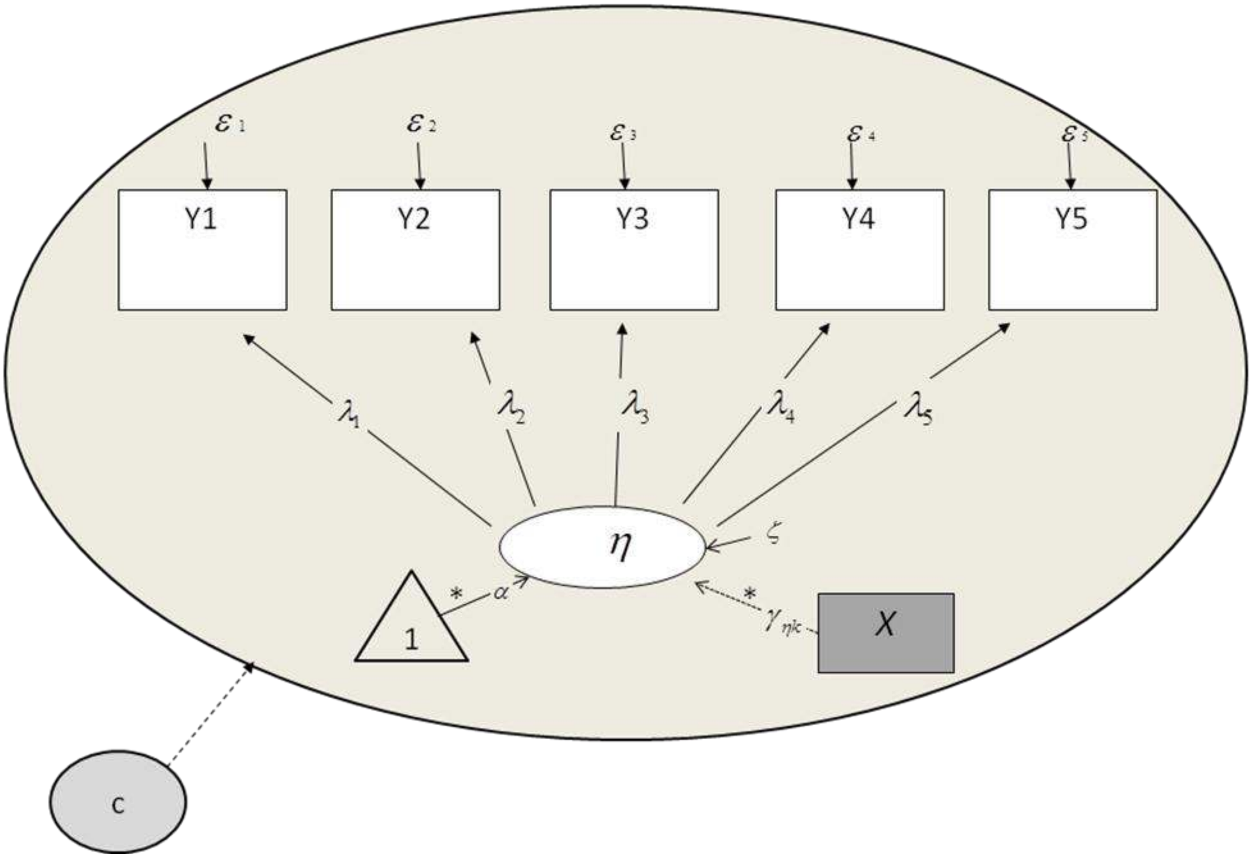

Figure 10. FMM with class-specific covariate effects

The FMM in which a latent continuous factor is regressed on a latent categorical class variable *c* is denoted as

$$\eta_{ik} = \alpha c_i + \gamma_{\eta k} X_i + \zeta_{ik} . \qquad [18]$$

Here, α represents a ($1\times K$) matrix containing factor intercepts or factor means for each latent class (k) , ζ_{ik} is a residual vector, and $\gamma_{\eta k}$ is defined as the effect of the covariate on the latent factor for each class, *k*.

FMM with both between-and class-specific covariate effects. Covariate variables (here, for example, ethnicity (*X1*) and gender (*X2*) in Figure 11) can be modeled as predicting latent class membership and explaining some variability in a latent variable (here, math ability) within each latent class. For example, males might be more likely to belong to class 1 (high math ability class) while females might be more likely to belong to class 2 (low math ability class). In addition, while there may be no differences between two ethnic groups within class 1 (high math ability class), the minority's math ability might be lower on average than the majority's math ability within class 2 (low math ability class). In this case, modeling between-class and class-specific covariate effects in FMM is needed, as shown in Figure 11.

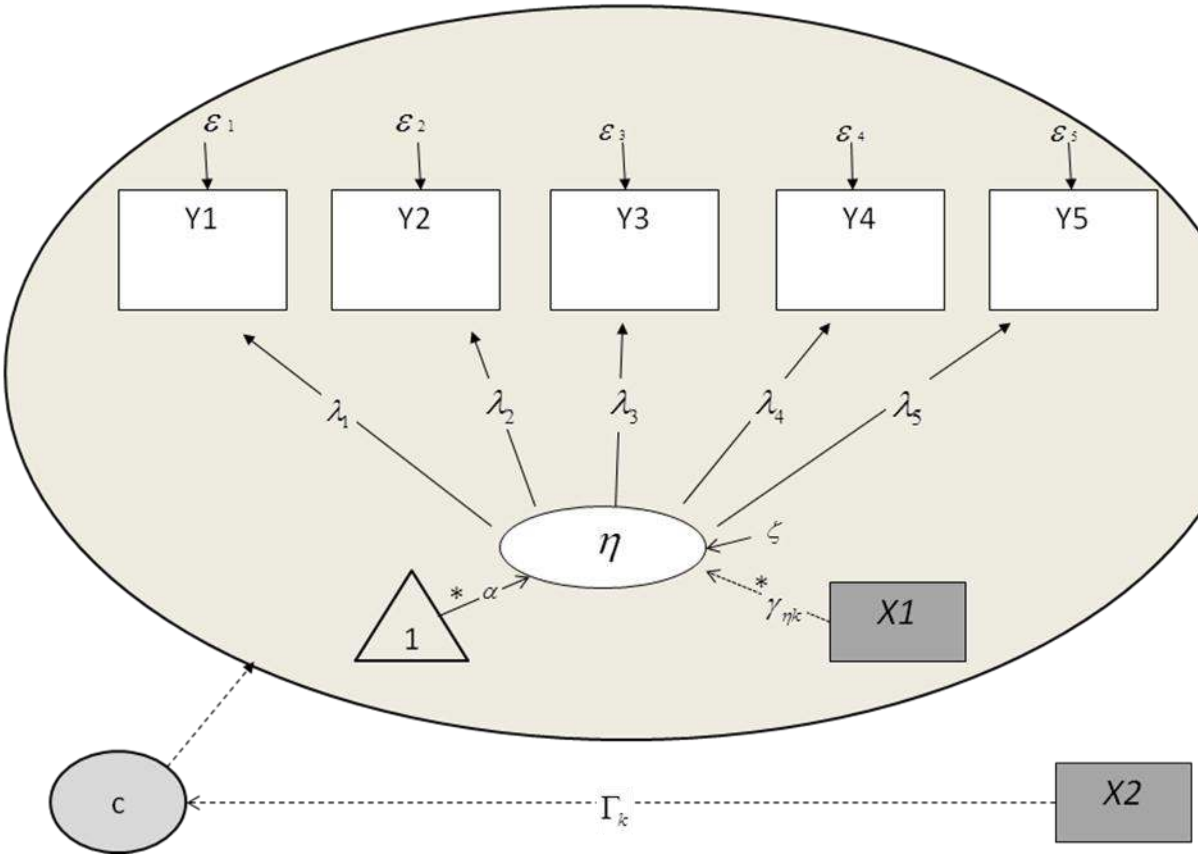

Figure 11. FMM with both between-class and class-specific covariates

Thus far, FMM with interval-scaled outcomes and combinations of covariates has been summarized. FMMs with binary outcomes, which have been the focus of the present research, have recently been explored in the context of DIF. Therefore, the next section contains a description of what has been found for FMMs with binary outcomes.

FMM with Binary Outcomes (Mixture IRT Model)

As noted earlier, the CFA model for dichotomous variables is the equivalent of dichotomous 1 and 2PL IRT models. In addition, an FMM with binary outcomes can be alternatively seen as a mixture IRT model (Asparouhov & Muthén, 2008). A mixture IRT model (that is, an FMM model with binary outcomes) involves the assumption that the probability of getting an item correct is conditional on ability level, but the assumption of an IRT model only holds within each latent class. Thus, in a mixture IRT model, it is assumed that respondents are collected from multiple populations and item parameters are the same for individuals within the same latent class but some parameters might differ across latent classes.

Model formulation. Muthén and Asparouhov (2002) presented latent variable models for categorical outcomes in two ways: postulating a conditional probability model and deriving a conditional probability model from a linear model for latent response variables under the assumption that observed outcomes are obtained by categorizing the latent response variables. In addition, Muthén and Asparouhov explained the equivalence of results between these two formulations of factor models with categorical outcomes (see Muthén and Asparouhov, 2002).

The first approach is to use a conditional probability formulation focusing on directly modeling the nonlinear relationship between the observed y and the latent continuous factor η:

$$P(y_i = 1 | \eta) = F[a_i (\eta - b_i)] \tag{19}$$

where a_i and b_i are the item discrimination and difficulty parameters for item *i*, respectively. The distributional function assumed for *F* is either a standard normal or logistic distribution function.

The second approach of dealing with categorical outcomes is to specify a latent response variable (LRV) formulation including the assumption that underlying each observed item response y is a continuous and normally distributed latent response variable y^*. The continuous latent variable can be considered a response tendency (Jackman, 2011). If an individual's response tendency exceeds a specific threshold, it is assumed that it is sufficiently high to answer an item correctly, and the individual will indeed answer the item correctly. On the other hand, if it falls below the threshold, then it is assumed that the individual will answer the item incorrectly. Based on the LRV formulation, the observed item responses can be considered to be a discrete categorization of the continuous latent variables. Thus, the relationship between two responses y and y^* can be represented by a nonlinear function:

$$y_i = \left\{ \begin{matrix} 1, if \quad y_i^* > \tau_i \\ 0, otherwise \end{matrix} \right\}, \tag{20}$$

where τ_i represents a threshold (difficulty) parameter of item *i*.

Because of the LRV formulation, the continuous response variable y^* is assumed to be unobserved, so the one-factor model for the continuous latent response variable can be respecified as

$$y_i^* = v + \lambda\eta_i + \varepsilon_i, \qquad [21]$$

where v is an intercept parameter, λ is a factor loading, η is a factor, and ε is a residual.

This leads to the conditional probability of a correct response as a function of η (Muthén and Asparouhov, 2002):

$$\begin{aligned} P(y_i = 1 | \eta_i) &= P(y_i^* > \tau | \eta_i) = 1 - P(y_i^* \le \tau | \eta_i) \\ &= 1 - F[(\tau - v - \lambda\eta_i)V(\varepsilon_i)^{-1/2}] \qquad [22] \\ &= F[-(\tau - v - \lambda\eta_i)V(\varepsilon_i)^{-1/2}], \end{aligned}$$

where F is typically a normal or a logistic function depending on the distributional assumption made for the residuals, ε.

When a covariate variable X is included in the model, Equation 21 can be extended to be

$$y_i^* = v + \lambda\eta_i + kX_i + \varepsilon_i \ , \qquad [23]$$

where k is the direct effect of the covariate X on indicator y^*. Using the conditional probability formulation (Muthén & Asparouhov, 2002), the conditional probability of a correct response as a function of η (Equation 23) can be extended by including the covariate variable X for a dichotomous item:

$$P(y_i = 1 | \eta_i, x_i) = F[-(\tau - \lambda\eta_i - kx_i)V(\varepsilon_i)^{-1/2}], \qquad [24]$$

where $\tau - kx_i$ is a new threshold (difficulty) parameter for the item, which varies across X values.

Model fit criteria. The literature includes studies of mixture modeling with binary outcomes that have used AIC, BIC, aBIC, CAIC, and likelihood ratio tests to assess models' fit (e.g., Cohen & Bolt, 2005; Li, Cohen, Kim, & Cho, 2009; Jackman, 2011; Lau, 2009; Maij-de-Meij et al., 2008; 2011; Tay, Newman, & Vermunt, 2010). Similar to the performance of fit indices in FMM with interval-scaled outcomes, fit indices in FMM with binary outcomes showed little to no agreement when supporting selection of the model with the optimal number of latent classes (Jackman, 2011). That is, when the two-latent class model was the true model, the AIC tended to support better fit of the three-class model. The BIC consistently supported the one-class model's fit, while the aBIC performed relatively well in terms of choosing the model with the correct number of classes.

Li et al. (2009) also examined model fit indices, including AIC and BIC, along with other model fit indices obtained using Bayesian estimation for FMM with binary outcomes. They found that the BIC index performed well in terms of selecting the correct model under most simulated conditions, but the AIC index performed poorly. In their study, the performance of fit indices overall was heavily dependent on the simulation conditions, although the BIC index appeared to perform best with correct model selection for the FMM with binary outcomes. Similarly, Nylund et al. (2007) indicated that the AIC was not able to identify the correctly specified model, regardless of total sample size for FMM with binary outcomes. The BIC and CAIC indices were able to correctly identify the correct model in close to 100% of the replications, regardless of total sample size. The performance of aBIC improved as total sample size increased. Thus, applied studies using real data for DIF detection have encouraged use of the BIC, aBIC, or CAIC indices as the preferred model fit criterion for FMM with binary outcomes (see, for example, Cohen & Bolt, 2005; Maij-de-Meij et al., 2008; 2011).

On the other hand, Lubke and Muthén (2005) suggested that model fit could also be improved by relaxing within-class restrictions. The authors indicated that imposed restrictions within latent class might lead to inaccurate extraction of additional numbers of latent classes or might result in distorting a within-class measurement model Relaxing parameters to allow them to be freely estimated across latent classes can be interpreted as DIF in an IRT framework, so FMM with binary outcomes can be extended for DIF identification. Thus, the next section addresses the extension of FMM with binary outcomes to detect DIF.

FMM with Binary Outcomes for DIF Identification

As mentioned above, violation of the measurement invariance assumption can be referred to as DIF in an IRT framework. When items' intercepts are non-invariant across latent classes, unobserved (latent) uniform DIF occurs, and referred to, here, as between-class latent uniform DIF. When loadings are non-invariant across latent classes, unobserved non-uniform DIF occurs, and it is referred to as between-class latent non-uniform DIF. Figure 12 depicts a latent ability (here, math ability) measured by 27 items. Item 1 is assumed to have a non-invariant intercept (between-class latent uniform DIF) across latent classes as represented by the path from the latent class variable to item 1's intercept, while item 2 has a latent non-invariant loading (between-class latent non-uniform DIF) as represented by the path from the latent class variable to item 2's slope. Thus, in order to examine whether the measure has between-class latent DIF in difficulty for item 1, path A can be tested. To examine between-class latent DIF in discrimination for item 2, path B can be tested. If path A or path B is significantly different between latent classes, then it is said to exhibit between-class latent uniform or non-uniform DIF, respectively.

Let us suppose that a researcher might find that there are two latent classes, and members of the first class have higher math ability than members of the second class. The researcher might also find that the probability of endorsing item 1 is consistently higher for individuals in the high math ability class than for those in the low math ability class, conditional on math ability. In addition, item 2 might be found to be highly effective in discriminating examinees' math ability in the high math ability class, but the item might not be as effective at discriminating examinees' math ability in the low math ability class, conditional on math ability. Note, however, that testing the non-invariance of factor loadings (non-uniform DIF) is not of interest in the present study, so the remainder of the study will focus only on the testing of non-invariant intercepts (that is, on uniform DIF).

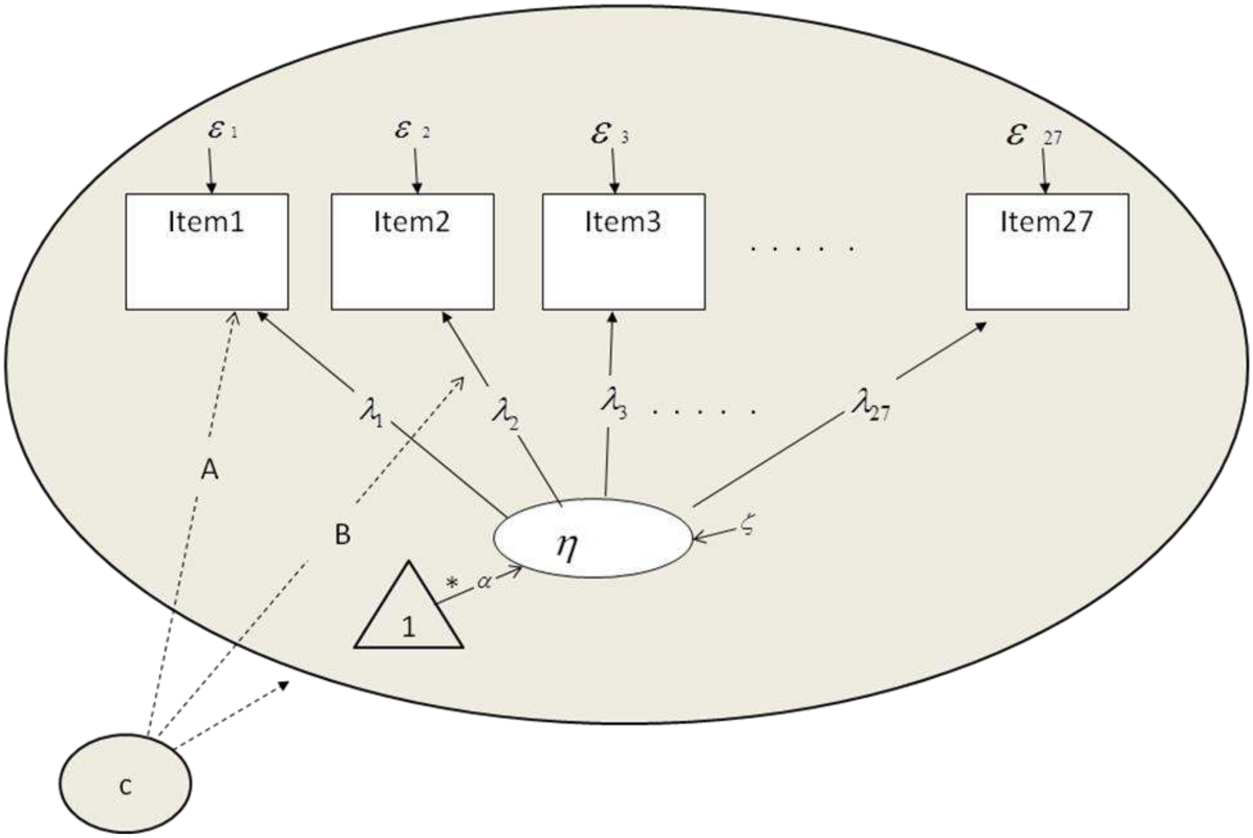

Figure 12. Between-class latent uniform and non-uniform DIF in a FMM

Traditional DIF detection methods vs. FMM with binary outcomes. Previous research has compared use of the FMM with binary outcomes versus more traditional DIF methods for DIF identification. For example, Samuelsen (2005) compared use of the Mantel-Haenszel test with that of the FMM with binary outcomes. The author manipulated different levels of overlap between an observed group and a latent class variable. *Overlap* indicates the proportion of cases having the same membership between observed groups and latent classes. For example, if all females belong to the latent class 1 and all males belong to the latent class 2, there is perfect overlap between the observed group and the latent class. In addition to different levels of overlap, the number of items exhibiting DIF, the DIF effect size, and the magnitude of the ability distribution means within the latent classes were also manipulated. Data were simulated for a twenty-item test. The outcomes that were assessed included the number of items correctly identified as having DIF (power to detect DIF), recovery of the DIF effect size, and the number of items falsely identified as DIF (Type I error).

As expected, when the M-H test was used, the power to correctly identify DIF items decreased for scenarios with less overlap between manifest groups and latent groups. However, under the conditions of equal class proportion, large sample size, large magnitude of DIF, and more than 80% of overlap between the latent and observed group, the M-H tests' power to correctly identify DIF items was relatively good. In addition, total sample sizes and the amount of overlap, and the observed group proportions influenced the Type I error rate when an M-H test was used.

On the other hand, in the case of FMM with binary outcomes, DIF was not accurately identified under the condition of a 60% overlap case with a small total sample size (500 examinees). However, with sample size increased to 2,000 examinees, all DIF items were correctly identified regardless of the degree of overlap between the observed and latent grouping variables. Under the condition of a 60% overlap with a large total sample size (2,000), only one of the non-DIF items was incorrectly identified as exhibiting DIF (a Type I error). Overall, Samuelsen (2005) suggested that FMM with binary outcomes is a better approach for determining a source of DIF that might be unobserved.

Unlike in the study by Samuelsen (2005), which compared FMM with binary outcomes with traditional DIF methods, Jackman (2010) focused on evaluating the performance of FMM with binary outcomes in the context of various simulation conditions using M*plus* software. The simulated data were generated to fit a CFA model with 15 dichotomous items for two latent classes with conditions that involved manipulating the total sample size, the magnitude of the DIF effect, values of item discrimination parameters, and the existence of the latent mean difference (representing impact). The simulation study assessed the recovery of the true number of latent classes using model fit indices such as AIC, BIC, and aBIC. As mentioned above in the discussion of model fit criteria in FMM with binary outcomes, many inconsistencies were observed in the fit indices' performance. The researcher assessed the overall Type I error rates for item difficulty parameters of invariant items and the overall power to correctly identify DIF items. The Type I error rate for each condition was assessed by computing the proportion of times that the nine DIF-free items out of fifteen items were incorrectly identified as displaying DIF (One item was the reference indicator, so the difficulty parameters for the item were constrained to be equal across latent classes). Likewise, the power for each condition was computed by dividing the total number of times any one of DIF items (five DIF items) was correctly identified by the total number of replications. The results indicated that the FMM with binary outcomes estimated in the study did not perform well in controlling the Type I error rate. That is, Type I error rates exceeding the nominal alpha level of .05 for invariant items were observed under all study conditions, especially when the total sample size and the magnitude of DIF were small. Regarding power, only under the conditions with a large sample and large DIF magnitude was an acceptable level of power (.80) achieved. However, Jackman's study included an issue related to data generation: there was 80 % overlap between an observed group and latent classes. That is, data were generated such that DIF was exhibited for only 80% of individuals in one of the latent classes (the focal group) without the inclusion of any observed group in the study. The author indicated that this might be one reason why the FMM with binary outcomes model performed poorly in the study.

DeMars and Lau (2011) assessed the recovery of latent class membership and the recovery of item parameters for both invariant and DIF items, with three factors: (1) existence of impact, (2) number of DIF items, and (3) number of invariant items. M*plus* software was used to estimate parameters for a two-latent class FMM with binary outcomes. The authors reported that the recovery of correct class membership was poor under all conditions tested, regardless of whether data were generated with impact or with no impact. Discrimination parameters for invariant items and for DIF items were estimated relatively well when data were generated with no impact. When data were generated with impact, discrimination parameters for DIF items were positively biased. However, under most simulation conditions, discrimination and difficulty parameters were estimated relatively well.

DeMars and Lau (2011) also assessed the bias in the estimation of the DIF effect (that is, estimation of the difference in difficulty parameters across latent classes). The study focused on the accuracy of estimates of DIF effect size, not on a test of each item as DIF or not DIF. When there were only four DIF items generated with no impact condition, DIF effect sizes were overestimated. However, as the number of DIF items was increased from 4 to 8, the DIF effect size was estimated well. Even though a small number of DIF items was generated (4 DIF items), as the number of invariant items was increased (from 10 to 20 invariant items) under the impact condition, estimates of DIF effect size improved.

In general, the results obtained from DeMars and Lau's (2011) study indicated that for a test with more DIF items, better estimates of DIF effect sizes will result. In addition, they found that invariant item parameters were estimated fairly well, but estimates of DIF items' parameters depended on the existence of impact. Correct latent class membership was not well-recovered under all simulation conditions tested. Therefore, they suggested that, based on previous studies' results (for example, Bandalos and Cohen, 2006; Lubke and Muthén, 2007; Meij-de Meij et al, 2008; 2010), including an observed variable that may contribute a piece of information in the mixture model can help improve the recovery of correct class membership and DIF item identification. The next section discusses inclusion of an observed variable in FMM with binary outcomes, and then describes inclusion of an observed group in FMM with binary outcomes for DIF identification.

Including an Observed Group in FMM with Binary Outcomes

An observed grouping variable can be included in FMM with binary outcomes as a between-class covariate effect. Smit, Kelderman, and van der Flier (1999) investigated the effect of including a dichotomous observed variable as a between-class covariate in FMM with binary outcomes. The results indicated that prediction of correct class membership as well as estimation of the standard errors of item parameter estimates could benefit from the inclusion of the between-class covariate. In particular, even when the total sample size was relatively small (500 examinees), including a between-class covariate that was associated with a latent class variable improved estimation of the item difficulties' standard errors as well as correct class membership identification. Because previous studies have indicated that parameter recovery and correct class membership in mixture models can benefit only if either a latent mean difference or a total sample size is relatively large (Kelderman & Macready, 1990; Lubke & Muthén, 2007), the results of this study have motivated applied researchers to include covariate effects in FMM with binary outcomes.

On the other hand, Clark (2011) focused on investigating how total sample size and entropy are related to the recovery of a between-class covariate effect parameter by looking at its mean squared error (MSE), confidence interval coverage, and power. Clark manipulated two total sample sizes (250 and 1,000) and four entropy values (.4, .6, .8, 1) using the intercept difference of observed indicators. The researcher generated two latent classes using ten dichotomous items. The researcher also set the generating value of the covariate effect parameter to 0.5. Generally, it was found that the greater the values of entropy, the smaller MSEs were and the better coverage rates were. More specifically, when the value of entropy reached .80, MSEs were much smaller, and the coverage was much better. MSEs of the covariate effect parameter were much smaller with a total sample size of 1,000 than with a total sample size of 250. When the value of entropy was .80 even when total sample size was 250, an acceptable level of power (.80) was achieved. In addition, the study found that even when the total sample size was 1,000 and the value of entropy was a little lower than .80 (.60), the power was greater than 80%. Therefore, the study suggested that the true value of a covariate effect could be well estimated if the value of entropy is .80 or .60 with a total sample size of 1,000.

In general, previous studies (Smit et al., 1999; Clark, 2011) have demonstrated that inclusion of an observed group(s) as the between-class covariate effect in FMM with binary outcomes have helped improve correct class membership identification and recovery of item parameters. In addition, a covariate effect parameter has typically been well estimated when using a total sample size of 1,000 or when the value of entropy has been 0.80.

The next section will describe whether inclusion of an observed group in FMM with binary outcomes has a benefit in identifying (observed and/or unobserved) sources of DIF.

Including an Observed Group in FMM with Binary Outcomes for DIF Identification

An observed group can be included in FMM with binary outcomes as a covariate effect for DIF identification. When the observed group predicts latent class membership (between-class covariate effect), correct class assignment can be improved, helping find between-class latent DIF. When some items function differently based on observed group membership within latent classes, and when the effect of the observed group on some items differs by latent classes, the result can be referred to as class-specific observed DIF.

Thus, this section focuses on discussion of including an observed grouping variable as a covariate in FMM with binary outcomes.

Helping find between-class latent DIF. Figure 13 depicts a model in which an observed grouping variable is included as a between-class covariate effect in FMM with binary outcomes. That is, including the observed group *X* can help identify correct class membership, resulting in improved detection of between-class latent DIF. As shown in Figure 13, difficulty parameters for item 1 are assumed to be different across latent classes, so path A can be tested to detect between-class latent uniform DIF.

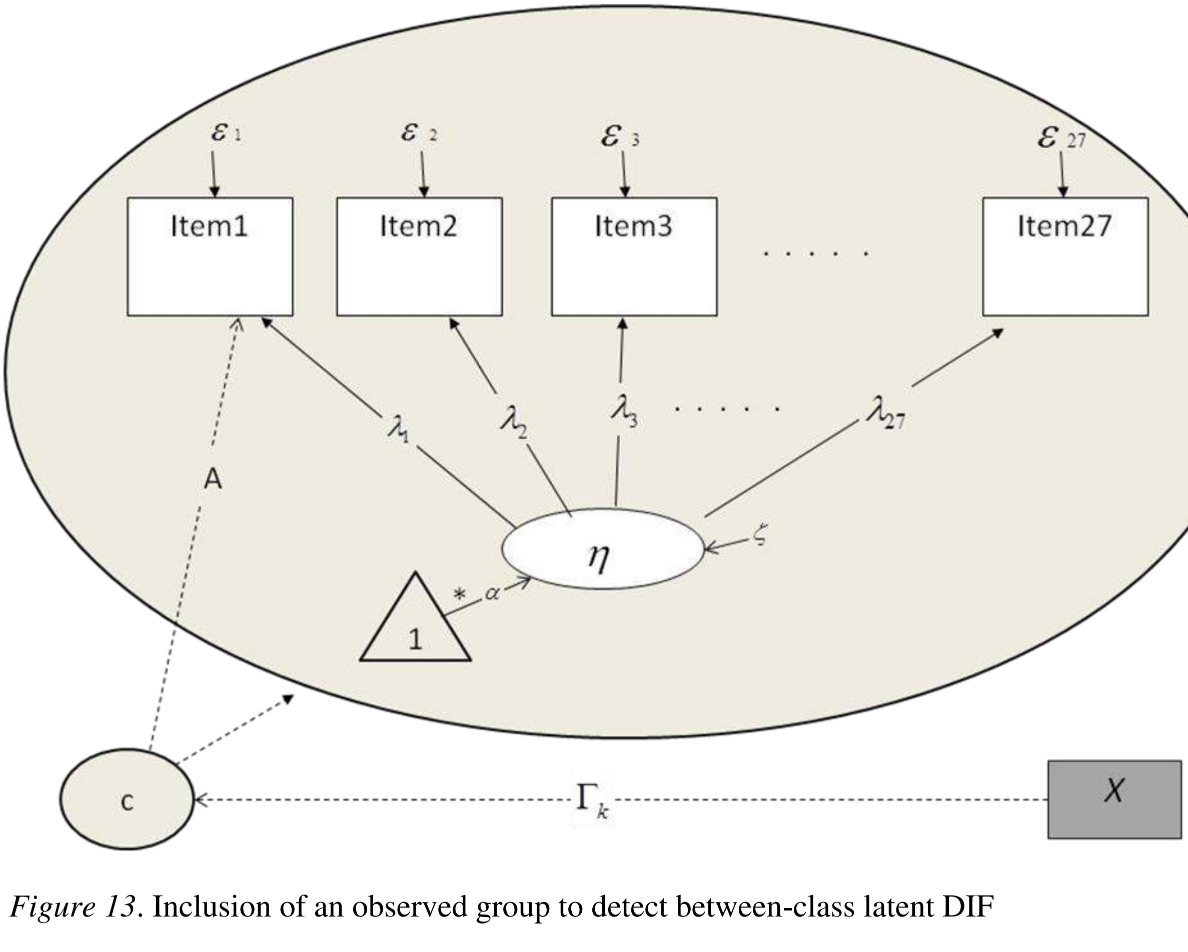

Figure 13. Inclusion of an observed group to detect between-class latent DIF

Maij-de Meij, Kelderman, and van der Flier (2011) used an observed variable that is related to between-class latent DIF to examine whether including the observed group can help in the detection of between-class latent DIF in an FMM with binary outcomes. That is, even if DIF is not fully explained by gender, a gender variable can be related to the source of DIF. So, in order to investigate the advantages of using a mixture model to detect uniform DIF rather than using typical DIF detection methods employing observed variable(s), the researchers conducted a simulation study to compare a traditional DIF method (here, Lord's chi-square test) with an FMM with binary outcomes including an observed group as a between-class covariate effect. To compare the two approaches directly, they also used Lord's chi-square test—which is used to test the differences between the difficulty parameters across two groups of examinees—for an FMM with binary outcomes. They manipulated four factors: the degree of overlap between latent class membership and the observed group (0%, 60%, 70%, 80%, 90% and 100%), total sample sizes (5000, 25000), levels of significance ($\alpha = .05$ or .01), and latent class proportions (equal: 50 vs. 50, unequal: 25 vs. 75). The researchers said that because including the observed group with even small effects in FMM with binary outcomes was preferred to FMM with binary outcomes without the observed group in terms of DIF detection, the study only compared FMM with binary outcomes including the observed group and the traditional DIF approach.

As expected, the results indicated that the identification rates of correct DIF for the traditional DIF approach that uses the observed group depended heavily on the associations between the observed group and the latent class. In contrast, FMM with binary outcomes when including the observed group performed well regardless of the degree of associations between the observed group and the latent class. In particular, when the association between the observed group and the latent class was low, FMM with binary outcomes including the observed group performed better in identifying between-class latent DIF in comparison to the traditional DIF approach. However, the two approaches performed equally well with respect to correct classification rates for DIF and invariant items when the association between the observed group and the latent class was high.

Studies previously mentioned (for example, Maij-de Meij et al., 2008; 2011; Smit et al., 1999) have used an observed group as a predictor of latent class membership (between-class covariate effect), meaning that they focused on whether inclusion of between-class covariate effects improved latent class membership or detection of between-class latent DIF. However, it is possible that using latent class variables to capture unobserved sources of DIF cannot capture all sources of DIF. That is, an observed source of DIF, such as gender, can exist within latent classes, and the effect of the observed source of DIF might be different between latent classes. Thus, it is possible to investigate whether modeling an observed group within latent classes improves parameter estimates or identification of items that exhibit DIF in FMM with binary outcomes.

Class-specific observed DIF within latent classes. As mentioned above, using either solely unobserved or observed groups may not capture all possible sources of DIF (Tay, Newman, & Vermunt, 2010). It is possible that some items are functioning differently based on latent class membership, while other items are functioning differently based on an observed group membership.

Figure 14 depicts a model in which an observed group is included within the shaded ellipse, and the effect of the observed group is class-specific as indicated by an asterisk * beside the path from the observed group *X* to item 27. The dashed arrow pointing to the intercept of item 27 indicates class-specific observed uniform DIF (path B). Thus, if path B is significantly different from zero (group mean difference), and the effect of the observed group on item 27 significantly differs across latent classes, the item is said to exhibit class-specific observed DIF. The advantage of the model in Figure 14 is that it enables the detection of both observed and unobserved sources of DIF. For example, a researcher might hypothesize that an unobserved source of DIF can exist in some items, but an observed source of DIF can also exist in other items. As shown in Figure 14, the researcher hypothesized that item 1exhibited between-class latent uniform DIF. The researcher also hypothesized that the difficulty parameter of item 27 differed by gender (male vs. female) in the low math ability class, controlling for math ability, but not in the high-math-ability class (that is, class-specific observed uniform DIF). Notice that the observed group is not a predictor of latent class membership in the model as shown in Figure 14, so it is assumed that the proportion of each observed group membership is the same across latent classes.

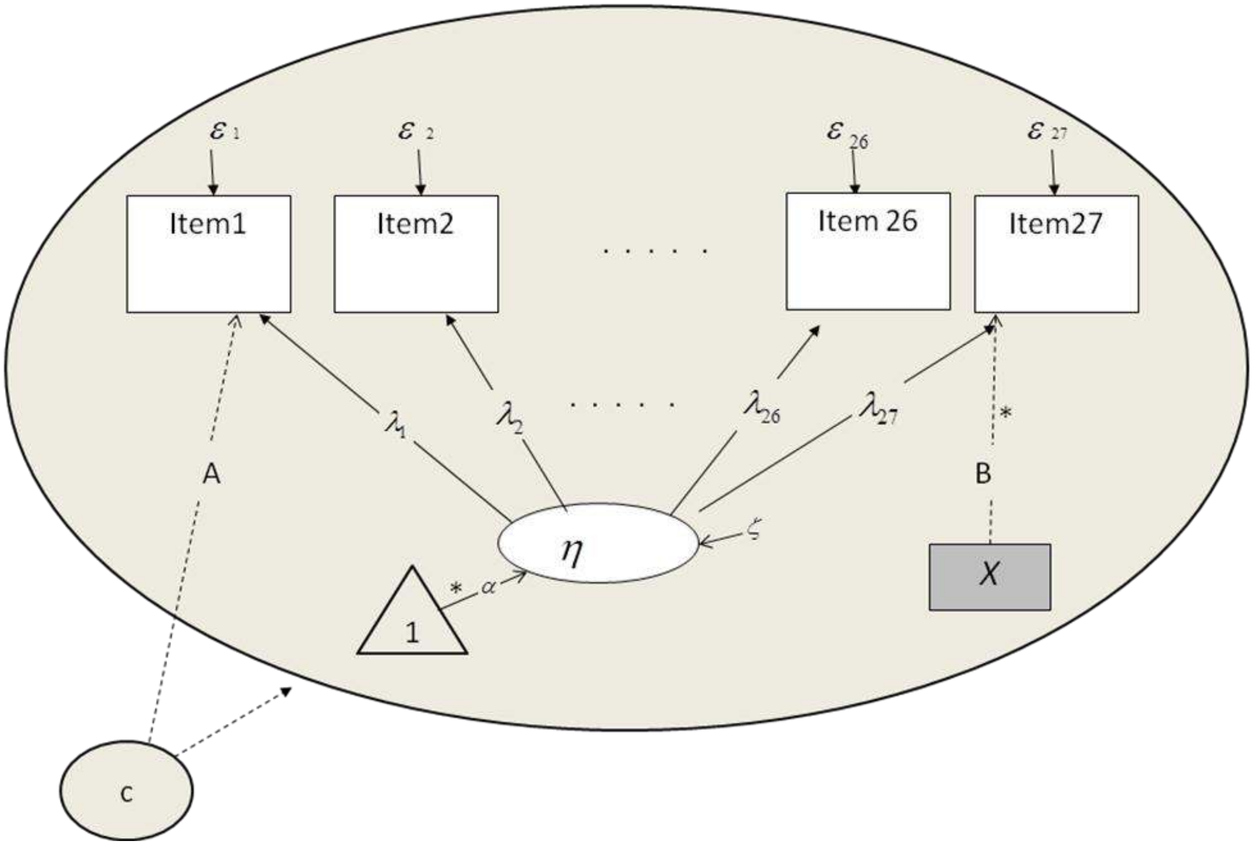

Figure 14. Inclusion of class-specific observed group in an FMM with binary outcomes

An empirical study using public school employee data conducted by Tay et al. (2010) included observed grouping variables—gender and work experience to detect whether between-class latent DIF and class-specific observed DIF existed in an 8-item union citizenship scale. The researchers specified the number of latent classes in the data, and initially they used the unconstrained model, which allows all item discrimination and difficulty parameters to be freely estimated. They found that the two-latent class model was a much better fit with the data than were the one- or three-latent-class models. In addition, they found that 68% of individuals belonged to latent class 1 (called *politicos* *class*) and 32% of individuals belonged to latent class 2 (called *non-politicos class*). Four items exhibited between-class DIF and one item exhibited class-specific observed DIF. Interestingly, one item (item 8) exhibited between-class latent DIF as well as class-specific observed DIF. That is, sources of DIF for item 8 were not fully captured by latent class, and males had a greater probability of endorsing item 8 within the non-politicos class, but there was no difference between males and females within the politicos class. The result that both unobserved and observed sources of DIF can be exhibited by a single item suggested that detecting the true source of DIF might be more complex than indicated in previous studies that used either traditional DIF detection methods that consider only observed source of DIF or conventional FMM with binary outcomes that considers only unobserved sources of DIF. However, until now, no single simulation study has investigated parameter recovery, correct class membership identification and DIF detection rates in models that include both between-class latent DIF and class-specific observed DIF simultaneously in FMM with binary outcomes. Therefore, it is necessary to investigate the effect of including an observed group in the performance of FMM with binary outcomes.

Incorrectly Specified Observed Group in FMM with Binary Outcomes

Almost all simulation studies investigating DIF in mixture models have produced estimates with the assumption of correctly specified models, to examine how well the parameters of DIF items are estimated and how well unobserved (*latent*) sources of DIF are detected. In addition, it is typically assumed that an observed group is correctly specified when an observed group is modeled in FMM with binary outcomes. That is, there has not been an investigation of whether a mis-specified observed group effect influences latent DIF detection rates, except in the study by Maij-de Meij et al. (2011). In that one study the researchers included an observed grouping variable as a predictor of latent class membership, but the observed group was not related to the latent class at all (zero correlation between the observed group and the latent class). Thus, including the observed group as the between-class covariate effect in the model being estimated made it a mis-specified model. Then, Maij-de Meij et al. compared a model including the observed grouping variable to a model excluding the observed grouping variable when there was zero correlation between the latent class and the observed group. In order to do this, they examined Type I and Type II errors with two sizes of total samples.

When the model was mis-specified by including the observed grouping variable, the outcome resulted in increasing Type II error rates when the class proportions were equal, but there was not much difference in the Type I error rates for non-DIF items between conditions when the observed grouping variable was correctly specified and when it was mis-specified. When the total sample size was increased by 25,000, correct classification rates of DIF items and invariant items were above 96% under almost all conditions tested. Interestingly, when latent class proportions were unequal, the Type II error rates for DIF items were even lower for the mis-specified model that included an observed group, but the Type I error rates for non-DIF items were higher for the mis-specified model than those for the correctly specified model. However, when the total sample size increased, the Type I error rates were reduced for the mis-specified model, resulting in no difference between conditions when the observed grouping variable was correctly specified and when it was mis-specified. Nevertheless, the Type II error rates for DIF items were still lower for the mis-specified model than for the correct model. Thus, it seems that, under the condition with a large total sample size, inclusion of an observed group as a between-class covariate effect does not negatively impact detection of between-class latent DIF; rather, it provides a benefit in helping find between-class latent DIF even though it is not relevant to latent class membership.

However, Maij-de Meij et al. (2011) only examined whether inclusion of an observed grouping variable that is not relevant to latent class membership impacted detection of between-class latent DIF. In a real-world scenario, it might be possible that an observed grouping variable is a potential source of DIF within latent classes. Yet it is assumed to be a predictor of latent class membership (between-class covariates) rather than being modeled as an observed source of DIF (class-specific observed DIF). For example, Samuelsen (2005) examined whether observed grouping variables--gender and ethnicity--were significant predictors of latent class membership in two-latent class FMM with binary outcomes. It was found that none of the observed grouping variables significantly influenced latent class membership. As a result, observed groups were excluded in the model. However, it was observed that Asian students in latent class 1 had significantly higher latent means than did Hispanic students, and the same pattern appeared for Asian and Hispanic students in latent class 2, Thus, ethnic group differences in factor means on the construct of interest (observed *impact*) might have existed in the data, or some of the items functioned differently *within* the latent class based on ethnicity.

However, it is unknown how modeling observed groups *within* a latent class in the model influenced the study's results.

The literature review has revealed that detecting observed source of DIF has not been of interest in DIF studies when using FMM with binary outcomes. That is, when FMM with binary outcomes has been specified to identify sources of DIF, only between-class latent DIF has been focused upon. Thus, an observed grouping variable has only been included in models to help find latent DIF. However, it is possible that class-specific observed DIF exists within latent classes. For example, if there is class-specific observed DIF but a researcher only examines whether between-class latent DIF exists, the researcher would mis-specify a model by excluding an observed group. In addition, if a researcher assumes that there is class-specific observed DIF, but only between-class latent DIF exists, the researcher would mis-specify a model by including an observed group. Likewise, there are many possibilities for mis-specifying FMM with binary outcomes to identify sources of DIF.

However, no study has been specifically designed to investigate whether a mis-specified source of DIF (unobserved/observed sources) can impact the performance of FMM with binary outcomes in terms of model identification, parameter recovery for DIF items correctly or incorrectly identified, and detection of when items are correctly or incorrectly identified as displaying DIF. Thus, it is appropriate to further examine whether mis-specified between-class DIF items and class-specific observed DIF items influence the performance of FMM with binary outcomes.

Statement of Purpose

Research has already indicated that FMM with binary outcomes that uses latent variable(s) has outperformed traditional DIF detection methods that use observed variable(s) when the source of DIF is not fully explained by an observed variable. Most studies using FMM with binary outcomes have focused on how well between-class latent DIF is detected under various simulation conditions including sample size, latent class membership proportion, magnitude of DIF effect, and number of invariant and DIF items.

The results obtained from such studies have generally indicated that a large total sample size and a large factor mean difference between latent classes can help identify class membership correctly in mixture models (Kelderman & Macready, 1990; Lubke & Muthén, 2007). In addition, a combination of large sample size (such as 1,000 or 5,000) and high magnitude of DIF best controlled Type I error rates for non-DIF items and achieved an adequate power level to correctly identify DIF items (e.g., De Mars & Lau, 2011; Jackman, 2011). Furthermore, it has been found that inclusion of an observed group(s) as a predictor of latent class membership has been found to improve the recovery of latent class membership, the recovery of estimates of item parameters (Smit et al., 1999), and between-class latent DIF detection rates (Maij-de Meij et al., 2011).

Most simulation studies have used an observed group as a predictor of latent class membership to help find between-class latent DIF, but they have not used the observed group to examine whether an observed source of DIF exists within and varies across latent classes (class-specific observed DIF). It is possible that all sources of DIF cannot be captured by a latent class variable, but that an observed variable, such as gender, which has a specific effect on some items within each latent class, might provide an additional source of DIF. Therefore, one of the purposes of the present study is to examine how well the following models perform in terms of parameter recovery for DIF items correctly identified, class membership identification, and DIF identification: a model that has both between-class latent DIF and class-specific observed DIF, a model that only has between-class latent DIF, and a model that only has class-specific observed DIF.

In addition, while most studies using FMM with binary outcomes to identify DIF items have used correctly specified models, only Maij-de Meij et al. (2011) examined whether including one observed but incorrectly specified variable would impact between-class latent DIF detection. However, the researchers only focused on investigating whether inclusion of the observed group variable helps find between-class latent DIF. Because a correctly specified model is assumed by applied researchers when they model real data, it is important to consider how a mis-specification of where the observed group's effect has influence might impact DIF identification, class membership identification, and the performance of fit indices by comparing correctly specified models. For example, an applied researcher might use a model including only an observed group as a predictor of latent class membership to help in the detection of between-class latent DIF, but it could turn out that there exists class-specific observed DIF, so the true model should include the class-specific observed grouping variable within latent classes. In this scenario, it would be necessary to investigate how the incorrectly specified observed group in FMM with binary outcomes might impact the performance of FMM with binary outcomes for DIF identification.

The purpose of the present study, therefore, was to investigate whether models correctly specified in terms of between-class latent DIF and/or class-specific observed DIF influence the performance of model fit indices, and class membership identification, as compared to models incorrectly specified in terms of either between-class latent or class-specific observed DIF. In addition, the present study examined the recovery of item difficulty parameters as well as investigating the proportion of replications in which items are correctly or incorrectly identified as displaying DIF. The simulation study manipulated the degree of the between-class and class-specific DIF effect sizes and the latent class proportion used to generate the data.

Chapter 3: Method

Overview

The simulation study was designed to investigate estimation of between-class latent DIF and class-specific observed DIF in FMMs with binary outcomes in scenarios in which the FMM was either correctly or incorrectly specified. In this study, the only part of the model that might be incorrectly specified was either or both of the between-class latent DIF (that is LDIF) and class-specific observed DIF (that is, ODIF). Data were generated to fit a two-class FMM with binary outcomes, and included a single latent continuous factor measured by 27 dichotomous items. Only one observed, dichotomous grouping variable was included. Half of the simulated sample was in one group with the other half randomly assigned to the other group in all models that were generated. Such an arrangement reflects a scenario in which a dataset contains half females and half males. Latent class probability (equal vs. unequal), between-class latent DIF effect size (small vs. large), and/or class-specific observed DIF effect size (small vs. large) were manipulated in the simulation study. The evaluation was focused on the accuracy of item difficulty estimates for a specific subset of test items and on the recovery of class assignment. In addition, the performances of fit indices (specifically, the AIC, BIC, aBIC, and CAIC) were compared to assess which fit indices exhibit the highest proportion of success in supporting the fit of the correctly specified model over the fit of the incorrectly specified models. In addition, the present study evaluated whether items are correctly or incorrectly identified as exhibiting DIF (i.e., Type I and power were also assessed).

Fixed Conditions

Total sample size. The convergence rates of mixture models are influenced by total sample size (Lubke, 2006). In addition, when the total sample size is large enough, item parameters are more accurately estimated and better identification of DIF has been found (Maij-de Meij et al., 2011). Previous empirical studies using real data have most frequently used total sample sizes of approximately 2,000 (Clark, 2011; Cohen & Bolt, 2005; De Ayala et al., 2002; Maij-de Meij et al., 2008; Samuelsen, 2005; Tay et al., 2011). On the other hand, methodological studies have examined various total sample sizes from 500 to 15,000 to examine the effect of total sample size on detection of DIF in mixture models, and the findings have indicated that the performance of fit indices depends on total sample size (De Ayala et al., 2002; Jackman, 2011; Maij-de Meij et al., 2011; Rost, 1990). That is, fit indices were more likely to support the fit of a model with a correct number of latent classes as total sample size increased. In addition, Samuelsen (2005) reported that a small sample size (for example, 500 examinees) required a large magnitude of DIF and a large overlap between the class membership and the observed group (greater than 80%) to correctly identify DIF items.

Based on such findings, a total sample size of 2,000 was used for the present study in order to reflect real-world testing practice as well as provide a sample size that should provide reasonable parameter estimates (Dai, 2009).

Test Length. Applied DIF studies have assessed DIF for tests of various lengths ranging from 8 to 50 items (Cohen & Bolt, 2005; Dai, 2009; De Ayala et al., 2002; Finch, 2005; Jackman, 2011; Maij-de Meij et al., 2011; Samuelsen, 2005; Shih & Wang, 2009;

Tay et al., 2011). However, a majority of methodological studies using real data have typically used around 20 to 30 items for a test (for example, Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005). Simulation studies that have investigated detection of DIF in FMM with binary outcomes have used various test lengths ranging from 20 to 30 items to reflect real testing scenarios with reasonable minimum test lengths (Cohen & Bolt, 2005; De Ayala et al., 2002; DeMars & Lau, 2011; Maij-de Meij et al., 2011; Samuelsen, 2005).

DeMars and Lau (2011) found that test length did not seriously impact recovery of item parameters, DIF effect size, and latent class membership using FMM with binary outcomes. When they manipulated test lengths using various combinations of invariant and DIF items, the number of invariant items (either 10 or 20 invariant items) did not make much difference in the recovery of item parameters, DIF effect size (in the case of a no impact condition), and latent class membership. However, Li et al. (2009) reported that, for test lengths of 30 items, the percentage of correct latent class membership classifications increased to above 96% for 1PL and 2PL models, the RMSEs for item parameters decreased, and the BIC index was particularly effective at selecting the correctly specified model. Because consensus is still lacking in terms of the appropriate test length for using the FMM for DIF detection, data for the present study were generated using a total test length of 27 items to reflect real testing scenarios. In addition, item parameters based on a simulation data analysis in Maij-de Meij et al. (2011) that entailed 27 test items were used, here.

Number of DIF items. DeMars and Lau (2011) observed that about 10% of items were regarded as DIF items when real data are used (10.7% DIF items in a study of disabled students with/without testing accommodations, Finch, Barton, & Meyer, 2009; 4.4% DIF items in a study of paper/computerized tests, Keng, McClarty, & Davis, 2008; and 13.6% DIF items in a study of racial group, Puhan, Moses, Yu, & Dorans, 2009). On the other hand, Shih and Wang (2009) reported that real tests may have 20% or more DIF items (27% of items showed gender DIF when measuring attitudes from 23 real data sets, Dodeen & Johanson, 2003; nearly 40% of the items had DIF when investigating cross-cultural measurement equivalence of items in the English language version of the NEO Personality Inventory, Huang, Church, & Katigbak, 1997; and 42% of 149 items from the Multidimensional Self Concept Scale exhibited gender DIF, Young and Sudweeks, 2005). Thus, no consensus is apparent in the published literature concerning how many DIF items typically comprise real tests.

Interestingly, DIF studies using surveys of actual tests (Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005) have found that about 15 to 20% of items exhibit DIF when traditional DIF detection methods are used, such as an M-H test and a likelihood ratio test. On the other hand, when FMM with binary outcomes has been used for detection of DIF with the same actual tests, more than 20% (and at most 50%) of items have exhibited DIF. That is, mixture models were more likely to identify items as displaying DIF than were traditional detection methods. However, it was not known how many items actually exhibited DIF, because the studies used surveys of actual tests. Therefore, many DIF studies using mixture models have simulated conditions with the higher proportion of DIF items (typically about 20-30%, but 50% for some simulation conditions) (Cho 2007; De Ayala et al., 2002; DeMars & Lau, 2011; Jackman, 2011; Maij-de Meij et al., 2011; Samuelsen, 2005).

Therefore, the present study simulated conditions in which 15% and 30% of items exhibit DIF. Specifically, when data were generated to fit a model having both between-class latent DIF and class-specific observed DIF, four DIF items were generated to exhibit between-class latent DIF, and four additional DIF items were generated to exhibit class-specific observed DIF (30% DIF). When data were generated to fit a model containing either between-class latent DIF or class-specific observed DIF, only four of the 27 items per dataset were generated to exhibit DIF (15% DIF).

Impact. DeMars and Lau (2011) reported that estimated DIF effect sizes were more likely to be accurate when data were generated with impact rather than without impact. Two additional studies have also found that, in conditions with a larger mean difference between latent classes, correct class membership was better identified (Lubke & Muthén, 2007), and the accuracy of item parameter estimates was better (Lubke & Muthén, 2007; Lu & Jiao, 2009). Jackman (2012) indicated that a large degree of impact occurs when the mean for the reference group is one standard deviation higher than the mean of the focal group.

Therefore, in the present study latent impact was simulated with the first latent class having a latent ability mean that is one standard deviation higher than the mean of the second latent class (namely, the reference group's ability distribution was θ_R ~N(1, 1) for latent class 1, and the focal group's ability distribution was θ_F ~N(0, 1) for latent class 2).

Simulated Conditions

Class probability. Methodological research has consistently indicated that in scenarios with larger numbers of DIF items, larger DIF effects, and equivalently-sized latent class membership, item parameter values are recovered well (DeMars & Lau, 2011; Lu & Jiao, 2009). However, groups (such as focal and reference groups) are not often of equal sizes. For example, a DIF study using real data indicated that about 20% of respondents belonged to the focal group and 80% of respondents belonged to the reference group (Samuelsen, 2005). Other empirical research has found that 64% of the sample belonged to the first latent class and 36% belonged to the second latent class (De Ayala et al., 2002). Simulated research has used various group size proportions: 25%:75% (Maij-de Meij et al., 2011) as well as 15%:85% and 30%:70% (Dai, 2009). Even though using unequal class sizes leads to less accurately estimated item parameters for both DIF and invariant items (Dai, 2009; Lu & Jiao, 2009), it seems that unequal class sizes best reflect real testing scenarios. Therefore, the present study compared results under an optimal scenario with equally-sized latent classes versus a scenario with unequally-sized classes (specifically, 70%: 30%).

DIF effect size. There is no single simulation study that has investigated both between-class latent DIF and class-specific observed DIF using FMM. However, most of the simulation studies that have investigated uniform DIF have manipulated the degree of uniform DIF, using values ranging from .3 to 1.5 (see, for example, Camilli & Shepard, 1987; De Ayala et al., 2002; DeMars & Lau, 2011; Maij-de Meij et al., 2011; Samuelsen, 2005). Jackman (2011) found that with mixture models, DIF is only identified accurately in scenarios with large magnitudes of DIF. In addition, Lubke and Muthén (2007) used the value of 0.5 as the small effect size and the value of 1.5 as the large effect size. Therefore, for the present study the degree of between-class latent DIF and class-specific observed DIF investigated with 0.5 as a small effect and 1.5 as a large effect. Table 1 contains a summary of the simulation conditions.

Table 1. *Simulation Conditions*

Factor	Value
Class probability	Equal: 50%:50% Unequal: 70%:30%
Between-class latent DIF effect size (LDIF)	Small effect: a difference of 0.5 in item difficulty parameters across the two latent classes Large effect: a difference of 1.5 in item difficulty parameters across the two latent classes
Class-specific observed DIF effect size (ODIF)	Small effect: a difference of 0.5 in item difficulty parameters across the two observed groups Large effect: a difference of 1.5 in the item difficulty parameters across the two observed groups

Study Design Overview

For the present study, three types of models were estimated for each generated data set (Table 2) where correctly and incorrectly specified FMMs with binary outcomes were specified.

Table 2. *Combinations of correctly and incorrectly specified models*

Generating Model	Estimating Model	Specification
Model with LDIF and ODIF	Model with LDIF and ODIF	Correctly specified
	Model with LDIF	Under specified
	Model with ODIF	Under specified
	Model with no DIF	Under specified
Model with LDIF	Model with LDIF and ODIF	Over specified
	Model with LDIF	Correctly specified
	Model with ODIF	Mis-specified
	Model with no DIF	Under specified
Model with ODIF	Model with LDIF and ODIF	Over specified
	Model with LDIF	Differently specified
	Model with ODIF	Correctly specified
	Model with no DIF	Under specified

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF.

As shown in Table 2, three types of data were generated including data that fit a model with both between-class latent DIF and class-specific observed DIF, data that fit a model with only between-class latent DIF, and data that fit a model with only class-specific observed DIF. The correctly specified models were always estimated. In addition, some of the models that were estimated were incorrectly specified. More specifically, three models were estimated using each generated dataset, under each of four scenarios: over-specified, under-specified, mis-specified, and differently specified. In an over-specification scenario, models were estimated with both between-class latent DIF and class-specific observed DIF when generating data fitted to a model either with between-class latent DIF or with class-specific observed DIF. In an under-specification scenario, models were estimated with only between-class latent DIF or with only class-specific observed DIF when data have been generated to fit a model with both between-class latent DIF and class-specific observed DIF. In addition, models with no DIF were always under specified. A mis-specified model results when the model with only class-specific observed DIF was fit to a model with only between-class latent DIF. A "differently specified" model results when a model with only between-class latent DIF was fit to a model generated to have class-specific observed DIF.

The simulation conditions included class probability (50%:50% and 30%:70%) and DIF effect size (small effect .5 and large effect 1.5) for between-class latent DIF and/or class-specific observed DIF. When data were generated to fit a model having both between-class latent DIF and class-specific observed DIF, the simulation conditions were class probabilities (equal vs. unequal) and DIF effect sizes for both between-class latent DIF (small vs. large) and class-specific observed DIF (small vs. large), resulting in eight conditions for the one correctly specified and three incorrectly specified models. Therefore, thirty-two models were estimated. When data were generated to fit a model having between-class latent DIF, class probability (equal vs. unequal) and between-class latent DIF effect size (small vs. large) were simulated, resulting in four simulation conditions for the one correctly specified and three incorrectly specified models. Therefore, sixteen models were estimated. When data were generated to fit a model having class-specific observed DIF, class probability (equal vs. unequal) and class-specific observed DIF (small vs. large) were simulated, resulting in four simulation conditions for the one correctly specified and three incorrectly specified models. Therefore, sixteen models were estimated. As a result, a total of 64 models was estimated. Each of the 64 models was evaluated in terms of correct model-identification rates using fit indices (AIC, BIC, aBIC, and CAIC), class assignment, DIF detection rates, and parameter recovery for correctly identified DIF items. Table 3 shows a summary of the resulting simulated conditions.

Table 3. *Resulting Simulated conditions*

Generating Model	Estimating Model	Simulated Condition	Resulting simulated conditions
LDIF & ODIF	LDIF & ODIF LDIF ODIF No DIF	Class Proportion (50:50; 70:30) Between-class latent DIF (.5; 1.5) Class-specific observed DIF (.5; 1.5)	32
LDIF	LDIF & ODIF LDIF ODIF No DIF	Class Proportion (50:50; 70:30) Between-class latent DIF (.5; 1.5)	16
ODIF	LDIF & ODIF LDIF ODIF No DIF	Class Proportion (50:50; 70:30) Class-specific observed DIF (.5; 1.5)	16

Note. LDIF represents between-class latent DIF; ODIF represents class-specific observed DIF

Data Generation

The difficulty parameters used in the present study were adapted from the simulation study conducted by Maij-de Meij et al. (2011). The researchers used a 27-item test with difficulty parameters ranging from -.99 to .99 with a mean of 0 for latent class 1 (reference group). In the present study, the difficulty parameters were varied by adding .5 (small effect) or 1.5 (large effect) to four or eight items that were generated so as to have between-class latent DIF and/or class-specific observed DIF for the second latent class (focal group). This resulted in a maximum value of 2.21 for the difficulty parameter for item 4 (in the large DIF effect size condition). When data were generated to fit a model exhibiting both between-class and class-specific observed DIF, the between-class latent uniform DIF was simulated for items 1, 2, 3, and 4 (in boldfaced font in Table 4), and class-specific observed uniform DIF was simulated for items 5, 6, 7 and 8 (in italic and underlined font in Table 4). The values of the item parameters in a model with both between-class latent DIF and class-specific observed DIF are presented in Table 4.

Table 4. *Difficulty parameters of the simulation condition of small DIF effects in a model with both between-class and class-specific observed DIF*

	Latent class 1		Latent class 2	
Item	*X*=0 (e.g., male)	*X*=1 (e.g., female)	*X*=0 (e.g., male)	*X*=1 (e.g., female)
1	**-0.89**	**-0.89**	**-0.39**	**-0.39**
2	**-0.33**	**-0.33**	**0.17**	**0.17**
3	**0.27**	**0.27**	**0.77**	**0.77**
4	**0.71**	**0.71**	**1.21**	**1.21**
5	*-0.71*	*-0.71*	*-0.71*	*-0.21*
6	*-0.2*	*-0.2*	*-0.2*	*0.30*
7	*0.13*	*0.13*	*0.13*	*0.63*
8	*0.63*	*0.63*	*0.63*	*1.13*
9	-0.99	-0.99	-0.99	-0.99
10	-0.80	-0.80	-0.80	-0.80
11	-0.63	-0.63	-0.63	-0.63
12	-0.55	-0.55	-0.55	-0.55
13	-0.48	-0.48	-0.48	-0.48
14	-0.4	-0.4	-0.4	-0.4
15	-0.27	-0.27	-0.27	-0.27
16	-0.13	-0.13	-0.13	-0.13
17	-0.07	-0.07	-0.07	-0.07
18	0.1	0.1	0.1	0.1
19	0.07	0.07	0.07	0.07
20	0.2	0.2	0.2	0.2
21	0.33	0.33	0.33	0.33
22	0.4	0.4	0.4	0.4
23	0.48	0.48	0.48	0.48
24	0.55	0.55	0.55	0.55
25	0.8	0.8	0.8	0.8
26	0.89	0.89	0.89	0.89
27	0.99	0.99	0.99	0.99

Note. Difficulty values presented in boldfaced font identify items with between-class latent uniform DIF; difficulty values presented in italics and underlined font identify items with class-specific observed DIF.

The example provided in Table 4 holds for the conditions in which difficulty parameters for items 1, 2, 3, and 4 differ by 0.5 (a small effect) between the two latent classes. The direction of the generated between-class latent DIF effect was consistent across the four items such that difficulty parameters were always higher for the DIF items in latent class 2 than for the ones in latent class 1. In addition, it should be noted that difficulty parameters for DIF items in latent class 1 were the same across the two observed groups for items 5, 6, 7, and 8, meaning that there was no observed source of DIF within latent class 1. However, this was not the case for latent class 2. Item difficulties for items 5, 6, 7 and 8 differed by observed group membership within latent class 2. The difference in difficulty parameters for the observed group in latent class 2 was also 0.5 for the small class-specific observed DIF effect conditions. As with the latent DIF condition, the pattern of the differences in difficulty parameters for the four class-specific observed DIF scenarios was consistent such that the item is always harder for members of the $X = 1$ group than for members of the $X = 0$ group.

When data were generated to fit a model having only between-class latent DIF, items 1, 2, 3, and 4 were identified as displaying between-class latent DIF (matching the pattern of values in Table 4), and difficulty parameters for items 5 to 8 as well as for the rest of the items were generated to be the same across the two latent classes and two observed groups (matching the values when $X = 0$ for latent class 1 in Table 4). When data were generated to fit a model having only class-specific observed DIF, items 5, 6, 7 and 8 were identified as displaying the pattern of class-specific observed DIF shown in Table 5. Difficulty parameters for items 1 to 4 as well as for the rest of the items were generated to be the same across both latent classes and observed groups (matching the values when $X = 0$ for latent class 1 in Table 4).

Dichotomous item responses for each of replication datasets per combination of conditions were generated using the 1PL IRT model for each latent class using the IRTGEN SAS macro (Whittaker, Fitzpatrick, Williams, & Dodd, 2003). The ability parameters for the first latent class were drawn from a standard normal distribution with a mean of one and a standard deviation of one. For the second latent class, the ability parameters were drawn from a standard normal distribution with a mean of zero and a standard deviation of one.

Fifty replications were generated for each combination of simulation conditions, because estimating FMMs with binary outcomes requires considerable computer time. Previous methodological DIF studies involving use of M*plus* software have suggested that thirty or fifty replications were enough to explore the relative contributions of the varying conditions (DeMars & Lau, 2011; Jackman, 2011).

Model Estimation

The item parameters were estimated using M*plus* V6 (L. Muthén & Muthén, 1998-2010) with robust maximum likelihood estimation using the expectation-maximization (EM) algorithm. Maximum likelihood estimation is an iterative procedure, so that the log-likelihood function monotonically increases until it reaches one final maximum. However, sometimes it converges to a local rather than a global maximum. Thus, one recommended approach is to use multiple starting values (L. Muthén & Muthén, 1998-2010). Although the default in M*plus* is 10 random starting values with the

2 best sets to be used for final optimization, M*plus* allows users to increase the number of starting values for final optimization. But as the number of starting values increases, the estimation time increases considerably. The present study used 50 sets of starting values, with 10 solutions with highest log likelihood retained and iterated until the convergence criterion is reached (see, for example, Lubke & Muthén, 2007).

Data Analysis

Four outcome measures were summarized and compared across conditions: the relative parameter bias and standard error bias of items' difficulties and of DIF effects as well as the performance of information criteria in correct model selection. In addition, the average entropy values were summarized and compared across conditions.

Relative parameter bias. The accuracy of parameter recovery for items was evaluated using relative parameter bias ($RPB(\hat{\theta}_i)$). The bias of the parameter estimates was evaluated for between-class latent DIF items (1, 2, 3, and 4) and/or class-specific observed DIF items (5, 6, 7, and 8) when DIF items are correctly specified in models.

For item *i*, relative parameter bias in the difficulty parameter is defined as

$$RPB(\hat{\theta}_i) = \frac{(\bar{\hat{\theta}}_i - \theta_i)}{\theta_i} \qquad [25]$$

where θ_i is the generating true value of the *i*th parameter and $\bar{\hat{\theta}}_i$ is the mean estimate of the *i*th parameter across the 50 converged replication datasets (Hoogland & Boomsma, 1998). Hoogland and Boomsma (1998) defined acceptable parameter estimate bias as $|B(\hat{\theta}_i)| < .05$.

Relative standard error (SE) bias. The relative standard error bias between the empirical standard error and the average *SE* estimate across each condition's 50 estimates is defined as

$$B(S\hat{E}_{\hat{\theta}_i}) = \frac{\overline{S\hat{E}}_{\hat{\theta}_i} - S\hat{E}_{\hat{\theta}_i}}{S\hat{E}_{\hat{\theta}_i}} \qquad [26]$$

where $\overline{S\hat{E}}_{\hat{\theta}_i}$ is the mean estimated standard error of the parameter estimate $\hat{\theta}$ for item *i* across the estimated standard errors of the converged solutions, and $S\hat{E}_{\hat{\theta}_i}$ is the estimated population standard error value of $\hat{\theta}$ for item *i.* Hoogland and Boomsma (1998) defined acceptable bias of the standard error estimates as $|B(S\hat{E}_{\hat{\theta}})| < .10$.

Relative parameter bias of DIF effects. The relative parameter bias will be calculated for the difference between item's difficulties (that is, DIF effect). The bias of DIF effect estimates will be evaluated for between-class latent DIF items (1, 2, 3, and 4) and/or class-specific observed DIF items (5, 6, 7, and 8) when DIF items are correctly specified in models. Because the true DIF effect will be zero for incorrectly specified DIF items (either between-class latent DIF or class-specific observed DIF) for incorrectly specified models, use of $RPB(\hat{\theta}_D)$ will not be appropriate. Instead the average bias for the difference between items' difficulties will be calculated for these items.

For the DIF effect of item *i*, relative parameter bias in the difference between difficulties for item *i* will be defined as

$$RPB(\hat{\theta}_D) = \frac{(\bar{\hat{\theta}}_{(i1-i2)} - \theta_{(i1-i2)})}{\theta_{(i1-i2)}} \qquad [27]$$

where $\theta_{(i1-i2)}$ is the generating true DIF effect of the *i*th parameter and $\bar{\hat{\theta}}_{(i1-i2)}$ is the mean estimate of the DIF effect for the *i*th parameter across the 50 converged replication datasets. The subscripts *i*1 and *i*2 represent item *i* for each latent class 1 and 2, respectively.

Fit indices. For each of the four models estimated per dataset, AIC, BIC, aBIC, and CAIC were estimated and compared across models. For each of the fit indices, the lowest value for each of the fit indices for each dataset across the four models estimated was identified and tallied. The proportion of replications in which each of the fit indices led to selection of the correct model was compared across simulation conditions and fit indices.

DIF detection. The present study used a chi-square statistic described by Lord (1980) to test for both unobserved and observed sources of uniform DIF (for example, Maij-de Meij et al., 2011). A chi-square statistic was used to assess the statistical significance of the mean differences in DIF items' difficulties between latent classes (or observed groups) for each replicated data set. The differences between the difficulty parameters across two different groups of examinees for an item, say $\hat{b}_1$ and $\hat{b}_2$, is an estimated item parameter for each group, respectively, that can be examined using

$$\chi_i^2 = \frac{(\hat{b}_{i1} - \hat{b}_{i2})^2}{\sigma_{\hat{b}_{i1}}^2 + \sigma_{\hat{b}_{12}}^2} \qquad [28]$$

where each $\sigma^2_{\hat{b}_{i1}}$ and $\sigma^2_{\hat{b}_{i2}}$ is the variance of difficulty parameter of the item *i* for each group. For each condition and replication, the number of times an item is identified as exhibiting DIF was recorded.

Latent class membership. Correct class assignment was computed as the proportion of subjects for whom the highest posterior class probability is equal to the true class probability, based on their highest posterior class probability. Entropy is closely related to the mean of each individual's highest class probability across individuals. When models were correctly specified, entropy was assessed to evaluate how it is affected by the simulation conditions examined for the present study. In addition, when comparing the correctly specified models to incorrectly specified models, entropy was assessed to evaluate how well the correctly specified models perform in assigning individuals to their true classes.

Chapter 4: Results

Convergence Rates

The present study investigated proper convergence for models estimated using data generated in each replication. For each condition, of the 50 replications attempted, 2% to 4% of models being estimated did not converge (see Table 5). Similarly, in Jackman's (2012) study there were minimal convergence problems: in general, model estimation for 96% of replications successfully converged. For the present study, new data sets were generated so that 100% of the results being analyzed were based on converged solutions for each model and condition.

Table 5. *Percentage of Convergence Rates for 1st 50 Replications for Each Condition and Generating Model*

		Equal class probability condition				Unequal class probability condition			
Generating Model	Estimating Model	Small DIF	Large DIF	Large LDIF& Small ODIF	Small LDIF & Large ODIF	Small DIF	Large DIF	Large LDIF& Small ODIF	Small LDIF & Large ODIF
Generating LDIF & ODIF	LDIF&ODIF	100%	100%	100%	100%	100%	100%	100%	100%
	LDIF	100%	100%	100%	100%	100%	100%	100%	100%
	ODIF	100%	100%	100%	100%	100%	100%	100%	100%
	No DIF	100%	100%	100%	100%	100%	100%	100%	100%
Generating LDIF	LDIF&ODIF	100%	100%	100%	100%	100%	100%	100%	100%
	LDIF	100%	100%	100%	100%	100%	100%	100%	100%
	ODIF	100%	100%	100%	100%	*98%*	100%	100%	*98%*
	No DIF	100%	100%	100%	100%	100%	100%	100%	100%
Generating ODIF	LDIF&ODIF	*98%*	100%	100%	100%	*98%*	100%	100%	100%
	LDIF	*98%*	100%	100%	100%	100%	100%	100%	100%
	ODIF	*98%*	100%	100%	100%	100%	100%	100%	100%
	No DIF	100%	100%	100%	100%	100%	100%	*96%*	100%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF.

Performance of the Fit Indices

The performance of the most commonly used IC indices–AIC, BIC, aBIC, and CAIC–was evaluated to identify which of the indices most frequently supported the better fit of the correctly versus incorrectly specified models. The performance of all fit indices was evaluated as a function of DIF effect sizes and class probability. The proportion of replications in which each of the AIC, BIC, aBIC, and CAIC indices led to selection of the correctly specified model was recorded for each condition. The results are presented in Tables 6 to 9.

Small DIF effect size and equal class probability. As shown in Table 6, when data were generated to fit a model having both between-class latent DIF (LDIF) and class-specific observed DIF (ODIF), the AIC index performed better in selecting the correct model, doing so for 98% of the replications. The aBIC index's corresponding rate was 52%. The BIC index performed poorly in supporting the fit of the correct model, producing correct model identification rates of only 2%, and the CAIC index never selected the correct model. The BIC and CAIC indices selected the under-specified model estimating only LDIF effects when data were generated to fit a model having both types of DIF effects for 98% of the replications. However, when data were generated to fit a model with LDIF effects, the BIC, aBIC, and CAIC indices led to selection of the correct model for 100% of the replications. The AIC index led to selection of the correct model for 98% of the replications. When data were generated to fit a model having ODIF, the AIC index led to selection of the correct model for 86% of the replications, and the aBIC did so for 40% of replications.

Combination of DIF effect sizes and equal class probability. As shown in Table 6, when data were generated to fit a model having large LDIF and small ODIF effects, the AIC index led to selection of the correct model for 100% of the replications, and the aBIC index did so for 72%. The BIC and CAIC indices led to selection of the under-specified model estimating only LDIF effects for 100% of the replications. When data were generated to fit a model having small LDIF and large ODIF, the AIC, BIC, and aBIC indices led to selection of the correct model for 100% of the replications, but the CAIC index led to selection of the under-specified model estimating only LDIF effects for 100% of the replications.

Large DIF effect size and equal class probability. When data were generated to fit a model having both large LDIF and ODIF effects, all fit indices led to selection of the correct model for 100% of the replications (see Table 7). In addition, when data were generated to fit a model having LDIF effects, all fit indices performed perfectly by selecting the correct model for 100% of the replications. However, for data generated to fit a model having ODIF effects, the BIC, aBIC, and CAIC indices led to selection of the correct model for 100% of the replications, but the AIC index selected the correct model for 93% of the replications.

Small DIF effect size and unequal class probability. As shown in Table 8, for data generated to fit a model having both types of DIF effects under the condition small DIF effect size with unequal class probability, the AIC index performed best, correctly identifying the model for 82% of the replications and the aBIC index selected the correct model for 2% of the replications. The BIC and CAIC indices selected the under-specified model with no DIF for 82% and 100% of replications, respectively. The aBIC index led to selection of the under-specified model estimating only LDIF effects and the under-specified model with no DIF for 80% and for 18% of the replications, respectively.

When data were generated to fit a model having LDIF effects, the AIC index led to selection of the correct model for 88% of the replications, and the aBIC led to selection of the correct model for 94% of the replications. For data generated to fit a model having ODIF effects, only the AIC index selected the correct model for 78% of the replications, and other fit indices, the BIC and aBIC selected the under-specified model with no DIF for 100% of the replications. The CAIC selected the under-specified model with no DIF for 96% of the replications.

Combination of DIF effect sizes and unequal class probability. When data were generated to fit a model having large LDIF and small ODIF, the AIC index led to selection of the correct model for 86% of the replications. On the other hand, the aBIC index led to selection of the correct model for only 4% of the replications, and other fit indices never selected the correct model (see Table 8). Instead, other fit indices, including the BIC, aBIC, and CAIC, led to selection of the under-specified model estimating only LDIF effects for 82% to 86% of the replications. However, when data were generated to fit a model having small LDIF and large ODIF, the AIC and aBIC indices led to selection of the correct model for 100% of the replications. The BIC and CAIC led to selection of the correct model for 54% and 36% of the replications, respectively.

Large DIF effect size and unequal class probability. As shown in Table 9, when data were generated to fit a model having both LDIF and ODIF effects with large DIF effect sizes and unequal class probabilities, all fit indices led to selection of the correct model for 100% of the replications. Likewise, when data were generated to fit a model having large LDIF effects, all fit indices led to selection of the correct model for 100% of the replications. On the other hand, when data were generated to fit a model having large ODIF effects, the BIC, aBIC, and CAIC indices performed perfectly by selecting the correct model for 100% of the replications, but the AIC index selected the correct model for 96% of the replications.

Table 6. *Fit indices for Generating and Estimating Models under the Conditions of Small DIF Effects and Equal Class Probability*

Generating Model	Estimating Model	AIC	BIC	aBIC	CAIC
LDIF & ODIF	LDIF & ODIF	**98%**	**2%**	**52%**	**0%**
	LDIF	2%	98%	48%	98%
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	2%
LDIF	LDIF & ODIF	0%	0%	0%	0%
	LDIF	**98%**	**100%**	**100%**	**100%**
	O DIF	2%	0%	0%	0%
	No DIF	0%	0%	0%	0%
ODIF	LDIF & ODIF	4%	0%	0%	0%
	LDIF	8%	4%	18%	0%
	O DIF	**86%**	**4%**	**40%**	**0%**
	No DIF	2%	92%	42%	100%
Large LDIF & Small ODIF	LDIF & ODIF	**100%**	**0%**	**72%**	**0%**
	LDIF	0%	100%	28%	100%
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	0%
Small LDIF & Large ODIF	LDIF & ODIF	**100%**	**100%**	**100%**	**0%**
	LDIF	0%	0%	0%	100%
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	0%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF.

Table 7. *Fit indices for Generating and Estimating Models under the Conditions of Large DIF Effects and Equal Class Probability*

Generating Model	Estimating Model	AIC	BIC	aBIC	CAIC
LDIF & ODIF	LDIF & ODIF	**100%**	**100%**	**100%**	**100%**
	LDIF	0%	0%	0%	0%
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	0%
LDIF	LDIF & ODIF	0%	0%	0%	0%
	LDIF	**100%**	**100%**	**100%**	**100%**
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	0%
ODIF	LDIF & ODIF	0%	0%	0%	0%
	LDIF	7%	0%	0%	0%
	O DIF	**93%**	**100%**	**100%**	**100%**
	No DIF	0%	0%	0%	0%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF.

Table 8. *Fit indices for Generating and Estimating Models under the Conditions of Small DIF Effects and Unequal Class Probability*

Generating Model	Estimating Model	AIC	BIC	aBIC	CAIC
LDIF & ODIF	LDIF & ODIF	**82%**	**0%**	**2%**	**0%**
	LDIF	18%	18%	80%	0%
	O DIF	0%	0%	0%	0%
	No DIF	0%	82%	18%	100%
LDIF	LDIF & ODIF	4%	0%	0%	0%
	LDIF	**88%**	**36%**	**94%**	**16%**
	O DIF	8%	0%	0%	0%
	No DIF	0%	64%	6%	84%
ODIF	LDIF & ODIF	0%	0%	0%	0%
	LDIF	0%	0%	0%	0%
	O DIF	**78%**	**0%**	**0%**	**4%**
	No DIF	22%	100%	100%	96%
Large LDIF & Small ODIF	LDIF & ODIF	**86%**	**0%**	**4%**	**0%**
	LDIF	4%	86%	82%	86%
	O DIF	0%	0%	0%	0%
	No DIF	10%	14%	14%	14%
Small LDIF & Large ODIF	LDIF & ODIF	**100%**	**54%**	**100%**	**36%**
	LDIF	0%	0%	0%	0%
	O DIF	0%	46%	0%	64%
	No DIF	0%	0%	0%	0%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF.

Table 9. *Fit indices for Generating and Estimating Models under the Conditions of Large DIF Effects and Unequal Class Probability*

Generating Model	Estimating Model	AIC	BIC	aBIC	CAIC
LDIF & ODIF	LDIF & ODIF	**100%**	**100%**	**100%**	**100%**
	LDIF	0%	0%	0%	0%
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	0%
LDIF	LDIF & ODIF	0%	0%	0%	0%
	LDIF	**100%**	**100%**	**100%**	**100%**
	O DIF	0%	0%	0%	0%
	No DIF	0%	0%	0%	0%
ODIF	LDIF & ODIF	4%	0%	0%	0%
	LDIF	0%	0%	0%	0%
	O DIF	**96%**	**100%**	**100%**	**100%**
	No DIF	0%	0%	0%	0%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF.

Power for Detecting DIF Effects

Power was assessed based on the proportion of times out of the 50 replications in which the DIF items were accurately identified as having DIF. Typically, a power of 80% is regarded as relatively accurate in correctly identifying items with DIF, based on the standard set by Cohen's study in 1988 (Jackman, 2011; Samuelsen, 2005). Results for the power analysis are shown in Tables 10 and 11.

Small DIF effect size with equal class probability. The overall accuracy of LDIF detection ranged from 66.50% to 72.50% when a DIF effect size was small with equal class probability (see Table 10). On the other hand, the overall accuracy of ODIF detection was substantially lower than the overall accuracy of LDIF detection, ranging from 27.50% to 31.00%. More specifically, when data were generated to fit a model having both types of DIF effects, the power to detect LDIF for the correct model estimating both LDIF and ODIF effects was 72.50%, but the power to detect ODIF was 31.00%. For the under-specified model estimating only LDIF effects when data were generated to fit a model having both types of DIF effects, the power to detect LDIF was marginally reduced to 66.50%. Likewise, the power to detect ODIF was also reduced to 27.50% for the under-specified model estimating only ODIF effects, compared to the detection rates under the correct model. When data were generated to fit a model having LDIF effects, the accuracy of detecting LDIF effects was not much different between the over-specified model estimating both types of DIF effects and correctly specified models. That is, the power to detect LDIF was 71.50% for the correct model estimating LDIF effects and 67.50% for the over-specified model estimating both effects. On the other hand, when data were generated to fit a model having ODIF, the power to detect ODIF was 28.50% under the correct model estimating ODIF effects, but the power was reduced to 27.50% under the over-specified model estimating both types of DIF effects.

Large DIF effect size with equal class probability. When the DIF effect size was large with equal class probability, all power for detecting both LDIF and ODIF effects was acceptable, ranging from 94% to 100% (see Table 10). Specifically, the power for detecting LDIF was 100% across the correct model and the under-specified model estimating only LDIF effects, but the power for detecting ODIF was reduced to 94% for the under-specified model estimating only ODIF effects. When data were generated to fit a model having large LDIF effects, power for detecting LDIF effects was 100% across the correctly specified and the over-specified models. However, when data were generated to fit a model having large ODIF effects, all power for detecting ODIF effects was 95.50% across correctly specified model and over-specified model estimating both types of DIF effects.

Combination of different magnitudes of DIF effect sizes with equal class probability. When data were generated to fit a model having a combination of different magnitudes of DIF effect sizes, the power for detecting LDIF and ODIF was different (see Table 10). Specifically, for the correct model estimating large LDIF and small ODIF effects, the power to detect LDIF was 100%, while the power for detecting ODIF was 47%. In addition, the power to detect LDIF was 100% even though the model was under-specified with only LDIF effects. However, the power to detect ODIF with the under-specified model estimating only ODIF effects was reduced to 31.50%.

On the other hand, when a model had a different pattern of effect sizes for LDIF and ODIF (i.e., small LDIF and large ODIF were generated) then the power to detect LDIF was acceptable (86.50%), and the power rate for detecting ODIF was as high as 96.50%. However, for the under-specified model estimating only LDIF effects, the power for detecting LDIF decreased to 63%. The power for detecting ODIF with the under-specified model estimating only ODIF effects was also 96.50%.

Small DIF effect size with unequal class probability. As shown in Table 11, the overall accuracy of LDIF detection ranged from 55% to 64% when the DIF effect size was small with unequal class probability (see Table 11). On the other hand, the overall accuracy of ODIF detection ranged from 23% to 25.50%. Specifically, when data were generated to fit a model having both types of DIF effects, the power for detecting LDIF effects was 58% and the power for detecting ODIF effects was 23%. Similarly, the power for detecting DIF effects were 55.50% and 25.50% for the under-specified models estimating only LDIF and only ODIF effects, respectively. When data were generated to fit a model having LDIF effects, the power for detecting LDIF was 64% for the correct model correctly estimating LDIF effects and 55% for the over-specified model estimating both types of DIF effects. Likewise, the power for detecting ODIF under the correct model estimating ODIF effects was 24.50%, but the power for detecting ODIF under the over-specified model estimating both types of DIF effects decreased slightly to 23%.

Large DIF effect size with unequal class probability. When the DIF effect size was large with unequal class probability, the power to detect both LDIF and ODIF effects was acceptable, ranging from 89% to 100% (see Table 11). Under the correct model estimating both types of DIF effects, the power was 100% for detection of LDIF and ODIF effects. Under incorrectly specified models, the power was still high at 100% and 94.50% for LDIF and ODIF effects, respectively. For data generated to fit a model having LDIF effects, the power to detect LDIF was 100% for both the correct model estimating LDIF effects and the over-specified model estimating both types of DIF effects. On the other hand, when data were generated to fit a model having ODIF effects, the power to detect ODIF was lower at 89% across the correctly specified and the over-specified models.

Combination of different magnitudes of DIF effect sizes with unequal class probability. For the correct model estimating large LDIF and small ODIF effects, the power to detect LDIF and ODIF effects were 100% and 40%, respectively (see Table 11). When models were under-specified, the power to detect LDIF with the under-specified model estimating only LDIF effects was still 100% while the power to detect ODIF with the under-specified model estimating only ODIF effects was lower at 26%. On the other hand, when data were generated to fit a model having a different pattern of DIF effect sizes (i.e., small LDIF and large ODIF) then the power to detect LDIF under the correctly specified model was lower at 75.70%, and the power to detect LDIF with the under-specified model estimating only LDIF effects decreased to 50%. In contrast to the effect on power for detecting LDIF, the power for detecting ODIF increased substantially to 90.50%. In addition, with the under-specified model estimating only ODIF effects, the power was even slightly higher at 92%.

Table 10. *Power for Identifying DIF by Generating and Estimating Models under the Conditions of Equal Class Probability*

Generating Model	Estimating Model	Small DIF Effect		Large DIF Effect	
		Power (LDIF)	Power (ODIF)	Power (LDIF)	Power (ODIF)
LDIF & ODIF	LDIF & ODIF	72.50%	31.00%	100%	100%
	LDIF	66.50%	-	100%	-
	O DIF	-	27.50%	-	94.00%
LDIF	LDIF & ODIF	67.50%	-	100%	-
	LDIF	71.50%	-	100%	-
ODIF	LDIF & ODIF	-	27.50%	-	95.50%
	O DIF	-	28.50%	-	95.50%
Large LDIF & Small ODIF	LDIF & ODIF	-	47.00%	100%	-
	LDIF	-	-	100%	-
	O DIF	-	31.50%	-	-
Small LDIF & Large ODIF	LDIF & ODIF	86.50%	-	-	96.50%
	LDIF	63.00%	-	-	-
	O DIF	-	-	-	96.50%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF; − = not estimated.

Table 11. *Power for Identifying DIF by Generating and Estimating Models under the Conditions of Unequal Class Probability*

Generating Model	Estimating Model	Small DIF Effect		Large DIF Effect	
		Power (LDIF)	Power (ODIF)	Power (LDIF)	Power (ODIF)
LDIF & ODIF	LDIF & ODIF	58.00%	23.00%	100%	100%
	LDIF	55.50%	-	100%	-
	O DIF	-	25.50%		94.50%
LDIF	LDIF & ODIF	55.00%	-	100%	-
	LDIF	64.00%	-	100%	-
ODIF	LDIF & ODIF	-	23.00%	-	89.00%
	O DIF	-	24.50%	-	89.00%
Large LDIF & Small ODIF	LDIF & ODIF	-	40.00%	100%	-
	LDIF	-	-	100%	-
	O DIF	-	26.00%	-	-
Small LDIF & Large LDIF	LDIF & ODIF	75.70%	-	-	90.50%
	LDIF	50.00%	-	-	-
	O DIF	-	-	-	92.00%

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF; − = not estimated.

Type I Error Rates

When data were generated to fit a model having LDIF effects but estimated both types of DIF effects (that is, over-specified ODIF effects) or ODIF (that is, mis-specified ODIF effects), the Type I error rates were assessed for incorrect DIF identification (that is, ODIF effects). In addition, when data were generated to fit a model having ODIF but estimated both types of DIF (that is, over-specified LDIF effects) or LDIF (that is, differently specified LDIF effects), the Type I error rates were assessed for incorrect DIF identification (that is, LDIF effects). The Type I error rates are shown in Table 12.

Equal class probability. As shown in Table 12, when data were generated to fit a model having LDIF effects but estimated both LDIF and ODIF effects under small DIF effect size conditions, the Type I error rate of incorrectly identified ODIF effects was 5.50%, but the Type I error rate of incorrectly identified ODIF effects slightly increased to 6.5% under the large LDIF effect size condition. On the other hand, when data were generated to fit a model having LDIF effects but the estimating model included only ODIF effects (mis-specification), the Type I error rate of mis-specified ODIF effects was 7% under the condition of small DIF effect size, but this rate increased to 11.50% as the true DIF effect size increased.

When data were generated to fit a model having ODIF effects but the estimating model included both LDIF and ODIF effects, the Type I error rates for the incorrectly specified LDIF effects were 8% and 11% for the small and large DIF effect size conditions, respectively. However, when data were generated to fit a model having ODIF effects, but the estimating model included LDIF effects (that is, different specification), the Type I error rates were on average 22.50%, and increased substantially to 99.75% as the true DIF effect size increased.

Unequal class probability. As shown in Table 12, when data were generated to fit a model having LDIF effects but estimated both LDIF and ODIF effects under the small DIF effect condition, the Type I error rate for incorrectly specified ODIF effects was 10.50%. The Type I error rate for incorrectly specified ODIF effects decreased slightly to 9% under large DIF effect size condition. On the other hand, when data were generated to fit a model having LDIF effects but estimated ODIF effects (mis-specification) under the small DIF effect size condition, the Type I error rate for mis-specified ODIF effects was 8.50%. However, the Type I error rate for mis-specified ODIF decreased to 6% as the true DIF effect size increased.

When a model was generated to have only ODIF effects but estimated both LDIF and ODIF effects, the Type I error rates for incorrectly specified LDIF effects were 7.50% across small and large DIF effect size conditions. When data were generated to fit a model having ODIF effects, but estimated LDIF effects (different specification), the Type I error rate for differently specified LDIF effects was 23%, and this rate increased substantially to 98% as the true DIF effect size increased.

Table 12. *Type I Error Rates for Incorrect DIF Identification by Generating and Estimating Models across Simulation Conditions*

Generating Model	Estimating Model	Equal Class Probability		Unequal Class Probability	
		Small DIF Effect	Large DIF Effect	Small DIF Effect	Large DIF Effect
LDIF	LDIF & ODIF	5.50%	6.50%	10.50%	9.00%
	O DIF	7.00%	11.50%	8.50%	6.00%
ODIF	LDIF & ODIF	8.00%	11.00%	7.50%	7.50%
	LDIF	22.50%	99.75%	23.00%	98.00%

Note. LDIF = between-class latent DIF; ODIF = class-specific observed DIF.

The percents appearing in the table were the Type I error rates for the relevant DIF that was mis-specified for the estimating model.

Entropy

The entropy values averaged across 50 replications for correctly specified and incorrectly specified models (under-, over-, mis-, or different specification) were compared across model specifications and simulation conditions. As mentioned above, the entropy value was calculated based on the estimated posterior probability for an individual in an estimated class. That is, the entropy value indicates how well individuals were classified into their estimated classes. It is presented in Table 13.

Equal class probability. Under the condition of equal class probability, class probabilities were simulated to be 50% versus 50% for each latent class and the estimated class probabilities were almost equal across latent classes under the correctly specified models. That is, individuals were classified equally into each latent class. Factor means were used to identify the two classes. As shown in Table 13, entropy values on average were not large, ranging from .280 to .537 even for the correctly specified models. More specifically, when data were generated to fit a model having both types of DIF effects under small DIF effect size conditions, the entropy value was on average .280. Unexpectedly, the average entropy values were higher for the under-specified models estimating only LDIF effects (.311) or only ODIF effects (.344) even though the average entropy value was relatively low under the under-specified model with no DIF (.243) in comparison to that for the correctly specified model.

Under the correct model estimating LDIF effects, the entropy values were on average .294. Even though models were incorrectly specified (that is, over-specification and mis-specification), the average entropy values were not much different than those for the correct model estimating LDIF effects, with an average entropy value of .292 for the over-specified model estimating both types of DIF effects and .332 for the mis-specified model with ODIF effects. But the average entropy value were lower (.236) for the under-specified model with no DIF. Similarly, when data were generated to fit a model having ODIF effects, the entropy values were on average .261 for the correctly specified model and the over-specified model estimating both types of DIF effects and .253 for the model with LDIF effects. However, the average entropy value was lower at .220 for the model with no DIF. This result indicated that the average entropy values were not much different between the correctly and incorrectly specified models, except for the model with no DIF.

When data were generated to fit a model having both types of DIF effects under the large DIF effect size conditions, the average entropy values improved to .537 (see Table 13). As expected, the average entropy values were higher for the correctly specified model than for the incorrectly specified models (under-specified models estimating only LDIF, ODIF effects or no DIF). For example, the entropy value was on average .484 and .395 for the under-specified model estimating only LDIF effects and the under-specified model estimating only ODIF effect, respectively. For data generated to fit a model having LDIF, the entropy values were on average .505 and .507 for the correctly specified and over-specified models, respectively. However, the entropy was slightly higher at .541 for the mis-specified model in which the model with ODIF was fit to the model with LDIF. However, the average entropy value under the under-specified model with no DIF were relatively lower. On the other hand, under the correct model estimating ODIF effects, the entropy value was on average .326 for the correctly specified model and the over-specified model estimating both types of DIF and .325 for the differently specified model. The result indicated that the average entropy values and were not much different between the correctly specified and incorrectly specified models.

When data were generated to fit a model having a combination of large LDIF and small ODIF effects, the entropy value was on average .479 for the correct model and .476 for the under-specified model with only LDIF (large DIF effect size). However, the average entropy values for the under-specified model with ODIF effects (small DIF effect size) or for the under-specified model with no DIF were relatively low (.296 and .289, respectively) compared to that for the correctly specified model. When data were generated to fit a model having different magnitudes of DIF effect sizes (that is, small LDIF and large ODIF effect sizes), the entropy value was on average .384 for the correct model and .347 for the under-specified model with only ODIF effects (large DIF effect size). However, the average entropy values for the under-specified model with only LDIF effects (small DIF effect size) and for the under-specified model with no DIF were relatively low (.319 and .290, respectively) compared to those for the correctly specified models.

Unequal class probability. Under the condition of unequal class probability, class probabilities were simulated to be 70% versus 30%, in the present study. Estimated probabilities were 60% to 77% for the reference class and 23% to 40% for the focal class across model specifications and simulation conditions. In general, average entropy values ranged from .360 to .643 for correctly specified models and from .308 to .618 for corresponding incorrectly specified models. The average entropy values under unequal class probability conditions were slightly higher than those under equal class probability conditions. More specifically, when the DIF effect size was small, entropy values were on average .405 and the average entropy value was slightly higher for the correct model than for incorrectly specified models. However, when data were generated to fit a model estimating LDIF effects, the entropy value was on average .360, and the average entropy values did not differ between correctly and incorrectly specified models. Similarly, for the correct model estimating ODIF effects, the entropy value was on average .370. When the correct model was compared to incorrectly specified models, the average entropy values did not differ.

As the DIF effect size increased, the entropy values also increased. In addition, the difference in entropy values among correctly specified models increased. For example, the entropy values increased to .643 for the correct model estimating both types of DIF effects, to .551 for the correct model estimating LDIF effects, and to .432 for the correct model estimating ODIF effects. That is, the quality of class assignment for the correct model estimating both types of DIF effects was larger than those for either the correct model estimating LDIF or ODIF effects.

When correctly specified models were compared with incorrectly specified models, for data generated to fit a model having both types of DIF, the average entropy value for the correctly specified model was .643, and it decreased to .521 for the under-specified model estimating only LDIF effects and to .396 for the under-specified model estimating ODIF effects. When data were generated to fit a model having LDIF effects, the entropy value was on average .551 under the correctly specified model, it was .568 for the over-specified model estimating both types of DIF effects, and it was .618 for the mis-specified model. That is, they were slightly larger for the mis-specified model than were those for the correctly specified model. Likewise, when data were generated to fit a model having ODIF, the average entropy values were similar, .432 and .454, under the correctly specified model and incorrectly specified models. However, the average entropy values were lower, at .322 for the under-specified model with no DIF.

For data generated to produce a combination of different magnitudes of effect sizes, similar to the results for equal class probability, the average entropy values for the correctly specified models were larger than were those for under-specified models. For example, the entropy values were on average .533 for the correct model estimating large LDIF and small ODIF effects, but the range of the average entropy values was from .308 to .499 for under-specified models. Likewise, the entropy value was on average .455 for the correct model estimating small LDIF and large ODIF effects, but the range of the average entropy values was from .318 and .418 for the under-specified models.

Table 13. *Average Entropy Values for Generating and Estimating Models under Simulation Conditions*

		Equal Class Probability				Unequal Class Probability			
Generating Model	Estimating Model	Small DIF	Large DIF	Large LDIF &Small ODIF	Small LDIF & Large ODIF	Small DIF	Large DIF	Large LDIF & Small ODIF	Small LDIF & Large ODIF
LDIF & ODIF	LDIF & ODIF	.280	.537	.479	.384	.405	.643	.533	.455
	LDIF	.331	.484	.476	.319	.371	.521	.499	.374
	O DIF	.344	.395	.296	.347	.356	.396	.323	.418
	No DIF	.243	.367	.289	.290	.322	.313	.308	.318
LDIF	LDIF & ODIF	.292	.507	-	-	.378	.568	-	-
	LDIF	.294	.505	-	-	.360	.551	-	-
	O DIF	.332	.541	-	-	.361	.618	-	-
	No DIF	.236	.340	-	-	.325	.414	-	-
ODIF	LDIF & ODIF	.261	.326	-	-	.376	.429	-	-
	LDIF	.253	.325	-	-	.359	.454	-	-
	O DIF	.261	.326	-	-	.370	.432	-	-
	No DIF	.220	.237	-	-	.333	.322	-	-

Note. LDIF=between-class latent DIF; ODIF=class-specific observed DIF; − = not estimated.

Relative Parameter Bias

The relative parameter bias (*RPB*) of item difficulty estimates for each latent class was computed for only correctly specified models. The values are presented in Tables 14 through 15. Substantial parameter estimate bias was identified for any estimate when the magnitude of the parameter bias was greater than .05 (Hoogland & Boomsma, 1998). An item difficulty parameter value can be negative or positive. In the present study, difficulty parameters for items 1, 2, 5, and 6 were negative in the reference class. When an item difficulty parameter is positive and positive bias is observed, it means that the item's difficulty is over-estimated. When an item difficulty is positive and the bias is negative then the difficulty is under-estimated. If, on the other hand, the item difficulty is negative and positive bias is observed, this means that the item's difficulty parameter is under-estimated. And when an item difficulty parameter is negative and negative bias is observed, it means that the item's difficulty is over-estimated. For example, if the true value is −.2 but the estimated value is −.3, the relative parameter bias is positive, but the difficulty parameter is under-estimated.

Equal class probability. As shown in Table 14, the relative parameter bias for item difficulty estimates under the correct model estimating both types of DIF effects ranged from −.268 to .060 for the reference class and from −.057 to .091 for the focal class under the condition of equal class probability with small DIF effect size. That is, the magnitude of the relative parameter bias of difficulty estimates was larger for the reference class than for the focal class. On the other hand, the majority of the values of relative parameter bias in difficulty estimates was acceptable under the correct model estimating LDIF effects, but two item difficulties were under-estimated, and the range of the relative parameter bias was from −.086 to .037. However, for the correct model estimating ODIF effects, the relative parameter bias in item difficulty estimates for each latent class was excessive, ranging from −.541 to .119. That is, the majority of item difficulty values were under-estimated, and the magnitude of the relative parameter bias of difficulty estimates for the correct model estimating ODIF effects was relatively large in comparison to estimates for both the correct model estimating both types of DIF effects and the correct model estimating LDIF effects.

As the DIF effect size increased, acceptable relative parameter bias was found in the item difficulty parameter estimates for the focal class across all correctly specified models. However, relative parameter bias of difficulty estimates for the reference class was observed across all correctly specified models. More specifically, the relative parameter bias of difficulty estimates for the reference class was not excessive, ranging from −.092 to .061 for the correct model estimating both types of DIF effects and from −.060 to .075 for the correct model estimating LDIF effects. That is, two or three item difficulties were under-estimated even though the majority of item difficulties were well-estimated for the correct model estimating both types of DIF effects and for the correct model estimating LDIF effects. In addition, for the correct model estimating ODIF effects, the relative parameter bias of difficulty estimates for the reference class ranged from .051 to .078, with the exception of the difficulty estimate for item 6 ($RPB = -.246$). When data were generated to fit a model having different magnitudes of both types of DIF effects, all item difficulty values were well-estimated across latent classes with exception of the difficulty estimate for item 5 for the focal class (RPB=$-$.132) under the correct model estimating large LDIF and small ODIF. In contrast, the majority of item difficulty values were biased under the correct model estimating small LDIF and large ODIF. That is, the range of relative parameter bias values was from −.132 to .036 for the correct model estimating large LDIF and small ODIF effects and from −.092to .062 for the correct model estimating small LDIF and large ODIF effects.

Unequal class probability. Table 15 shows the relative parameter bias in difficulty estimates in conditions with unequal class probabilities with small DIF effect sizes. Similar to the relative parameter bias results under the equal class conditions, the relative parameter bias of difficulty estimates was unacceptable (ranging from −.113 to .092) for both classes under the correct model estimating both types of DIF effect sizes. In addition, it was observed that there was positive relative parameter bias of difficulty estimates under the correct model estimating LDIF effects and negative relative parameter bias under the correct model estimating ODIF effects. That is, the ranges of relative parameter bias were from −.013 to .129 for the correct model estimating LDIF effects and from −.087 to .039 for the correct model estimating ODIF effects. Even though the DIF effect size increased, the relative parameter bias found was similar with those values under the small DIF effect size conditions. The majority of relative parameter bias values in difficulty estimates for the focal class were acceptable under the correct model estimating both types of DIF effects and the correct model estimating LDIF effects. However, unacceptable but not excessive relative parameter bias was found in difficulty estimates for both latent classes. For example, the ranges of relative parameter bias were from−.161 to .089 for the reference class and from −.008 to .051 for the focal class under the correct model estimating both types of DIF effects. In addition, the ranges of relative parameter bias were from −.004 to .102 for the reference class and from .004 to .050 for the focal class under the correct model estimating LDIF effects. However, unacceptable relative parameter bias was found for both classes for the correct model estimating ODIF effects, ranging from −.138 to .041 for the reference class and from −.064 to .053 for the focal class.

When data were generated to fit a model estimating different magnitudes of DIF effect sizes, unacceptable relative parameter bias of difficulty estimates was found. These results were not consistent with those under the equal class probability conditions. The relative parameter bias of difficulty estimates for the correct model estimating large LDIF and small ODIF effects ranged from −.174 to .086 for the reference class and from −.094 to .067 for the focal class. In addition, the relative parameter bias of difficulty estimates for the correct model estimating small LDIF and large ODIF effects ranged from −.178 to .076 for the reference class and from −.035 to .099 for the focal class. That is, the relative parameter bias was greater for the reference class than for the focal class.

Table 14. *Relative Bias of Estimated Item Difficulty Parameters by Correctly Estimating Models under Equal Class Probability Conditions*

		Small DIF Effect			Large DIF effect			Combination of DIF effect sizes	
	Item	Est. LDIF &ODIF	Est. LDIF	Est. ODIF	Est. LDIF &ODIF	Est. LDIF	Est. ODIF	Est. Large LDIF& Small ODIF	Est. Small LDIF& Large ODIF
Reference Class	i1	**0.060**	0.037	-	**0.061**	**0.057**	-	0.036	0.046
	i2	-0.038	0.035	-	-0.027	**0.075**	-	-0.019	-0.016
	i3	**-0.077**	**-0.086**	-	-0.031	**-0.060**	-	0.000	**-0.086**
	i4	-0.010	-0.002	-	-0.014	0.002	-	-0.006	-0.004
	i5	0.038	-	**0.059**	0.018	-	**0.051**	0.022	**0.050**
	i6	**-0.268**	-	**-0.297**	-0.035	-	**-0.246**	-0.027	**-0.092**
	i7	**-0.178**	-	**-0.541**	**-0.092**	-	**0.078**	-0.032	**0.062**
	i8	0.037	-	-0.037	-0.029	-	-0.012	-0.020	-0.043
Focal Class	i1	-0.012	0.004	-	0.014	-0.006	-	0.022	**-0.050**
	i2	-0.019	**-0.063**	-	-0.005	0.018	-	-0.019	0.009
	i3	0.025	0.016	-	-0.011	-0.002	-	0.009	0.038
	i4	0.006	0.003	-	0.001	-0.022	-	0.012	-0.008
	i5	**-0.057**	-	**0.119**	-0.034	-	-0.034	**-0.132**	-0.000
	i6	**0.091**	-	**0.106**	0.003	-	0.027	-0.010	0.028
	i7	0.010	-	**0.088**	0.010	-	0.023	0.023	-0.004
	i8	-0.049	-	-0.012	-0.018	-	-0.010	-0.012	-0.009

Note. The boldfaced font indicates that the parameter estimate was substantially biased (Hoogland & Boomsma, 1998); – = not estimated.

Table 15. *Relative Bias of Estimated Item Difficulty Parameters by Correctly Estimating Models under Unequal Class Probability Conditions*

		Small DIF Effect			Large DIF effect			Combination of DIF effect sizes	
	Item	Est. LDIF &ODIF	Est. LDIF	Est. ODIF	Est. LDIF &ODIF	Est. LDIF	Est. ODIF	Est. Large LDIF& Small ODIF	Est. Small LDIF & Large ODIF
Reference Class	i1	0.027	0.017	-	0.036	0.028	-	0.027	0.025
	i2	0.027	**0.061**	-	0.036	0.031	-	0.021	0.010
	i3	**0.092**	**0.061**	-	**0.089**	**0.102**	-	**0.086**	**0.076**
	i4	-0.012	-0.013	-	-0.007	-0.004	-	-0.014	-0.015
	i5	0.031	-	0.039	0.005	-	0.041	0.041	**0.068**
	i6	**-0.054**	-	0.017	-0.039	-	**-0.110**	-0.036	**-0.151**
	i7	**-0.113**	-	**-0.076**	**-0.161**	-	**-0.138**	**-0.174**	**-0.178**
	i8	0.019	-	0.011	0.010	-	0.003	0.003	-0.005
Focal Class	i1	0.002	0.011	-	**0.051**	**0.050**	-	**0.067**	0.022
	i2	0.029	**0.129**	-	0.025	0.039	-	0.035	**0.099**
	i3	-0.007	0.002	-	-0.020	0.004	-	0.011	-0.001
	i4	0.023	0.025	-	-0.007	0.004	-	0.030	0.002
	i5	**0.072**	-	**-0.087**	0.028	-	**-0.064**	**-0.057**	-0.035
	i6	-0.032	-	-0.022	-0.008	-	0.004	**-0.094**	0.004
	i7	-0.007	-	-0.022	0.009	-	**0.053**	0.040	0.014
	i8	-0.013	-	-0.024	-0.008	-	0.041	-0.007	0.005

Note. The boldfaced font indicates that the parameter estimate was substantially biased (Hoogland & Boomsma, 1998); − = not estimated.

Relative Standard Error Bias (*RSEB*)

The relative bias of the standard error estimates of item difficulty parameters for each latent class was computed for the correctly and incorrectly specified models. The values are presented in Tables 16 through 17. Standard error estimation bias was considered to be substantial when any absolute standard error exceeded .10 (Hoogland & Boomsma, 1998).

Equal class probability. As shown in Table 16, in conditions when the DIF effect size was small with equal class probability, most of the values of the relative standard error bias (*RSEB*) of difficulty parameter estimates were acceptable for each latent class under correctly specified model estimating both types of DIF effects. That is, the relative standard error bias ranged from −.159 to .113 under the correct model estimating both types of DIF effects. When data were generated to fit a model having LDIF effects, all relative standard error bias values for difficulty estimates for the reference class were acceptable, ranging from −.050 to .033, but the relative standard error bias in item difficulty parameter estimates for the focal class was larger than the criterion, ranging from −.051 to .177. In contrast, when data were generated to fit a model having ODIF effects, acceptable relative standard error bias was found for the focal class (ranging from −.046 to .084), but the relative standard error bias of difficulty estimates for the reference class was not acceptable, ranging from −.140 to .107. As the DIF effect size increased, the relative standard error bias found was greater than under the small DIF effect size conditions. For example, the relative standard error bias of difficulty estimates for the reference class under the correct model estimating both types of DIF effects was substantial, ranging from −.078 to .207. However, acceptable relative standard error bias of difficulty estimates was found for the focal class, except for item 3 for the focal class (*RSEB*=.340), under the correct model estimating both types of DIF effects. On the other hand, the relative standard error bias of difficulty estimates for both classes was observed for the correct model estimating LDIF effects, ranging from −.089 to .269. Likewise, the relative standard error bias of difficulty estimates for the reference class was unacceptable for the correct model estimating ODIF effects (ranging from .061 to .132). However, acceptable relative standard error bias was found for the focal class (ranging from −.029 to .053). The magnitude of the relative standard error bias was larger for the large DIF effect size conditions than for the small DIF effect size conditions. When data were generated to fit a model having a different combination of DIF effect sizes, the relative standard error bias of difficulty estimates was found. The relative standard error bias of difficulty estimates for the correct model estimating large LDIF and small ODIF effects ranged from −.057 to .281 for the reference class and ranged from −.140 to .173 for the focal class. Likewise, the relative standard error bias of difficulty estimates ranged from −.027 to .258 for the reference class and ranged from −.096 to .216 for the focal class.

Unequal class probability. Table 17 contains the relative standard error bias in unequal class probability conditions. A similar pattern was found for both equal and unequal class probability conditions, but the relative standard error bias found was greater under the equal class probability conditions. When the DIF effect size was small with unequal class probabilities, the relative standard error bias ranged from −.172 to .153 for the correct model estimating both types of DIF effects. On the other hand, acceptable relative standard error bias was found across both classes for the correct model estimating LDIF effects. The range of the relative standard error bias was from −.059 to .100. However, the relative standard error bias of difficulty estimate for item 2 for the focal class was relatively large (.237). When data were generated to fit a model having ODIF effects, values ranged from −.015 to .249. In particular, the relative standard error bias of difficulty of item 7 for both classes was observed (*RSEB* = .204 and .249 for the reference and focal classes, respectively).

As the DIF effect size increased, the relative standard error bias was greater than under the small DIF effect size conditions. Excessive relative standard error bias was found for both classes across the correctly specified models. That is, the relative standard error bias ranged from −.094 to .395 for the correct model estimating both types of DIF effects and from −.085 to .257 for the correct model estimating LDIF effects and from −.193 to .208 for the correct model estimating ODIF effects. Likewise, the relative standard error bias was substantial across both classes under the correct model estimating different magnitudes of DIF effect sizes. For example, the relative standard error bias ranged from −.115 to .303 for the correct model estimating large LDIF and small ODIF effects and from −.148 to .335 for the correct model estimating small LDIF and large ODIF effects. That is, the relative standard error bias was greater for the correct model estimating small LDIF and large ODIF effects than for the correct model estimating large LDIF and small ODIF effects. In addition, the relative standard error bias was greater for the focal class than for the reference class across correctly specified models when class probabilities were unequal.

Table 16. *Relative Standard Error Bias of Estimated Item Difficulty Parameters by Correctly Estimating Models under Equal Class Probability Conditions*

		Small DIF Effect			Large DIF effect			Combination of DIF effect sizes	
	Item	LDIF & ODIF	LDIF	ODIF	LDIF & ODIF	LDIF	ODIF	Large LDIF& Small ODIF	Small ODIF & Large LDIF
Reference Class	i1	-0.089	-0.001	-	**0.150**	0.027	-	0.088	**0.132**
	i2	-0.020	-0.014	-	**0.150**	**0.211**	-	0.037	0.033
	i3	-0.047	0.033	-	**0.146**	**0.145**	-	**0.180**	**0.228**
	i4	**0.113**	-0.050	-	-0.078	-0.089	-	-0.057	-0.027
	i5	0.039	-	**-0.140**	-0.028	-	**0.127**	-0.041	0.032
	i6	0.034	-	**0.107**	**0.207**	-	0.061	**0.114**	0.095
	i7	**-0.156**	-	0.053	**0.162**	-	**0.132**	0.016	**0.258**
	i8	-0.001	-	0.089	**0.101**	-	**0.123**	**0.281**	-0.023
Focal Class	i1	0.007	**0.177**	-	0.073	0.038	-	**-0.140**	**0.216**
	i2	-0.071	-0.045	-	0.048	0.000	-	0.016	**0.124**
	i3	**-0.159**	**0.149**	-	**0.340**	-0.077	-	0.062	-0.068
	i4	-0.026	-0.051	-	-0.033	**0.269**	-	-0.015	0.005
	i5	-0.063	-	-0.020	0.097	-	0.020	**0.105**	-0.096
	i6	0.031	-	-0.046	-0.031	-	-0.004	0.009	**0.213**
	i7	**-0.113**	-	0.084	0.035	-	0.053	**0.173**	**0.189**
	i8	0.079	-	0.072	-0.047	-	-0.029	**-0.118**	-0.027

Note. The boldfaced font indicates that the standard error estimate was substantially biased (Hoogland & Boomsma, 1998); − = not estimated.

Table 17. *Relative Standard Error Bias of Estimated Item Difficulty Parameters by Correctly Estimating Models under Unequal Class Probability Conditions*

		Small DIF Effect			Large DIF effect			Combination of DIF effect sizes	
	Item	LDIF & ODIF	LDIF	ODIF	LDIF & ODIF	LDIF	ODIF	Large LDIF& Small ODIF	Small ODIF & Large LDIF
Reference Class	i1	-0.035	0.019	-	**0.192**	0.054	-	0.060	**0.168**
	i2	**-0.144**	0.034	-	0.058	0.032	-	-0.065	-0.075
	i3	0.075	-0.010	-	-0.094	-0.080	-	-0.007	-0.044
	i4	**0.153**	**0.100**	-	**0.395**	**0.257**	-	**0.261**	**0.219**
	i5	0.092	-	0.059	-0.039	-	0.057	-0.100	-0.078
	i6	-0.058	-	-0.015	0.046	-	0.040	**-0.113**	**0.112**
	i7	0.087	-	**0.204**	-0.054	-	0.018	-0.091	-0.093
	i8	**-0.109**	-	0.019	-0.013	-	-0.050	-0.089	**-0.121**
Focal Class	i1	-0.040	-0.040	-	0.030	-0.085	-	**0.303**	**0.335**
	i2	0.043	**0.237**	-	**0.247**	**0.103**	-	-0.029	**0.111**
	i3	**0.121**	0.038	-	**0.122**	0.029	-	0.060	**0.124**
	i4	-0.024	-0.059	-	**0.231**	0.019	-	**0.250**	**-0.148**
	i5	**0.138**	-	0.012	**0.144**	-	**0.208**	**-0.115**	**0.117**
	i6	**-0.105**	-	-0.001	0.031	-	**0.111**	-0.027	0.017
	i7	**0.110**	-	**0.249**	0.030	-	**0.191**	0.065	-0.058
	i8	**-0.172**	-	0.067	-0.090	-	**-0.193**	-0.094	-0.020

Note. The boldfaced font indicates that the standard error estimate was substantially biased (Hoogland & Boomsma, 1998); − = not estimated.

Recovery of DIF Effect

The LDIF effect was determined by obtaining the difference between difficulty estimates across two latent classes of individuals for an item, and the ODIF effect was calculated by obtaining the difference between path parameters from observed groups to an items across latent classes. Thus, the relative parameter bias in DIF effect estimates was calculated by subtracting the true DIF effect value between the reference and focal classes from mean estimates of the DIF effect between the reference and focal classes. Substantial parameter estimation bias was identified for any estimate when the magnitude of the parameter bias was greater than .05 (Hoogland & Boomsma, 1998). The relative parameter bias in DIF effect estimates under the correctly specified models are presented in Tables 18 through 19.

Equal class probability. When the DIF effect size was small, the relative parameter bias in DIF effect estimates was substantial under the correct model estimating both types of DIF effects, ranging from −.170 to .163 (see Table 18). Even though the relative parameter bias in DIF effect estimates was not excessive for the correct model estimating LDIF effects, substantial positive relative parameter bias was found, ranging from .002 to .071. With the correct model estimating ODIF effects, the relative parameter bias in the DIF effect estimate ranged from −.143 to .171. However, the magnitude of the DIF effect size generated had a large impact on the bias in DIF effect estimates (see Table 18). The relative parameter bias in DIF effect estimates was acceptable across correctly specified models as the DIF effect size increased. That is, the relative parameter bias in DIF effect estimates ranged from −.049 to .050 across correctly specified models. When data were generated to fit a model estimating different magnitudes of DIF effect sizes, the relative parameter bias in small DIF effect estimates was observed. That is, the relative parameter bias in ODIF effect estimates (small DIF effect size) was found for the correct model estimating large LDIF and small ODIF effects (ranging from −.034 to .131). Likewise, the relative parameter bias in LDIF effect estimates (small DIF effect size) was found for the correct model estimating small LDIF and large ODIF effects (ranging from −.013 to .121).

Unequal class probability. As shown in Table 19, a similar pattern was observed to those found under equal class probability conditions. When the DIF effect size was small, the relative parameter bias in DIF effect estimates ranged from −.124 to .072 for the correct model estimating both types of DIF effects. When data were generated to fit a model estimating LDIF effects, the positive relative parameter bias was found, ranging from −.030 to .084. In contrast, for the correct model estimating ODIF effects, the negative relative parameter bias in the DIF effect estimates occurred, ranging from −.140 to .000. However, under conditions with a large DIF effect size with unequal class probability, as shown in Table 19, acceptable relative parameter bias was observed across correctly specified models (ranging from −.044 to .042).

When data were generated to fit a model estimating large LDIF and small ODIF effects, the relative parameter bias in DIF effect estimates was observed. That is, substantial relative parameter bias in the small DIF effect estimate for item 8 was observed (*RPB*=−.166) and that in the large DIF effect estimate for item 4 was .051. However, acceptable relative parameter bias was observed for the correct model estimating small LDIF and large ODIF effects (ranging from −.046 to .040).

Table 18. *Relative Parameter Bias of DIF Effects by Generating and Estimating Models under the Conditions of Equal Class Probability*

	Small DIF effect			Large DIF effect			Combination of DIF effect sizes	
Item	LDIF&ODIF	LDIF	ODIF	LDIF & ODIF	LDIF	ODIF	Large LDIF& Small ODIF	Small LDIF & Large ODIF
i1	**0.117**	**0.064**	-	0.042	0.032	-	0.030	**0.121**
i2	-0.032	0.002	-	-0.010	0.030	-	-0.019	-0.008
i3	**0.079**	**0.071**	-	-0.007	0.008	-	0.010	**0.104**
i4	0.030	0.011	-	0.008	-0.033	-	0.021	-0.013
i5	0.031	-	-0.020	0.025	-	0.010	**0.054**	-0.005
i6	**-0.162**	-	**-0.143**	-0.014	-	-0.049	0.042	-0.012
i7	**0.163**	-	**0.171**	**0.050**	-	0.015	**0.131**	0.037
i8	**-0.170**	-	0.023	-0.037	-	0.033	-0.034	0.031

Note. The boldfaced font indicates that the parameter estimate was substantially biased (Hoogland & Boomsma, 1998); − = not estimated.

Table 19. *Relative Bias of DIF Effects by Generating and Estimating Models under the Condition of Unequal Class Probability*

	Small DIF effect			Large DIF effect			Combination of DIF effect sizes	
Item	LDIF &ODIF	LDIF	ODIF	LDIF & ODIF	LDIF	ODIF	Large LDIF& Small ODIF	Small LDIF & Large ODIF
i1	0.046	0.022	-	0.042	0.037	-	0.043	0.028
i2	0.028	**0.084**	-	0.028	0.037	-	0.032	0.040
i3	**-0.060**	-0.030	-	-0.040	-0.014	-	-0.002	-0.043
i4	**0.072**	**0.078**	-	-0.007	0.008	-	**0.051**	0.025
i5	-0.022	-	0.000	-0.003	-	0.035	0.039	0.008
i6	0.021	-	-0.045	0.001	-	-0.017	0.039	-0.046
i7	**-0.124**	-	**-0.057**	-0.017	-	-0.044	0.019	0.001
i8	-0.084	-	**-0.140**	-0.007	-	-0.027	**-0.166**	-0.002

Note. The boldfaced font indicates that the parameter estimate was substantially biased (Hoogland & Boomsma, 1998); − = not estimated.

Chapter 5: Discussion

This chapter contains three sections. The first section summarizes results and discusses findings. The second section addresses limitations of the present study with suggestions for future research. The last section presents conclusions.

Summary of the Results and Discussions

The present study was designed to assess the performance of fit indices, entropy values, and the significance of LDIF and ODIF effects by comparing correctly specified models with incorrectly specified models. Additionally, this study was intended to investigate parameter and standard error bias in difficulty parameter estimates and parameter bias in DIF effect estimates for LDIF and ODIF effects. Discussion of the results will be followed by a summary of the performance of the AIC, BIC, aBIC, and CAIC indices, DIF detection rates (power and Type I error), entropy values, bias in item difficulty parameters, standard error bias in item difficulty parameters, and bias in DIF effect size.

Fit indices. The performance of fit indices was evaluated to identify the simulation conditions and model specifications under which the fit indices performed well. In addition, an evaluation was conducted to assess which fit index more often supported the better fit of the correct model. When the fit indices were compared across the simulation conditions, their performance varied under conditions with the smaller DIF effect size. However, in the conditions with the larger DIF effect size, all of the fit indices performed well regardless of class probability. All of the fit indices performed relatively poorly under the unequal class probability conditions, except when the DIF effect size was large. When data were generated to fit a model having a combination of different magnitudes of DIF effect sizes, the fit indices performed better for the model with small LDIF and large ODIF effects than for the model with large LDIF and small ODIF effects, regardless of class probability. That is, LDIF effects had more influence on the performance of fit indices than did ODIF effects, and mis-specifying small ODIF effects by excluding them from the model did not make any differences in fit index values between the correctly and incorrectly specified models. However, mis-specifying even small LDIF effects by excluding them from the model had a greater impact on the performance of fit indices, so that the fit indices frequently supported better fit of the correct model in comparison to incorrectly specified models (that is, for under-specified models). This is because the LDIF effects that were generated in this study influenced a larger proportion of individuals as compared to the sample size of those affected by ODIF effects.

The AIC index was found to perform better than the other fit indices, followed by the aBIC index for the model with both types of DIF effects as well as for the model with ODIF effects under small DIF effect size and equal class probability conditions. In addition, in the unequal class probability and small DIF effect size conditions, the AIC index supported the correct model a relatively high proportion of the time (82% and 88%) under the model with both types of DIF effects and the model with LDIF effects, respectively. In addition, the AIC index supported the correct model with a slightly lower rate (78%) under the model with ODIF effects. However, in the same set of conditions, other fit indices never resulted in correct selection of the model with both types of DIF effects nor the model estimating ODIF. As the DIF effect size increased, the BIC, aBIC, and CAIC indices performed as well as if not better than the AIC index. For example, under the model with ODIF effects, the AIC index supported the correct model for 96% of the replications, while other fit indices led to selection of the correct model for 100% of the replications. In addition, under the model with large LDIF and small ODIF effects, the AIC index performed better than other fit indices, regardless of class probability. However, under the model with small LDIF and large ODIF effects, the performance of fit indices differed between the conditions of equal and unequal class probability. More specifically, under the equal class probability conditions, all fit indices performed perfectly in selecting the correct model, except for the CAIC index. However, only the AIC and aBIC indices performed perfectly in selecting the correct model under the unequal class probability conditions.

In sum, it was found in the present study that the AIC index generally performed better than (or as well as) other fit indices, followed by the aBIC index, under smaller DIF effect size conditions. These results are not consistent with previous studies (for example, Nylund et al., 2007; Li et al., 2009), in which the AIC index performed poorly compared to other fit indices. In addition, there is little consensus in previous mixture modeling research about the performance of the BIC in terms of correct mixture model identification. Some previous research has found that the BIC index, in general, performed better than other fit indices (Li et al, 2009; Nylund et al., 2007) while other studies have found that the BIC performed poorly (Jackman, 2012; Tofighi & Enders, 2008). The performance of the BIC index under the small DIF effect condition in the present study matched the results found in the latter set of studies in which the BIC was found to perform poorly relative to other information criteria.

There are two possible reasons why the performance of fit indices in the present study differed from what was found in previous studies. Previous, related studies examined the performance of fit indices under mixture models with only LDIF effects. The current study also examined conditions with ODIF effects. When only LDIF effects were included in the present study, results matched those found in previous studies in that the AIC index performed more poorly than did the other fit indices.

The second reason is that previous studies investigated which fit index performed well in supporting better fit of a model in terms of the optimal number of latent classes (Allua, 2007; Leite & Cooper, 2010; Lubke & Neale, 2005; McLachlan & Peel, 2000, Nylund et al., 2007; Tofighi & Enders, 2008). However, the present study investigated which fit index performed well in supporting better fit of a correctly versus incorrectly specified models with all models assuming the same and correct number of latent classes. One previous, related study did find that the AIC index performed at least as well as if not better than the BIC and CAIC indices (followed by the aBIC index) when used to compare model fit for mixture models differently parameterized although using the same number of latent classes (Lee & Beretvas, 2011). In that study, the performance of fit indices was compared for the cases when covariate effects were correctly versus incorrectly specified.

Inspection of the equations used to calculate fit indices (see Equations 13 through 16) reveals that under the small DIF effect conditions, log likelihood values were not much different between correctly and incorrectly specified models, and so the distinctions in fit index values is largely a function of the latter terms in the equations, which entailed a function of total sample size and the number of free parameters. Therefore, the BIC and CAIC indices which are more influenced by the latter terms than the AIC and aBIC indices resulted in supporting fit of a model with fewer free parameters under the small DIF effect size conditions. However, as the true DIF effect size value increased, the difference in log likelihood values between correctly and incorrectly specified models also increased, and so the latter term in the equations had little impact on the performance of all fit indices evaluated, with a few exceptions. As a result, all fit indices evaluated for the present study performed relatively well in supporting the correct model against the incorrectly specified models.

Power. The present study also evaluated the power for DIF detection. The power was substantially different between LDIF and ODIF effects. Whereas the power for LDIF effects ranged from 66.50% to 72.50%, the power for ODIF detection ranged from 27.50% to 31% under small DIF effect size and equal class probability conditions. That is, using a minimum cutoff of 80% as representing acceptable power, the power for both LDIF and ODIF detection were not acceptable under the equal class probability and small DIF effect size conditions examined here. In particular, the detection rate for ODIF was considerably lower, because ODIF effects were under-estimated, compared to the estimation of LDIF effects. In addition, the sample size for each type of DIF was different. ODIF effects were exhibited based on observed group membership (for example, male vs. female) within latent classes, so the number of individuals with ODIF effects was much smaller than those with LDIF effects. This would clearly lead to differing power levels for LDIF versus ODIF.

When class probabilities were unequal, the power for LDIF detection slightly decreased to about 60%. The power for ODIF detection also decreased to about 20%. That is, power for DIF detection was slightly influenced by class probability, because the number of individuals in focal classes was generated to be smaller (in this study) and standard errors of item difficulty estimates were larger under the unequal class probability conditions than under the equal class probability conditions. However, consistent with the results of previous studies (Jackman, 2012; Lu & Jiao, 2009; Maij-de-Meij et al., 2012; Samuelsen, 2005), in general, the power for both LDIF and ODIF detection were acceptable (90.50% to 100%) when the DIF effect size was large, regardless of class probability.

When data were generated to fit a model having large LDIF and small ODIF effects under the equal class probability conditions, the power for LDIF detection was 100% regardless of class probabilities, and the power for ODIF detection was consistently lower. The lower power for ODIF detection is not unexpected. As noted, the size of the sample for which ODIF was generated was smaller than that for the sample for which LDIF was generated. And obviously, smaller sample sizes will lead to less power. For the same datasets (for which large LDIF and small ODIF effects were generated) the power to detect ODIF was 47% for the correct model (modeling both LDIF and ODIF) and 31.50% for the under-specified model with only ODIF effects. This difference likely results from the omission of true, large LDIF in the latter under-specified model (estimating ODIF when true LDIF and ODIF exist) resulting in more error and thus, likely larger variances for item difficulty parameter estimates. Under the unequal class probability conditions, the rates for ODIF detection marginally decreased, but the rates for LDIF detection were still high (100%) for the model with large LDIF and small ODIF effects.

On the other hand, for data generated to fit a model with small LDIF and large ODIF under the equal class probability conditions, the power for LDIF detection as well as for ODIF detection were acceptable (86.50% and 96.50%) under the correctly specified model. However, the power for LDIF detection decreased to 75.70% under unequal class probability conditions. The power for ODIF detection also decreased to 90.50%, although it was still acceptable. This pattern of differences is expected again due to the sample sizes involved in the unbalanced latent class sample size (unequal class probabilities) conditions. Thus, the results of the present study are consistent with results of previous studies, in which DIF was more accurately identified when the DIF effect size was large and when there were more DIF items. In addition, the present study indicates that LDIF effects were more accurately identified when balanced sample sizes across latent classes were used.

Type I error rates. The present study evaluated the Type I error rates of over-specified ODIF effects and the Type I error rates of mis-specified ODIF effects when data were generated to fit a model having LDIF effects. The Type I error rates were also evaluated for over-specified LDIF effects (that is, both types of DIF effects were estimated for ODIF-generated data) and for (differently specified) LDIF effects when data were generated to fit a model having ODIF. Generally, consistent with Jackman's (2012) findings of inflated Type I error rates across all simulation conditions, the present study also found inflated Type I error rates across simulation conditions. Jackman found that the Type I error rates for (truly invariant) item difficulty were 10% under large DIF effect size conditions (1.5) with large sample size (N = 1,000). Likewise, in the present study the over-specified ODIF or LDIF effects yielded inflated error rates ranging from 5.50% to 10.50% across simulation conditions. Similar to the findings of Maij-de-Meij et al. (2011) that Type I error rates increased when class probabilities became unequal, the present study found that the Type I error rates of over-specified ODIF effects were larger under the unequal class probability conditions than under the equal class probability conditions. On the other hand, the DIF effect size influenced the Type I error rates of over-specified LDIF effects. That is, the Type I error rates of over-specified LDIF effects increased as the DIF effect size increased under the equal class probability conditions, but this was not the case under the unequal class probability conditions. For example, approximately 8% of invariant items were detected as displaying LDIF under the condition of equal class probability with the small DIF effect size as well as under the conditions of unequal class probabilities. Under the condition of equal class probability with the large DIF effect size, 11% of invariant items were detected as displaying LDIF. That is, the magnitude of bias in item difficulties across latent classes increased as the DIF effect size increased for the over-specified model estimating both types of DIF effects when data were generated to fit a model estimating ODIF under the equal class probability condition.

In addition to addressing over-specified DIF effects, the present study evaluated the Type I error rates of the mis-specified DIF effects. When data were generated to fit a model with LDIF but ODIF was estimated, then the Type I error rates for the mis-specified ODIF effects were 7% and 11% for the small and large DIF effect size conditions, respectively. In contrast, when class probabilities were unequal, the Type I error rates of the mis-specified ODIF effects were 8.50% and 6% for the small and large DIF effect size conditions, respectively. Thus, the pattern of the effects for the true LDIF effect size on ODIF Type I error rates is reversed under unequal versus equal sample size conditions. Given the current study did not examine the parameter and standard error bias for incorrectly specified models, it is unclear exactly why this reversal occurred. However, future research should examine this interaction effect more closely to help understand its source.

Additionally, detection rates for differently specified LDIF effects were examined when data were generated to fit a model having ODIF but modeled LDIF. The detection rate of differently specified LDIF effects was 22.50% under small DIF effect size conditions. However, as the true DIF effect size increased, 99.75% of ODIF effects were captured by specifying LDIF effects. When class probabilities were unequal, detection rates of differently specified LDIF effects were similar to those found for equal class probabilities with values of 23% and 98% for the small and large DIF effect size conditions, respectively. Because no previous study has investigated how well an LDIF model captures ODIF effects, supporting or opposing evidence for the present findings was lacking. Nevertheless, results of the present study suggest that, not surprisingly, the magnitude of the true DIF effect size had a substantial impact on the detection of DIF. In particular, when the true DIF effect is sufficiently large, ODIF effects can be fully captured by specifying LDIF effects.

Entropy. Lubke and Muthén (2007) found that entropy values were low (about 0.43) when a single factor was estimated with a mean difference of 1.5 across latent classes. The present study also found that average entropy values, in general, were low under the correctly specified models, ranging from .280 to .405 in small DIF effect size conditions, from .326 to .643 under large DIF effect size conditions, and from .384 to .533 under conditions with different magnitudes of DIF effect sizes. Consistent with the finding that average entropy values increase when class separation increases (Lubke & Muthén, 2007), the present study also found that the average entropy values were larger under large DIF effect size conditions than under small DIF effect size conditions. In addition, under conditions with a combination of different magnitudes of DIF effect sizes, the average entropy values were larger than those under the model with both small DIF effect sizes, but the rates were smaller than those under the model with both large DIF effect sizes, regardless of simulation conditions and model specifications. The average entropy values for the model with large LDIF and small ODIF effects were larger than those for the model with small LDIF and large ODIF effects. As mentioned earlier, fit indices more frequently supported better fit of the correctly versus incorrectly specified models under the model with small LDIF and large ODIF effects than under the model with large LDIF and small ODIF effects. In other words, the results of entropy values were not consistent with the performance of fit indices when comparing correct model identification for models with large LDIF and small ODIF versus the model with small LDIF and large ODIF. Future research should explore the relationships between entropy and the performance of fit indices.

The average entropy values under the correct model estimating both types of DIF effects were larger than those under the correct model estimating either LDIF or ODIF effects. In addition, the average entropy values under the correct model estimating LDIF effects were larger than those under the model correctly estimating ODIF effects across simulation conditions. This is likely because sample sizes differed for models in which LDIF versus ODIF effects were generated. For example, when a model was correctly specified under the equal class probability condition, a large number of individuals (1,000) differed by LDIF effects while a relatively smaller number of individuals (500) differed by ODIF effects within each latent class.

When correctly specified and incorrectly specified models were compared, the average entropy values were larger under the correct model estimating both types of DIF effects than those under the incorrectly specified models (under-specified LDIF or ODIF models). The average entropy was related to class separation, and entropy increased as class separation increased matching previous research findings about entropy (for example, Lubke & Muthén, 2007). More DIF items increases class separation, so entropy values were larger for the correct model with both types of DIF effects compared to the incorrectly specified models. However, the average entropy values under the correct model estimating LDIF effects or under the correct model estimating ODIF effects were very similar to those under the incorrectly specified models.

Bias and standard error bias. The present study obtained the relative parameter bias and standard error bias for the sets of four or eight items for each latent class under each condition for correctly specified models. Under conditions when the DIF effect size was small with equal class probability, the relative bias in difficulty estimates that had LDIF effects was substantial under the correct model estimating both types of DIF effects and the correct model estimating LDIF effects. Consistent with the findings of DeMars and Lau (2011) who found that the grand mean bias averaged across difficulty estimates was acceptable, the present study also found that the grand mean bias averaged across item difficulty estimates was acceptable for the correct model estimating both types of DIF and for the correct model estimating LDIF effects. However, values of relative bias in difficulty estimates that had ODIF effects were excessive. As the DIF effect size value increased, the relative bias in difficulty estimates decreased across the correctly specified models. Consistently, when data were generated to fit a model estimating a combination of different magnitudes of effect sizes under the equal class probability conditions, the majority of the relative bias in difficulty estimates was acceptable for the correct model with large LDIF and small ODIF, but the relative bias was substantial for the correct model with small LDIF and large ODIF effects under the equal class probability conditions. On the other hand, under the unequal class probability conditions, the relative bias in item difficulties for the focal class was larger for the model with large LDIF and small ODIF than for the model with small LDIF and large ODIF. To help understand the reason for this finding, future research could capture correct latent class membership rates to assess whether the source of this greater bias resulted from higher misclassification rates when data were generated to fit a model with different combinations of DIF effect sizes.

The standard error bias in difficulty estimates was substantial across correctly specified models and simulation conditions. The degree of standard error bias in difficulty estimates was larger for the focal class under the unequal class probability conditions than under the equal class probability conditions, because a relatively small number of individuals belonged to the focal class under the unequal class probability conditions. Most previous studies (for example, DeMars & Lau, 2011; Lu & Jiao, 2009; Samuelsen, 2005) investigating LDIF effects have examined the recovery of DIF effects through the assessment of differences in items' difficulties and have not focused on the recovery of item difficulties for each latent class. The recovery of item difficulty values for each latent class is important to explore why the recovery of DIF effects is poor. Thus, future research should explore not only the recovery of DIF effect sizes but also the recovery of item difficulties for each latent class.

Bias in DIF effect estimation. When the true DIF effect was small, substantial relative bias in estimated DIF effects was observed across correctly specified models. Even though positive and negative bias in ODIF effect estimates was observed, only positive relative bias was found in LDIF effect estimates. For larger true DIF effect size conditions, the relative bias in DIF effect estimates was acceptable across correctly specified models. Likewise, when data were generated to fit a model estimating a combination of different magnitudes of DIF effects, unacceptable bias was found under small DIF effect size conditions, while acceptable bias was found under large DIF effect size conditions. In addition, as expected, when DIF effect sizes were small, the magnitude of relative bias in ODIF effects was larger than the relative bias in LDIF effects, regardless of class probabilities. The results of the recovery of DIF effects were consistent with the results of the recovery of item difficulty estimates, entropy, and power for DIF detection. As mentioned earlier, no methodological studies have investigated both LDIF and ODIF effects under FMMs with binary outcomes, so future research is necessary to examine how large a sample size might be necessary to achieve reasonable recovery of ODIF effects.

Implications and Recommendations

If the sources of population heterogeneity are unobserved or unmeasured (i.e., latent), then conventional DIF analysis procedures cannot be used to identify the latent sources of DIF. In such cases, mixture modeling, which introduces latent categorical variables (that is, latent classes), as sources of heterogeneity can be used. Numerous applied and simulation studies have investigated the performance of mixture models with interval-scaled outcomes in terms of fit indices, latent class assignment, and the recovery of parameter estimates.

However, mixture models with binary outcomes (that is, mixture IRT models) for detecting DIF have not been thoroughly explored. Several studies have found that typical DIF detection methods—which identify differences among *manifest* groups formed by such characteristics as age, gender and ethnicity—have not performed well in fully explaining potential DIF (Cohen & Bolt, 2005; De Ayala et al., 2002). That is, membership in a *manifest* group defined by characteristics such as gender and ethnicity does not always explain all the heterogeneity identified in item scores even though such characteristics may be somewhat related to actual cause(s) of DIF.

Thus, to replace traditional DIF detection methods, factor mixture models have been suggested as a means to identify groups on the basis of unobserved characteristics (latent class), such as personality traits, unmeasured socioeconomic status, or educational background. Under such models it is assumed that latent class membership accounts for DIF. However, studies based on such models have overlooked that sources of DIF might be more complex. That is, one or some actual causes of DIF might be observed, but one or some other sources of DIF might be unobserved (latent class membership). As examined here, an observed source of DIF may exist distinguishing observed groups in one of several latent classes but not in other classes.

The present study demonstrated how both LDIF and ODIF effects were recovered under various conditions of class probability and DIF effect sizes by comparing correctly specified models with incorrectly specified models. To address the absence of methodological research investigating both types of DIF effects under correctly specified models as well as under incorrectly specified models, the present study investigated the implications of alternatives to typical manifest DIF methods and between-latent class DIF methods.

It was found that the performance of fit indices varied as a function of DIF effect size and class probability. The AIC index was the best indicator when comparing models with the same number of latent classes. As DIF effect size increased, differences in the performance of fit indices were negligible. Thus, applied researchers should consider the

AIC and aBIC indices first if the DIF effect size is small, when they are comparing models within the same number of latent classes. As Lubke and Muthén (2007) indicated, results of the present study indicated that entropy increased as the magnitude of covariate effects (DIF effect) increased. Likewise, the performance of fit indices, the recovery of item difficulty estimates, and the recovery of DIF effect estimates improved for larger true DIF effect size conditions. When comparison was performed of correct model identification between models with large LDIF and small ODIF effects and models with small LDIF and large ODIF effects, fit indices performed better for the models with small LDIF and large ODIF effects, while other measures (that is, entropy and the recovery of item difficulty estimates) were better for the models with large LDIF and small ODIF effects. Thus, when applied researchers assume that there exists both types of DIF effects with different magnitudes of DIF effect sizes in data, they should consider not only the performance of fit indices, but also entropy values and difficulty estimates across and within-latent classes.

While Type I errors were prevalent across all model specifications and simulation conditions evaluated, the magnitude of Type I error rates varied across model specifications and simulation conditions. This means that under the condition of FMM with binary outcomes, it frequently happens that an item will be incorrectly identified as displaying DIF. Thus, applied researchers should be careful to note when examining items for DIF effect that approximately 6% to 11% of invariant items might be incorrectly flagged as DIF. In addition, given how well estimation of LDIF models recovered true ODIF, applied researchers might start their exploration of potential item bias by first estimating the more general LDIF model. If evidence is found supporting potential LDIF, then applied researchers should consider estimating models that test potential observed sources of the DIF that was found.

Limitations and Suggestions for Future Research

The present study evaluated both LDIF and ODIF effects together, which has not been investigated in previous methodological research. However, the present study considered only two DIF effect sizes (small/large) and two class probabilities (equal/unequal). In addition, the present study only examined uniform LDIF and ODIF effects and did not investigate recovery of non-uniform DIF. Moreover, the probabilities of dichotomous observed group membership (for example, female and male) were the same, which may not necessarily reflect realistic applied conditions. Future research should incorporate varying levels of DIF effect size, latent class probability, and observed group membership probability. In addition, future research should extend the model and test for nonuniform LDIF and ODIF by allowing the item discrimination (factor loading) parameters to vary across both or either of latent classes and observed grouping variables.

At the time of the present study, insufficient research was found on the performance of fit indices in terms of their use in selecting the correctly specified models amongst a set of DIF assuming the same number of (two) latent classes. Therefore, a future study might compare various information criteria under various simulation conditions by manipulating combinations of LDIF and ODIF effects. In addition, the present study only assessed information criteria and did not examine the performance of likelihood-based tests, such as aLRT or BLRT. Previous studies have found that aLRT or

BLRT performed better compared to information criteria in terms of correctly identifying a correctly specified model (for example, Li and Hser, 2010; Nylund et al., 2007). Thus, a future study might investigate how well likelihood-based tests perform in correctly identifying the correct model under various simulation conditions. A future study might investigate how incorrectly specified covariate effects (DIF) impact the selection of numbers of latent classes based on the performance of aLRT and BLRT.

The present study considered only the case in which the DIF effect size was the same for every item estimated. That is, the DIF effect size was fixed at .5 for the small effect and at 1.5 for the large effect across all items estimated. Maij-de Meij et al. (2012) considered various DIF effect sizes by manipulating various differences between focal and reference item difficulty parameters. Thus, a future study might manipulate various DIF effect sizes across item difficulty parameters.

Conclusions

The present study investigated class-specific observed DIF (ODIF) as well as between-latent class DIF (LDIF) under various model specifications and simulation conditions. In addition, the present study compared the performance of correctly specified models with incorrectly specified models in terms of fit indices, entropy values, and the power for identifying DIF. Furthermore, the recovery of item difficulty parameters and of DIF effect size was investigated under the correctly specified models.

Generally, findings of the present study are consistent with those of previous studies that a large number of DIF items, large DIF effect size, and equally distributed proportions of each latent class improve detection rates of DIF effects, recovery of DIF effects, and class assignment rates. Further, the present study indicates that the AIC index is the most accurate among the set of indices investigated under the small DIF effect condition. However, as DIF effect size increased, the AIC, BIC, aBIC, and CAIC indices performed well in selecting the correct model. In addition, previous studies have found that FMM with binary outcomes performed more poorly in recovering item difficulty parameters under conditions with unequal class probabilities, because the item difficulties for the focal class were poorly recovered due to potential misclassification of individuals who belonged to the focal class and to the presence of a smaller number of individuals in the focal class. However, as DIF effect size increased, item difficulties and DIF effects were recovered well, resulting in high detection rates of DIF effects. However, a relatively small number of simulation conditions were assessed here for LDIF and ODIF effects, and limitations exist in the present study. Thus, more research should be performed to assess recovery of LDIF and ODIF effects under more extensive simulation conditions.

Practical Importance

When applied researchers estimate DIF using a conventional DIF detection method such as the M-H test, they have to address several issues. Typically, gender and ethnicity variables have been widely used as sources of potential DIF; however, the homogeneity within female and within male examinees is questionable (see Cohen & Bolt, 2005). And, for example, merging Filipinos, Koreans, Indonesians, Taiwanese, and Asian Indians into an Asian American group is problematic (DeAyala et al., 2002), because these peoples are culturally distinct. In addition, Skaggs and Lissitz (1992, pg.

239) commented that "*Black* is not a cognitively meaningful dimension and not even a well-defined one for that matter."

Thus, even though an item does not exhibit DIF with respect to gender or ethnicity, that does not mean that there is no DIF within the item set. Rather, cultural and curricular differences across countries (Sadeghi, 2009) or differences among individuals' response styles (Bolt & Johnson, 2009) might have a significant impact on the equivalence of test items. DeAyala et al. commented, "The selection of manifest grouping variables is based on political not psychometric considerations." (2002, p. 274)

Beyond the difficulties described with using observed grouping variables as potential sources of DIF, use of a latent class approach has practical advantages even though it might not present an easy approach for DIF detection. It might be the case that the source of DIF is not observable. Detecting individual differences in human behavior or identifying potentially meaningful dimensions (unobserved sources) instead of using convenient external, directly observable characteristics can support discovery of latent constructs not originally hypothesized to underlie test and item scores. So, even though DIF based on observed characteristics like gender or ethnicity may not be found using a conventional DIF detection test (for example, M-H test), the item might be flagged for DIF under a mixture modeling approach.

For example, similar to findings in the present study when simulating ODIF effects, findings from applied research may indicate that DIF occurred between males and females within one latent class but not within other latent classes in the data. In such cases, identifying latent classes can be an important source of insight for understanding the nature of observed characteristics (here, gender) (Tay et al., 201).

Eid and Diener (2001) found that the majority of individuals across countries belonged to one latent class, but only a smaller number of individuals from one specific country belonged to another latent class. In such a case, the focus is not only on whether DIF occurs but also on for whom does DIF occur. This latter question leads to another practical concern that is commonly associated with detection of DIF. Detection of DIF does not necessarily imply that an item score is biased. Inferences about item or test score bias are value judgments that are made in a grander context of construct validity. Test developers must call upon relevant experts in the construct being measured (e.g., mathematics or reading, etc.) who can make sense of whether the source of the DIF is a necessary additional dimension that is an inevitable – crucial even - part of the fuller construct of interest or whether the dimension interferes with the validity of what is being measured. For example, use of a FMM with binary outcomes might identify that there are two latent classes of respondents to a set of mathematics achievement items. Upon further analysis of the members of the two classes, the test developer might realize that reading ability distinguishes the two classes. For example, students with low reading ability constituted the first class while students with better reading ability might be most likely to be members of the other class. It would be up to the test developers and the experts on math achievement to decide whether reading ability is an essential component of mathematics achievement or whether the intention of the test score was to distinguish examinees solely on their pure mathematics achievement.

Use of FMM with binary outcome variables provides a more flexible model to identify potential subpopulations of respondents and ultimately measurement non-invariance. Use of conventional observed DIF detection procedures can be used given the researcher (or test developer) has a priori hypotheses about sources of DIF. Assessment of LDIF might lead to identification of more items with DIF (than if using ODIF analyses) and of items for which the source of the DIF cannot be explained using observed variables. However, even when DIF is identified for observed groups, it is not always the case that a reasonable explanation for the DIF can be found (e.g., expected cultural differences). Regardless, it behooves test developers and researchers to try and identify whether there might be DIF and hope that a reasonable source of the DIF can be found that will then allow the user to identify whether the DIF reflects some kind of bias or not. And results of this dissertation seem to indicate that estimation of LDIF and of ODIF within latent classes provides a more sensitive method for finding DIF. Guidelines for the interpretation of DIF as bias or not is beyond the scope of this dissertation although it is a crucial issue that must be considered to make sense of DIF results.

References

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

Allua, S. (2007) *Evaluation of single- and multilevel factor mixture model estimation*. Unpublished doctoral dissertation, University of Texas, Austin.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–332.

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

Asparouhov, T. & Muthén, B. (2008). Multilevel mixture models. In Hancock, G. R., & Samuelsen, K. M. (Eds.), Advances in latent variable mixture models, pp. 27-51. Charlotte, NC: Information Age Publishing, Inc.

Bandalos, D. L., & Cohen, A. S. (2006, April). *Using factor mixture models to identify differentially functioning test items*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bentler, P.M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107 (2), 238-46.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores (chapters 17-20). Reading, MA: Addison-Wesley.

Bozdogman, H. (1987). Model selection and Akaike's information criteria (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

Camilli, G. (1993). The case against item bias techniques based on internal criteria: Do item bias procedures obscure test fairness issues? The use of differential item functioning statistics:

A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Cho, H., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a Mixture IRT model. *Applied Measurement in Education*, 25, 281-304.

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

Clark (2011). *Mixture modeling with behavioral data.* Unpublished doctoral dissertation, University of California, Los Angeles.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42, 133-148.

Dai, Y. (2009) *A mixture Rasch model with a covariate: a simulation study via Bayesian MCMC estimation*. Unpublished doctoral dissertation, University of Maryland, College Park.

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2, 243-276.

Dodeen, H., & Johanson, G. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment and Evaluation in Higher Education*, 28, 129-134.

Embretson, S. E. (2006). Mixed Rasch models for measurement in cognitive psychology. In von Davier, M. & Carstensen C. H. Multivariate and Mixture Distribution Rasch Models: Extensions and Applications. New York, NY: Springer Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman & Hall.

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. Journal of Personality and Social Psychology, 81, 869-885.

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

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences*, 57B(5), S275-S283.

Gagné, P. (2004). *Generalized confirmatory factor mixture models: A toll for assessing factorial invariance across unspecified populations*. Unpublished doctoral dissertation. University of Maryland.

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

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

Hagtvet, K. A., & Sipos, K. (2004). Measuring anxiety by ordered categorical items in data with subgroup structure: The case of the Hungarian version of the Trait Anxiety Scale of the

State-Trait Anxiety Inventory for Children (STAIC-H). *Anxiety, Stress, and Coping*, 17, 49–67.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, 30, 91-105.

Holland, P. W., & Thayer, D. T. (1988). *Differential item functioning and the Mantel-Haenszel procedure*. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Erlbaum.

Holland, P.W, & Wainer H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329-367.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117-144.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits:Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, 28,192-218.

Jackman, M.G. (2011). *A Monte Carlo investigation of the performance of factor mixture modeling in the detection of differential item functioning*. Unpublished doctoral dissertation. University of Florida.

Jedidi, K., Ramaswarmy, V., DeSarbo, W. S., & Wedel, M. (1996). On estimating finite mixtures of multivariate regression and simultaneous equation models. *Structural Equation Modeling, 3,* 266–289.

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

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, 27 (4), 307–327.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education*, 21, 207-226.

Kim, S., Beretvas, S, & Sherry, A (2010). A validation of the factor structure of OQ-45 scores using factor mixture modeling. *Measurement and Evaluation in Counseling and Development*, 42(4), 275-295.

Lazarsfeld, P., & Henry, N. (1968). *Latent structure analysis*. New York: Houghton-Mifflin.

Leite, Walter L. & Cooper, Lou Ann. (2010). Detecting social desirability bias using factor mixture models. *Multivariate behavioral research*, 45: 2, 271-293.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33, 353-373.

Li, L & Hser, Y. (2011). On inclusion of covariates for class enumeration of growth mixture models. *Multivariate behavioral research*, 46, 266-302.

Lo, Y., Mendell, N., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767-778.

Lord, F.M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.

Lu, R., & Jiao, H. (2009, April). *Detecting DIF using the mixture Rasch model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lubke, G.H. (2010). *Latent variable mixture modeling*. In GR Hancock & RO Mueller (Eds.), The Reviewer's Guide to Quantitative Methods in the Social Sciences. New York, NY: Routledge.

Lubke, G. & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, 14(1), 26–47.

Lubke, G.H. & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10, 21-39.

Lubke, G.H., & Neale, M.C. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, 43:592-620.

Lubke, G.H. & Spies, J. (2008). *Choosing a correct factor mixture model: Power, limitations, and graphical data exploration*. In G. R. Hancock & K. M. Samuelsen (Eds.), Advances in latent variable mixture models. Charlotte, NC: Information Age Publishing, 343-362.

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

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, 32, 611-631.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, 45, 975-999.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525-543.

Muthén, B. O. & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Mplus Web Note No. 4, at http://www.statmodel.com/mplus/examples/webnote.html.

Muthén, B., Asparouhov, T. & Rebollo, I. (2006). Advances in behavioral genetics modeling using Mplus: Applications of factor mixture modeling to twin data. *Twin Research and Human Genetics*, 9, 313-324.

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

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

Muthén, B. & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463-469.

Nylund, K.L., Asparouhov, T., & MuthénMuthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.

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

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, 45, 247-270.

Penfield, R. D., & Camilli, G. (2007). *Differential item functioning and item bias*. In S. Sinharay & C.R. Rao (Eds.), Handbook of Statistics, Volume 26: Psychometrics (pp. 125-167). New York: Elsevier.

Rasch, G. (1960*). Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raju, N.S. (1988). The area between two item characteristic curves. Psychometrika, 54, 495-502.

Raju, N.S., Bond, R.K., & Larsen, V.S. (1989). An empirical assessment of the Mantel-Haenszel statistic to detect differential item functioning. *Applied Measurement in Education*, 2, 1-13.

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

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

Samuelsen, K. M. (2008). *Examining differential item functioning from a latent mixture perspective*. In G. R. Hancock & K. M. Samuelsen (Eds.), Advances in latent variable mixture models (pp. 177-197). Charlotte, NC: Information Age.

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

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.

Schwartz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.

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

Skaggs, G., & Lissitz, R.W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement*, 29(3), 227-242.

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed rasch models. *Methods of Psychological Research*, 4 (3), 19–32.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical Statistics in Psychology*, 27, 229-239.

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

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

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, 14, 147-176.

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

Tofighi, D., & Enders, C. K. (2007). *Identifying the correct number of classes in a growth mixture model*. In G. R. Hancock (Ed.), Mixture models in latent variable research (pp. 317–341). Greenwich, CT: Information Age.

Tucker, J. S., Orlando, M., & Ellickson, P. L. (2003). Patterns and correlates of binge drinking trajectories from early adolescence to young adulthood. *Health Psychology*, 22, 79-87.

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.

Van Nijlen, D., & Janssen, R. (2008, March). *Mixture IRT-models as a means of DIF-detection: Modelling spelling in different grades of primary school*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, 34, 166-180.

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

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 30*, 299-300.

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

Wright, B.D. (1968). *Sample-free test calibration and person measurement*. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.