ADAPTING MIXTURE MODELS TO TAKE INTO ACCOUNT MEASUREMENT NON-INVARIANCE

Veronica T. Cole

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for a Doctor of Philosophy degree in the Department of Psychology (Quantitative).

Chapel Hill
2017

Approved by:

Daniel Bauer

Kenneth Bollen

Patrick Curran

Kathleen Gates

Donglin Zeng

ABSTRACT

Veronica T. Cole: Adapting Mixture Models to Take Into Account Measurement Non-Invariance
(Under the direction of Daniel J. Bauer)

Researchers in the social sciences often use finite mixture models to find clusters of individuals on the basis of patterns of indicators. Though covariates are often incorporated in mixture models, it is most often assumed that these covariates exclusively affect class membership, rather than directly impacting the indicators themselves. Violation of this assumption indicates that the measurement of the latent classes by a given indicator is not constant across all individuals. Such violations, known as differential item functioning (DIF), have been well-studied in models for continuous latent variables, but virtually unexamined in models for categorical latent variables.

The current study extends the analytic and testing framework developed in continuous latent variable models to the case of latent class analysis. First, a Monte Carlo simulation systematically examined the effects of omitted DIF on mixture model results, as well as the performance of tests to detect DIF. In the presence of DIF in the data-generating model, the omission of these effects in the fitted model was associated with overestimation of the number of classes, as well as biased estimates of covariate effects on class membership and model-implied endorsement probabilities, particularly when classes were poorly separated and DIF was large. Including DIF in the model, even if the nature of this DIF was misspecified, mitigated this bias considerably. Standard model-based procedures drawn from the continuous latent variable modeling literature were shown to detect DIF with high sensitivity and specificity. Finally, DIF

was examined in an application of latent class analysis to alcohol use disorder (AUD) diagnostic criteria in an undergraduate sample. Researchers are advised to test comprehensively for DIF in applications of mixture models, in order to ensure that the results obtained are truly applicable to all individuals under study.

To Marion, with love and gratitude.

ACKNOWLEDGEMENTS

First and foremost, thanks are owed to my mentor, Daniel Bauer. Your willingness to take a chance on me set me on the right track; your support, wisdom, and patience have kept me there.

I am grateful to everyone in the L.L. Thurstone Psychometric Laboratory for their camaraderie and support. Additionally, I am grateful to my committee for their continuous feedback, which greatly strengthened this dissertation.

I would like to thank the data collection team from the REAL-U study (PI: Daniel J. Bauer) for generously providing the data used in Chapter 3. This work was supported by National Institutes of Health Grant F31 DA040334.

I am grateful to my parents, Paul Cole and Rita Braverman Cole, and my sister, Molly Cole, for their unyielding support. And finally, to my wonderful wife, Marion Johnson: there aren't enough words to express my gratitude, so I will thank you by being quiet for once.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1 1

AN INTEGRATED VIEW OF MEASUREMENT INVARIANCE IN CATEGORICAL AND CONTINUOUS LATENT VARIABLE MODELS 1

A general framework of latent variable models 4

Continuous latent variable models 4

Categorical latent variable models 6

Measurement invariance in continuous and categorical latent variable models 8

The effects of covariates in continuous latent variable models 8

The effects of covariates in categorical latent variable models 13

Tests of measurement invariance in continuous and categorical latent variable models 20

Summary and research questions 26

Chapter 2 26

Chapter 3 27

CHAPTER 2 29

A MONTE CARLO SIMULATION OF OMITTED DIF IN MIXTURE MODELS 29

Question A 29

Hypotheses 30

Data-generating model 31

Model fitting 35

Outcomes of interest 37

Results 42
Discussion 49
Question B 52
Hypotheses 52
Data-generating model 52
Model fitting 53
Outcomes of interest 54
Results 56
Discussion 58
CHAPTER 3 61
DIFFERENTIAL ITEM FUNCTIONING IN THE MEASUREMENT OF ALCOHOL USE DISORDER 61
Method 64
Study and sample 64
Data analytic strategy 66
Results 69
The unconditional model 69
Model-building strategy using itemwise tests 71
The final model 71
Comparing the impact-only and final models 73
Discussion 74
CHAPTER 4 77
DISCUSSION 77

LIST OF TABLES

Table 1. Demographic characteristics in a sampling of AUD studies using mixture models. 95

Table 2. Findings in a sampling of AUD studies using mixture models. 96

Table 3. Item parameters in data-generating model under varying levels of class separation and DIF. 97

Table 4. Class-specific logits in data-generating models containing both intercept and loading DIF. 98

Table 5. Class enumeration results for unconditional models. 99

Table 6. Class enumeration results for conditional models. 100

Table 7. Class membership parameters under the impact-only and intercept DIF models. 101

Table 8. Effects of covariates on items under the intercept DIF model, given loading DIF in the data-generating model. 102

Table 9. Baseline endorsement probabilities under the impact-only model for all data-generating models. 103

Table 10. Baseline endorsement probabilities under the intercept DIF model for all data-generating models. 104

Table 11. Adjusted Rand Index (ARI) statistics comparing true and estimated class membership under all data-generating and fitted models. 105

Table 12. Percent of replications with improper solutions, sensitivity, and specificity in the model-based and posthoc testing procedures. 106

Table 13. Alcohol Use Disorder (AUD) criteria used in the current study. 107

Table 14. Class enumeration statistics for unconditional models in both samples. 108

Table 15. Itemwise model-based DIF test results in both samples. 109

Table 16. Class membership and item parameters in the final model. 110

LIST OF FIGURES

Figure 1. Summary of all models fitted in Chapter 2. 111

Figure 2. Logit parameter estimates in classes 1 and 2 across all models. 112

Figure 3. Average baseline endorsement probabilities in large DIF conditions. 113

Figure 4. Average estimates of individual predicted probabilities in large DIF conditions. 114

Figure 5. Distribution of randomly-selected individual predicted probabilities in large DIF conditions. 115

Figure 6. Estimates of prevalence of membership to class 1 across all models. 116

Figure 7. Model-implied endorsement probabilities for both the impact-only and full models. 117

Figure 8. Model-implied endorsement probabilities for male and female subjects under the full model. 118

Figure 9. Model-implied endorsement probabilities for subjects of different ages under the full model. 119

Figure 10. Model-implied endorsement probabilities for subjects seen at visit 1 and visit 2 under the full model. 120

CHAPTER 1
AN INTEGRATED VIEW OF MEASUREMENT INVARIANCE IN CATEGORICAL AND CONTINUOUS LATENT VARIABLE MODELS

In behavioral research, it is often of interest to form homogeneous groupings of people based on some pattern of variables. These groupings may be regarded as approximations of more complex patterns in the population (Nagin, 1999), or they may be directly interpreted as scientifically or clinically meaningful categories (Meehl, 1992; 2004). In research into the developmental psychopathology of substance use, it has been of interest to determine whether there are any number of qualitatively different groups of individuals who endorse different patterns of alcohol use disorder (AUD) or substance use disorder (SUD) symptoms in various populations. The clinical utility of this endeavor may lie in finding different patterns of symptom endorsement among individuals meeting criteria for diagnosis in order to tailor treatment interventions (e.g., Chung and Martin, 2001). Alternatively, particularly in studies of young adults, the goal may be to identify subthreshold patterns of symptoms which predict poor outcomes including risky drinking behavior, use of other drugs, or transition to full-blown alcohol or substance use disorder (e.g., O'Connor and Colder, 2005).

Mixture models (McLachlan and Peel, 2000), a broad class of models which decompose a population into homogeneous categories, have frequently been used to form empirically-derived subgroups in the service of these goals. Commonly used models include latent class analysis (LCA; Lazarsfeld and Henry, 1968), latent profile analysis (LPA; Gibson, 1959), and factor mixture models (FMM; Lubke and Muthén, 2005; 2007), each of which impose different models on items within a given class; the general class of models is referred to interchangeably as mixture models and categorical latent variable models going forward. A number of different types of mixture models have been used to form homogeneous groupings of adolescents based

the use of alcohol, (Chung and Martin, 2001; Reboussin et al, 2006; Beseler et al., 2012) , marijuana (Schulenberg et al., 2005; Windle and Wiesner, 2004), and tobacco (Karp et al., 2005; Henry and Muthén, 2010), among other substances.

However, evidence across different applications of mixture models to AUD and SUD symptoms has been highly inconsistent, with different numbers and patterns of subgroups being found from study to study. Tables 1 and 2 illustrate the indeterminacy of these findings in mixture models of AUD symptoms by reviewing a small sample of studies using mixture models to find subgroups of individuals based on DSM-IV and DSM-V diagnosis items[1]. As shown in Table 1, these studies examine AUD symptoms in a wide diversity of samples, from heavy drinking undergraduates (Rinker and Neighbors, 2015), to a community sample of middle- and high school students (Mancha, Hulbert, and Latimer, 2011). Table 2 shows the discrepancies in findings across studies, with anywhere between 2 and 5 classes being identified across samples. Moreover, class solutions differ widely from one another across studies, with different prevalence rates in similar classes; further, while most arrange studies along a continuum of AUD symptom severity, three (Lynskey et al., 2005; Beseler et al., 2012; Jackson et al., 2014) find additional classes characterized by patterns of symptoms falling outside this continuum.

On the one hand, in light of the view of mixture models as a data reduction strategy which provides an imperfect approximation to reality, it is unreasonable to expect that any one, true configuration of AUD subgroups holds across samples and measurement contexts (Titterington, Smith, and Makov, 1985; Nagin, 1999). In this view, mixture models are not applied in an explanatory capacity -- as in, they are not intended to uncover meaningful classes which represent behavioral processes that exist in the real world. Rather, they may be employed descriptively or predictively, in order to help to describe patterns in the data or make predictions about new observations, regardless of whether classes themselves are directly interpretable.

[1] Though the set of items used to diagnose AUD and SUD changed from DSM-IV to DSM-V, there is sufficient overlap between the two sets of items that they are compared here; the issue of subtle differences in item sets used to diagnose AUD and SUD is elaborated upon in Study 2.

Under this view, it is arguably not necessary that mixture modeling results agree across studies, so long as each individual application results in a useful description of the data.

On the other, it may be the case that the lack of generalizability of findings owes to subtle differences in measurement properties of test items across samples and measurement contexts. The latter possibility represents the violation of a critical assumption in latent variable models known as measurement invariance (Meredith, 1993; Millsap, 2006). In the most general terms, measurement invariance represents the assumption that the measurement of some latent variable is the same across subjects, so that differences across subjects in item responses are solely a function of the latent variables they measure (Meredith, 1993). More specifically, measurement invariance involves the idea that the implied distribution of a vector of items is related to covariates only through its relationship with the latent variable. Thus, given observed item responses $\mathbf{y}_i$ measuring latent variables $\boldsymbol{\eta}_i$, and covariate values $\mathbf{x}_i$, measurement invariance can be written as:

$$f\left(\mathbf{y}_i \mid \boldsymbol{\eta}_i, \mathbf{x}_i\right) = f\left(\mathbf{y}_i \mid \boldsymbol{\eta}_i\right) \qquad (1)$$

If $\mathbf{x}_i$ has no effect on $\mathbf{y_i}$ over and above its effect on $\boldsymbol{\eta}_i$, then $\mathbf{y_i}$ shows measurement invariance with respect to $\mathbf{x}_i$ (Mellenbergh, 1989). Violations of measurement invariance occur when subjects differ systematically on the basis of a grouping variable in their responses to an item, despite having similar levels of the underlying construct that item measures. For example, in one study of an alcohol-related consequences index among college students, female participants were much less likely to endorse the item, "I spent too much money on alcohol" than male participants (Neal, Corbin, and Fromme, 2006). The authors hypothesized that this effect occurred because male subjects were more likely to pay for alcohol in mixed-gender settings than female subjects; thus, even if a male and female subject had the same level of the underlying construct, drinking consequences, the female subject would be much less likely to endorse this item, leading to a potentially biased estimate of this subject's problem drinking.

Measurement invariance has been explored extensively for the past several decades in models for continuous latent variables, such as the confirmatory factor analysis (CFA; Joreskog,

1967) and item response theory (IRT; Lord, 1980) models. However, in models for categorical latent variables, there has been very little systematic study of measurement invariance, so that there exists virtually no consensus on the consequences of unmodeled measurement non-invariance, as well as the detection and accommodation of measurement non-invariance, in mixture models.

This dissertation seeks to extend the evidence about measurement invariance from the continuous latent variable case to the categorical latent variable case in several ways. In Chapter 1, measurement invariance is described in the context of both continuous and categorical latent variables, with attention paid to the interpretation of measurement invariance-related findings. Additionally, procedures for identifying and directly modeling measurement non-invariance in the continuous and categorical latent variable cases are introduced. In Chapter 2, a computer simulation study will is conducted in order to assess (a) the robustness of categorical latent variable-related findings to measurement invariance, and (b) the efficacy of the proposed tests in the identification of measurement non-invariance in categorical latent variables. In Chapter 3, measurement non-invariance within categorical latent variables is examined in an empirical setting using a laboratory analog study of the measurement of alcohol and substance use in college students, in order to gauge to what extent measurement non-invariance may bias the inferences drawn from substantive results.

The continuous and categorical latent variable models will first be introduced, before proceeding to a general treatment of measurement invariance; finally, the extant testing procedures will be reviewed.

A general framework of latent variable models

Continuous latent variable models

Though the common factor model was originally formulated for continuous, normally distributed variables, generalization to the case of mixture models, in which indicators of all scale types are common, is facilitated by considering a more general approach which allows for a wide range of response distributions. The generalized linear factor analysis (GLFA;

Bartholomew, Knott, and Moustaki, 2011) is a flexible modeling framework which allow for the measurement of M continuous, normally distributed underlying latent variables ($m = 1, \ldots, M$) using a set of J items ($j = 1, \ldots, J$) and N subjects ($i = 1, \ldots, N$). As a subset of generalized linear models (GLM; McCullagh and Nelder, 1989; Agresti, 2007), the GLFA allows for the measurement of latent variables η_{im} by individual items y_{ij} using three basic components: a linear composite of latent variables η_{im} known as the linear predictor; a link function which translates the linear predictor to the expected value of y_{ij}; and a distributional specification for the random component of y_{ij}. In the GLFA the linear predictor, here denoted ω_{ij}, follows the form of the common factor model:

$$\omega_{ij} = \nu_j + \sum_{m=1}^{M} \lambda_{jm} \eta_{im} \tag{2}$$

The effects of η_{im} are transmitted through λ_{jm}, a factor loading which represents the linear effect of η_{im} on ω_{ij}; individual values of λ_{jm} are arranged in a $J \times M$ matrix of factor loadings $\mathbf{\Lambda}$. The item intercept ν_j represents the value of ω_{ij} when $\eta_{im} = 0$; individual values of ν_j are arranged in a $J \times 1$ vector $\boldsymbol{\nu}$. For each subject, individual values of latent variables η_{im} are arranged in an $M \times 1$ vector $\boldsymbol{\eta}_i$. The distribution of $\boldsymbol{\eta}_i$ is given by $\boldsymbol{\eta}_i \sim N_M(\boldsymbol{\kappa}, \boldsymbol{\Phi})$, where $\boldsymbol{\kappa}$ is an $M \times 1$ vector of factor means and $\boldsymbol{\Phi}$ is an $M \times M$ covariance matrix.

The expected value of y_{ij}, denoted μ_{ij}, is related to the linear predictor ω_{ij} through the link function, $g(x)$:

$$g\left(E\left(y_{ij} \mid \boldsymbol{\eta}_i\right)\right) = g\left(\mu_{ij}\right) = \omega_{ij} \tag{3}$$

where all of a subject i's expected values μ_{ij} may be arranged in a $J \times 1$ vector $\boldsymbol{\mu}_i$. Given this expected value, the random component of y_{ij} is then modeled by specifying the conditional distribution of $\mathbf{y}_i$, the $J \times 1$ vector of observed responses y_{ij} given latent variables $\boldsymbol{\eta}_i$. This conditional distribution may be any distribution in the exponential family (e.g., normal, binomial, Poisson, or gamma).

Many models commonly used in psychometrics may be considered specific parameterizations of the GLFA (Takane and De Leeuw, 1987; Wirth and Edwards, 2007; Bauer and Hussong, 2009). For instance, the normal common factor model is obtained when $g\left(\mu_{ij}\right)$ is

the identity link, and the conditional distribution of $\mathbf{y}_i$ is specified as multivariate normal with $J \times J$ covariance matrix $\boldsymbol{\Psi}$ Similarly, the two-parameter logistic (2-PL; Birnbaum, 1968) model may be parameterized by choosing a logistic link, so that

$$\omega_{ij} = g\left(\mu_{ij}\right) = \ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) \tag{4}$$

and specifying the conditional distribution of $\mathbf{y}_i$ as a p-variate vector of independent Bernoulli trials The choice of a logistic link function and a Bernoulli distribution yields the familiar equation for the probability of individual *i* endorsing item *j*:

$$\mu_{ij} = P\left(y_{ij} = 1 \mid \boldsymbol{\eta}_i\right) = \frac{1}{1+\exp\left(-\left(\sum_{m=1}^{M} \lambda_{jm}\eta_{im} + \nu_j\right)\right)} \tag{5}$$

The intercept parameter ν_j, often denoted c_j in the 2-PL model, represents the log-odds of endorsing item *j* given a score of 0 on η_{im} The parameter λ_{jm} (often denoted a_{jm} in the 2-PL model), referred to here as the loading parameter, is the predicted increment in log-odds of endorsing item *j* given a one-unit increase in η_{im} .

Categorical latent variable models

Finite mixture models (McLachlan and Peel, 2000) express observed variables $\mathbf{y}_i$ as a function of a given subject's membership to one of *K* latent subgroups, each of which is governed by its own subgroup-specific set of parameters. We consider finite mixtures with arbitrary response distributions of $\mathbf{y}_i$ as generalizations of the GLFA above, in which the latent variable $\boldsymbol{\eta}_i$ represents membership to a given latent class. This concordance between finite mixture models and the GLFA will later allow us to draw parallels between the two models in terms of measurement invariance.

As in the continuous latent variable case, $\boldsymbol{\eta}_i$ here denotes a vector of latent variables for subject *i*; here, however, $\boldsymbol{\eta}_i$ is a $K \times 1$ vector of latent variables η_{ik} which take a value of 1 if subject *i* is in class *k* and 0 otherwise. Class membership $\boldsymbol{\eta}_i$ is distributed according to a multinomial distribution with endorsement probabilities given by the mixing probability vector $\boldsymbol{\pi}$

, with individual elements π_k, where, $\sum_{k=1}^{K}\pi_k = 1$. Each class is governed by its own set of parameters, which produce a class-specific implied distribution of $\mathbf{y}_i$, $f\left(\mathbf{y_i} \mid \eta_{ik} = 1\right)$; these distributions are weighted by the mixing probabilities $\boldsymbol{\pi}$ to yield the marginal distribution of observed variables:

$$f(\mathbf{y_i}) = \sum_{k=1}^{K}\pi_k f\left(\mathbf{y_i} \mid \eta_{ik} = 1\right) \tag{6}$$

Each individual may also be characterized by a *K*×1 vector of posterior probabilities, $\boldsymbol{\tau}_i$, whose individual elements τ_{ik} represent the probability, given $\mathbf{y_i}$, that individual *i* is a member of class *k.* These posterior probabilities are given by Bayes' Rule as:

$$\tau_{ik} = \frac{\pi_k f\left(\mathbf{y_i} \mid \eta_{ik} = 1\right)}{\sum_{h=1}^{K}\pi_h f\left(\mathbf{y_i} \mid \eta_{ih} = 1\right)} \tag{7}$$

The within-class specification of $\mathbf{y}_i$ may take a number of different forms. Here we consider latent class (LCA; Lazarsfeld and Henry, 1968) and latent profile (LPA; Gibson, 1969) models, which specify a conditional independence relationship between items y_{ij}, which may be categorical (LCA), continuous (LPA), or any combination thereof. As in the continuous latent variable framework, we model the expected values of the items using a linear predictor, a link function $g\left(\mu_{ij}\right)$, and a conditional distribution.

Define subject *i*'s linear predictor for item *j* given membership to class *k* as ω_{ijk}. In an LPA or LCA, no explicit model for values of ω_{ijk} is invoked within-class; more complex models (e.g., factor mixture models) impose a parameterization similar to Equations 2-3 in the GLFA within-class. In an LCA or LPA model without covariates, ω_{ijk} depends exclusively on class membership; thus,

$$\omega_{ijk} = \delta_{jk} \tag{8}$$

where δ_{jk} is the value of the linear predictor for item *j* that is assigned to all members of class *k.* The values ω_{ijk} and δ_{jk} may each be arranged in *J*×1 vectors $\boldsymbol{\omega}_{ik}$ and $\boldsymbol{\delta}_k$ respectively.

The expected value of y_{ij} given membership to class *k* is denoted μ_{ijk} and is related to ω_{ijk} through the link function. Because ω_{ijk} shows no variance within a given latent class, neither does μ_{ijk}, i.e.:

$$\mu_{ijk} = E\left(y_{ij} \mid \eta_{ik} = 1\right) = g^{-1}\left(\omega_{ijk}\right) = g^{-1}\left(\delta_{jk}\right) \tag{9}$$

In the case of an LCA for a binary latent variable, we may wish to use a logit link as shown in Equation 5, and specify a Bernoulli distribution for y_{ij} with class-specific endorsement probabilities μ_{ijk} given by:

$$\mu_{ijk} = P\left(y_{ij} = 1 \mid \eta_{ik} = 1\right) = \frac{1}{1 + \exp(-\delta_{jk})} \tag{10}$$

The continuous and categorical latent variable models above do not take into account the effects of covariates that might generate measurement non-invariance; we will now extend the models accordingly, in order to formally introduce measurement invariance.

Measurement invariance in continuous and categorical latent variable models

Equation 1 indicates that measurement invariance rests on determining conditional independence between covariates and items. In order to formally make this determination, we must distinguish between two sorts of covariate relationships: impact of covariates on the latent variable, and direct effects of covariate on the items after controlling for the latent variable. Whereas the former indicates true differences between subjects on the underlying latent variable, the latter represents violations of the assumption of measurement invariance.

<u>The effects of covariates in continuous latent variable models</u>

Ultimately, the relationship between covariates, latent variables, and items may be considered in the context of an arbitrary number of covariates of any scale. However, because impact and measurement invariance were originally formulated in the context of differences across groups (Meredith, 1993; Reise, Widaman, and Pugh, 1993; Widaman and Reise, 1997), these basic concepts will be presented in terms of a multiple groups factor model (Joreskog, 1971) first; definitions will then be generalized to covariates of any scale.

Multiple group approaches. Given some number of groups G, define $G - 1$ binary indicator variables X_g ($g = 1, \ldots, G$), which take a value of 1 if subject i is a member of group g and 0 otherwise. Given this formulation, the GLFA is fit simultaneously to each of the g groups, allowing all item parameters to potentially differ across groups (subject to identification constraints), so that the expression for the linear predictor of $\mathbf{y}_i$ becomes:

$$\omega_{ijg} = \nu_{jg} + \sum_{m=1}^{M} \lambda_{jmg} \eta_{img} \tag{11}$$

Note that all terms are now subscripted by g; ω_{ijg} is now the expression for the linear predictor given membership to group g, λ_{jmg} is the loading for the m^{th} latent variable on the j^{th} item for group g; and ν_{jg} is the measurement intercept for the j^{th} item for group g. These terms may be arranged into $J \times M$ matrix $\mathbf{\Lambda}_g$ and the $J \times 1$ vector $\boldsymbol{\nu}_g$, respectively. The latent variable η_{img} may be arranged in a subject-specific, group-specific $M \times 1$ vector $\boldsymbol{\eta}_{ig}$, which is distributed according to a group-specific mean and variance, i.e., $\boldsymbol{\eta}_{ig} \sim N_M\left(\boldsymbol{\kappa}_g, \mathbf{\Phi}_g\right)$. Typically for identification purposes, the mean and variance of the latent variable are set to 0 and 1 for identification in one group. Differences across groups in $\boldsymbol{\kappa}_g$ and $\mathbf{\Phi}_g$ indicate impact of group membership on the distribution of the latent variable. This impact does not violate any model assumptions; indeed, it is often expected that there will be differences between groups in these aspects of the latent variable distribution (Vandenberg and Lance, 2000).

By contrast, measurement parameters are expected to be the same across groups; this is the assumption of measurement invariance. The most fundamental form of invariance is configural (or "pattern") invariance, which generally holds if the same factors account for the same items across groups (Steenkamp and Baumgartner, 1998; Meredith, 1993; Widaman and Reise, 1997; Horn and McArdle, 1992). Configural invariance thus requires that the number of latent variables, and thus the dimension of the $\mathbf{\Lambda}_g$ and $\mathbf{\Phi}_g$ matrices, must be identical across groups. Further, it requires that the pattern of zero loadings in $\mathbf{\Lambda}_g$ be identical across groups, so that the general pattern of relationships between $\boldsymbol{\eta}_{ig}$ and $\mathbf{y}_i$ is identical across groups, regardless of the magnitude of these effects. Configural invariance is critical to establish before further

invariance testing, for without it one cannot have confidence that $\mathbf{y}_i$ represents the same construct across subjects (Vandenberg and Lance, 2000).

Having established configural invariance, one must then examine weak, strong, and strict metric invariance (Meredith, 1993; Vandenberg and Lance, 2000; Steenkamp and Baumgartner, 1998). Under weak metric invariance, the factor loadings $\mathbf{\Lambda}$ must be equal across groups. Whereas configural invariance lacks a direct mathematical expression, we may consider weak invariance as a simple equality relationship across subjects in $\mathbf{\Lambda}$. Weak metric invariance holds when: $\mathbf{\Lambda}_g = \mathbf{\Lambda}$ for all g. In general, if researchers wish to make inferences about covariance structure which hold across different subjects or groups thereof, weak invariance must hold. Under strong metric invariance, weak invariance must hold and, additionally, measurement intercepts $\mathbf{\nu}$ be equal across groups. Considered again in the context of binary grouping variable X_g, strong metric invariance holds when $\mathbf{\nu}_g = \mathbf{\nu}$.

If researchers wish to compare factor means between groups, strong invariance must hold, as group-specific factor means will be biased in the event that differences in measurement intercepts are not accounted for (Bollen, 1989; Joreskog and Sorbom, 1996; Millsap, 2011). Finally, in the normal factor model, in which the conditional distribution of the indicators is multivariate normal with covariance matrix $\mathbf{\Psi}$, the equivalence of this covariance matrix across groups may be tested. Thus, under strict metric invariance (Meredith, 1993), both weak and strong measurement invariance must hold, but the unique factor covariance matrix $\mathbf{\Psi}$ must also be equal across all groups, i.e., $\mathbf{\Psi}_g = \mathbf{\Psi}$.

The form of measurement invariance expressed by Equation 1, which requires that the entire distribution of $\mathbf{y}_i$ depend only on $\mathbf{\eta}_i$, requires strict invariance. However, this is an extremely stringent condition which generally does not hold in practice. The presence of weak and strong measurement invariance yields what is referred to as first-order invariance (Millsap, 2006; p. 50)- meaning that, while the entire conditional distribution of $\mathbf{y}_i$ may not be equal across all values of g, the expected values are – i.e.,

$$E\left(\mathbf{y}_i \mid \mathbf{\eta}_i, \mathbf{x}_i\right) = E\left(\mathbf{y}_i \mid \mathbf{\eta}_i\right) \tag{12}$$

Furthermore, even though strict invariance of $\boldsymbol{\Psi}$ matrices is often untenable in practice, it is theoretically necessary if one wants to compare score estimates $\hat{\boldsymbol{\eta}}_i$ derived from test $\mathbf{y}_i$ across groups in the normal factor model (Millsap, 1997; Millsap, 2006). However, within the GLFA setting, the dispersion of $\mathbf{y}_i$ given $\boldsymbol{\mu}_i$ is defined not always freely estimated but may be determined by the distributional specification chosen by the researcher. Thus, in cases such as the 2-PL IRT model in which a Bernoulli distribution relates $\boldsymbol{\mu}_i$ to $\mathbf{y}_i$, strict measurement invariance is undefined and strong invariance implies full invariance – and thus equivalence of loadings and intercepts are sufficient to ensure that estimates $\hat{\boldsymbol{\eta}}_i$ are unbiased.

Under the above formulation, the respective forms of invariance (weak, strong, and strict) refer to entire matrices ($\boldsymbol{\Lambda}_g$, $\boldsymbol{\nu}_g$, and $\boldsymbol{\Psi}_g$) that must be invariant across groups. By contrast, "partial measurement invariance" (Byrne, Shavelson, and Muthén, 1989) refers to a condition in which some, but not all, items are shown to have measurement parameters which vary across groups; thus, partial weak, strong, or strict non-invariance would occur when $\lambda_{jmg} \neq \lambda_{jm}$, $\nu_{jg} \neq \nu_j$, or $\psi_{jhg} \neq \psi_{jh}$ for some *h, j, or m*.

The amount of partial non-invariance that is tolerable in a given model depends largely on the goals of the researcher. Byrne, Shavelson, and Muthén (1989) argue that even in the case of some factor loadings being non-invariant, thus causing the researcher to reject the hypothesis of complete weak invariance across all items in test $\mathbf{y}_i$, some cross-group comparisons may be made as long as non-invariant items are in the minority. The authors, as well as others in later work (Steenkamp and Baumgartner, 1998), further argue that if cross-validation work supports a test's validity, a partial lack of metric invariance can often be considered an artifact of the sample.

Importantly, measurement invariance is typically assessed at the item level in IRT as differential item functioning (DIF). These item level analyses are synonymous with detection of partial non-invariance. Differences in λ_{jm} are generally termed *non-uniform DIF* in IRT; when non-uniform DIF is present, differences across groups in the endorsement probability vary across levels of the latent variable. Differences in ν_{jm} within the IRT setting are typically termed

uniform DIF; when uniform DIF is present, differences across subjects in endorsement probability do not vary across levels of the latent variable. Importantly, while non-uniform and uniform DIF are related to differences in the same parameters as in weak and strong measurement invariance, the two approaches differ in that IRT considers uniform and non-uniform DIF independently of one another, whereas in the CFA framework one form of invariance is implied by successive levels (e.g., strong metric invariance necessarily implies weak metric invariance). However, when speaking about the topic in general terms (i.e., not considering any particular sort of measurement invariance or DIF), the terms DIF and measurement non-invariance will generally be used interchangeably in the current work.

Regression-based approaches. Though most of the original formulations of measurement invariance consider it in terms of differences across groups in parameters, it is clear that this does not capture the full range of covariates across which the conditional distribution of $\mathbf{y}_i$ may differ. In order to consider measurement invariance in terms of an arbitrary number of $\mathbf{x}_i$ variables of any scales, it is helpful to consider the multiple-groups expression as a special case of a general model which allows parameter differences across covariates $\mathbf{x}_i$. Such models include the multiple item multiple cause (MIMIC; Joreskog and Goldberger, 1975), and the more general moderated nonlinear factor analysis (MNLFA; Bauer and Hussong, 2009), which incorporate covariates directly into the expressions for impact and measurement parameters; as the MIMIC is a special case of the MNLFA, the latter will be presented here. Given an $N \times P$ design matrix of covariates $\mathbf{X}$, rows $\mathbf{x}_i$ represent individual i's value on each of p covariates, each individually denoted x_{ip}, the MNLFA considers the mean and variance of the latent variable, $\boldsymbol{\kappa}(\mathbf{x}_i)$ and $\boldsymbol{\Phi}(\mathbf{x}_i)$, as well as measurement parameters $\boldsymbol{\Lambda}(\mathbf{x}_i)$ and $\boldsymbol{\nu}(\mathbf{x}_i)$, as functions of each subject's $p \times 1$ vector of covariates $\mathbf{x}_i$. The $M \times 1$ vector of latent variable means, $\boldsymbol{\kappa}(\mathbf{x}_i)$, and the M×M factor covariance matrix $\boldsymbol{\Phi}(\mathbf{x}_i)$ have individual elements $\kappa_m(\mathbf{x}_i)$ and $\phi_{hm}(\mathbf{x}_i)$, given by:

$$\kappa_m(\mathbf{x}_i) = \kappa_{0m} + \sum_{p=1}^{P} \kappa_{pm} x_{ip}$$
$$\phi_{mm}(\mathbf{x}_i) = \phi_{0mm} \exp\left(\sum_{p=1}^{P} \phi_{mm} x_{ip} \right) \quad (13)$$

where $\kappa_m(\mathbf{x}_i)$ is the m^{th} element of $\kappa_m(\mathbf{x}_i)$ and $\phi_{hm}(\mathbf{x}_i)$ is the *h*, m^{th} element of $\mathbf{\Phi}(\mathbf{x}_i)$. Here, κ_{0m} and ϕ_{0hm} are the intercept values corresponding subject *i*'s predicted mean for factor *m* and covariance between factors *m* and *h*, respectively at values of zero on all covariates. The coefficients κ_{pm} and ϕ_{phm} transmit the effect of the p^{th} covariate onto these quantities.

Measurement parameters are allowed to differ across subjects, with $\lambda_{jm}(\mathbf{x}_i)$ and $\nu_{ij}(\mathbf{x}_i)$ representing the loading and intercept parameter implied by an individual's covariates $\mathbf{x}_i$. These values are given by:

$$\lambda_{jm}(\mathbf{x}_i) = \lambda_{0jm} + \sum_{p=1}^{P} \lambda_{pjm} x_{ip}$$
$$\nu_j(\mathbf{x}_i) = \nu_{0j} + \sum_{p=1}^{P} \nu_{pj} x_{ip} \quad (14)$$

Under the above formulation coefficients λ_{pjm} and ν_{pj} transmit the effect of covariates $\mathbf{x}_i$ onto subject i's predicted factor loading and intercept parameters for item *j* and latent factor *m*, respectively. A nonzero value of λ_{pjm} indicates that $\lambda_{jm}(\mathbf{x}_i)$ will vary across levels of x_{ip}; a nonzero value of ν_{pj} indicates that $\nu_{jm}(\mathbf{x}_i)$ will vary across levels of x_{ip}. Thus, weak metric invariance holds when $\lambda_{pjm} = 0$ for all p, and thus $\lambda_{jm}(\mathbf{x}_i) = \lambda_{0jm}$ for all individuals, and strong metric invariance holds when, additionally, ν_{pj} for all p and thus $\nu_{jm}(\mathbf{x}_i) = \nu_{0jm}$ for all individuals. Note that the multiple-groups formulation may be obtained by considering binary grouping variable X_g as the only element of $\mathbf{x}_i$.

The expected value of y_{ij}, conditional on covariates $\mathbf{x}_i$, may be denoted $\mu_{ij}(\mathbf{x}_i)$, and is given by:

$$\mu_{ij}(\mathbf{x}_i) = g^{-1}\left(\nu_j(\mathbf{x}_i) + \sum_{m=1}^{M} \lambda_{jm}(\mathbf{x}_i) \eta_{im} \right) \quad (15)$$

where item parameters $\nu_j(\mathbf{x}_i)$ and $\lambda_{jm}(\mathbf{x}_i)$ now incorporate the effects of covariates $\mathbf{x}_i$.

<u>The effects of covariates in categorical latent variable models</u>

Though measurement invariance has been explored much less extensively in the categorical latent variable framework, a number of parallels to the continuous latent variable approach, including the distinction between the multiple-groups and model-based strategy, may be drawn. These will now be explored.

Multiple group analysis. As in the GLFA scenario, measurement non-invariance may be examined in terms of a multiple groups LCA (Clogg and Goodman, 1985; McCutcheon, 2000; Collins and Lanza, 2010; pp. 113-148). As before, we define a categorical grouping variable X_g which takes a value of 1 if subject *i* is a member of group *g* and 0 otherwise. Prior probabilities of class membership are now conditional on group membership, so that $\boldsymbol{\pi}_g$ is the probability of membership to class *k* given membership to group *g*. Similarly, the class-specific function for the distribution of $\mathbf{y}_i$ is now estimated conditional on membership to group *g*, and is denoted $f\left(\mathbf{y}_i \mid \eta_{ik}=1, X_g=1\right)$. The marginal distribution of $\mathbf{y}_i$ within group *g* is now given by:

$$f(\mathbf{y}_i \mid X_g=1)=\sum_{k=1}^{K}\pi_{kg} f\left(\mathbf{y}_i \mid \eta_{ik}=1, X_g=1\right) \tag{16}$$

Within the multiple-groups setting, differences across groups in $\boldsymbol{\pi}_g$ may be considered as a form of impact: as these are the class endorsement rates for group *g*, the finding $\boldsymbol{\pi}_g$ differs across groups (i.e., $\boldsymbol{\pi}_g \neq \boldsymbol{\pi}$) indicates that the distribution of $\boldsymbol{\eta}_i$ is different for members of group *g* than nonmembers.

Given that the vast majority of multiple groups applications focus on the LCA case, we discuss this case here and thus assume that $f\left(\mathbf{y}_i \mid \eta_{ik}=1, X_g=1\right)$ specifies a multivariate Bernoulli distribution. Define the endorsement probability for y_{ij} for individual *i,* given that this individual belongs to class *k* and group *g*, as μ_{ijkg}, which is given by

$$\mu_{ijkg}=P\left(y_{ij}=1 \mid \eta_{ik}=1, X_g=1\right)=\frac{1}{1+\exp(-\delta_{jkg})} \tag{17}$$

Note that, in this case, the linear predictor also differs based on membership to group *g*; i.e., $\omega_{ijkg}=\delta_{jkg}$. In the context of multiple group LCA, Collins and Lanza (2010) have drawn a few comparisons to the continuous latent variable case to begin to define measurement invariance. First, configural invariance may be considered to hold when the same number of

classes K holds across groups (i.e., $K_g = K$ for all g) and, additionally, the general pattern of differences across classes in endorsement probabilities is the same across groups. Beyond this comparison, however, weak and strong measurement invariance are not generally defined; rather, measurement invariance with respect to group is said to hold if

$$\delta_{jkg} = \delta_{jk} \Leftrightarrow \omega_{ijkg} = \omega_{ijk} \Leftrightarrow \mu_{ijkg} = \mu_{ijk} \tag{18}$$

for all g. In other words, measurement invariance holds if, given membership to a given class, individuals in group g have the same probability of endorsing items y_{ij} than those who are not in group g.

Regression-based approaches**.** As in the continuous latent variable framework, covariate effects may be directly modeled in categorical latent variable models. Huang and Bandeen-Roche (2004) presented an approach which directly models covariate effects in LCA; in allowing for the impact of covariates on both item responses and underlying latent variables, this model is very similar to the MNLFA presented above, but for categorical latent variables. Reboussin et al. (2008) propose a similar model with added constraints to accommodate local dependence between item pairs, but because estimation of this model is somewhat complicated and requires second-order estimating equations, we present the original formulation here.

In the presence of covariates, the prior probability of class membership π_{ik} becomes $\pi_{ik}(\mathbf{x}_i)$, which is related to covariates through a multinomial logistic regression equation as follows:

$$\pi_{ik}(\mathbf{x}_i) = P(\eta_{ik} = 1 | \mathbf{x}_i) = \frac{\exp(\alpha_{0k} + \sum_{p=1}^{P} \alpha_{pk} x_{ip})}{\sum_{h=1}^{K} \exp(\alpha_{0h} + \sum_{p=1}^{P} \alpha_{ph} x_{ip})} \tag{19}$$

Here, α_{0k} is an intercept representing the baseline log-odds of membership to class k, and α_{pk} is a coefficient transmitting the effect of x_{ip} on $\pi_{ik}(\mathbf{x}_i)$. Covariates' effects on class membership may be considered as conceptually similar to impact in the GLFA setting, for they represent the effects of covariates on the underlying latent variable. As in the GLFA, we often

expect differences between subjects in their class membership probabilities, and the presence of such effects is not a violation of any sort of assumption.

Within the LCA, effects of covariates on the items after controlling for $\boldsymbol{\eta}_i$ are expressed by considering the expected value of $\mathbf{y}_i$ given membership to class *k* as $\boldsymbol{\mu}_k\left(\mathbf{x}_i\right)$ with individual elements $\mu_{jk}\left(\mathbf{x}_i\right)$, given by:

$$\mu_{jk}\left(\mathbf{x}_i\right)=\frac{1}{1+\exp\left(-(\delta_{0\,jk}+\sum_{p=1}^{P}\delta_{pjk}x_{ip})\right)} \tag{20}$$

Here the intercept value $\delta_{0\,jk}$ represents the log-odds of endorsing item *j* within class *k* when all covariates are zero and δ_{pjk} transmits the effect of covariate x_{ip} on this log-odds. As in the GLFA, here measurement invariance holds when δ_{pjk} for all *p*.

Though the model was originally proposed in terms of an LCA, it has been extended to include measures of all different scales (Muthén 2002) and can be considered in the GLM format with a linear predictor given by:

$$\omega_{ijk}\left(\mathbf{x}_i\right)=\delta_{0\,jk}+\sum_{p=1}^{P}\delta_{pjk}x_{ip} \tag{21}$$

Note that, unlike in Equation 8 in the model without covariates, the linear predictor $\omega_{ijk}\left(\mathbf{x}_i\right)$ is no longer dependent only on class membership, but also incorporates information about covariates $\mathbf{x}_i$. The formulation in Equation 20 for binary items may be extended to any link function, as in Equation 9:

$$\mu_{ijk}\left(\mathbf{x}_i\right)=g^{-1}\left(\omega_{ijk}\left(\mathbf{x}_i\right)\right) \tag{22}$$

Huang and Bandeen-Roche (2004) advise against estimating δ_{pjk} across classes, and advocate instead for constraining δ_{pjk} to be equal to some δ_{pj0} across all classes, citing the possibility of under-identification; in particular, they caution that the parameters δ_{pjk} may be linearly dependent with α_{pk}, so that covariate effects on item endorsement probability cannot be disentangled from their effects on class membership in the LCA. However, Wang and Zhou (2014) offer evidence against this concern by showing conditions for local and global identifiability of freely estimated δ_{pjk} parameters in a general class of finite mixture models

which subsume LCA. In particular, for local identifiability of parameters, they show a number of conditions which are sufficient to ensure that the Jacobian matrix of the likelihood function is full rank. Furthermore, they extend the results of Kruskal (1977) and Allman (2008) to show that sufficient conditions for global identifiability, while difficult to show in practice, are actually easier to meet in a model with covariates than one without covariates.

However, even when the model in Equations 19, 21, and 22 is identified, the meaning of class-varying δ_{pjk} parameters is not fully straightforward: how does the interpretation of item responses under class-varying δ_{pjk} differ from the case in which $\delta_{pjk} = \delta_{pj0}$ for all classes? While much of the extant literature examining measurement non-invariance imposes this constraint (e.g., Asparouhov and Muthén, 2014; Reboussin et al., 2006), others allow for class-varying δ_{pjk} parameters (Muthén, 2004); however, by and large these approaches do not directly interpret this distinction. One exception is the factor mixture model (FMM; Lubke and Muthén, 2005; Lubke and Muthén, 2007), which relates $\mathbf{y}_i$ to a combination of categorical and continuous latent variables $\boldsymbol{\eta}_i$. This framework allows for class-varying direct effects δ_{pjk} and interprets them as evidence of non-invariance, but caution that "although the possibility of specifying different class-specific effects is clearly an advantage, it is also obvious that the interpretability of a model can rapidly decrease with an increasing number of effects" (Lubke and Muthén, 2007, p. 29).

One potential interpretation of δ_{pjk} arises from expressing the generalized mixture model in Equations 29, 21, and 22 in the form of a GLFA. Bartholomew, Knott, and Moustaki (2011; pp. 157-191) point out that the categorical latent variable model shown above and the normal GLFA are special cases of the same general latent variable model; they differ only in the prior distribution of the latent variable, which is multivariate normal in the continuous latent variable case and a degenerate distribution with probability π_k where $\eta_{ik} = 1$ and 0 elsewhere in the categorical latent variable case. They further show that, given the same response distribution linking $\mathbf{y}_i$ to $\boldsymbol{\mu}_i$, a finite mixture model with *K* classes yields the same values of $\boldsymbol{\mu}_i$ as a GLFA with *K*-1 factors. This concordance is particularly useful because, if we reformulate the generalized mixture model as a GLFA, the within-class expected values of the linear predictor

$\omega_{ijk}(\mathbf{x}_i)$ may be recast as a function of an intercept and a loading relating class membership to the overall expected value of y_{ij}, permitting mixture models to be considered within the standard framework of measurement invariance.

To see the concordance between these two models, note that the expectation of y_{ij} conditional on the latent variable $\boldsymbol{\eta}_i$ is:

$$\mu_{ij}(\mathbf{x}_i) = E(y_{ij} \mid \boldsymbol{\eta}_i, \mathbf{x}_i) = \sum_{k=1}^{K} \eta_{ik} \mu_{ijk}(\mathbf{x}_i) = \sum_{k=1}^{K} \eta_{ik} g^{-1}\left(\delta_{0jk} + \sum_{p=1}^{P} \delta_{pjk} x_{ip}\right) \tag{23}$$

Because $\boldsymbol{\eta}_i$ is an indicator variable which takes a value of 1 when $\eta_{ik} = 1$ and 0 otherwise,

$$\begin{aligned} \sum_{k=1}^{K} \eta_{ik} g^{-1}\left(\delta_{0jk} + \sum_{p=1}^{P} \delta_{pjk} x_{ip}\right) &= g^{-1}\left(\sum_{k=1}^{K} \eta_{ik}\left(\delta_{0jk} + \sum_{p=1}^{P} \delta_{pjk} x_{ip}\right)\right) \\ &= g^{-1}\left(\sum_{k=1}^{K}\left(\eta_{ik}\delta_{0jk} + \eta_{ik}\sum_{p=1}^{P} \delta_{pjk} x_{ip}\right)\right) \end{aligned} \tag{24}$$

Given an LCA or LPA formulation, we may then decompose δ_{0jk} to be consistent with the expression of effects in a GLFA, as follows:

$$\delta_{0jk} = \lambda_{0jk} + \nu_{0j} \tag{25}$$

Here ν_{0j} is an intercept term, which does not vary across classes. It may be coded to represent either the predicted value of δ_{0jk} in a reference class or a weighted or unweighted mean value of δ_{0jk} across groups. The term λ_{0jk} is a loading representing the deviation from ν_j associated with membership to class *k*. We can similarly decompose δ_{pjk} as follows:

$$\delta_{pjk} = \lambda_{pjk} + \nu_{pj} \tag{26}$$

where ν_{pj} is an intercept term, which may be coded to represent either the effect of covariates $\mathbf{x}_i$ on δ_{jk} in a reference class or a weighted or unweighted mean effect of $\mathbf{x}_i$ on δ_{jk} across classes. The coefficient λ_{pjk} represents the deviations from ν_{pj} associated with membership to class *k*. Substitution of the above terms into Equation 24 and rearrangement yield the following expression for $\mu_{ij}(\mathbf{x}_i)$:

$$\begin{aligned}\mu_{ij}\left(\mathbf{x}_i\right) &= g^{-1}\left(\sum_{k=1}^{K}\left(\eta_{ik}\left(\lambda_{0jk}+\nu_{0j}\right)+\eta_{ik}\sum_{p=1}^{P}\left(\lambda_{pjk}+\nu_{pj}\right)x_{ip}\right)\right)\\ &= g^{-1}\left(\sum_{k=1}^{K}\eta_{ik}\left(\lambda_{0jk}+\sum_{p=1}^{P}\lambda_{pjk}x_{ip}\right)+\sum_{k=1}^{K}\eta_{ik}\left(\nu_{0j}+\sum_{p=1}^{P}\nu_{pj}x_{ip}\right)\right)\end{aligned} \tag{27}$$

Because values of ν_{pj} and ν_{pj} do not vary across class, the summation in the second term may be eliminated:

$$g^{-1}\left(\sum_{k=1}^{K}\eta_{ik}\left(\lambda_{0jk}+\sum_{p=1}^{P}\lambda_{pjk}x_{ip}\right)+\left(\nu_{0j}+\sum_{p=1}^{P}\nu_{pj}x_{ip}\right)\right) \tag{28}$$

We may define the first and second terms in parentheses, respectively, as:

$$\begin{aligned}\lambda_{jk}\left(\mathbf{x}_i\right) &= \lambda_{0jk}+\sum_{p=1}^{P}\lambda_{pjk}x_{ip}\\ \nu_{j}\left(\mathbf{x}_i\right) &= \nu_{0j}+\sum_{p=1}^{P}\nu_{pj}x_{ip}\end{aligned} \tag{29}$$

Equation 27 becomes the familiar expression which is equivalent to Equation 15 in the continuous latent variable case:

$$\mu_{ij}\left(\mathbf{x}_i\right) = g^{-1}\left(\nu_j\left(\mathbf{x}_i\right)+\sum_{k=1}^{K}\lambda_{jk}\left(\mathbf{x}_i\right)\eta_{ik}\right) \tag{30}$$

The formulation presented in Equations 29-30 allows for an intuitive extension of the levels of invariance described for continuous latent variable models to categorical latent variable models: when $\lambda_{jk}\left(\mathbf{x}_i\right)=\lambda_{0jk}$ for all *j*, weak metric invariance exists with respect to covariates $\mathbf{x}_i$; when $\nu_j\left(\mathbf{x}_i\right)=\nu_{0j}$ for all *j*, strong metric invariance exists with respect to covariate $\mathbf{x}_i$. Currently, mixture models are not fit using this formulation, leading to a relative lack of interpretability of mixtures as a general measurement model. However, reparameterization of the basic mixture model in Equations 19, 21, and 22 to be consistent with this formulation can be accomplished through the use of non-linear constraints. By estimating the model this way, one is able to disaggregate levels of measurement invariance with categorical latent variables.

Having established a theoretical parallel between levels of invariance in continuous and categorical latent variable models, I now turn to the issue of how best to test measurement invariance in the categorical latent variable framework. Given the parallels established here, I

begin by considering existing testing procedures, which are well-developed for continuous latent variable models but have received scant attention for categorical latent variable models.

Tests of measurement invariance in continuous and categorical latent variable models

Within continuous latent variable methods, the most common method for testing measurement invariance is to conduct a multiple groups analysis as outlined in Equation 11. In the most general terms, the strategy is to initially set some combination of tem parameters $\mathbf{\Lambda}$ and $\mathbf{v}$ to be equal across groups, and to remove equality constraints across groups one at a time; if removing an equality constraint on a parameter results in a significant improvement in model fit, that parameter is non-invariant across classes. Items whose measurement parameters λ_{jm} and ν_j are held to equality across classes while other items' measurement parameters are tested for non-invariance are termed anchor items.

There exists significant controversy over virtually every aspect of the general procedure described above, including the choice of anchor items (Yoon and Millsap, 2007), whether to test invariance of multiple items at a time (Stark, Chernyshenko, and Drasgow, 2006), the order in which to test different types of parameters (i.e., whether λ_{jm} and ν_j should be tested at the same time; Vandenberg and Lance, 2000), and the optimal choice of baseline model (i.e., whether each successive model should be compared to a minimally or maximally constrained model; Reise, Widaman, and Pugh, 1993). One of the most widely used of these procedures is the likelihood ratio test algorithm (IRT-LR-DIF; Thissen, 2001), which is formulated as follows in the case of a single latent variable:

1. Set the mean and variance of η_i in the reference group *G*, κ_g and Φ_g, to 0 and 1 respectively; allow the corresponding mean and variance in the other groups to be estimated freely.
2. Allow the loading and intercept for item *j* to differ between groups – $\lambda_{jmg} \neq \lambda_{jm}$ and $\nu_{jg} \neq \nu_j$.
3. Set item parameters equal between groups for all other items – i.e., set $\lambda_{hmg} = \lambda_{hm}$, $\nu_{hg} = \nu_h, h \neq j$.

The model defined by these constraints is tested against a baseline model. Within the IRT literature, a likelihood ratio test comparing the difference in log-likelihoods to a χ^2 distribution is typically performed[2]. In the CFA framework with normally distributed items, there are a number of other fit indices to consult, with the CFI, RMSEA, and SRMR suggested as being particularly sensitive to measurement non-invariance (Cheung and Rensvold, 2002; Chen, 2007)[3]. A significant test statistic indicates that a model with item *j*'s item parameters allowed to freely vary across groups is a better fit to the data than a model in which item *j*'s parameters are invariant – i.e., a significant result indicates that item *j* shows DIF. This test is repeated for all items, which necessitates controlling for multiple comparisons; this may be done using the Benjamini-Hochberg procedure (Thissen, Steinberg, and Kuang, 2002).

Some additional strategies, which do not require that item parameters be tested iteratively but instead use a rank-based strategy – i.e., ranking items based on the amount of DIF found in an initial model, and choosing anchor items based on those which have the smallest amount of DIF (e.g., Woods, 2009) – have also shown strong performance in the MIMIC model setting.

The logic of the multiple groups testing procedure above has been partially extended to the testing of measurement invariance within categorical latent variables, although considerably less evidence exists as to which choices yield optimal detection of non-invariance. Collins and Lanza (2010) describe a general procedure for invariance testing within an LCA, which involves first determining that the number and general pattern of item endorsements is similar; this is generally done by conducting separate LCAs by group and using standard fit indices (e.g., BIC, AIC, likelihood ratio test) to determine the optimal configuration of classes within each group. After this initial step, testing generally proceeds as in the continuous latent variable case described above, by successively testing constraints in a multiple-group LCA. There is some

[2] It should be noted, however, that one of the biggest concerns about the IRT-LR-DIF procedure is that the χ^2 will yield biased results if the baseline model is misspecified (Maydeu-Olivares and Cai, 2006).
[3] However, Chen (2007) points out that the SRMR might be more sensitive to differences in factor loadings than intercepts or residual covariance matrices

evidence (Finch, 2015) that, within this general procedure, comparisons to the baseline model should be made using a fit index which makes minimal assumptions, such as the bootstrap likelihood ratio test (BLRT; Nylund, Asparouhov, and Muthén, 2007). However, considerably more research is needed in order to determine best practices for assessing invariance in the multiple-groups LCA setting.

Regression-based approaches for assessing measurement non-invariance rest on the successive significance testing of the corresponding parameters. Within the continuous latent variable framework, strategies for successively testing for impact and DIF are similar to the multiple-group case: estimate some subset of impact parameters κ_{pm} and ϕ_{phm}, as well as some measurement non-invariance parameters ν_{pj} and λ_{pjm}, while constraining others to zero. In the MIMIC model, it has been suggested (Muthén, 1989) that testing proceed in the same order as the IRT-LR-DIF algorithm, starting from a model with no measurement non-invariance and then freely estimating parameters ν_{pj} and λ_{pjm} for each item one by one. However, less research exists on the proper order to in which to test these parameters than in the multiple group case, and there is certainly no consensus on best practices (Finch, 2005; Woods, 2009).

Furthermore, there is no known work on the proper algorithm for testing for measurement non-invariance in the categorical latent variable case. Even setting aside concerns as to the robustness of DIF tests to the misspecification of the baseline model (Maydeu-Olivares and Cai, 2006; Yuan and Bentler, 2004), there are reasons to question whether this robustness will hold for categorical latent variables. In particular, fitting a baseline model which erroneously assumes no DIF on any items other than the one under study may lead to biased estimates of $\pi_k\left(\mathbf{x}_i\right)$; the extent of this bias is not yet known (and will be addressed as part of the current work), but given that misspecification in the within-class structure of a mixture model may often lead to bias in estimates of between-class parameters (Bauer and Curran, 2003; 2004), it is hypothesized that the potential for bias in $\pi_k\left(\mathbf{x}_i\right)$ based on misspecification of DIF is significant. Because tests for DIF in continuous latent variables have been shown to be biased when the distribution of the latent variable is misspecified (Woods, 2008), it stands to reason that bias in the estimated class

probabilities $\pi_k(\mathbf{x}_i)$, which govern the distribution of categorical latent variables, is likely to lead to bias in DIF tests.

In general, virtually all methods for testing for measurement invariance for either type of model face one significant hurdle: because failing to account for differences across $\mathbf{x}_i$ in one parameter leads to bias when assessing the effect of $\mathbf{x}_i$ on another, significant care must be taken in determining whether either latent variable impact or measurement non-invariance is present. This requires consultation of substantive theory before conducting such tests, and also consulting testing procedures which are robust to some degree of model misspecification.

Post-hoc tests for DIF. The multiple-groups and regression-based testing procedures for testing measurement invariance involve the incorporation of either a single grouping variable or multiple covariates in the estimation of the model for the latent variables. However, particularly within the IRT setting, there are a number of testing procedures which indirectly test for DIF outside of model estimation. While many are no longer in use (e.g., Holland and Thayer, 1988), one is discussed here due to its potential for extension to the categorical latent variable case. First, in applications of IRT with a single latent variable η_i, the logistic regression procedure (Swaminathan and Rogers, 1990) is a post-hoc test for both uniform and non-uniform DIF in binary items. After an IRT model presuming no impact or DIF is estimated, estimated values $\hat{\eta}_i$ are obtained and treated as a regressor in a logistic regression equation, which also includes a single covariate x_{ip}, as follows:

$$\text{logit}(\mu_{ij}) = \gamma_0 + \gamma_1\hat{\eta}_i + \gamma_2 x_{ip} + \gamma_3 x_{ip}\hat{\eta}_i \tag{31}$$

The logic of this test is that, should there be no DIF between items based on x_{ip}, there should be no association between y_{ij} and x_{ip}, or the interaction between x_{ip} and $\hat{\eta}_i$, after controlling for $\hat{\eta}_i$. A significant main effect of x_{ip} indicates uniform DIF, whereas a significant interaction effect indicates non-uniform DIF.

The above procedure can be generalized to y_{ij} variables by modeling an arbitrary link function of the expected value of y_{ij}, i.e., $g^{-1}(\mu_{ij})$, rather than a logit. Additionally, more independent variables may be added in, as well as their interactions, to test for DIF by *P*

covariates based on *M* latent variables on a given item. Thus, a more general expression is given by:

$$g^{-1}\left(\mu_{ij}\right) = \boldsymbol{\Gamma}\tilde{\mathbf{x}} \qquad (32)$$

where $\boldsymbol{\Gamma}$ is a $1\times\left(KP+K\right)$ vector of coefficients, and all main effects of $\hat{\boldsymbol{\eta}}_i$ and $\mathbf{x}_i$, along with all two-way interaction terms between $\hat{\boldsymbol{\eta}}_i$ and $\mathbf{x}_i$, are collapsed into the $\left(1+M+P+MP\right)\times 1$ vector $\tilde{\mathbf{x}}$, i.e.,

$$\tilde{\mathbf{x}} = \left[1,\left(\hat{\eta}_{i1},\cdots,\hat{\eta}_{iM}\right),\left(x_{i1},\ldots,x_{iP}\right),\left(\left(\hat{\eta}_{i1}x_{i1},\ldots,\hat{\eta}_{i1}x_{iP}\right),\ldots,\left(\hat{\eta}_{iK-1}x_{i1},\ldots,\hat{\eta}_{iM}x_{iP}\right)\right)\right].$$

The extensibility of this procedure to the categorical latent variable case is immediately apparent, given that estimates of class membership $\hat{\eta}_{ik}$ may be used post-estimation in a regression identical to the one given by Equation 32. Class membership may be estimated by, among other ways, "proportional" or "modal" assignment (Goodman, 1974; Dias and Vermunt, 2008). In proportional assignment, the vector of posterior probabilities of class membership $\boldsymbol{\tau}_i$ is used as the estimate of $\hat{\boldsymbol{\eta}}_i$. In modal assignment, individual *i* is simply assigned to the class for which their posterior probability $\boldsymbol{\tau}_i$ is the highest. Define the *K*×1 vector $\tilde{\boldsymbol{\eta}}_i$ of these assignments, whose individual values $\tilde{\eta}_{ik}$ are coded as 1 for the element corresponding to the class with the highest value of τ_{ik} and 0 for all other classes. Within the LCA framework, Garrett and Zeger (2000) propose a set of post-hoc graphical diagnostics which, while it does not involve fitting a logistic regression model, follows the same general logic by plotting expected and observed endorsement probabilities based on values of covariates, as well as modal assignments $\tilde{\boldsymbol{\eta}}_i$. Under the logic of this procedure, plotted endorsement probabilities conditional on $\tilde{\boldsymbol{\eta}}_i$ should not differ between covariates; such a difference is suggestive of DIF. A full consideration of the post-hoc procedure follows.

An adjusted post-hoc regression testing procedure. The post-hoc regression procedures to test for the effects of $\mathbf{x}_i$ over and above the effect of $\boldsymbol{\eta}_i$ suffer from a number of flaws. First, because $\tilde{\boldsymbol{\eta}}_i$ is an estimate, it is necessary to account for uncertainty in its measurement, which the current procedure does not do, instead treating this modal class estimate as subject *i*'s true class membership. Second, the effects of $\tilde{\boldsymbol{\eta}}_i$ on each item y_{ij} are tested item-wise, using an estimate of

$\boldsymbol{\eta}_i$ from a model without DIF; thus, the test of $\mathbf{x}_i$'s effects on y_{ij} controlling for $\tilde{\boldsymbol{\eta}}_i$ necessarily assume no DIF on other items y_{ih}, $h \neq j$. A number of methods for accounting for measurement error in estimates of $\tilde{\boldsymbol{\eta}}_i$ have been proposed (e.g., Wang et al, 2005; Lanza et al., 2013; Bolck, Croon, and Hagenaars, 2004; Vermunt, 2010; Asparouhov and Muthén, 2014). Here, we will use the method of Bolck, Croon, and Hagenaars (2004) and Vermunt (2010), originally formulated in order to consider simple relationships between modal classifications $\tilde{\boldsymbol{\eta}}_i$ and categorical outcome variables, which was generalized to a wider variety of models by Vermunt (2010) and Bakk, Tekle, and Vermunt (2013). In this strategy, the conditional distribution $P\left(\tilde{\boldsymbol{\eta}}_i \mid \boldsymbol{\eta}_i\right)$ is represented by a $K \times K$ table of probabilities, denoted **D**, with the k, h^{th} element given by:

$$D_{kh} = P\left(\tilde{\eta}_{ik} = 1 \mid \eta_{ih} = 1\right) = \frac{\sum_{i=1}^{N} \tau_{ih} \tilde{\eta}_{ik}}{\pi_h} \tag{33}$$

Diagonal elements of this matrix D_{kk} represent the estimated probability that any given individual is correctly classified into class k - i.e., the probability that their modal class estimate is class k, given that they are truly a member of class k. Likewise, off-diagonal elements D_{kh} (where $k \neq h$) represent the estimated probability that any given individual is misclassified into class k given that they are truly a member of class h. The relative size of diagonal elements D_{kk} relative to off-diagonal elements D_{kh} represents the general accuracy of measurement of $\boldsymbol{\eta}_i$.

Given the matrix **D**, one may fit the regression model in Equation 32 using maximum likelihood with modal estimates of the latent variable $\tilde{\boldsymbol{\eta}}_i$ in place of $\boldsymbol{\eta}_i$, weighting each individual's contribution of the log-likelihood function using the rows of **D** as follows:

$$\log L = \sum_{i=1}^{N} \log \sum_{h=1}^{K} P\left(y_{ij} \mid \tilde{\eta}_{ih} = 1, \mathbf{x}_i; \boldsymbol{\omega}\right) D_{kh} \tag{34}$$

Note that in this method, the loglikelihood for each individual is calculated under each of K classes in Equation 34, resulting in the weighted loglikelihood being maximized on an expanded dataset with $K \times N$ rows. The form of the loglikelihood will differ based on the scale of items y_{ij}, but the general premise of the same regardless: the contribution of individual i to

the overall log-likelihood will be adjusted to account for the extent of the certainty with which they have been classified using modal classification. This method has been found to produce unbiased estimates of regression parameters and their standard errors (Vermunt, 2010), as well as generally greater power than resampling and multiple imputation-based methods for accounting for uncertainty in the measurement of $\boldsymbol{\eta}_i$.

While this method mitigates the first concern about the post-hoc regression test – error in the measurement of $\boldsymbol{\eta}_i$ – the problem of item-wise testing, which is that estimates $\tilde{\boldsymbol{\eta}}_i$ are free of DIF from other items when testing the effect of $\boldsymbol{\eta}_i$ on y_{ij}, remains. There are results to suggest that, in the absence of extreme non-compensatory DIF across multiple items (i.e., DIF in the same direction on all items based on a given covariate), item-wise tests are generally unbiased in the model-based (i.e., IRT-LR-DIF) testing framework even if there is some DIF from $\mathbf{x}_i$ on other items (Cohen, Kim, and Wollack, 1996; Bolt, 2002; Kim and Cohen, 1998). However, it is unclear to what extent this unbiasedness holds when estimates $\tilde{\boldsymbol{\eta}}_i$ are used post-hoc. Thus, the performance of this test on the basis of the number of items with DIF and the magnitude of this DIF will be investigated in the proposed work.

Summary and research questions

There is currently a paucity of evidence as to the extent of mixture models' robustness to measurement non-invariance or differential item functioning (DIF), as well as which strategies for testing DIF generalize most successfully from the continuous to the categorical latent variable case. Additionally, the extent to which these results make a practical difference in the conclusions drawn from mixture model results is not yet known; it may be the case that some of the discrepancies between findings from various applications of mixture models may be reconciled by accounting for differences in measurement across studies.

As such, the current work systematically investigated measurement invariance in the categorical latent variable framework through two studies, which are described in Chapters 2 and 3.

Chapter 2

Chapter 2 consists of a Monte Carlo simulation study of DIF in mixture models, in which uniform and nonuniform DIF were simulated under a number of different conditions. Monte Carlo simulation was pursued to answer Questions A and B because, due to the use of the expectation-maximization (EM; Dempster, Laird, and Rubin, 1977) algorithm to estimate mixture model parameters, an analytic solution would be intractable. In particular, Chapter 2 investigated two questions:

Question A: What are the consequences of ignoring DIF in mixture models? In particular, what is the effect of unmodeled DIF on estimated values of model parameters, as well as individual-level quantities such as class membership estimates?

Question B: What is the ideal way to test for DIF in the mixture model setting? In particular, how sensitive are post-hoc tests to the presence of DIF relative to strategies which model DIF directly?

Both of the above two questions were addressed using the same data, which was generated from a two-class latent class analysis with varying types of DIF. The bias associated with omitted DIF effects (Question A), as well as the performance of DIF tests (Question B), were examined across four factors: class prevalence; class separation; DIF magnitude; and DIF type. The number of cases and items, as well as the overall pattern of DIF, were all held constant.

Chapter 3

Chapter 3 is a secondary data analysis of real data derived from a unique laboratory study of the measurement of alcohol use disorder (AUD) in a college sample. Specifically, potential DIF according to demographic covariates, as well as prior exposure to AUD items, was investigated in a computerized diagnostic battery for AUD. The goals of this chapter are twofold. First, whereas Chapter 2 systematically investigates the effects of omitted DIF and the performance of DIF tests in simulated data, Chapter 3 provides an example of how DIF testing and modeling may be done in real data and the sorts of questions which may be answered by investigating DIF in LCA.

The second goal of Chapter 3 is to use DIF analyses to determine whether measurement bias on the basis of background variables may be partially to blame for discrepancies in mixture model findings across different studies of AUD symptoms. Numerous studies (e.g., Lynskey et al., 2005; Beseler et al., 2012; Jackson et al., 2014, Rinker and Neighbors, 2015) have been conducted with the aim of finding latent classes of AUD symptoms, and many have found different numbers and configurations of classes. The reason for these discrepancies are unknown, but recent research (e.g., Cole, Bauer, Hussong, and Giordano, 2017) suggests that LCA results may be highly sensitive to minor changes in measurement. Strong findings of DIF on the basis of background variables would suggest that the measurement properties of tests for AUD may indeed by inconsistent across populations and testing situations.

Chapters 2 and 3 are motivated by the complementary goals described above, and there is much overlap between the studies in terms of the nature of the questions they address. As mentioned earlier, the goals of applying a mixture model may be explanatory, descriptive, or predictive. The goals of the simulation study in Chapter 2 are explicitly explanatory: by simulating a mixture model with DIF and either omitting this DIF (Question A) or using testing procedures to locate DIF empirically (Question B), we aim to create conditions which allow for causal statements about the effects of DIF in mixtures to be made. Moreover, the simulation study in Chapter 2 seeks to assess mixture models' explanatory capacity, by determining how frequently mixture modeling procedures arrive at the correct number and configuration of classes and DIF effects in the presence of DIF. However, as mixture models are so frequently applied descriptively or predictively (Nagin, 1999), both Chapters 2 and 3 will aim to also assess how well mixture models with DIF approximate the data at both the aggregate and individual levels. In particular, the simulation study in Chapter 2 assesses the accuracy of predicted endorsement probabilities, and Chapter 3 assesses how predicted endorsement probabilities change based on the extent to which DIF is included in the fitted model.

CHAPTER 2
A MONTE CARLO SIMULATION OF OMITTED DIF IN MIXTURE MODELS

Chapter 2 systematically investigated the following two questions:

Question A: What are the consequences of ignoring DIF in mixture models? In particular, what is the effect of unmodeled DIF on estimated values of model parameters, as well as individual-level quantities such as class membership estimates?

Question B: What is the ideal way to test for DIF in the mixture model setting? In particular, how sensitive are post-hoc tests to the presence of DIF relative to strategies which model DIF directly?

Both of the above two questions were addressed using the same data, which was simulated from a two-class LCA. The bias associated with omitted DIF effects (Question A), as well as the performance of DIF tests (Question B), were examined across four factors: class prevalence (2 levels: equal or unequal); class separation (2 levels: small or large); DIF magnitude (2 levels: small or large); and DIF type (3 levels: intercept, loading, or both intercept and loading). A summary of the models fitted in both Questions A and B are shown in Figure 1. The procedures by which Questions A and B were addressed, as well as the results of these two separate but related inquiries, are now discussed.

Question A

To address the question of bias when DIF is not accounted for in latent class analysis (LCA), a series of misspecified models were fit to data generated from a two-class LCA with DIF. First, in order to determine whether unmodeled DIF may impact class enumeration, models with $K = 1$ to $K = 4$ classes were fit to the data both in models with and without covariates effects on class membership. Second, even when the correct number of classes is selected, unmodeled DIF may have effects on other quantities of interest, including class prevalence rates, covariate

effects, endorsement probabilities, and individual class membership assignments. Thus, bias in these quantities was assessed in a model with K=2 with only covariate effects on class membership (the impact-only model), only covariate effects on item intercepts (the intercept-DIF model), and the fully specified model, which contains covariate effects on item intercepts and loadings (the intercept-and-loading DIF model).

Hypotheses

Class enumeration. In studies which have examined the sensitivity of mixture models to changes in measurement in empirical data (Jackson and Sher, 2005; Cole, Bauer, Hussong, and Giordano, 2017), class enumeration was generally stable even across drastically different measurement conditions. Thus, it was predicted that the correct number of classes (K = 2) would be favored by fit indices in most experimental conditions when covariates were not included in the model. However, it was predicted that problems in class enumeration might arise when covariates with unmodeled DIF were included as predictors of class membership. This follows the findings of Kim et al. (2016) and Nylund-Gibson and Masyn (2016) that, in the presence of omitted DIF effects, models with covariate effects on class membership tend to overextract classes. Thus, it may be the case that spurious classes are estimated which merely capture DIF effects.

Parameter bias. It was predicted that estimated covariate effects on class membership, as well as estimated item parameters, would absorb omitted DIF effects. For example, given a class with higher levels of item endorsements (a "high-endorsement" class), and given a covariate which increases the overall item endorsement probability (i.e., intercept DIF), failure to model this DIF may produce an upwardly biased estimate of the covariate's effect on membership to the high-endorsement class. This prediction is based on the results of Chen (2008) in the multiple group factor analysis context, omitted DIF effects from a grouping variable to an item often manifested as spurious group differences in the underlying latent factor.

Individual class assignments. It was predicted that individual class assignments would be less biased than parameter estimates, even in the case where DIF was omitted entirely. This

follows from the finding of Curran et al. (2016), who report in the continuous latent variable case that factor scores are often relatively unbiased in the event that DIF is omitted, as long as all relevant mean effects of covariates on the underlying latent factor were included. As such, it was predicted here that, as long as all relevant effects of covariates on latent classes were included, misclassification would generally be low.

Data-generating model

Data were generated from a two-class latent class analysis model. Data generation followed a three-step process. First, data were generated on covariates. Then, the binary latent class variable was generated as a function of these covariates. Finally, data were generated for binary items, conditional on both covariates and latent class membership. Each step will be described in turn.

Class membership and item endorsement are affected by a subject-specific vector of four covariates $\mathbf{x}_i = \left(x_{i1}, x_{i2}, x_{i3}, x_{i4}\right)$, where *p* indexes covariates $\left(p = 1, \ldots, P\right)$ and *i* indexes individuals $\left(i = 1, \ldots, N\right)$. Of these, x_{i1}, x_{i2}, and x_{i3} affect class membership; x_{i1}, x_{i2}, and x_{i4} generate DIF. As such, x_{i3} has impact but no DIF, whereas x_{i4} has DIF but no impact. These covariates are characterized by a multivariate normal distribution, with weak correlations between all variables:

$$\begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \end{bmatrix} \sim N_4 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .3 & .3 & .3 \\ .3 & 1 & .3 & .3 \\ .3 & .3 & 1 & .3 \\ .3 & .3 & .3 & 1 \end{bmatrix} \right). \tag{35}$$

Define a 2×1 vector of latent class indicator variables $\boldsymbol{\eta}_i$, whose individual elements η_{ik} take on a value of 1 if subject i is a member of class k and 0 otherwise. The probability of membership to class 1, conditional on all covariates $\mathbf{x}_i$, $\pi_{i1}\left(\mathbf{x}_i\right)$, is affected by covariates through a typical logistic regression equation, as specified more generally in Equation 19:

$$\pi_{i1}\left(\mathbf{x}_i\right) = P\left(\eta_{i1} = 1 \mid \mathbf{x}_i\right) = \frac{1}{1 + \exp\left(-\left(\alpha_{01} + \alpha_{11} x_{i1} + \alpha_{21} x_{i2} + \alpha_{31} x_{i3}\right)\right)} \tag{36}$$

Class membership was coded such that any increase in the probability of membership to Class 1 leads to a corresponding decrease in the probability of membership to Class 2. Thus $\alpha_{p2} = -\alpha_{p1}$ for all *p*.

For each case, *J* = 10 binary items, denoted y_{ij} $(i = 1,\ldots,N)$, were generated. The probability of endorsing item *j*, conditional on class membership and covariates $\mathbf{x}_i$, is denoted $\mu_{ij1}(\mathbf{x}_i)$ and $\mu_{ij2}(\mathbf{x}_i)$ for Classes 1 and 2 respectively. These probabilities are given by class-specific logistic regressions, as expressed more generally in Equation 20:

$$\begin{aligned}\mu_{ij1}(\mathbf{x}_i) &= \frac{1}{1+\exp\left(-\left[\delta_{0j1}+\delta_{1j1}x_{i1}+\delta_{2j1}x_{i2}+\delta_{4j1}x_{i4}\right]\right)}\\ \mu_{ij2}(\mathbf{x}_i) &= \frac{1}{1+\exp\left(-\left[\delta_{0j2}+\delta_{1j1}x_{i1}+\delta_{2j1}x_{i2}+\delta_{4j1}x_{i4}\right]\right)}\end{aligned} \tag{37}$$

where δ_{0jk} represents the log-odds of endorsing item *j* within class *k* when all covariates are zero and δ_{pjk} transmits the effect of covariate x_{ip} on this log-odds.

Recall from Equations 23-30 that, although this is the original and most common parameterization of direct effects in LCA, the study of DIF is facilitated by decomposing each logit parameter δ_{pjk} into an intercept parameter ν_{pj} and class-specific loadings λ_{pjk}. As explained in Equations 25-26, the model is parameterized such that ν_{pj} is an unweighted mean of δ_{pjk} across classes, and λ_{pjk} represents Class *k*'s deviation in δ_{pjk} from this unweighted mean. Thus, any increase in δ_{pjk} in Class 1 must produce a corresponding decrease in Class 2. Therefore, logits $\delta_{qj1} = \nu_{qj} + \lambda_{qj1}$ and $\delta_{qj2} = \nu_{qj} - \lambda_{qj1}$ for all *q*, and class-specific endorsement probabilities may be written as:

$$\begin{aligned}\mu_{ij1}(\mathbf{x}_i) &= \frac{1}{1+\exp\left(-\left(\left[\nu_{0j}+\nu_{1j}x_{i1}+\nu_{2j}x_{i2}+\nu_{4j}x_{i4}\right]+\left[\lambda_{0j1}+\lambda_{1j1}x_{i1}+\lambda_{2j1}x_{i2}+\lambda_{4j1}x_{i4}\right]\right)\right)}\\ \mu_{ij2}(\mathbf{x}_i) &= \frac{1}{1+\exp\left(-\left(\left[\nu_{0j}+\nu_{1j}x_{i1}+\nu_{2j}x_{i2}+\nu_{4j}x_{i4}\right]-\left[\lambda_{0j1}+\lambda_{1j1}x_{i1}+\lambda_{2j1}x_{i2}+\lambda_{4j1}x_{i4}\right]\right)\right)}\end{aligned} \tag{38}$$

Though data were generated by manipulating values of item parameters ν_{pj} and λ_{pjk}, a number of outcomes (e.g., class-specific covariate effects) are considered in terms of traditional logits δ_{pjk}. In the presence of exclusively intercept DIF, $\delta_{pj1} = \nu_{pj}$ for $k = 1$ and $k = 2$. In the presence of exclusively loading DIF, $\delta_{pj1} = \lambda_{pj1}$ and $\delta_{pj2} = -\lambda_{pj1}$. For a model containing both intercept and loading DIF, values of item parameters ν_{pj} and λ_{pjk} in Table 3 are translated to the corresponding class-specific logits δ_{pjk} in Table 4 for comparison.

The values of parameters in Equations 36-38 define the differences between experimental conditions. Details of how parameter values were chosen are given below.

Parameters held constant across cells. A number of factors are held constant across all cells. These include the magnitude of covariate effects on class membership, number of items, baseline item parameters, overall pattern of DIF, and overall sample size.

Number of items. The number of items J is held constant at $J = 10$. This value was chosen on the basis of yielding adequate power to detect the optimal number of classes in the absence of DIF in LCA (Nylund, Asparouhov, and Muthén, 2007).

Magnitude of covariate effects. The magnitude of α_{11}, α_{21} and α_{31} is held constant at values of 0.7, 0.7, and 0.7 so that a one-unit increase in covariates x_{i1}, x_{i2}, and x_{i3} is associated with a roughly 2-fold increase in the odds of membership to class 1, relative to class 2. These effect sizes are consistent with those used in previous simulation work (Asparouhov and Muthén, 2014; Vermunt, 2010), and are expected to produce meaningful differences in class membership without driving class membership probabilities to zero or one.

Number and pattern of DIF items. Each variable affects 2/10 items. As described below (under "baseline class separation"), there was some heterogeneity in magnitude among λ_{0j1} coefficients; thus, DIF effects were spread evenly across items.

The values of the DIF coefficients are shown in Table 3. For covariates x_{i1} and x_{i4}, both λ_{pj1} or ν_{pj} effects are positive. For covariate x_{i2}, both λ_{pj1} or ν_{pj} effects are negative. Thus, covariates x_{i1} and x_{i4} only ever increase the difference between classes in their endorsement probabilities (through their effect on loadings) and only ever increase the overall item

endorsement probability (through their effect on intercepts). Covariate x_{i2} only ever decreases the difference between classes in their endorsement probabilities (through its effect on loadings) and only ever decreases the overall item endorsement probability (through its effect on intercepts). This is done in order to create the realistic, compensatory pattern of measurement non-invariance often seen in practice, in which measurement non-invariance effects in opposite directions may cancel one another out.

Total sample size. Total sample size was held constant across cells at *N=500*, a sample size consistent with generally good power in mixture models (Nylund, Asparouhov, and Muthén, 2007).

Between-cell factors. Four factors are manipulated: overall class membership probabilities (2 levels); baseline class separation (2 levels); magnitude of DIF (2 levels); and type of DIF (3 levels). This yields 24 unique cells. For each cell, *R* = 500 replications are evaluated.

Baseline class membership probability. Class membership probability is manipulated between-cells with 2 levels: either classes are of equal size, or classes 1 and 2 contain approximately 80% and 20% of the sample respectively. Given covariate effects, these prevalences correspond to $\alpha_{01} = 0$ and $\alpha_{01} = -1.8$.

Baseline class separation. Class separation is manipulated between cells, with 2 levels: small (average value of $\lambda_{0j1} = .75$ across items); and large (average value of $\lambda_{0j1} = 1.2$ across items). All baseline values of item parameters are listed in Table 3. Note that all values of ν_{0j} are held to zero.

After generating baseline values of λ_{0j1}, the corresponding values of entropy were determined by simulating 1000 datasets, each with *N* = 500 cases; this was done for either equally-sized classes (i.e., $\alpha_{01} = 0$) or unequally-sized classes (i.e., $\alpha_{01} = -1.8$), without any covariate effects on class membership or item endorsement. For each of these cases, the average value of entropy was calculated. For equally-sized classes, entropy was .66 and .91 in the low and high class separation, respectively; for unequally-sized classes, entropy was .78 and .94 for low and high class separation, respectively. Entropy is necessarily lower for equally-sized classes

than unequally-sized classes (Celeux and Soromenho, 1996), and thus the differences between the cases of equally and unequally sized classes were expected.

Type of DIF. The type of DIF is manipulated between-cells with 3 levels: DIF due exclusively to differences in intercepts (the intercept-DIF generating model), DIF due exclusively to differences in loadings (the loading-DIF generating model), and DIF due to differences in both (the intercept-and-loading DIF generating model). Given the findings from the continuous latent variable literature that loading DIF may in some cases be more difficult to detect (e.g., Stark, Chernyshenko, and Drasgow, 2006), and may exert a strong biasing influence on impact parameters when it is not included in the model (e.g., Chen, 2008), it was of interest to consider these three instances separately.

Magnitude of DIF. The absolute value of DIF parameters ν_{pj} and λ_{pj1} is manipulated between cells, with 2 levels: small ($\nu_{pj} = 0.8$ or $\lambda_{pj1} = 0.4$) and large ($\nu_{pj} = 1.6$ or $\lambda_{pj1} = 0.8$). Importantly, the covariates x_{i1}, x_{i2}, and x_{i4} are coded so that the values $\left(\nu_{1j}, \lambda_{1j1}\right)$, $\left(\nu_{2j}, \lambda_{2j1}\right)$, and $\left(\nu_{4j}, \lambda_{4j1}\right)$ represent the difference in measurement parameter values between individuals at ± 1 SD on x_{i1}, x_{i2}, and x_{i4} controlling for the influence of all other covariates. Table 3 shows these parameters, as well as the items they affect.

Model fitting

All mixture models were estimated using the accelerated expectation-maximization (EM; Dempster, Laird, and Rubin, 1977) algorithm as implemented in Mplus version 7.2 (Muthén and Muthén, 2014). This implementation uses multiple random seed values (here 100 were used) and a user-defined number of maximum iterations for a given random seed value (here 1000 were used).

Data generation and management was completed using R. Additionally, all of the process flow involving the automation of model-based DIF testing in Mplus was completed using the MplusAutomation() (Hallquist and Wiley, 2011) package in R, which allows for automatic writing, execution, and reading of code to be evaluated by Mplus.

Class enumeration. Both of the models in the left-hand side of the top panel of Figure 1 were first fit iteratively to the data for increasing numbers of classes from $K = 1$ to $K = 4$. This included the *unconditional model*, which assumes no covariate effects on either the latent variable or the items as well as the *impact-only model*, which assumes covariate effects on the latent class variable but not the items.

In all of these models, the values of a number of fit statistics and likelihood ratio test results (Akaike, 1973; Schwarz, 1978; Vuong, 1989; Lo, Mendell, and Rubin, 2001; McLachlan and Peel, 2000), further described below, were recorded. Regardless of the values of K favored by this set of fit statistics, subsequent models were fit on the correct number of classes, i.e., $K = 2$. This strategy permitted the problem of correct class enumeration in the presence of measurement invariance to be disentangled from the question of whether DIF can be accurately identified when the correct number of classes is in fact chosen or known.8

It is well-known that decisions about the number of classes are sensitive to the inclusion of all relevant covariates (Tofighi and Enders, 2008). In particular, when DIF-generating covariates are included exclusively as covariates affecting class membership (as in the impact-only model), recent results (Kim et al., 2016; Nylund-Gibson and Masyn, 2016) suggest that spurious classes may be detected. Therefore, it was important to establish the frequency with which either the unconditional or impact-only models – which researchers are likely to fit as a first step – failed to identify the correct number of classes, $K = 2$, relative to the correctly-specified nonuniform DIF model. If a researcher were to proceed with DIF testing in a model with the incorrect number of classes, a potentially nonsensical pattern of DIF effects might emerge.

Bias in parameters and individual-level estimates at K=2. In addition to the two models described above, for the correct number of classes $K = 2$, models 2, 3, and 4 in Figure 1 were also fit to the data. This included: the impact-only model, which assumes covariate effects on the latent class variable but not the items (model 2 in Figure 1); the *uniform DIF fitted model*, which includes DIF on all DIF-containing items (i.e., all items which truly have DIF on $\nu_j(\mathbf{x}_i), \lambda_{j1}(\mathbf{x}_i)$,

or both) but only on the intercept parameters (model 3 in Figure 1);[4] and the *non-uniform DIF fitted model*, which is the full model containing both intercept and loading DIF (model 4 in Figure 1). As it was well-established that item endorsement probabilities would be severely biased in an unconditional LCA (i.e., model 1 in Figure 1), the unconditional model was not considered for this aspect of question A.

It is of interest to establish whether the inclusion of any covariates, even those whose effects have been misspecified (as they are in the impact-only model for all cells, and in the uniform DIF fitted model for all cells with loading DIF), may absorb omitted DIF effects to produce relatively unbiased estimates of model parameters and scores. Curran et al. (2016) have found some estimates of scores to be highly accurate in the analogous case (i.e., impact included but DIF omitted) in continuous latent variable models. The uniform DIF fitted model was included in order to determine whether simply including intercept DIF would absorb the biasing effects of covariates $\mathbf{x}_i$ on both intercepts and loadings; this is particularly important given that the model with intercept DIF only is easier to identify (Huang and Bandeen-Roche, 2004), and that researchers might favor this model, despite its being misspecified, for ease of estimation.

Outcomes of interest

Outcomes of interest whose quality may be affected by omission of DIF effects included the number of classes selected, as well as a number of quantities assessed at the correct number of classes. These quantities are divided into three general types. First, we assessed the accuracy of model parameter estimates, including covariate effects on class membership and covariate effects on endorsement probabilities. Second, we assessed the accuracy of quantities pertaining to class membership, including the prevalence of Class 1, as well as the classification accuracy of individuals. Third, we assessed the accuracy of average item endorsement probabilities within

[4] Note that, whereas this model is misspecified for the loading DIF and loading-and-intercept DIF condition, it is the true model for the intercept DIF condition. Thus, while it will be referred to as a misspecified model from here forward, this is only the case for 2/3 DIF type conditions.

a given class, which may be calculated a number of ways. These quantities, as well as how they are assessed, are described below.

Class enumeration. Solutions with K=1 to K=4 classes were obtained for all replications in all cells. For each cell, the values of several widely-used fit indices were recorded under each value of K. The Bayesian Information Criterion (BIC; Schwarz, 1978) and Akaike Information Criterion (AIC; Akaike, 1973) two information criteria which weigh the loglikelihood of a given model against the number of parameters, balancing fit and parsimony. Lower values of BIC and AIC represent better correspondence between model and data. Additionally, two likelihood ratio tests, the Lo-Mendell-Rubin likelihood ratio test (LMR; Vuong, 1989; Lo, Mendell, Rubin, 2001), and the bootstrap likelihood ratio test (BLRT; McLachlan and Peel, 2000) were consulted. These two tests compare a model with K classes to one with $K - 1$ classes, and differ in how they approximate the distribution of the test statistic. The LMR and BLRT are applied to models with increasing numbers of classes, and the chosen solution is the one with the highest value of K for which the fit of a $K - 1$-class model is significantly worse than that of a K-class model.

Model parameters. The estimated values of all model parameters under $K = 2$ were recorded in all cells under all data-generating models; this includes class membership parameters, baseline endorsement probabilities for each class, and covariate effects on endorsement probabilities. Standardized bias (SB) was computed comparing all estimated parameters to their true values:

$$SB = \frac{\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r - \theta}{SE\left(\hat{\theta}_r\right)} \tag{39}$$

Here, $\hat{\theta}_r$ is the estimate of some quantity θ under replication r, where R is the total number of replications, and $SE\left(\hat{\theta}_r\right)$ is calculated as the standard deviation of all parameter estimates. Standardized bias puts parameter bias in the unit of standard errors for a given parameter. It is preferable to relative bias because it may be used to calculate bias in estimates where true values are 0. Collins, Schaefer, and Kim (2001) note that standardized bias values of

40%, corresponding to a difference in 40% of a standard error between true and estimated values, are ever enough to be practically significant.

For parameters governing class membership, α_{p1}, intercept parameter $\hat{\alpha}_{01}$ was compared to a true value of $\alpha_{01} = 0$ for equally-sized classes and $\alpha_{01} = -1.8$ for unequally-sized classes; covariate effects were compared to true values of $\alpha_{p1} = 0.7$ for p = 1, 2, and 3 and $\alpha_{p1} = 0$ for p = 4.

Finally, for the intercept-DIF and intercept-and-loading-DIF models, standardized bias was assessed in logit parameters transmitting covariate DIF effects on items. Rather than considering bias in $\hat{\nu}_{pj}$ and $\hat{\lambda}_{pj1}$, which were not estimated for every model (i.e., $\hat{\lambda}_{pj1}$ was not estimated in the uniform DIF model), we considered standardized bias in $\hat{\delta}_{1jk}$, $\hat{\delta}_{2jk}$, and $\hat{\delta}_{4jk}$, as these are the parameters governing item endorsement within each class that researchers may obtain without reparameterizing the model.

Class membership. The whole-sample prevalence of classes may be calculated a number of ways, but arguably the most common is to take the average of posterior probabilities across cases. Denote the set of estimated parameters for a given replication $\hat{\boldsymbol{\Theta}}$, and the true model parameters $\boldsymbol{\Theta}$. Posterior probabilities of class membership were calculated under the estimated model and true models, using $\hat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}$ respectively:

$$\begin{aligned}\hat{\tau}_{i1} &= P\left(\eta_{i1}=1 \mid \mathbf{x}_i, \mathbf{y}_i; \hat{\boldsymbol{\Theta}}\right) \\ &= \frac{P\left(\mathbf{y}_i \mid \eta_{i1}=1, \mathbf{x}_i; \hat{\boldsymbol{\Theta}}\right) P\left(\eta_{i1}=1 \mid \mathbf{x}_i; \hat{\boldsymbol{\Theta}}\right)}{P\left(\mathbf{y}_i \mid \eta_{i1}=1, \mathbf{x}_i; \hat{\boldsymbol{\Theta}}\right) P\left(\eta_{i1}=1 \mid \mathbf{x}_i; \hat{\boldsymbol{\Theta}}\right) + P\left(\mathbf{y}_i \mid \eta_{i2}=1, \mathbf{x}_i; \hat{\boldsymbol{\Theta}}\right) P\left(\eta_{i2}=1 \mid \mathbf{x}_i; \hat{\boldsymbol{\Theta}}\right)}\end{aligned} \tag{40}$$

The posterior probability of membership to class 1 under true model parameters, τ_{i1}, is given by:

$$\begin{aligned}\tau_{i1} &= P\left(\eta_{i1}=1 \mid \mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\Theta}\right) \\ &= \frac{P\left(\mathbf{y}_i \mid \eta_{i1}=1, \mathbf{x}_i; \boldsymbol{\Theta}\right) P\left(\eta_{i1}=1 \mid \mathbf{x}_i; \boldsymbol{\Theta}\right)}{P\left(\mathbf{y}_i \mid \eta_{i1}=1, \mathbf{x}_i; \boldsymbol{\Theta}\right) P\left(\eta_{i1}=1 \mid \mathbf{x}_i; \boldsymbol{\Theta}\right) + P\left(\mathbf{y}_i \mid \eta_{i2}=1, \mathbf{x}_i; \boldsymbol{\Theta}\right) P\left(\eta_{i2}=1 \mid \mathbf{x}_i; \boldsymbol{\Theta}\right)}\end{aligned} \tag{41}$$

Then, the prevalence of class 1 in each replication, denoted $\hat{\Pr}$ under estimated model parameters and $\Pr$ under true model parameters is calculated as:

$$\hat{\Pr} = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_{i1} \tag{42}$$

for estimated model parameters, and

$$\Pr = \frac{1}{N}\sum_{i=1}^{N}\tau_{i1} \tag{43}$$

for true model parameters, where $\hat{\tau}_{i1}$ and τ_{i1} obtained from Equations 40 and 41, respectively. The distributions of $\hat{\Pr}$ were then compared to true values $\Pr$.

In addition to overall prevalence of Class 1, the accuracy of individual classifications into latent classes was also assessed. For each individual, modal class assignments, $\tilde{\eta}_{ik}$, were generated for each subject by choosing the class to which subject *i* has the highest value of $\hat{\tau}_{ik}$ following Equation 40. Because true class memberships η_{ik} were retained in the data-generating process, binary modal classifications $\tilde{\eta}_{ik}$ were compared to true class memberships η_{ik} using the Adjusted Rand Index (ARI; Hubert and Arabie, 1985), which is an index measuring the concordance between two partitions, adjusting for chance. Given the $A \times B$ contingency table with rows indexed by a $(a = 1, \ldots, A)$ and columns indexed by b $(b = 1, \ldots, B)$, the cell count n_{ab} represents the number of subjects classified into category *a* by one classification and category *b* by the other. Then the ARI is computed as follows:

$$ARI = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\binom{n_{ab}}{2} - \frac{\sum_{a=1}^{A}\binom{n_{a.}}{2}\sum_{b=1}^{B}\binom{n_{.b}}{2}}{\binom{n}{2}}}{\frac{\sum_{a=1}^{A}\binom{n_{a.}}{2}\sum_{b=1}^{B}\binom{n_{.b}}{2}}{2} - \frac{\sum_{a=1}^{A}\binom{n_{a.}}{2}\sum_{b=1}^{B}\binom{n_{.b}}{2}}{\binom{n}{2}}} \tag{44}$$

The ARI ranges from -1 to 1, with values closer to 1 indicating greater agreement between the two classifications. Values of ARI are averaged across cells, yielding a total of 24 average ARI values.[5]

[5] In addition to modal classifications, the accuracy of posterior probabilities was assessed using an extension of the ARI which uses cosine similarity to assess the agreement between two fuzzy partitions (Brouwer, 2009). However,

Class-specific endorsement probabilities. DIF in mixture models presents researchers with a challenging problem from an interpretational standpoint: because individual covariates affect the expected values of indicators, class-specific expected values are not dependent exclusively on class membership. This complicates the traditional practice of showing and interpreting differences between classes in expected values of class membership.

There are (at least) two potential ways researchers may choose to summarize the expected values of indicators within a given class, and we assessed bias in both of these quantities. The first, which we denote the *baseline endorsement probability*, represents the probability of endorsing item y_{ij} when all covariates are zero. It is calculated as a function of item parameter intercepts, as

$$\begin{aligned}\hat{\mu}_{0j1} &= \frac{1}{1+\exp\left(-\left[\hat{\nu}_{0j}+\hat{\lambda}_{0j1}\right]\right)} \\ \hat{\mu}_{0j2} &= \frac{1}{1+\exp\left(-\left[\hat{\nu}_{0j}-\hat{\lambda}_{0j1}\right]\right)}\end{aligned} \tag{45}$$

for estimated parameters, and

$$\begin{aligned}\mu_{0j1} &= \frac{1}{1+\exp\left(-\left[\nu_{0j}+\lambda_{0j1}\right]\right)} \\ \mu_{0j2} &= \frac{1}{1+\exp\left(-\left[\nu_{0j}-\lambda_{0j1}\right]\right)}\end{aligned} \tag{46}$$

for true parameters. Standardized bias was assessed comparing endorsement probabilities generated by $\hat{\nu}_{0j}$ and $\hat{\lambda}_{0j1}$ were compared to those generated by true values ν_{0j} and λ_{0j1}.

Alternatively, one may calculate the expected value of individual conditional endorsement probabilities across all values of covariates. These values, here denoted *marginal endorsement probabilities*, are obtained as sums of class-specific, individual-specific endorsement probabilities, $\hat{\mu}_{ij1}$ and $\hat{\mu}_{ij2}$, weighted by class membership probabilities. As with

---

no meaningful differences were found between these results and the ARI examining hard partitions, and so the simpler ARI results are presented.

posterior probabilities of class membership, predicted endorsement probabilities were calculated under the estimated model and true models, using $\hat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}$ respectively. For true parameter values $\boldsymbol{\Theta}$, values of μ_{ijk} are calculated using Equation 38. For estimated parameter values $\hat{\boldsymbol{\Theta}}$, predicted endorsement probabilities under each of the two classes are calculated as:

$$\begin{aligned}\hat{\mu}_{ij1} &= P\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1; \hat{\boldsymbol{\Theta}}\right) \\ &= \frac{1}{1+\exp\left(\left[\hat{\nu}_{0j} + \hat{\nu}_{1j}x_{i1} + \hat{\nu}_{2j}x_{i2} + \hat{\nu}_{4j}x_{i4}\right] + \left[\hat{\lambda}_{0j1} + \hat{\lambda}_{1j1}x_{i1} + \hat{\lambda}_{2j1}x_{i2} + \hat{\lambda}_{4j1}x_{i4}\right]\right)} \\ \hat{\mu}_{ij2} &= P\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1; \hat{\boldsymbol{\Theta}}\right) \\ &= \frac{1}{1+\exp\left(\left[\hat{\nu}_{0j} + \hat{\nu}_{1j}x_{i1} + \hat{\nu}_{2j}x_{i2} + \hat{\nu}_{4j}x_{i4}\right] - \left[\hat{\lambda}_{0j1} + \hat{\lambda}_{1j1}x_{i1} + \hat{\lambda}_{2j1}x_{i2} + \hat{\lambda}_{4j1}x_{i4}\right]\right)}\end{aligned} \tag{47}$$

Marginal endorsement probabilities $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{ik} = 1\right)$ are then calculated as averages of these individual predicted values, weighted by individual posterior probabilities of class membership, as

$$E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1; \hat{\boldsymbol{\Theta}}\right) = \sum_{i=1}^{N} \hat{\tau}_{i1}\hat{\mu}_{ij1} \qquad E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1; \hat{\boldsymbol{\Theta}}\right) = \sum_{i=1}^{N} \hat{\tau}_{i2}\hat{\mu}_{ij2} \tag{48}$$

for estimated model parameters, and

$$E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1; \boldsymbol{\Theta}\right) = \sum_{i=1}^{N} \tau_{i1}\mu_{ij1} \qquad E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1; \boldsymbol{\Theta}\right) = \sum_{i=1}^{N} \tau_{i2}\mu_{ij2} \tag{49}$$

for true model parameters. Because of the fact that there are direct effects of covariates on items over and above class membership, the predicted endorsement probabilities obtained by baseline item parameters in Equation 45 may be biased, even when within-class endorsement probabilities are not. Relative bias in average posterior probability, comparing cross-replication averages of $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{ik} = 1; \hat{\boldsymbol{\Theta}}\right)$ to $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{ik} = 1; \boldsymbol{\Theta}\right)$, was then calculated.

Note that, in the impact-only model and in all indicators with no DIF, the probability of endorsing y_{ij} does not depend on covariates; therefore $\mu_{ijk} = \mu_{0jk}$ under the impact-only model.

Results

Class enumeration. Rates at which all fit indices chose the correct 2-class solution, as well as the average number of classes, are shown in Tables 5 and 6 for unconditional and conditional models respectively. In unconditional models, the BIC almost always correctly chose the 2-class model, and only very rarely models with more than three classes; only in the presence of large DIF and poorly-separated, unequally-sized classes, did the BIC select the 2-class solution less than 95% of the time. The AIC selected the 2-class solution less than 50% of the time in all cases and typically overestimated the number of classes. The performance of the two likelihood ratio tests, the Vuong Lo-Mendell-Rubin test (LMR) and the bootstrap likelihood ratio test (BLRT), was considerably more varied. For both these likelihood ratio tests, the tests' ability to choose the correct number of classes was most degraded when DIF was severe, rather than when classes were poorly separated. For both the LMR and the BLRT, the presence of intercept DIF (i.e., in the intercept-DIF and intercept-and-loading DIF generating models) was associated with lower levels of the correct 2-class solution being chosen. This effect was considerably larger in the BLRT, which performed very well in the presence of only loading DIF but selected the correct solution under 40% of the time when large intercept DIF was present.

In conditional models, rates of correctly choosing the 2-class solution were generally quite low for all fit indices. Strong differences according to DIF type emerged with regard to the BIC, which chose the correct number of classes almost all of the time when exclusively loading DIF was present, with the exception of large DIF and unequally-sized classes. By contrast, the BIC overestimated the number of classes virtually all of the time when intercept DIF was present. The correct model was chosen virtually none of the time by the AIC and BLRT, which consistently overestimated the number of classes. The LMR performed only slightly better, but chose the 2-class solution less than half of the time in most cells; though still poor, performance was better in the presence of exclusively loading DIF in the generating model.

Model parameters.***Covariate effects on class membership.*** Table 7 shows standardized bias in covariate effects on class membership in the impact-only and uniform DIF fitted models. Standardized bias was under 40% when models were correctly specified, i.e., the nonuniform

DIF model for all cells, and the uniform DIF model in the presence of exclusively intercept DIF. Therefore, bias is only presented for misspecified models.

The pattern of standardized bias shown in Table 7 demonstrates three primary findings. First, bias was greatest under the impact-only fitted model, whereas for most cells under the uniform DIF fitted model bias was fairly small. Second, low class separation (small λ) was associated with considerably more bias across all other conditions. Finally, there was virtually no bias in the effect of x_{i3}, which has no DIF, on class membership (α_{13}); however, the effects of all other covariates, including the null effect of x_{14}, showed varying degrees of bias across cells and fitted models.

Interestingly, the pattern of bias differed between the impact-only and uniform DIF fitted models, with respect to differences between both cells and covariates. In the impact-only model, bias was greatest in the presence of intercept DIF (i.e., in the intercept DIF and intercept-and-loading DIF models) and was virtually absent in the loading-DIF only model. In this case, covariate effects α_{11} and α_{14} were negatively biased, whereas the covariate effect α_{12} was positively biased. By contrast, in the uniform DIF fitted model, bias was greatest in the presence of loading DIF (i.e., in the loading DIF and intercept-and-loading DIF models). Here, covariate effects α_{11} and α_{14} were positively biased, whereas the covariate effect α_{12} was negatively biased. This is of particular interest given that α_{11} and α_{14} are positive in the population, thus increasing the likelihood of membership to Class 1, whereas α_{12} is negative in the population, thus decreasing the likelihood of membership to Class 1.

Covariate effects on items. Table 8 shows standardized bias in covariate effects on items in the uniform DIF fitted model. Note that these effects are not estimated in the impact-only fitted model, and so are not tabulated. Additionally, as with covariate effects on classes, standardized bias was minimal under the nonuniform DIF model for all cells, and under the uniform DIF model in the presence of exclusively intercept DIF; therefore, standardized bias is not presented here. Table 8 includes all of the covariate effects included in the data-generating

model: the regressions of y_{i3} on x_{i1} ($\hat{\delta}_{31k}$), y_{i3} on x_{i4} ($\hat{\delta}_{34k}$), y_{i4} on x_{i2} ($\hat{\delta}_{42k}$), y_{i4} on x_{i4} ($\hat{\delta}_{44k}$), y_{i8} on x_{i1} ($\hat{\delta}_{81k}$), and y_{i8} on x_{i2} ($\hat{\delta}_{31k}$).

As shown in Table 8, when the uniform DIF model was fitted to data with unmodeled loading DIF in the data-generating model, bias was (1) very severe across all cells; and (2) of opposite signs in Class 1 and Class 2. The distributions of estimates deviated around the true parameter value (i.e., absolute bias) are shown for one representative example, the effect of covariate x_{i1} on y_{i8}, $\hat{\delta}_{81k}$, in Figure 2, which provide a clearer picture of the opposing patterns of bias between classes observed in the uniform DIF model. Here, the uniform DIF fitted model is shown in green and the nonuniform DIF fitted model is shown in blue. Note that nonuniform DIF fitted model, shown in the blue boxplots, was associated with minimal bias. Thus, standardized bias in the nonuniform DIF model is not tabulated in Table 8, but the distribution of estimates is shown here for comparison to highlight bias in the uniform DIF fitted model. Note also that in the uniform DIF fitted model, the estimates of covariate effects are the same across classes; thus, the distributions of these effects shown in Figure 2 are the same between both the upper and lower panels. For items with positive loading DIF effects in the data-generating model, including all effects of x_{i1} and x_{i4}, regression coefficients (i.e., $\hat{\delta}_{31k}$, $\hat{\delta}_{34k}$, $\hat{\delta}_{44k}$, and $\hat{\delta}_{81k}$) are positively biased in Class 1 and negatively biased in Class 2. For items with negative loading DIF effects in the data-generating model, including all effects of x_{i2}, regression coefficients (i.e., $\hat{\delta}_{42k}$ and $\hat{\delta}_{82k}$ are negatively biased in Class 1 and positively biased in Class 2.

Endorsement probabilities. ***Baseline endorsement probabilities.*** Tables 9 and 10 show standardized bias in baseline endorsement probabilities $\hat{\mu}_{0j1}$ and $\hat{\mu}_{0j2}$ for the three DIF items (items y_{i3}, y_{i4}, and y_{i8}), as well as a non-DIF item, Item y_{i5}, for comparison, under the impact-only and uniform DIF fitted models. Though y_{i5} is chosen arbitrarily as a non-DIF comparison item, patterns of bias were similar across all non-DIF items. Because standardized bias was under 40% in almost all cells under the nonuniform DIF model, standardized bias is not tabulated for the nonuniform DIF model. Additionally, because the uniform DIF model is the correct

model in the presence of exclusively intercept DIF, Table 10 does not show standardized bias for the intercept DIF data-generating model.

Two similarities emerged between results in the impact-only and uniform DIF models. First, under both fitted models bias was more severe for class-specific endorsement probabilities of item y_{i3} ($\hat{\mu}_{03k}$), which had noncompensatory DIF, than for items y_{i4} and y_{i8} ($\hat{\mu}_{04k}$ and $\hat{\mu}_{08k}$), which had compensatory DIF. Second, in both fitted models bias was more pronounced for cells with large DIF than small DIF. However, patterns of bias were different across the impact-only and uniform DIF models. In the impact-only fitted model, bias was similar between the intercept DIF and intercept-and-loading DIF data-generating models: $\hat{\mu}_{0j1}$ tended to be positively biased, and $\hat{\mu}_{0j2}$ negatively biased. By contrast, when exclusively loading DIF was present in the data-generating model, bias was generally less severe, and was uniformly negative across both classes. This set of findings stands in contrast to the uniform DIF fitted model, in which bias was almost exclusively negative across classes in either model with loading DIF, regardless of whether concomitant intercept DIF was also present.

A clearer picture of the differences across fitted models is shown in Figure 3, where average values of μ_{0jk} for both classes are shown for the impact-only model (shown by the solid and dashed red lines), the uniform DIF model (green lines), and the non-uniform DIF model (blue lines).[6] Because bias increased uniformly with large DIF, profile plots are shown only for the large DIF conditions. In the presence of intercept DIF, the impact-only fitted model gives the erroneous impression that class-specific endorsement probabilities are closer together than they truly are. Given low class separation, this bias extends even to items without DIF. By contrast, in the uniform DIF fitted model, bias is negative and confined to items with DIF.

[6] Note that in this and all subsequent figures for Question A, plots are shown for cases in which the data-generating and fitted models are the same – including the uniform fitted DIF model in the presence of exclusively intercept DIF in the data-generating model, and the nonuniform DIF fitted model in the presence of loading DIF – even though standardized bias is not tabulated for these cases.

Marginal predictions of endorsement probabilities. Average estimated endorsement probabilities for all items within Classes 1 and 2, $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1; \hat{\boldsymbol{\Theta}}\right)$ and $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1; \hat{\boldsymbol{\Theta}}\right)$ respectively, are shown in Figure 4. Recall that these profiles, which are weighted sums of each individual's predicted endorsement probabilities given both their class membership and DIF effects, represent average effects marginalizing over all levels of the covariates. Here, bias is considerably less severe than in the corresponding estimates of bias in $\hat{\mu}_{0jk}$. Specifically, bias only ever occurred under the impact-only fitted model: even in the presence of loading DIF, the uniform DIF model yielded mostly unbiased class-averaged conditionally predicted trends. Even under the misspecified impact-only model, bias was only present given low class separation, and was less pronounced when classes were equally sized.

Because the marginal predicted endorsement probabilities $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1; \hat{\boldsymbol{\Theta}}\right)$ and $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1; \hat{\boldsymbol{\Theta}}\right)$ are aggregates of individual predicted endorsement probabilities, conditional on a given individual's configuration of covariates, it is of interest to examine the accuracy of these individual values that comprise aggregate-level estimates. The relationship between individual endorsement probability estimates for item 3, $\hat{\mu}_{i3}$, and the true values μ_{i3} are shown in Figure 5 for high-DIF cells. For each replication within each cell, one case was chosen at random, yielding *R* cases for a given cell; values of $\hat{\mu}_{i3}$ were plotted against true values μ_{i3} for the intercept DIF, loading DIF, and intercept-and-loading DIF data generating models. In the presence of loading DIF, the distribution of $\hat{\mu}_{i3}$ appears to be bimodal under the impact-only and uniform DIF fitted models, with values close to $\hat{\mu}_{03k}$ indicating that estimates of $\hat{\mu}_{ij}$ may be shrunken toward class means. For most cells, estimates $\hat{\mu}_{i3}$ under the impact-only model are quite inaccurate, hovering around .5 even as true values μ_{i3} approach 0 and 1. Notably, these individual values are inaccurate in all cells, even those in which class-specific average values were generally unbiased $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1; \hat{\boldsymbol{\Theta}}\right)$ and $E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1; \hat{\boldsymbol{\Theta}}\right)$.

Class membership.***Class prevalence.*** Figure 6 shows prevalence estimates for Class 1 under all data-generating and fitted models, deviated around their true values. Bias was greatest in the impact-only fitted model, shown in the red boxplots in all three panels. As with

misclassification rates, bias in prevalence estimates under the impact-only fitted model was most severe in the presence of intercept DIF in the data-generating model (i.e., in the intercept-DIF model and intercept-and-loading DIF models), particularly given low class separation and unequally-sized classes. Prevalence estimates under the uniform DIF and nonuniform DIF fitted models were largely unbiased. However, in the presence of loading DIF (i.e., in the loading-only and intercept-and-loading models), the uniform DIF fitted model slightly underestimated the prevalence of Class 1 given low class separation and unequally-sized classes.

Classification accuracy. Table 11 shows values of the Adjusted Rand Index (ARI) comparing modal classifications $\tilde{\eta}_{ik}$ and true values η_{ik}, which demonstrated three primary findings. First, individuals were only misclassified in the case of low class separation; for cells with high levels of class separation, values of ARI were close to 1 for most cells even under significant model misspecification. Second, among cells with low class separation, misclassification was greatest in the presence of large intercept DIF (i.e., large DIF in the intercept-DIF or intercept-and-loading DIF generating models), and was exacerbated by unequally-sized classes. Finally, in the presence of intercept DIF, classification accuracy mainly increased between the baseline and uniform DIF fitted models.

The difference in classification accuracy between baseline and uniform DIF fitted models was moderate in the presence of small DIF (mean $\Delta(ARI)$ across cells = .144) and substantial in the presence of large DIF (mean $\Delta(ARI)$ across cells = .469). In particular, when classes were unequally sized, the impact-only fitted model showed extremely poor classification accuracy in the presence of large intercept DIF and poorly separated classes; values of the ARI for these cells (ARI = .070 for intercept DIF; ARI = .076 for intercept-and-loading DIF) suggest chance levels of correspondence between estimated and true class membership. Importantly, even in the presence of both intercept and loading DIF, the greatest improvements in classification accuracy occurred between the baseline and intercept-DIF only models. By contrast, when exclusively loading DIF was present in the generating model, agreement between estimated and true classifications was considerably better than in the presence of intercept DIF.

Here classification accuracy increased mainly between the uniform DIF and nonuniform DIF fitted models. This difference was modest in the presence of small DIF (mean $\Delta(ARI)$ across cells = .036) and moderate in the presence of large DIF (mean $\Delta(ARI)$ across cells = .131).

Discussion

The goal of Question A was to determine whether and how omitted DIF effects bias class enumeration and parameter values in LCA. A general conclusion which may be drawn from this study is that simply fitting a model with covariates affecting class membership (i.e., an impact-only model) does not solve the problem of omitted DIF effects. Indeed, in the case of class enumeration, the correct number of classes was chosen much more frequently when misspecified covariate effects were excluded altogether. In the case of parameter bias, a more complicated picture emerges. Whereas omitting DIF in intercepts (i.e., direct effects of covariates which do not differ over classes) led to pervasive bias, omitting DIF in loadings (i.e., class-varying direct effects of covariates) did not.

A clear recommendation emerges from Question A's investigation of class enumeration: the number of classes should be decided using an unconditional model, as omitted DIF effects of any kind are likely to lead to the detection of spurious classes if a misspecified conditional (i.e., impact-only) model is fitted. The same conclusion was recently reached by Kim et al. (2016) in regression mixture models, as well as Nylund-Gibson and Masyn (2016) in latent class analysis. Unlike the current study, Nylund-Gibson and Masyn also investigated class enumeration given a fully-specified DIF model (i.e., they tested classes with varying values of K given direct effects of covariates on all relevant items), and found that the BIC, BLRT, and LMR all selected the correct number of classes a majority of the time. This comparison was not made here, which constitutes a significant limitation of the current work. The reason for this omission was that, in initial investigations, models with all relevant DIF effects and $K > 2$ were generally empirically underidentified, likely due to the pervasive nature of DIF in the population model. However, the results of Nylund-Gibson and Masyn suggest that if DIF is properly specified, class enumeration in a conditional model is at least as accurate as in the corresponding unconditional model.

The investigation of parameter bias in Question A raises an interesting question: does loading DIF matter at all, or is the inclusion of class-invariant direct effects (i.e., uniform DIF) sufficient to produce acceptably unbiased results? Indeed, these results give the impression that unmodeled loading DIF, particularly in the absence of co-occurring intercept DIF, is not problematic. Even when the impact-only model was fitted, only minimal bias in item endorsement probabilities and class membership was observed and cases were generally not misclassified at elevated rates. This general impression is complicated by noticing that bias associated with unmodeled loading DIF was still present, and in some cases exacerbated, when the uniform DIF model was fitted. In particular, both baseline endorsement probabilities (Figure 3) and covariate effects on items (Table 8; Figure 2) showed considerable bias under the uniform DIF model. However, marginal endorsement probabilities (Figure 4), which incorporate both baseline item endorsement parameters and covariate effects on items, were unbiased in almost all cases under the uniform DIF fitted model, suggesting that these two sources of bias effectively cancel one another out. Similarly, even though class membership effects were biased under the uniform DIF model (Table 7), overall prevalence rates were not (Figure 6) and classification accuracy was still high (Table 11). Thus results suggest that, in the event that a misspecified uniform DIF model is fitted in the presence of loading DIF, individual-level estimates and marginal trajectories may still be trusted even if parameter estimates cannot.

However, while omitting loading DIF may not yield biased estimates in all cases, omitting intercept DIF -- i.e., fitting the impact-only model in the presence of intercept DIF -- was considerably more problematic. In the presence of intercept DIF, there were virtually no cases -- across all levels of class separation, class size, or DIF magnitude -- in which the impact-only baseline model provided an acceptable level of accuracy in either model parameters or individual-level quantities. In particular, DIF effects appear to be absorbed by covariate effects on class membership (Table 7), which are over- or under-estimated depending on the sign of the omitted DIF effect. Indeed, particularly in the case of x_{i4}, which has no effect on class membership in the population, a spurious negative effect on class membership was frequently

identified. This compounds the increasing body of evidence (Asparouhov and Muthén, 2014; Nylund-Gibson and Masyn, 2016) that, when omitted from the model entirely, DIF effects may present as covariate effects on class membership. As in the case of continuous latent variables (e.g., Chen et al., 2007), the implications for researchers seeking to link covariates to latent classes are severe: a DIF effect of one covariate on one or more items, which is essentially a nuisance, may manifest as a seemingly meaningful but ultimately spurious covariate effect on class membership. The consequences for researchers are equally significant with respect to item endorsement probabilities. The configuration of endorsement probabilities across classes (Tables 9 and 10; Figures 2 and 3) is particularly biased in the presence of poorly-separated classes; here, endorsement probabilities are biased even for items without DIF. Because cases are more likely to be misclassified when class separation is low (Table 11), the effect of misclassified cases likely compounds with the effect of omitted DIF to produce incorrect estimates of item endorsements. Critically, in the impact-only fitted model, there are no individually-varying endorsement probabilities -- baseline endorsement probabilities, which are severely biased here, are all a researcher has to interpret. Therefore, a recommendation resulting from Question A is to include all suspected uniform DIF effects in LCA, as the cost of omitting DIF entirely may be considerable.

However, in order to include all relevant DIF effects, researchers must know that these DIF effects exist. In the absence of strong prior theory suggesting that a DIF effect is likely to be present for a given covariate on a given item, researchers require tools for identifying whether and where DIF is present. The question of how best to identify DIF effects is now addressed in Question B.

Question B

The question of whether and how DIF may be detected was addressed by assessing the performance of two DIF-testing testing procedures on the same data as Question A. The first is the model-based procedure, an adaptation of the aMNLFA procedure for continuous latent variable models (Gottfredson, in preparation; described in Chapter 1), which involves repeatedly comparing an impact-only LCA to a model with loading and intercept DIF parameters for one item at a time. The second is the post-hoc procedure described in Chapter 1, which uses covariates and uncertainty-adjusted modal class assignments from an impact-only LCA as regressors in a series of logistic regressions, each with one item treated as the outcome. The sensitivity and specificity of each of these procedures was then assessed.

Hypotheses

The preponderance of evidence from simulation and empirical studies in the continuous latent variable case shows that testing procedures perform relatively well at detecting both intercept (e.g., Stark, Chernyshenko, and Drasgow, 2006) and loading (e.g., Chen, 2008; Woods and Grimm, 2011) DIF. However, the post-hoc approach has been found to have somewhat lower power than model-based approaches in the continuous latent variable case, particularly when testing for loading DIF (see, e.g., Woods, 2009). Therefore, lower power in post-hoc tests than model-based tests was hypothesized here. However, it was predicted that post-hoc tests would be more computationally efficient and also outperform the model-based test once DIF became sufficiently large (i.e., in the large magnitude condition) for two reasons. First, it was predicted that the power of post-hoc tests would catch up to that of model-based tests given sufficiently large effect size. Second, it was hypothesized that large DIF might lead to sparseness of certain items and response patterns, causing unstable parameter estimates in LCA. .

Data-generating model

Data-generating conditions used to investigate Question B are the same as in Question A. As such, data were generated from an LCA with $K = 2$ classes ($k = 1, \ldots, K$), $J = 10$ indicators ($j = 1, \ldots, J$), $P = 4$ covariates affecting class membership ($p = 1, \ldots, P$) and $N = 500$ cases (

$i = 1, \ldots, N$). Due to the computationally intensive nature of the DIF testing procedures tested in Question B, a total of $R = 250$, rather than $R = 500$, replications were tested here.

Cells differed from one another according to four factors: class prevalence (2 levels: equal or unequal); class separation (2 levels: small or large); DIF magnitude (2 levels: small or large); and DIF type (3 levels: intercept, loading, or both intercept and loading). This yielded a total of 24 unique cells. Overall sample size, as well as the number and pattern of covariate effects on class membership and items, were all held constant.

Model fitting

All data generation and management was conducted using R. The post-hoc regression procedure of DIF testing was evaluated in R as well, using Rcpp to evaluate the weighted likelihood function in Equation 34 in Chapter 1. The use of Rcpp allows for the writing and compilation of C++ code in R, which allowed for considerably faster execution of the regression procedure. Model-based tests were performed using Mplus 7.2, using the same specifications as Question A.

A series of model-based and posthoc tests for DIF was conducted for each of the 250 replications. The path diagrams corresponding to each of these sequentially-applied tests are shown in the bottom panel of Figure 1, and explained below.

Iterative model-based DIF testing. The general set of models in the bottom panel of Figure 1 were fit according to an adaptation of the automated MNLFA (aMNLFA; Gottfredson, in preparation) procedure for multiple covariates, again for the correct number of classes *K*=2. This model-based framework is altered for the LCA setting as follows. For each of *P* covariates and *J* items, a model was fit containing impact from covariate *p* on class membership, as well as DIF from covariate *p* to item *j*, as follows:

1. Allow impact of all *P* covariates on class membership – i.e., $\alpha_{p1} \neq 0$ for all p .
2. Allow the loading and intercept for item *j* to differ across levels of covariate *p* – i.e., $\lambda_{pj1} \neq 0$, $\nu_{pj1} \neq 0$.

3. Set item parameters equal across levels of covariate *p* for all other items – i.e., set $\lambda_{ph1} = 0$, $\nu_{ph} = 0, h \neq j$.

Then a penultimate model containing all of the significant effects found in this step (i.e., significant DIF effects of all items on all covariates) was fit. Finally, a final model was fit, in which any newly non-significant DIF effects were removed. Importantly, if any itemwise model yielded an improper solution, the corresponding item-by-covariate pair was excluded from the final model. Given $P = 4$ and $J = 10$, the above algorithm requires $(4 \times 10) + 2 = 42$ models be fit to each replication and was thus fairly computationally intensive. Estimates of the quantities of interest, described in subsequent sections, were obtained for the final model.

Iterative post-hoc DIF testing. The post-hoc regression test for DIF described by Equations 32-34 in Chapter 1 were used to detect DIF on the basis of all covariates, using the modal class estimates from an impact-only model with $K = 2$. For each item, the pattern of significant main effects of $\mathbf{x}_i$ (representing uniform DIF) and interactions between $\boldsymbol{\eta}_i$ and $\mathbf{x}_i$ (representing non-uniform DIF) was recorded. As in the model-based framework, a penultimate model containing all of the significant effects found in this step (i.e., significant DIF effects of all items on all covariates) was fit, followed by a final model with all newly nonsignificant effects trimmed. Given $P = 4$ and $J = 10$, the above algorithm required $(4 \times 10) = 40$ logistic regressions, but only two additional LCAs (one for the penultimate model, one for the final model), be fit to the data; thus, given that logistic regression is less computationally intensive than LCA, fitting the post-hoc procedure was predicted to be considerably faster than model-based procedure described above. Importantly, as in the model-based procedure above, any item-by-covariate combination whose logistic regression yielded an improper result was excluded from the final model.

Estimates of the quantities of interest, described below, were obtained for the final model.

Outcomes of interest

Computation time. The average per-replication computation time in seconds for each of the model-based and posthoc procedures was computed by dividing the total elapsed time by the number of replications ($R = 250$).

Improper solutions**.** In both the posthoc and model-based testing procedures, an improper solution for any item-by-covariate test (of which there are $P \times J = 40$ in this case) results in that item-by-covariate test not being included in the final model. In theory, improper solutions encompass a wide range of problems with a model (e.g., parameters being fixed at boundary values, saddle points, empirical underidentification). However, in both posthoc and model-based tests, the only type of improper solutions which ever occurred were non-positive definite Hessian matrices, likely associated with complete or quasi-complete separation. For both procedures, the number of times such cases occurred was tabulated.

Sensitivity and specificity. Sensitivity is given by the number of non-zero DIF effects accurately detected in a given replication, divided by true the number of non-zero DIF effects – i.e., it is the probability of correctly identifying a DIF effect. Specificity is given by the number of null DIF effects correctly detected in a given replication, divided by the true number of null DIF effects – i.e., it is the probability of correctly failing to find DIF. Sensitivity and specificity are generally given by:

$$\begin{aligned} Sensitivity &= \frac{true\ positive}{(true\ positive\ +\ false\ negative)} \\ Specificity &= \frac{true\ negative}{(true\ negative\ +\ false\ positive)} \end{aligned} \tag{50}$$

True positives, false positives, true negatives, and false negatives were assessed in the final model. A positive result here is a DIF effect being flagged as significant in the final model; a negative result is a DIF effect not being flagged as significant in the final model. As described in Table 3, six DIF parameters are different from zero in cells with intercept DIF $(\nu_{31}, \nu_{34}, \nu_{42}, \nu_{44}, \nu_{81}, \nu_{82})$; six DIF parameters are different from zero in cells with loading DIF $(\lambda_{311}, \lambda_{341}, \lambda_{421}, \lambda_{441}, \lambda_{811}, \lambda_{821})$; and twelve DIF parameters are different from zero in cells with intercept and loading DIF $(\nu_{31}, \lambda_{311}, \nu_{34}, \lambda_{341}, \nu_{42}, \lambda_{421}, \nu_{44}, \lambda_{441}, \nu_{81}, \lambda_{811}, \nu_{82}, \lambda_{821})$.

Thus, a true positive (*TP*) is defined as one of these parameters being identified as significantly different from zero. A false negative (*FN*) is defined as one of these parameters failing to be identified as significantly different from zero. Therefore, sensitivity is defined in cells with exclusively intercept DIF and cells with exclusively loading DIF as:

$$Sensitivity = \frac{TP}{6} \tag{51}$$

and sensitivity is defined in cells with both intercept and loading DIF as:

$$Sensitivity = \frac{TP}{12} \tag{52}$$

A true negative (TN) is defined as a null effect not being identified. Because there are 4 covariates, 10 items, and 2 types of DIF (loading and intercept), a total of $4 \times 2 \times 10 = 80$ tests are conducted. Therefore, in models with exclusively intercept DIF or exclusively loading DIF, there are 80-6 = 74 null effects; in models with both loading and intercept DIF, there are 80-12=68 null effects. Thus, specificity is defined in cells with exclusively intercept DIF and cells with exclusively loading DIF as:

$$Specificity = \frac{TN}{74} \tag{53}$$

and specificity is defined in cells with both intercept and loading DIF as:

$$Specificity = \frac{TN}{68} \tag{54}$$

Results

Computation time. The posthoc testing procedure was almost eight times faster than the model-based procedure. Note that the posthoc procedure consists of two stages: first, a set of weighted logistic regressions and, second, and a final model consisting of all effects which were significant in the logistic regressions. For a given replication, the set of logistic regressions took an average of 1.1 seconds to run, and the writing and running of the final model took an average of 19.1 seconds to run. Therefore, for a given replication, the posthoc test was completed in an average of 20.2 seconds. By contrast, the model-based procedure was completed in an average of 144.1 seconds (or 2.4 minutes) for a given replication.

Improper solutions. While the total proportion of improper solutions in itemwise tests was extremely low (under 2%) for both tests, the two procedures differed greatly in terms of the overall number of replications with at least one improper solution (out of 40). Table 12 shows the percentage of replications with at least one improper solution in itemwise tests for posthoc and model-based procedures. Improper solutions were a considerably more serious issue with the posthoc testing procedure than the model-based testing procedure. While the model-based procedure had a maximum of 6% of tests with improper solutions (in the case of large intercept DIF with poorly separated, unequally sized classes), the percentage of improper solutions ranged from 9% to 31% in the posthoc procedure when classes were poorly separated and unequally sized.

Sensitivity and specificity. Table 12 shows sensitivity and specificity for model-based and posthoc procedures under all data-generating conditions. Several general findings emerged. First, both the model-based and posthoc procedures showed high levels of specificity, only rarely identifying false positives. Second, model-based tests were generally more sensitive than posthoc tests. Particularly at low levels of class separation, the posthoc procedure frequently failed to identify DIF effects of all types. By contrast, the only time the model-based procedure failed to identify DIF effects was in the case of small loading DIF.

Model-based testing procedures showed high sensitivity in almost all cases. In the case of large DIF, sensitivity was uniformly high for detecting both intercept and loading DIF. However, given small DIF effects, detection of loading DIF was generally fairly poor; in models with exclusively small loading DIF effects, sensitivity dropped to between .48 and .71. In models with both intercept and loading DIF, sensitivity to small DIF was slightly better, ranging from .76 to .88. Because these latter figures include both intercept and loading DIF effects, true positives were disaggregated by DIF type, showing that intercept effects were generally detected far more frequently than loading effects in these models. Given equally-sized classes, an average of 5.912 out of the 6 possible intercept effects were detected, contrasted with 4.232 of 6 possible loading effects, for low class separation; when class separation was high, these rates increased to 5.973/6

intercept DIF effects vs. 4.620/6 intercept DIF effects. This discrepancy between intercept and loading DIF detection rates for small DIF increased for unequally-sized classes, with 5.852/6 intercept effects vs. 3.356/6 loading effects detected given low class separation, and 5.960/6 intercepts vs. 3.664/6 loading effects detected given high class separation.

Posthoc testing procedures were less sensitive to DIF effects than the model-based procedure. Given high levels of class separation, this difference in sensitivity was negligible in most cases. However, when classes were poorly separated, the sensitivity of posthoc tests suffered considerably more. This difference was particularly pronounced in the case of intercept DIF; in the case of exclusively loading DIF, the model-based and posthoc procedures performed similarly. Interestingly, the cases in which the sensitivity of the posthoc procedure was most compromised relative to the model-based procedure, particularly those with large intercept DIF and low class separation, were those with high levels of misclassification in Question A.

Discussion

Question B investigated two strategies for detecting DIF effects in LCA: a posthoc test using modal classifications from an impact-only model, and a model-based strategy. By almost every metric, model-based testing performed better than the posthoc strategy, which lacked power to detect both intercept and loading DIF.

The causes and ramifications of the two procedures' differentially good performance are interesting to consider in light of the results of Question A. The posthoc test was generally less sensitive than the model-based strategy (Table 12), which mirrors results in the continuous latent variable setting (Woods, 2009). Interestingly, however, the cases in which the posthoc test's sensitivity was the lowest were those in which misclassification was high in Question A: those with low levels of class separation and large DIF. This may be due to one of at least two things. First, inaccuracy in the class membership variable, which is controlled for in each itemwise logistic regression equation, may have led to downwardly biased coefficients corresponding to DIF effects. In order to investigate this possibility, relative bias in intercept and loading parameters under the posthoc testing procedure was computed for all cases where this effect was

nonzero in the population. Across all item-covariate pairs in all models, intercept effects were negatively biased by 16.7%. However, loading bias was considerably more extreme; across all item-covariate pairs in all models, loading effects were negatively biased by -158.9%.

Second, because misclassification probabilities are used to weight the likelihood function in Equation 34, high levels of misclassification may have led to inflated standard errors. Even though Vermunt (2010) showed standard errors calculated using the weighted likelihood estimation in Equations 32-34 to be unbiased, this finding was in the context of class membership being an outcome, as opposed to a predictor as it is here. In adaptations of this procedure, Bakk, Tekle, and Vermunt (2013) and Bakk, Oberski, and Vermunt (2014) have proposed a number of adaptations to these standard errors. Interestingly, in the posthoc regression procedure, standard errors were severely positively biased for intercept parameters. Across all item-covariate pairs in all models, estimated standard errors of regression parameters were on average 2.02 times as large as the standard deviation of parameter estimates across replications. For loading effects, the picture was considerably more varied with severe downward bias in some cells and upward bias in others. More research is needed in order to determine how biases in parameters and standard errors combine to produce low power in the posthoc testing procedure.

However, while this question is interesting in its own right, the model-based test performed sufficiently well that researchers are well-advised to use it to detect DIF in LCA rather than attempting the posthoc test. Another viable option was recently presented by Masyn (2017), who proposed a posthoc test based on MIMIC modeling strategies. As with the posthoc procedure presented here, this procedure starts with modal classifications from a model without DIF, and weights subsequent tests using misclassification probabilities. However, rather than testing DIF in one item at a time, this procedure tests DIF on the basis of a given covariate in all items at once. This strategy has the strong advantage of not assuming non-invariant anchor items (often an untenable assumption) while performing individual DIF tests, as both the model-based and posthoc tests proposed here necessarily do. Future work may focus on comparing this

approach with the model-based testing strategy. Additionally, the superior performance of the model-based procedure to the posthoc procedure was only established here in the case of two classes; future work must determine whether this is still the case with more than two classes. In the meantime, however, particularly given the potentially biasing effects shown in Question A, researchers are well-advised to employ either strategy to test for DIF in LCA and include any relevant effects in their final model.

CHAPTER 3
DIFFERENTIAL ITEM FUNCTIONING IN THE MEASUREMENT OF ALCOHOL USE DISORDER

Chapter 2 used artificial data to determine that omitting DIF in mixture models leads to bias in results (Question A), and that DIF may be easily tested for and modeled (Question B). In this chapter, we explore the question of DIF in mixture models of alcohol use disorder (AUD) symptoms. The dual goals of this analysis are to demonstrate the use of mixture models with DIF in a real dataset, and to use mixture models to enhance the assessment and diagnosis of AUD.

The two most commonly used sets of diagnostic criteria for AUD come from the International Statistical Classification of Diseases (ICD-10; WHO, 1992) and the Diagnostic and Statistical Manual-5 (APA, 2013). The publication of DSM-5 saw a departure from the previous edition (DSM-IV; APA, 1994) in terms of the proposed underlying structure of AUD. Though the set of items used to measure AUD is almost identical across editions, the DSM-IV considers AUDs as two separate disorders, alcohol abuse and alcohol dependence, whereas DSM-5 considers AUD as one single disorder. Additionally DSM-5 removes legal problems, which had been a criterion for alcohol abuse in DSM-IV, and adds in a criterion assessing craving. However, aside from these two changes, the two sets of criteria are the same between editions.

These criteria may be considered as binary items which intend to measure some categorical latent variable representing AUD diagnosis. One could conduct a two-class LCA, for instance, which seeks to find AUD diagnosis (i.e., class 1 = healthy; class 2 = AUD); such diagnostic studies are common in the absence of a "gold standard" diagnostic instrument (Rindskopf, 1986); in this case, it would be expected that the prevalence of the AUD class would correspond to the prevalence of AUD in the population.

Alternatively, many researchers have used mixture models not to form a binary diagnosis variable, but to attempt to find some other empirically-determined number of clinically

meaningful subgroups based on DSM-IV and DSM-5 criteria. A sample of such studies is shown in Tables 1 and 2. Though these sets of criteria used to measure AUD are almost completely overlapping across these studies, the results are clearly inconsistent with one another. As shown in Tables 1, these studies focus on widely varying populations: gender balance differs widely across studies, with the proportion of male participants ranging from 0% (LaFlair et al., 2012; LaFlair et al, 2013) to over 60% (Chung and Martin, 2001); age groups range from adolescence (Mancha, Hulbert, and Latimer, 2011; Wells, Horwood, and Ferguson, 2004) to adulthood up to age 60 (Jackson et al., 2014); and racial breakdown ranges from over 80% European American (Beseler et al., 2012; Chung and Martin, 2001) to over 95% Hispanic (Mancha, Hulbert, and Latimer, 2011). Table 2 summarizes the mixture model results from all of these studies, which find anywhere between 2 and 5 classes differing from one another in a number of ways. First, while a majority generally find differences between classes in terms of level, with classes generally increasing monotonically in severity of symptoms, the proportion of individuals at each level of severity differs widely across studies. Additionally, a few studies find classes which fall outside of this continuum, with one report of a small, exclusively male class with only abuse symptoms (Lynskey et al., 2005), another report of a moderately-sized class with subthreshold dependence symptoms only (Jackson et al., 2014), and another report of a class of individuals who exclusively endorse tolerance and drinking more than intended (Beseler et al., 2012).

There are two general possibilities which may explain the incongruity among these results. The first is that differences between groups in these studies actually represent population-level differences between these groups in the latent structure of AUD. For instance, Jackson et al. (2014), who examine AUD symptoms in a nationally representative sample covering a wide range of ages (18-60), find four classes and place anywhere between 27% and 36%[7] of the sample in classes characterized by either full AUD or some degree of dependence. This is a

[7] This figure is based on summing the membership proportions of the AUD disordered class and the minimally dependent class at either time point.

vastly different finding from that obtained by Mancha, Hulbert, and Latimer (2011), who find three classes in their sample of young, predominantly Hispanic adolescents, with the majority of the sample (86%) characterized by a very low level of symptoms. However, it seems implausible to expect that two populations differing this widely in age and ethnicity would show the same prevalence of AUD – as such, it may be the case that the latent structure of AUD is not measured differently across these studies, but that the two populations are differently located within that latent structure.

Another possibility, however, is that the indices used to assess AUD show some degree of measurement non-invariance across these studies. It may be the case that the differences in age, gender, race, ethnicity, and type of sample (e.g., college students versus individuals in treatment for AUD) are associated with differences in the measurement properties of AUD indices at either the item or test level. Indirect evidence to support this proposition comes from two general sources. First, individual configurations of items within class are different across studies. In particular, – some studies find classes which fall outside of a continuum of severity, whereas others do not. These inconsistent findings suggest the possibility that the items most responsible for across-study differences are functioning differently (have DIF) based on background characteristics that vary between studies. For instance, the finding of Beseler et al. (2012) of a class of "diagnostic orphans," who show elevated probabilities of endorsing tolerance and drinking more than intended, may be suggestive of some degree of DIF in these items based on the characteristics of the study. In particular, given that this study focuses on heavy-drinking undergraduates, it may be the case that the normative culture of drinking in college gives rise to an increased probability of developing alcohol tolerance, as well as a greater density of occasions to drink more than intended, even among individuals who do not suffer from AUD.

Second, in studies treating AUD as a continuous latent variable, a number of DSM-IV and DSM-5 criteria have shown DIF on the basis of demographic characteristics, with intercept DIF being the most common finding. For instance, in a nationally representative sample, tolerance was found to show age-related intercept DIF such that older participants were more

likely to endorse this criterion than younger participants (Kahler and Strong, 2006). DIF on the basis of gender is another potential problem; in a community sample of adolescents, Martin, Chung, Kirisci, and Langenbucher (2006) found four criteria to show intercept DIF based on gender, including that female adolescents were less likely to endorse legal problems and hazardous use than male adolescents, even at the same level of AUD. Findings on racial and ethnic differences have been somewhat less conclusive; drinking was shown to have a somewhat complicated pattern of intercept DIF with respect to race in one study of DSM-5 (Casey, Adamson, Shevlin, and McKinney, 2012) such that Black participants were more likely and Hispanic participants were less likely to endorse several criteria than White participants. Given these and other (e.g., Neal, Corbin, and Fromme, 2006; Harford et al. 2009) findings of DIF among groups when treating AUD as a continuous latent variable, it stands to reason that at least some of these problems will persist – and indeed, may even be exacerbated – when AUD is treated as categorical.

The goal of this chapter is thus to apply LCA with DIF to AUD symptoms, in order to determine whether and how the measurement of AUD differs across demographic covariates when AUD is treated as a categorical latent variable. The presence of DIF effects may provide evidence that the discrepancies in findings across different studies owe to measurement bias, rather than differences across participants in underlying levels of AUD symptoms.

Method

Study and sample

Study design. Data come from the Real Experiences and Lives in the University (REAL-U) study, a study of the measurement of AUD, SUD, and related constructs in an undergraduate sample. One of the principal goals of the REAL-U study is to test the biasing effects of subtle differences across studies in measurement, and to assess how well data harmonization methods such as integrative data analysis (IDA; Bauer and Hussong 2009; Curran and Hussong, 2009; Hussong, Curran, and Bauer, 2013) can detect and mitigate this bias.

The study consisted of two visits, separated by a period of two weeks. At each of these visits, participants answered one of two batteries, denoted Battery A and Battery B. As the goal of the REAL-U study is to investigate the extent to which differences in measurement may bias results in AUD studies, the two batteries contained slightly different items intending to measure the same constructs. Participants were randomized to Battery A at visit 1 and Battery B at visit 2 (A/B), Battery B at visit 1 and Battery A at visit 2 (B/A), Battery A at both time points (A/A), or Battery B at both time points (B/B).

Sample. Participants (N = 854; 46.7% male) were undergraduates at a large southeastern research university, recruited by email. Student contact information was obtained from the university Registrar's office, and African American students (the largest ethnic minority group on this campus) and men (given that 57% of the undergraduate population on this campus were women) were oversampled. A total of 6,000 students received an initial email inviting their participation (and for many, several follow-up emails), yielding a total of 854 study participants. In order to be included in the study, subjects must have been between 18-23 years of age and consumed alcohol in the past year. Though participants were not all of legal drinking age, having consumed alcohol in the past year was an inclusion criterion.

The sample was relatively ethnically diverse, with 58.7%, 22.1%, 10.5%, 0.5% of participants respectively identifying as White, Black, Asian, and American Indian/Native American; 6.1% and 2.9% of participants identified as more than one race or some other race. Of these participants, 3.0% identified as Hispanic or Latino. The mean age of alcohol initiation was 17.26 years (SD =1.814). In addition, 28.6% of the participants were first year students, 20.5% were sophomores, 20% were juniors, 28.9% were seniors, and 2% were non-students, did not specify or were graduate students.

In order to assess and enhance the generalizability of results, analyses were conducted on two equally sized partially overlapping subsamples, denoted sample 1 (N = 419) and sample 2 (N = 411). These samples were formed by taking data from visits 1 and 2 in groups A/A, A/B, and B/A described above. The measure of AUD used in the current analysis, described below, was

only administered in Battery A, so subjects in group B/B were not included. Sample 1 consists of individuals from group A/A at visit 1, and group B/A at visit 2; sample 2 consists of individuals from group A/B at visit 1, and group A/A at visit 2. The two samples thus each included Battery A measures of AUD, with roughly equal balance between visit 1 and 2. The two samples are not independent replications, since half of the individuals in Sample 1 were also in Sample 2, but this overlap provides the advantage of allowing an assessment of the stability of classification for individuals in both samples.

Measures. In both batteries, lifetime AUD was measured using the 12 DSM criteria listed in Table 13. This list of criteria includes all items used in either DSM-IV or DSM-5, and thus both legal trouble (Item 3; discarded in DSM-5) and craving (Item 12; new to DSM-5) were initially included. However, Items 3 (legal trouble) and 8 (gave up activities for drinking) showed sufficiently low endorsement probabilities in both samples that they were not retained, given the sensitivity of LCA to extremely low-endorsement items at small sample sizes. Thus, the final dataset consisted of only ten items.

AUD criteria was assessed using a computerized adaptation of the Structured Clinical Interview for Diagnosis (SCID; APA, 2015), in which each criterion is endorsed on the basis of a subject's answers to a number of other sub-questions. For instance, Item 1 is endorsed if a subject answers in the affirmative to any two of the following three questions: "Did you ever miss work or school because you were intoxicated, high, or hung over?", "Did you ever do a bad job at work or fail courses at school because of your drinking?", and "Did you ever have trouble with your housing situation because of your drinking (e.g., forgetting to pay rent or bills or not keeping your place clean)?"

Data analytic strategy

Model testing procedure**.** For both samples 1 and 2, a number of separate LCAs, described in greater detail below, were fit. First, drawing on the results obtained in Chapter 2 (Question A), an unconditional model without any effects of covariates on either class membership was fit with varying numbers of classes, in order to determine the optimal number

of classes for subsequent analyses and establish a comparison point for models including DIF. Then, an impact-only model, containing only the effects of covariates on class membership, was fit using the number of classes chosen at the previous step. A series of itemwise tests, in which loading and intercept DIF was tested for each item, was then conducted, according to the model-based testing procedure which showed superior performance in Chapter 2 (Question B). Finally, based on the results of these itemwise tests, final models including all relevant DIF effects were fitted to the data.

The covariates included were age (centered around age 21), gender (dummy coded; 1 = male and 0 = female), race (dummy coded; 1 = white; 0 = all other races), and study visit (dummy coded; 1 = visit 2; 0 = visit 1). Age was of interest for a number of reasons, including widely-reported normative increases in drinking in emerging adulthood (Brown et al., 2008; Chan et al., 2007). Gender and race were critical to consider because of the common finding of gender- and race-based DIF on a number of diagnostic criteria for AUD (Martin, Chung, Kirisci, and Langenbucher 2006; Harford et al. 2009; Casey, Adamson, Shevlin, and McKinney, 2012). Finally, visit (i.e., visit 1 vs. visit 2) was considered as a source of DIF in order to account for any possible retest effects.

The fitted LCA. Latent class analyses with some unknown number of classes *K* were fit to the data. Class membership is represented by the *K*-variate vector $\boldsymbol{\eta}_i$ with individual elements η_{ik} which take a value of 1 if subject *i* is in class *k* and 0 otherwise. Each individual *i*'s probability of membership to class *k* $(k = 1,\ldots,K)$ and endorsement of item *j* $(j = 1,\ldots,10)$ potentially affected by a subject-specific vector of four covariates $\mathbf{x}_i = (gender_i, age_i, white_i, visit_i)$. The prior probability of class membership $\pi_{ik}(\mathbf{x}_i)$ is related to covariates through a multinomial logistic regression equation as follows:

$$\pi_k(\mathbf{x}_i) = \frac{\exp(\alpha_{0k} + \alpha_{1k} gender_i + \alpha_{2k} age_i + \alpha_{3k} white_i + \alpha_{4k} visit_i)}{\sum_{h=1}^{K} \exp(\alpha_{0h} + \alpha_{1h} gender_i + \alpha_{2h} age_i + \alpha_{3h} white_i + \alpha_{4h} visit_i)} \quad (55)$$

where $\sum_{k=1}^{K}\pi_k(\mathbf{x}_i)=1$ for each individual *i* and class *K* is a reference class, for which parameters are not estimated. Allowing for the possibility of DIF, the probability that subject *i* endorses item *j* given class membership and covariates is denoted μ_{ij}, and is given by:

$$\mu_{ij}=\frac{1}{1+\exp\left(-\left(\begin{array}{l}\left[\nu_{0j}+\nu_{1j}gender_i+\nu_{2j}age_i+\nu_{3j}white_i+\nu_{4j}visit_i\right]+\\ \sum_{k=1}^{K}\left[\lambda_{0jk}+\lambda_{1jk}gender_i+\lambda_{2jk}age_i+\lambda_{3jk}white_i+\lambda_{4jk}visit_i\right]\eta_{ik}\end{array}\right)\right)} \tag{56}$$

Two things are of note about Equation 56. First, the value of DIF effects ν_{pj} and λ_{pj} may be zero for any covariate *p* and item *j*, if there is no DIF for that covariate-item combination. Second, latent classes are treated as effects-coded variables here, as initially outlined in Equation 24 . Thus, ν_{pj} represents the unweighted mean of the effects of covariate *p* on item *j* across all classes, and λ_{pj} represents the deviation for class *k* from this unweighted mean. Note that this effect is not estimable in reference class *K*, and thus $\lambda_{pjK}=-\sum_{k}^{K-1}\lambda_{pjk}$ for all *p*. Given class membership and covariate effects, items are assumed locally independent. As in Chapter 2, class-specific predicted endorsement probabilities can be calculated using Equations 25-26. Individual *i*'s probabilities of endorsing item *j* conditional on class membership and covariates, denoted μ_{ij1} and μ_{ij2} for classes 1 and 2 respectively, are given by:

$$\begin{aligned}\mu_{ij1}(\mathbf{x}_i)&=\frac{1}{1+\exp\left(-\left(\begin{array}{l}\left[\nu_{0j}+\nu_{1j}gender+\nu_{2j}age+\nu_{3j}white+\nu_{4j}visit\right]+\\ \left[\lambda_{0j1}+\lambda_{1j}gender+\lambda_{2j}age+\lambda_{3j}white+\lambda_{4j}visit\right]\end{array}\right)\right)}\\ \mu_{ij2}(\mathbf{x}_i)&=\frac{1}{1+\exp\left(-\left(\begin{array}{l}\left[\nu_{0j}+\nu_{1j}gender+\nu_{2j}age+\nu_{3j}white+\nu_{4j}visit\right]-\\ \left[\lambda_{0j1}+\lambda_{1j}gender+\lambda_{2j}age+\lambda_{3j}white+\lambda_{4j}visit\right]\end{array}\right)\right)}\end{aligned} \tag{57}$$

Each individual is characterized by a posterior probability of membership to each class, conditional on both covariates $\mathbf{x}_i$ and items $\mathbf{y}_i$, denoted τ_{ik}:

$$\tau_{ik}=P(\eta_{ik}=1|\mathbf{x}_i,\mathbf{y}_i)=\frac{P(\mathbf{y}_i|\eta_{ik}=1,\mathbf{x}_i)P(\eta_{ik}=1|\mathbf{x}_i)}{\sum_{k=1}^{K}P(\mathbf{y}_i|\eta_{ik}=1,\mathbf{x}_i)P(\eta_{ik}=1|\mathbf{x}_i)} \tag{58}$$

where $\sum_{k=1}^{K} \tau_{ik} = 1$ for each individual *i*.

Finally, given posterior probabilities τ_{ik} and class-specific trajectories μ_{ijk}, the average endorsement probabilities for each class are given by:

$$E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i1} = 1\right) = \frac{\sum_{i=1}^{N} \tau_{i1} \mu_{ij1}}{\sum_{i=1}^{N} \tau_{i1}} \qquad E\left(y_{ij} \mid \mathbf{x}_i, \eta_{i2} = 1\right) = \frac{\sum_{i=1}^{N} \tau_{i2} \mu_{ij2}}{\sum_{i=1}^{N} \tau_{i2}} \tag{59}$$

Results

The unconditional model

For both samples 1 and 2, an impact-only model without any covariate effects on either class membership or items was first fit. This model is described by Equations 55 and 56, but with all class membership parameters aside from α_{0k}, and all item parameters aside from ν_{0j} and λ_{0jk} set to 0.

Models with between *K* = 1 and *K* = 5 classes were fit to the data; however, the 5-class solution was empirically underidentified in both samples. Table 14 shows fit indices for different numbers of classes in both samples. The Akaike Information Criterion (AIC; Akaike, 1973) and bootstrap likelihood ratio test (BLRT; McLachlan and Peel, 2000) favored a 4-class model, which consisted in both samples of one class with fewer than 10 cases. In both samples, a 2-class model was favored by both the Bayesian Information Criterion (BIC; Schwarz, 1978) and the Vuong-Lo-Mendell-Rubin likelihood ratio test (VLMR; Vuong, 1989; Lo, Mendell, and Rubin, 2001). Given that the BIC reliably chose the correct number of classes in Chapter 2, its support of a 2-class solution here was given extra weight when adjudicating between fit indices. Thus, *K* =2 was retained for subsequent analyses.

Model-building

Following the fitting of the unconditional model, which favored a 2-class solution, an impact-only model with *K* = 2 was fit in both samples. This model, which contains effects of covariates only on class membership, is described by Equations 55 and 56, but with all item

parameters aside from ν_{0j} and λ_{0jk} set to 0. In sample 2, the intercept parameter ν_{04} had to be fixed at -15, denoting a boundary condition in which the probability of endorsing Item 4 is effectively zero.

Item endorsement patterns in the two-class solution are shown in Figure 7 for the impact-only model. In both samples, the majority of the sample (73.5% in sample 1; 71.3% in sample 2) fell into a class (the "low-symptoms" class) characterized by low levels of AUD symptom endorsement. The only symptoms which were endorsed more than 5% of the time in this class were Item 5 (uncontrolled drinking), which was endorsed by roughly 40% of individuals in this class, and Item 10 (tolerance), which was endorsed by 20% of individuals in this class. In both samples, the remainder of the sample (26.5% in sample 1; 28.7% in sample 2) fell into a class (the "high-symptoms" class) characterized by higher levels of AUD symptom endorsement. In particular, a majority of individuals in the high-symptoms class endorsed Items 5 (uncontrolled drinking) and 10 (tolerance),which were endorsed by roughly 95% and 70% of individuals in this class respectively. Additionally, roughly half the members of this class endorsed Items 1 (role impairment), 2 (drinking in dangerous situations), and 9 (continued drinking despite health or psychological problems) in both samples, and between one quarter and one third of members in this class endorsed Items 6 (unsuccessful quit attempts) and 7 (spent a lot of time drinking). Given cutoffs of 2, 5, and 8 item endorsements required for mild, moderate, and severe AUD diagnoses respectively, the modal member of this class met criteria for a least mild AUD.

In both samples all covariate effects on class membership were significantly different from zero in this impact-only model, with the exception of the effect of gender in sample 1. In sample 2, male participants were more likely to be in the high-symptoms class than female participants (OR = 1.909, $z = 2.431$, $p = .015$). In both samples, white participants were more likely to be in the high-symptoms class (Sample 1: OR = 1.889, $z = 3.037$, $p = .002$; Sample 2: OR = 2.037, $z = 2.147$, $p = .032$), as were older participants (Sample 1: OR = 1.398, $z = 3.446$, $p = .001$; Sample 2: OR = 1.362, $z = 3.053$, $p = .002$). Finally, in both samples, participants were

less likely to be classified into the high-symptoms class at Visit 2 than Visit 1 (Sample 1: OR = 0.406, z = -3.037, p = .002; Sample 2: OR = 0.325, z = 3.953, p < .001).

Model-building strategy using itemwise tests

In both samples, the model-based testing algorithm described in Chapter 2 was conducted in order to determine which items, if any, had DIF on the basis of gender, age, race, or visit. Due to the superiority of model-based procedures to posthoc tests in Chapter 2 in terms of sensitivity and specificity, posthoc tests were not considered here.

Table 15 shows all significant results for itemwise tests in Samples 1 and 2. For a given item-covariate pair, ● denotes that DIF was found. Recall that, during itemwise DIF tests, both intercept DIF effects ν_{jp} and loading DIF effects λ_{pjk} were tested for each covariate-item pair, and a pair was flagged if either ν_{jp} or λ_{pjk} were significantly different from zero. Notably, the results of itemwise tests were inconsistent across the two samples, sharing only the finding of age DIF on Item 10 (tolerance) and visit DIF on Items 5 (uncontrolled drinking) and 9 (continued use despite health or psychological problems) in common. Following itemwise tests, a penultimate model was fit containing all effects that were significantly different from zero in itemwise models. Finally, in the last step of the MNLFA procedure, all effects that were rendered nonsignificant in this penultimate model were pruned, resulting in the final model presented below.

The final model

Parameter estimates for the final model in each sample are shown in Table 16. In general, class prevalence and average endorsement probabilities were very similar to those obtained under the impact-only model. This is evident from Figure 7, which shows average class-specific endorsement probabilities for the impact-only model and the final model; the latter are given by Equation 57. In both samples, the majority of individuals (71.9% in sample 1; 71.7% in sample 2) fell into a large class (the "low-symptoms" class) characterized by generally low endorsement probabilities for all items, aside from Items 5 (uncontrolled drinking) and 10 (tolerance). The remainder of individuals (28.1% in sample 1; 28.3% in sample 2) fell into a class (the "high-

symptoms" class) characterized by considerably higher endorsement probabilities for most items, particularly Items 1 (role impairment), 2 (used in dangerous situations), 5 (uncontrolled drinking), 9 (continued use despite health problems), and 10 (tolerance). Covariate effects on class membership were very similar between the two samples in magnitude, direction, and statistical significance. In sample 1, all effects on class membership were significantly different from zero, whereas in sample 2 all class membership effects aside from age were significantly different from zero.

Figures 8-10 show patterns of model-implied endorsement probabilities within each class for participants of different genders (Figure 8), ages (Figure 9), and study visits (Figure 10). No DIF effects were found on the basis of race in either sample. The strongest DIF effect in both samples was the uniform DIF effect of visit on item 9, which assesses continued us despite health or psychological problems; in both samples 1 and 2, subjects were less likely to endorse this item at visit 2 than at visit 1, regardless of class membership. There was an additional uniform DIF effect of visit on item 5, which assesses uncontrolled drinking; here too subjects were less likely to endorse this item at visit 2 than at visit 1 in both samples. Most other DIF effects were inconsistent across samples. In both samples at least one uniform DIF effect of gender emerged, but these effects were on different items and in different directions. In sample 1 the effects of gender on items 1 (role impairment) and 7 (spending a great deal of time drinking) were positive, so that being male increased the probability of endorsement. By contrast, in sample 2 the only effect of gender was on item 5 (uncontrolled drinking), and this effect was negative, so that being female increased the probability of endorsement. Age effects were similarly inconsistent. In sample 1, age showed uniform DIF effects on items 1, 10, and 12, and a nonuniform DIF effect on item 4; In sample 2, age showed a uniform effects on item 2, and a nonuniform effect on item 10. Interestingly, however, despite the fact that the effect of age on item 10 (tolerance) was uniform in sample 1 but nonuniform in sample 2, Figure 9 shows that the effects of age on model-implied endorsement probability are similar across the two samples. In particular, among members of the high-symptom class in both samples, older participants were

more likely to endorse experiencing tolerance than younger participants. This finding underscores the fact that different sets of DIF parameters may produce similar model-implied values, particularly given a nonlinear relationship between the latent variable and the observed variable.

The majority of DIF effects did not agree between the two samples and thus should not be interpreted as substantively meaningful findings. The primary interest in modeling DIF is often not to make inferences about the nature of items and their relationship to background variables, but to account for the bias that these background variables may cause in the estimation of the latent variable. However, given their instability here, it is possible and indeed probable that a number of these effects would not generalize to other samples, and should not be regarded as systematic bias but rather as indications of error-prone measurement. The most consistent effects between the two samples were the differences between visits 1 and 2, which suggest effects of prior exposure to items 5 and 9 on their ability to detect AUD.

Comparing the impact-only and final models

The average endorsement probabilities for both classes in both samples are shown in Figure 7, along with the class-implied endorsement probabilities under the impact-only model. As shown in the figure, these sets of probabilities are virtually identical, suggesting that the addition of DIF effects does little to change the implied item endorsement probabilities at the aggregate level, except for in item 9.

In the impact-only model, almost all covariate effects on class membership were significantly different from zero in both samples. When DIF effects were added in the final model, most of these effects remained significantly different from zero, with two exceptions. First, in sample 1 the effect of gender in sample 1, was not significantly different from zero in the impact-only model or in the final model, $z = .334, p = .738$. Second, also in sample 1, the effect of visit on class membership was now only marginally significant, $z = -1.847, p = .065$. In both samples the effects of visit on class membership were attenuated once DIF effects from visit were added to the model.

It was of interest to see whether and how individual classifications, as well as the accuracy of these classifications, changed based on the addition of DIF. Modal class assignments, denoted $\tilde{\eta}_{ik}$, are generated by assigning individual *i* to the class *k* to which their estimated posterior probability $\hat{\tau}_{ik}$ is the highest. Agreement between modal classifications generated by the impact-only and final models was assessed using the Hubert-Arabie Adjusted Rand Index (ARI; Hubert and Arabie, 1985; Steinley, 2004), which is shown in Equation 44 in Chapter 2. The ARI comparing modal classifications under the impact-only and final models was 0.901 in sample 1, and 0.844 in sample 2, suggesting high concordance between individual classifications generated by the two models. This impression is confirmed by examining classification rates, which show that individuals are grouped into the same class by the two models 96.1% of the time in sample 1 and 97.6% of the time in sample 2. Of the individuals who were assigned different modal classifications under the two models, all met either 2 or 3 diagnostic criteria, suggesting that these individuals represent borderline cases who would very narrowly meet the DSM-5 diagnostic criteria for AUD.

Discussion

In this chapter, the effects of demographic covariates on diagnostic criteria for AUD were investigated using LCA in two partially-overlapping samples. After fitting a model in which the effects of gender, age, race, and study visit only affected class membership (the impact-only model), a series of itemwise DIF tests were conducted in order to find a model with the optimal configuration of DIF effects (the final model), which was then fit. In both samples, a 2-class solution was the best fit to the data in the impact-only model, and thus this model was retained for all subsequent steps. Importantly, the addition of DIF did not change the bulk of aggregate-level findings. In both the impact-only and full models, roughly three quarters of the sample fell into a class (the "low-symptoms" class) with low levels of all symptoms aside from drinking more than intended and tolerance, with the remainder falling into a class (the "high-symptoms" class) in which most members were likely to meet criteria for at least mild AUD. The effects of covariates did not change drastically between the impact-only and final models, and whole-

sample model-implied endorsement probabilities were virtually identical. A minority of individual-level class assignments and posterior class membership probabilities changed between the two models. Perhaps unsurprisingly, differences in posterior probabilities of diagnosis (i.e., membership to the high-symptoms class) were often found in the presence of strong DIF effects. The most striking example of this finding was the effect of visit, which strongly impacted Item 9 such that subjects were less likely to endorse this item (continued drinking despite health problems) at visit 1 than visit 2. For subjects seen at visit 2, posterior probabilities of being in the high-symptoms class were considerably lower in the impact-only model than in the full model. As such, individual classifications into diagnostic categories might change based on whether DIF effects, such as the effect of visit in this study, are explicitly modeled.

Do the current findings help to reconcile the discrepancies between empirical findings between different studies of AUD symptoms? In some sense, particularly given the number and prevalence of classes was similar between models with and without DIF, it would appear that applications of LCA to AUD symptoms may be surprisingly robust to measurement bias. Moreover, because the presence of significant DIF effects based on age and gender was inconsistent across samples, there is no DIF effect which can be reliably identified as the cause of potential age- or gender-related bias in subsequent analyses. This finding is contextualized by previous work (Jackson and Sher, 2005; Cole, Bauer, Hussong, and Giordano, 2017), which suggests that even though the number of classes may be stable from one application of LCA to the next, patterns of AUD item endorsement may vary substantially based on even minor alterations to measurement. Importantly, measurement features of the items – e.g., the wording of item stems and response options – were not manipulated in the current study, and thus no measurement-related DIF effects were examined. Thus, it may be the case that AUD measures are indeed robust to measurement bias from demographic variables, but that differences across studies in the instruments used to measure AUD leads to some cross-study inconsistencies in results.

The findings of Chapter 3 must additionally be considered in light of those of Chapter 2. In some sense, a number of the differences – and lack thereof – between the impact-only and full models are consistent with those of Question A. As in Question A, while the addition of DIF effects did strongly affect the significance and magnitude of covariate effects, neither class prevalence nor model-implied endorsement probabilities changed much between the impact-only and full models. However, the DIF testing results – i.e., that the model-based DIF tests identified DIF effects inconsistently across the two samples – are slightly surprising, given the results of Question B, in which the model-based testing procedure used here showed high levels of sensitivity and specificity. These differences underscore the challenges of generalizing from simulation work to empirical data, and will be considered further in the concluding chapter.

CHAPTER 4
DISCUSSION

Recent years have seen widespread use of mixture models such as latent class analysis (LCA) in the social and behavioral sciences. One of the most attractive features of mixture models is their ability to incorporate covariate effects, both on class membership and on items themselves. However, while mixture models have long been able to accommodate direct effects of covariates on items (Huang and Bandeen-Roche, 2004; Muthén and Shedden, 1999), relatively few applications of mixture models include these direct effects, possibly owing to the challenges of interpreting them (e.g., Muthén, 2004; Lubke and Muthén, 2007). At the same time, increased attention has been paid to the goal of quantifying and interpreting differences between groups in mixture model results (Morin et al, 2016; Collins and Lanza, 2010; Finch, 2015). In the current study, we attempted to integrate these two lines of research by framing direct effects in mixture models as measurement non-invariance, and using this framework to investigate the ways in which mixture models may incorporate these effects as well as the consequences of omitting them.

More specifically, it was proposed here that measurement non-invariance in mixture models be considered, modeled, and tested for in a way similar to traditional methods in continuous latent variable models. In the current formulation, uniform DIF is synonymous with class-invariant direct effects of covariates on items, and nonuniform DIF is synonymous with class-varying direct effects of covariates on items. As in the continuous latent variable case, uniform DIF results from differences across levels of a covariate in an intercept parameter; nonuniform DIF results from differences in a loading parameter, with or without concomitant intercept DIF. The timeliness of this work is underscored by a recent paper by Masyn (2017), which also conceptualizes direct effects in mixture models as uniform and nonuniform DIF,

using the same formulation as Chapter 1 of the current work. This may reflect a growing sentiment that measurement noninvariance is a problem in categorical latent variable models.

After establishing that mixture models may easily accommodate DIF effects, the next step was to empirically determine the extent to which the inclusion of such effects is worthwhile. The first part of Chapter 2, Question A, addressed this question through a simulation assessing bias due to omitted DIF in LCA. As expected, bias was most severe when classes were not well-separated and when DIF effects were large. Somewhat more surprising, however, was the fact that omitted loading DIF led to much less severe consequences than omitted intercept DIF. The presence of omitted intercept DIF was associated with overextraction of latent classes, as well as pervasive bias in model parameters, item endorsement probabilities, and class membership at $K = 2$.

In the second portion of Chapter 2, Question B, two iterative procedures for testing for DIF were compared. These tests included a posthoc test, modeled after the posthoc logistic regression test in IRT (Swaminathan and Rogers, 1990), and a model-based itemwise test for DIF modeled after IRT-LR-DIF (Thissen, 2001) and automated moderated nonlinear factor analysis (aMNLFA; Gottfredson, in preparation). The model-based test showed superior performance, reliably identifying the correct set of DIF effects under most circumstances. Moreover, model-based tests are easy to implement using any standard software for fitting mixture models (e.g., Mplus and LatentGold) and, although less computationally efficient than the post-hoc procedure, are not excessive in their time requirements.

As such, the recommendations for researchers following from Chapter 2 are clear. In the absence of strong hypotheses about the number of classes or the presence of DIF effects, class enumeration should be performed using an unconditional model, given that the misspecification of covariate effects was found both here and elsewhere (Nylund-Gibson and Masyn, 2016) to lead to overextraction of classes. Once the number of classes has been decided, DIF should be tested using iterative model-based procedures. After DIF effects are located, a model including at least all relevant uniform DIF effects should be fit. One of the key findings of Question A in

Study 2 was that, even if the nature of DIF is misspecified (i.e., a uniform DIF model is fit when nonuniform DIF exists in the population), individual classifications and model-implied indicator values are accurate, even though parameter values are biased. Thus, if a researcher wishes to interpret parameter values such as covariate effects on class membership or item endorsement, they should include nonuniform DIF in the event that it is suspected to exist or revealed by DIF tests. However, if the sole purpose of the analysis is to obtain individual class assignments or predicted endorsement probabilities, the inclusion of uniform DIF is likely sufficient.

The simulation conducted in Chapter 2 provided strong evidence for the inclusion of DIF effects in LCA, but did not give a sense of the sorts of questions which can be empirically addressed by testing for DIF or how such tests would perform when implemented with real data. Thus, Chapter 3 motivated the study of measurement non-invariance in mixture models by showing how they may provide insight into discrepancies in findings across studies in alcohol use disorder (AUD) research. In this study, the effects of gender, race, age, and study visit on DSM criteria for AUD were investigated using LCA in two partially-overlapping samples. Here, the inclusion of DIF did not strongly affect aggregate-level findings, such as the number and prevalence of classes. Additionally, DIF effects were not reliably identified across the two samples, yielding little in the way of substantively interpretable relationships between covariates and items. However, varyingly large DIF effects of gender, race, age, and visit were found for a number of items, and posterior probabilities of class membership differed on the basis of the inclusion of these effects. Thus, even though these DIF effects are not substantively interpretable, the results of Chapter 3 underscore the sensitivity of LCA results to covariate effects.

The lack of consistent findings of DIF across samples in Chapter 3 is in some ways discrepant with the results of Question B in Chapter 2, in which the model-based testing procedure showed high levels of both sensitivity and specificity in finding DIF effects. Given these findings, the inconsistent results in the current chapter raise an important question: what is the cause of DIF tests' inconsistent performance, when the results of Question B indicated that

model-based testing procedures should detect DIF effects when they are present? The conditions in which DIF was tested here are similar in many ways to the data-generating conditions in Chapter 2. The number of covariates (P = 4 in both chapters), items (J = 10 in Chapter 2; J =12 in Chapter 3), and cases (N = 500 in Chapter 2; N = 419 for Sample 1 and N = 411 for Sample 2 in Chapter 3) were similar across the two studies. Moreover, the proportion of cases in each class (71.9% / 28.1% in Sample 1; 72.3 %/ 27.7% in Sample 2), as well as the degree of class separation (Entropy = .816 in Sample 1; Entropy = .827 in Sample 2) are consistent with unequally-sized, well-separated classes in Chapter 2. Thus, given the similarity of data-generating conditions between the two studies, the difference in results is perplexing.

Thus, it may be of interest to consider the ways in which the current dataset differs from the simulated one used in Chapter 2. One difference is that item endorsement probabilities were generally quite low in Chapter 3 (average across items = .180 in Sample 1, average across items = .195 in Sample 2). Perhaps model-based DIF testing procedures' performance is degraded when base rates are lower. On a more fundamental level, however, these differences underscore the limitations of generalizing from simulation results to "messy" real data. Whereas data in Chapter 2 were generated from a 2-class LCA, in Chapter 3 data likely possessed a number of unknown features that were not addressed by the model. For instance, in both the simulation in Chapter 2 and the empirical data in Chapter 3, the fitted model assumed conditional dependence of indicators given class membership and covariate effects. While data were generated to meet this assumption in Chapter 2, it may be the case that local dependence existed between items in Chpater 3, which would have been better accommodated by including continuous latent factors within a given class (Lubke and Muthén, 2005; 2007). Additionally, while a number of covariates not ultimately used in the final model were also considered, it may be the case that important covariate effects were omitted. As such, further simulation work must take into account model error (Cudeck and Henly, 1991; MacCallum and Tucker, 1991) by generating data for which the fitted model, even one with properly specified DIF, does not hold exactly. On a more fundamental level, the simulation study assumed that the data were truly generated from a

latent class model. Given the well-documented finding that multiple classes may be spuriously detected in homogeneous data given distributional misspecification (Bauer and Curran, 2003; Bauer and Curran, 2004), it is certainly plausible that a more exhaustive search of single-class models (e.g., IRT models with local dependence between indicators; models with causal indicators) might have revealed a better-fitting model for the data in Chapter 3.

Limitations and Future Directions

This study is characterized by a number of limitations which should be rectified in future work. Possibly the most substantial limitation of the current work is that, while Chapter 1 presents DIF in mixture models as a problem which may be explored with indicators of any scale, in both the simulation and empirical studies only mixture models with binary items (i.e., LCA) were considered. One of the draws of modern mixture modeling techniques is that they can accommodate multiple outcomes with a diversity of scales (Muthén, 2002). Thus, it is critical to empirically establish the effects of unmodeled DIF and the performance of DIF tests when items are ordinal, count, or continuous variables. In particular, it is of interest to extend the study of DIF in mixture models to ordinal items. This is because the definitions of uniform and nonuniform DIF change somewhat in continuous latent variable models when ordinal items are used, as each item is characterized not only by an intercept and loading but also individual thresholds for each endorsement category (Cohen and Kim, 1998). However, when ordinal variables are considered in mixture models, it is not always clear whether and how thresholds should be constrained across classes. Thus, considering ordinal indicators using a proportional odds model is likely to complicate the parallels established between uniform and nonuniform DIF in continuous and categorical latent variable models.

The simulation study presented in Chapter 2 considered a relatively narrow range of data-generating models, and could be extended in at least three ways. First, only two-class models were considered, and future work should consider models with three or more classes. Models with more than two classes are of interest because the nature of differences between classes may be considerably more complicated than they were here. For instance, item parameter values may

differ in one class relative to all other classes, but be similar between the remaining classes. The consequences of "uneven" patterns of DIF such as this remain to be seen, and cannot be addressed in a simulation with only two classes. The second obvious extension to the population models considered in Chapter 2 is adding more sample sizes. While the one sample size considered here, $N = 500$, is a common sample size associated with adequate power in mixture models (Nylund et al., 2007), it is of interest to see whether and how the performance of the DIF tests changes at lower and higher sample sizes. In particular, it is possible that posthoc tests may have greater power at higher sample sizes (e.g., $N = 1000$), or that the uniformly strong performance of the model-based tests may deteriorate at lower sample sizes (e.g., $N = 250$). Finally, the simulation in Chapter 2 only considered a data-generating model with complete local independence between indicators. However, a wide diversity of covariance relationships may exist between indicators in finite mixture models (McLachlan and Peel, 2000), from LCA models with only incidental local dependence (Rebousin et al., 2008) to factor mixture models (Lubke and Muthén, 2005; Lubke and Muthén, 2007) in which continuous latent variables govern the distribution of indicators within a given class. It remains to be seen how omitted DIF biases model parameters, as well as how DIF effects should be interpreted, in the presence of dependence between variables.

Similarly, the nature of the fitted models in Question A of Chapter 2 can and should be further extended in order to consider the effects of different types model misspecification when incorporating DIF effects. In particular, while the type of DIF (i.e., uniform vs. nonuniform DIF) was misspecified in a number of the fitted models, the location of DIF was not. In other words, if a data-generating model contained the effect of a given covariate on a given item, the model fitted to this dataset necessarily contained a DIF effect for this covariate-item pair. There were no models in which only a subset of the existing relationships between covariates and items was modeled, nor were any models with spurious DIF effects fit. However, it could be the case that as long as a certain percentage (e.g., 50%) of the existing DIF effects are modeled, bias in parameters and individual-level quantities is mitigated.

In a similar vein, a wider range of testing procedures may be evaluated in further work, extending Question B. In particular, the recently proposed MIMIC-LCA test (Masyn, 2017) may be compared to the IRT-LR-DIF strategy, which performed optimally in the current analyses. Unlike IRT-LR-DIF, in which each itemwise test compares a model with DIF on one item to a baseline model with no DIF, MIMIC-LCA compares each itemwise DIF model to a minimally-constrained baseline model. It may be the case that MIMIC-LCA outperforms the IRT-LR-DIF, particularly when sample size, the scale of items, and the nature of the relationships between indicators (i.e., local dependence) are manipulated. These hypotheses could not be tested by the current simulation design, but should certainly be addressed in follow-up work.

Finally, while Chapter 3 provided an interesting example of the potential effects of DIF in AUD research, results pertaining to discrepancies across studies in mixture models of AUD were inconclusive. In essence, the DIF effects found in Chapter 3 provide indirect evidence that categorical latent variables formed on the basis of AUD symptom items are not defined identically across levels of demographic covariates. However, particularly given that many DIF effects were not replicated across partially-overlapping samples, there is not a systematic relationship between any subset of AUD items and any demographic covariate. As such, no concrete recommendations for items to avoid due to systematic measurement non-invariance arise from Chapter 3. Thus, particularly given that this study was performed on a predominantly European American college sample, DIF analyses on samples which include wider ranges of demographic covariates should be performed in order to determine whether and to what extent measurement bias exists in mixture models of AUD symptoms.

REFERENCES

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory.* Budapest: Akademiai Kiado.

Agresti, A. (2007). *An introduction to categorical data analysis.* Hoboken: John Wiley and Sons.

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 3099-3132.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV*. Washington, DC: Author.

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329-341.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis, 22*(4).

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology, 43*(1), 272-311.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological methods, 8*(3), 338.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological methods, 9*(1), 3.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological methods, 14*(2), 101.

Beseler, C. L., Taylor, L. A., Kraemer, D. T., & Leeman, R. F. (2012). A Latent Class Analysis of DSM-IV Alcohol Use Disorder Criteria and Binge Drinking in Undergraduates. *Alcoholism: Clinical and experimental research, 36*(1), 153-161.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis, 12*(1), 3-27.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*(4), 331-348.

Brouwer, R. K. (2009). Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. *Journal of Intelligent Information Systems, 32*(3), 213-235.

Brown, S. A., McGue, M., Maggs, J., Schulenberg, J., Hingson, R., Swartzwelder, S., ... & Winters, K. C. (2008). *A developmental perspective on alcohol and youths 16 to 20 years of age. Pediatrics, 121(Supplement 4),* S290-S310.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological bulletin, 105*(3), 456.

Casey, M., Adamson, G., Shevlin, M., & McKinney, A. (2012). The role of craving in AUDs: dimensionality and differential functioning in the DSM-5. *Drug and alcohol dependence, 125*(1), 75-80.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification, 13*(2), 195-212.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2015). It Might Not Make a Big DIF Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement, 76*(1), 114-140.

Chan, K. K., Neighbors, C., Gilson, M., Larimer, M. E., & Marlatt, G. A. (2007). Epidemiological trends in drinking by age and gender: providing normative feedback to adults. *Addictive Behaviors, 32*(5), 967-976.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling, 14*(3), 464-504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, *95*(5), 1005.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Chung, T., & Martin, C. S. (2001). Classification and course of alcohol problems among adolescents in addictions treatment programs. *Alcoholism: Clinical and Experimental Research, 25*(12), 1734-1742.

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. *Sociological methodology, 15*, 81-110.

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

Cole, V. T., Bauer, D. J., Hussong, A. M., & Giordano, M. L. (2017). An Empirical Assessment of the Sensitivity of Mixture Models to Changes in Measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 159-179.

Collins, L. M., & Lanza, S. T. (2010*). Latent class and latent transition analysis.* Hoboken, NJ: Wiley

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods, 6*(4), 330.

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving Factor Score Estimation Through the Use of Observed Background Characteristics. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 827-844.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological methods, 14*(2), 81.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological),* 1-38.

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics, 23*(4), 643-659.

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*(4), 278-295.

Finch, H. (2015). A Comparison of Statistics for Assessing Model Invariance in Latent Class Analysis. *Open Journal of Statistics, 5*(3), 191.

Flora, D. B., Curran, P. J., Hussong, A. M., & Edwards, M. C. (2008). Incorporating measurement nonequivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling, 15*(4), 676-704.

Garrett, E. S., & Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics, 56*(4), 1055-1067.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24*(3), 229-252.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215-231.

Hallquist, M., & Wiley, J. (2011). MplusAutomation: Automating Mplus model estimation and interpretation.

Harford, T. C., Yi, H. Y., Faden, V. B., & Chen, C. M. (2009). The dimensionality of DSM-IV alcohol use disorders among adolescent and adult drinkers and symptom patterns by age, gender, and race/ethnicity. *Alcoholism: Clinical and Experimental Research, 33*(5), 868-878.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling, 17*(2), 193-215.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research, 18*(3), 117-144.

Huang, G. H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika, 69*(1), 5-32.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification, 2*(1), 193-218.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology, 9*, 61-89.

Jackson, N., Denny, S., Sheridan, J., Fleming, T., Clark, T., Teevale, T., & Ameratunga, S. (2014). Predictors of drinking patterns in adolescence: a latent class analysis. *Drug and Alcohol Dependence, 135*, 133-139.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: a methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors, 19*(4), 339.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*(4), 443-482.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351a), 631-639.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide. Scientific Software International.

Kahler, C. W., & Strong, D. R. (2006). A Rasch model analysis of DSM-IV alcohol abuse and dependence items in the national epidemiological survey on alcohol and related conditions. *Alcoholism: Clinical and Experimental Research, 30*(7), 1165-1175.

Karp, I., O'loughlin, J., Paradis, G., Hanley, J., & Difranza, J. (2005). Smoking trajectories of adolescent novice smokers in a longitudinal study of tobacco use. *Annals of epidemiology, 15*(6), 445-452.

Kim, S. H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 345-355.

Kim, M., Vermunt, J., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 601-614.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear algebra and its applications, 18(2), 95-138.

La Flair, L. N., Bradshaw, C. P., Storr, C. L., Green, K. M., Alvanzo, A. A., & Crum, R. M. (2012). Intimate partner violence and patterns of alcohol abuse and dependence criteria among women: a latent class analysis. *Journal of studies on alcohol and drugs, 73*(3), 351-360.

La Flair, L. N., Reboussin, B. A., Storr, C. L., Letourneau, E., Green, K. M., Mojtabai, R., ... & Crum, R. M. (2013). Childhood abuse and neglect and transitions in stages of alcohol involvement among women: a latent transition analysis approach. *Drug and alcohol dependence, 132*(3), 491-498.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural equation modeling: a multidisciplinary journal, 20*(1), 1-26.

Lazarsfeld, P. F., Henry, N. W., & Anderson, T. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767-778.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods, 10*(1), 21.

Lubke, G., & Muthén, B. O. (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.

Lynskey, M. T., Nelson, E. C., Neuman, R. J., Bucholz, K. K., Madden, P. A., Knopik, V. S., ... & Heath, A. C. (2005). Limitations of DSM-IV operationalizations of alcohol abuse and dependence in a sample of Australian twins. *Twin Research and Human Genetics, 8*(06), 574-584.

Mancha, B. E., Hulbert, A., & Latimer, W. W. (2012). A latent class analysis of alcohol abuse and dependence symptoms among Puerto Rican youth. *Substance use & misuse, 47*(4), 429-441.

Martin, C. S., Chung, T., Kirisci, L., & Langenbucher, J. W. (2006). Item response theory analysis of diagnostic criteria for alcohol and cannabis use disorders in adolescents: implications for DSM-V. *Journal of abnormal psychology, 115*(4), 807.

Masyn, K. E. (2017). Measurement Invariance and Differential Item Functioning in Latent Class Analysis With Stepwise Multiple Indicator Multiple Cause Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-18.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*(1), 55-64.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (Vol. 37).* CRC press.

McCutcheon, A. L. (2002). Basic concepts and procedures in single-and multiple-group latent class analysis. *Applied latent class analysis*, 56-88.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* John Wiley & Sons.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60*(1), 117-174.

Meehl, P. E. (2004). What's in a taxon?. *Journal of abnormal psychology, 113*(1), 39.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research, 13*(2), 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248.

Millsap, R. E. (2006). *Statistical approaches to measurement invariance.* Routledge.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-585.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117.

Muthén, B.O. (2004). *Latent variable analysis. The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications, 345-68.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods, 4*(2), 139.

Neal, D. J., Corbin, W. R., & Fromme, K. (2006). Measurement of alcohol-related consequences among high school and college students: application of item response models to the Rutgers Alcohol Problem Index. *Psychological assessment, 18*(4), 402.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling, 14*(4), 535-569.

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 782-797.

O'Connor, R. M., & Colder, C. R. (2005). Predicting alcohol patterns in first-year college students through motivational systems and reasons for drinking. *Psychology of Addictive Behaviors, 19*(1), 10.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

Reboussin, B. A., Song, E. Y., Shrestha, A., Lohman, K. K., & Wolfson, M. (2006). A latent class analysis of underage problem drinking: evidence from a community sample of 16–20 year olds. *Drug and alcohol dependence, 83*(3), 199-209.

Reboussin, B. A., Ip, E. H., & Wolfson, M. (2008). Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(4), 877-897.

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*(3), 552.

Rindskopf, D., & Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in medicine, 5*(1), 21-27.

Rinker, D. V., & Neighbors, C. (2015). Latent class analysis of DSM-5 alcohol use disorder criteria among heavy-drinking college students. *Journal of substance abuse treatment, 57*, 81-88.

Schulenberg, J. E., Merline, A. C., Johnston, L. D., O'Malley, P. M., Bachman, J. G., & Laetz, V. B. (2005). Trajectories of marijuana use during the transition to adulthood: The big picture based on national panel data. *Journal of Drug Issues, 35*(2), 255-280.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research, 25*(1), 78-90.

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological methods, 9*(3), 386.

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

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Documentation for computer program]. L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77-83.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). Statistical analysis of finite mixture models.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In *Advances in latent variable mixture models.* Charlotte, NC: Information Age Publishing, Inc.

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*(1), 4-70.

Vermunt, J. K. (2010). Latent Class Modeling with Covariates: Two Improved+ Three-Step Approaches. *Political analysis, 18*, 450-469.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307-333.

Wang, C. P., Hendricks Brown, C., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association, 100*(471), 1054-1076.

Wang, Z., & Zhou, X. H. (2014). Nonparametric Identifiability of Finite Mixture Models with Covariates for Estimating Error Rate without a Gold Standard. University of Washington Biostatistics Working Paper Series.

Wells, J. E., Horwood, L. J., & Fergusson, D. M. (2004). Drinking patterns in mid-adolescence and psychosocial outcomes in late adolescence and early adulthood. *Addiction, 99*(12), 1529-1541.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research,* 281-324.

Windle, M., & Wiesner, M. (2004). Trajectories of marijuana use from adolescence to young adulthood: Predictors and outcomes. *Development and psychopathology, 16*(04), 1007-1027.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods, 12*(1), 58.

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

Woods, C. M. (2008). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement, 68*(4), 571-586.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339-361.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling, 14*(3), 435-463.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological measurement, 64*(5), 737-757.

Table 1.

Demographic characteristics in a sampling of AUD studies using mixture models.

Study	Larger study (if applicable)	N	Inclusion criteria	% Male	Age	Race/Ethnicity	Analytic strategy	Scale of measures	Measurement
Jackson (K.) et al., 2014	NESARC	33644 T1 / 25186 T2	Current drinkers measured twice 3-year separation between waves)	51%	age 18-60	71% European American	FMM	Binary	AUDADIS-IV, 5 12-month alcohol consumption indicators
La Flair et al., 2012; La Flair et al., 2013	NESARC	11750	Female current drinkers age 18 or older	0%	young adults oversampled	African American, Hispanic oversampled	LTA: 2 time points, 3 years apart	Binary	AUDADIS-IV
Beseler et al., 2012	College sample	361	Adults with lifetime exposure to alcohol	26.70%	mean age = 19.1	84.8% European American (est.)	LCA	Binary	DSM-IV criteria
Rinker and Neighbors, 2015	College sample	394	Heavy-drinking undergraduates (at least one binge in past month)	48.50%		62.9% European American / 21.1% Hispanic / 3.6% Africal American, 1% Asian American	LCA	Binary	DSM-V criteria
Mancha, Hulbert, and Latimer, 2011	International Longitudinal Survey of Adolescent Health	622	Middle and high school students in San Juan, PR	41.60%		95.3% Hispanic	LCA	Binary; collapsed from 3: "never" vs. "once" or "two times or more"	DSM-IV abuse and dependence
Chung and Martin, 2001	Pittsburgh Alcohol Research Center treatment study	300	Adolescents recruited from outpatient treatment centers before and after treatment (1-year separation between waves)	62%	mean age 16.2	82% European American / 12% African American / 1% Hispanic or Asian Amerian / 5% other	LTA: 2 time points, 1 year apart (pre- and post-treatment)	Binary	DSM-IV SCID
Lynskey et al., 2005	Australian National Health and Medical Research Council Twin Panel (phone survey)	6285	Twins recruited from twin studies	44.70%	median age = 30		LCA	Binary	SSAGA for DSM-IV
Wells, Horwood, and Ferguson (2004)	Chirstchurch Health and Development Study	953	16-year-olds interviewed for longitudinal study	50.20%	16-year-olds		LPA	Continuous quantity, ordinal frequency, count measures for problems	Quantity in mL (continuous); frequency in past 3 months (ordinal) CIDI for DSM-IV binned into a count

Table 2.
Findings in a sampling of AUD studies using mixture models.

		Classes found						
		Classes along a continuum of severity						
Study	**Num. Classes**	None	Infrequent/ Unproblematic	Moderate	Frequent/ Problematic	Excessive / Highly Problematic / AUD Likely	**Classes outside of continuum**	**Covariate-related findings**
Jackson (K.) et al., 2014	4		Infrequent (52% T1 / 42% T2)	Regular moderate (23% T1 / 27% T2)		AUD disordered (9% T1 / 13% T2)	Minimally dependent (23% T1 / 18% T2)	
La Flair et al., 2012; La Flair et al., 2013	3		No problems (87.1%/83.9%)		Hazardous (11.3%/14.2%)	Severe (1.5%/1.9%)		Childhood abuse predicted transition to severe or hazardous classes; Intimate partner violence predicted membership to severe and hazardous classes
Beseler et al., 2012	3	Class 1: low symptom endorsement (60.1%)				Class 3: high symptom endorsements (8.3%)	Class 2: "Diagnostic orphans" (31.5%)	
Rinker and Neighbors, 2015	2	Less severe (86%)				More severe (14%)		Drinker identity predicted membership to severe class; drinking refusal self-efficacy predicted membership to less severe class
Mancha, Hulbert, and Latimer, 2011	3		Low severity (86.0%)		Moderate severity (11.7%)	High severity (2.3%)		Associations with a number of risky behaviors
Chung and Martin, 2001	3		Asymptomatic (8% baseline / 44% follow-up)	Mild (35% baseline / 43% follow-up)		High risk (57% baseline / 13% follow-up)		Transition to asymptomatic class was less likely for males and those with conduct disorder
Lynskey et al., 2005	4 (female) / 5 (male)	No problems (66.5% F / 43.7% M)		Heavy drinking (23.9% F / 34.9% M)	Moderate dependence (7.6% F / 12.5% M)	Severe dependence (2.0% F / 3.2% M)	Excessive drinking with abuse (5.7% M)	Membership to problem drinking classes was associated with depression and conduct disorder.
Wells, Horwood, and Ferguson (2004)	4	Class 1: nondrinkers (23.5%)	Class 2: Occasional drinkers, no problems	Class 3: Frequent drinkers, some problems	Class 4: Frequent drinkers, many problems			Membership related to alcohol problems at ages 21-25, as well as number of sexual partners and violence.

Table 3.
Item parameters in data-generating model under varying levels of class separation and DIF.

Low Class Separation

	Loadings							Intercepts						
		Small DIF			Large DIF				Small DIF			Large DIF		
Item	λ_{0j1}	λ_{1j1}	λ_{2j1}	λ_{4j1}	λ_{1j1}	λ_{2j1}	λ_{4j1}	Item	ν_{0j}	ν_{1j}	ν_{2j}	ν_{4j}	ν_{1j}	ν_{2j}	ν_{4j}
1	0.25	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0.5	0	0	0	0	0	0	2	0	0	0	0	0	0	0
3	0.75	0.4	0	-0.4	0.8	0	-0.8	3	0	0.8	0	-0.8	1.6	0	-1.6
4	1	0	-0.4	0.4	0	-0.8	0.8	4	0	0	-0.8	0.8	0	-1.6	1.6
5	1.25	0	0	0	0	0	0	5	0	0	0	0	0	0	0
6	0.25	0	0	0	0	0	0	6	0	0	0	0	0	0	0
7	0.5	0	0	0	0	0	0	7	0	0	0	0	0	0	0
8	0.75	0.4	-0.4	0	0.8	-0.8	0	8	0	0.8	-0.8	0	1.6	-1.6	0
9	1	0	0	0	0	0	0	9	0	0	0	0	0	0	0
10	1.25	0	0	0	0	0	0	10	0	0	0	0	0	0	0

High Class Separation

	Inter							Intercepts						
		Small DIF			Large DIF				Small DIF			Large DIF		
Item	λ_{0j1}	λ_{1j1}	λ_{2j1}	λ_{4j1}	λ_{1j1}	λ_{2j1}	λ_{4j1}	Item	ν_{0j}	ν_{1j}	ν_{2j}	ν_{4j}	ν_{1j}	ν_{2j}	ν_{4j}
1	0.4	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0.8	0	0	0	0	0	0	2	0	0	0	0	0	0	0
3	1.2	0.4	0	-0.4	0.8	0	-0.8	3	0	0.8	0	-0.8	1.6	0	-1.6
4	1.6	0	-0.4	0.4	0	-0.8	0.8	4	0	0	-0.8	0.8	0	-1.6	1.6
5	2	0	0	0	0	0	0	5	0	0	0	0	0	0	0
6	0.4	0	0	0	0	0	0	6	0	0	0	0	0	0	0
7	0.8	0	0	0	0	0	0	7	0	0	0	0	0	0	0
8	1.2	0.4	-0.4	0	0.8	-0.8	0	8	0	0.8	-0.8	0	1.6	-1.6	0
9	1.6	0	0	0	0	0	0	9	0	0	0	0	0	0	0
10	2	0	0	0	0	0	0	10	0	0	0	0	0	0	0

Table 4.
Class-specific logits in data-generating models containing both intercept and loading DIF.

	Low Class Separation													
			Small DIF						Large DIF					
	δ_{0jk}		δ_{1jk}		δ_{2jk}		δ_{4jk}		δ_{1jk}		δ_{2jk}		δ_{4jk}	
Item	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$
1	-0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0
2	-0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0
3	-0.75	0.75	0.4	1.2	0	0	-0.4	-1.2	0.8	2.4	0	0	-0.8	-2.4
4	-1	1	0	0	-0.4	-1.2	0.4	1.2	0	0	-0.8	-2.4	0.8	2.4
5	-1.25	1.25	0	0	0	0	0	0	0	0	0	0	0	0
6	-0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0
7	-0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0
8	-0.75	0.75	0.4	1.2	-0.4	-1.2	0	0	0.8	2.4	-0.8	-2.4	0	0
9	-1	1	0	0	0	0	0	0	0	0	0	0	0	0
10	-1.25	1.25	0	0	0	0	0	0	0	0	0	0	0	0

	High Class Separation													
			Small DIF						Large DIF					
	δ_{0jk}		δ_{1jk}		δ_{2jk}		δ_{4jk}		δ_{1jk}		δ_{2jk}		δ_{4jk}	
Item	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$	$k=1$	$k=2$
1	-0.4	0.4	0	0	0	0	0	0	0	0	0	0	0	0
2	-0.8	0.8	0	0	0	0	0	0	0	0	0	0	0	0
3	-1.2	1.2	0.4	1.2	0	0	-0.4	-1.2	0.8	2.4	0	0	-0.8	-2.4
4	-1.6	1.6	0	0	-0.4	-1.2	0.4	1.2	0	0	-0.8	-2.4	0.8	2.4
5	-2	2	0	0	0	0	0	0	0	0	0	0	0	0
6	-0.4	0.4	0	0	0	0	0	0	0	0	0	0	0	0
7	-0.8	0.8	0	0	0	0	0	0	0	0	0	0	0	0
8	-1.2	1.2	0.4	1.2	-0.4	-1.2	0	0	0.8	2.4	-0.8	-2.4	0	0
9	-1.6	1.6	0	0	0	0	0	0	0	0	0	0	0	0
10	-2	2	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.
Class enumeration results for unconditional models.

	BIC		AIC		Lo Mendell Rubin LRT		Bootstrap Likelihood	
	Mean # Classes	% Correct	Mean # Classes	% Correct	Mean # Classes	% Correct	Mean # Classes	% Correct
Intercept-only DIF								
Equal classes								
Small lambda								
Small DIF	2	1	3.05	0.27	2.2	0.87	2.15	0.85
Large DIF	2.01	0.99	3.67	0.01	2.4	0.66	2.93	0.17
Large lambda								
Small DIF	2	1	2.82	0.43	2.31	0.78	2.08	0.92
Large DIF	2.01	0.99	3.78	0.02	2.56	0.58	2.84	0.31
Unequal classes								
Small lambda								
Small DIF	2	1	2.98	0.3	2.15	0.89	2.07	0.93
Large DIF	2.02	0.98	3.68	0.03	2.37	0.7	2.89	0.19
Large lambda								
Small DIF	2	1	2.98	0.3	2.37	0.74	2.1	0.91
Large DIF	2.01	0.99	3.68	0.04	2.64	0.5	2.86	0.21
Loading-only DIF								
Equal classes								
Small lambda								
Small DIF	2	1	2.74	0.42	2.13	0.89	2.04	0.96
Large DIF	2	1	2.94	0.35	2.18	0.85	2.09	0.91
Large lambda								
Small DIF	2	1	2.72	0.51	2.31	0.78	2.01	0.99
Large DIF	2	1	2.82	0.44	2.27	0.8	2.07	0.93
Unequal classes								
Small lambda								
Small DIF	2	1	2.75	0.5	2.11	0.92	2.07	0.94
Large DIF	2	1	2.88	0.38	2.09	0.93	2.07	0.93
Large lambda								
Small DIF	2	1	2.62	0.52	2.19	0.87	2.03	0.97
Large DIF	2	1	3	0.33	2.24	0.81	2.09	0.93
Intercept-and-loading DIF								
Equal classes								
Small lambda								
Small DIF	2	1	3.11	0.24	2.3	0.8	2.26	0.77
Large DIF	2.02	0.98	3.58	0.02	2.5	0.62	2.86	0.18
Large lambda								
Small DIF	2	1	3.06	0.29	2.27	0.79	2.13	0.87
Large DIF	2.01	0.99	3.65	0	2.56	0.52	2.8	0.26
Unequal classes								
Small lambda								
Small DIF	2	1	3.38	0.09	2.3	0.77	2.54	0.48
Large DIF	2.17	0.83	3.56	0	2.59	0.41	2.96	0.06
Large lambda								
Small DIF	2	1	3.32	0.13	2.36	0.73	2.36	0.65
Large DIF	2.16	0.84	3.66	0	2.73	0.39	3.05	0.01

Table 6.
Class enumeration results for conditional models.

	BIC		AIC		Lo Mendell Rubin LRT		Bootstrap Likelihood	
	Mean # Classes	% Correct	Mean # Classes	% Correct	Mean # Classes	% Correct	Mean # Classes	% Correct
Intercept-only DIF								
Equal classes								
Small lambda								
Small DIF	2.61	0.41	4	0	2.74	0.44	4	0
Large DIF	4	0	4	0	3.4	0.06	4	0
Large lambda								
Small DIF	2.36	0.71	4	0	3.04	0.21	4	0
Large DIF	4	0	4	0	3.72	0.03	4	0
Unequal classes								
Small lambda								
Small DIF	2.98	0.02	4	0	3	0.24	3.97	0
Large DIF	3.96	0	4	0	3.35	0.13	4	0
Large lambda								
Small DIF	2.79	0.21	4	0	3.21	0.16	3.97	0
Large DIF	3.9	0	4	0	3.61	0	4	0
Loading-only DIF								
Equal classes								
Small lambda								
Small DIF	2	1	3.98	0	2.92	0.46	3.31	0.26
Large DIF	2.09	0.93	4	0	2.8	0.4	3.94	0
Large lambda								
Small DIF	2	1	3.99	0	2.75	0.56	3.44	0.18
Large DIF	2.12	0.91	4	0	3.14	0.25	3.99	0
Unequal classes								
Small lambda								
Small DIF	2	1	3.96	0	3.02	0.36	3.57	0.09
Large DIF	2.91	0.09	3.99	0	2.9	0.27	3.89	0
Large lambda								
Small DIF	2	1	3.98	0	2.75	0.55	3.22	0.24
Large DIF	2.88	0.13	4	0	3	0.23	3.93	0
Intercept-and-loading DIF								
Equal classes								
Small lambda								
Small DIF	2.86	0.15	4	0	3.11	0.22	4	0
Large DIF	3.96	0	4	0	3.42	0.09	4	0
Large lambda								
Small DIF	2.86	0.14	4	0	3.23	0.09	3.98	0
Large DIF	3.89	0	4	0	3.61	0.01	4	0
Unequal classes								
Small lambda								
Small DIF	3.39	0	4	0	3.4	0.09	4	0
Large DIF	4	0	4	0	3.7	0.04	4	0
Large lambda								
Small DIF	3.28	0	4	0	3.44	0.05	4	0
Large DIF	4	0	4	0	3.66	0	4	0

Table 7.
Class membership parameters under the impact-only and intercept DIF models.

	Impact-Only Model					Uniform DIF Model				
	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}
Intercept-only DIF										
Equal classes										
Small lambda										
Small DIF	-4.30	-75.45	**175.70**	18.38	**-143.10**					
Large DIF	-0.75	-65.88	**179.90**	-19.24	**-122.90**					
Large lambda										
Small DIF	6.42	-33.65	78.74	7.02	-48.57					
Large DIF	-3.76	-42.25	**110.70**	11.41	-63.26					
Unequal classes										
Small lambda										
Small DIF	21.67	-70.06	**154.90**	4.74	**-124.90**					
Large DIF	74.39	-39.98	28.19	-18.56	-26.69					
Large lambda										
Small DIF	-12.47	-35.23	70.31	11.81	-53.02					
Large DIF	-24.51	-22.69	**106.70**	10.01	-71.23					
Loading-only DIF										
Equal classes										
Small lambda										
Small DIF	-26.32	18.98	9.33	15.62	-4.99	-27.31	24.73	11.21	15.31	-1.99
Large DIF	-44.27	11.03	-12.34	5.78	17.06	-52.82	37.42	-11.61	5.50	35.44
Large lambda										
Small DIF	-11.29	3.14	4.60	13.14	6.56	-12.07	5.38	4.96	13.22	7.46
Large DIF	-9.59	7.18	13.99	18.03	-6.57	-12.81	13.26	15.80	18.47	-3.00
Unequal classes										
Small lambda										
Small DIF	-34.16	2.52	26.54	16.77	-9.02	-87.62	64.81	-19.26	19.89	60.46
Large DIF	-24.29	-5.09	34.79	5.72	-20.43	**-190.00**	**142.10**	-48.41	4.80	**124.00**
Large lambda										
Small DIF	-24.18	8.94	15.56	15.18	-1.36	-40.55	29.53	-0.80	16.84	21.12
Large DIF	-25.83	-0.69	4.93	5.14	4.05	-87.31	56.81	-29.22	9.81	62.55
Intercept-and-loading DIF										
Equal classes										
Small lambda										
Small DIF	55.55	-78.17	**176.20**	4.68	**-129.30**	-47.87	12.42	3.43	3.99	11.52
Large DIF	86.08	-69.78	**180.20**	-1.31	**-125.60**	-59.05	17.91	10.28	9.13	4.74
Large lambda										
Small DIF	15.43	-25.60	65.55	6.28	-52.38	-16.93	13.68	-5.77	8.29	3.51
Large DIF	19.84	-36.56	**107.50**	10.59	-75.67	-30.44	11.44	7.38	7.58	4.17
Unequal classes										
Small lambda										
Small DIF	**106.90**	**-114.20**	**183.60**	-28.58	**-148.40**	-84.97	40.40	-17.62	19.02	36.81
Large DIF	**192.60**	**-113.60**	**120.40**	**-116.60**	**-195.00**	**-102.60**	51.45	-34.22	14.57	43.97
Large lambda										
Small DIF	-4.25	-35.20	80.07	5.04	-65.14	-46.61	22.95	8.39	9.93	6.44
Large DIF	6.01	-40.41	85.86	2.97	-56.69	-60.74	27.15	-4.38	9.76	20.91

Note: Standardized bias with absolute value greater than 40 is denoted with light shading; standardized bias with absolute value greater than 100 is denoted with dark shading and bold text.

Table 8.
Effects of covariates on items under the intercept DIF model, given loading DIF in the data-generating model.

	Y3 on X1		Y3 on X4		Y4 on X2		Y4 on X4		Y8 on X1		Y8 on X2	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Loading-only DIF												
Equal classes												
Small lambda												
Small DIF	**358.8**	**-286.9**	**391.2**	**-338.1**	**-293.8**	**316.5**	**339.7**	**-348.6**	**330.8**	**-310.5**	**-350.5**	**363.5**
Large DIF	**526.9**	**-357.3**	**605.5**	**-463.7**	**-520.9**	**583.5**	**611.7**	**-620.5**	**636.8**	**-581.3**	**-648.9**	**662**
Large lambda												
Small DIF	**350.1**	**-276.3**	**358.4**	**-289.8**	**-285.2**	**304.7**	**291.7**	**-308.7**	**338.4**	**-324.4**	**-340.8**	**339.1**
Large DIF	**773.9**	**-575.5**	**756.2**	**-588.3**	**-508.9**	**588.3**	**529.7**	**-579.8**	**694.2**	**-673.1**	**-684.3**	**702.9**
Unequal classes												
Small lambda												
Small DIF	**582.3**	-60.86	**601.9**	-90.56	**-519.1**	**127.8**	**565.9**	**-137.4**	**546.3**	**-124.9**	**-558**	**149.7**
Large DIF	**910.8**	-58.11	**968.7**	**-100.6**	**-1031**	**310.8**	**1028**	**-279.9**	**989**	**-246.7**	**-1101**	**318**
Large lambda												
Small DIF	**549.3**	-91.74	**575.5**	**-106.2**	**-469.9**	**125.6**	**472**	**-116.1**	**530.2**	**-134.7**	**-499.6**	**134.8**
Large DIF	**1009**	**-150.5**	**1006**	**-173.2**	**-884.7**	**269.4**	**900.2**	**-262.5**	**995.2**	**-272.7**	**-983**	**297.4**
Intercept-and-loading DIF												
Equal classes												
Small lambda												
Small DIF	**242**	**-336.2**	**261.8**	**-365.3**	**-266**	**348.7**	**247.2**	**-302.8**	**254.7**	**-337.9**	**-273.9**	**361.6**
Large DIF	**262.1**	**-571.7**	**294.7**	**-657**	**-301**	**553.5**	**309.2**	**-552.9**	**293.1**	**-551.3**	**-312.8**	**608.4**
Large lambda												
Small DIF	**255.5**	**-329.7**	**271**	**-373.6**	**-244.8**	**256.3**	**265.6**	**-285.1**	**257.6**	**-300.2**	**-277.7**	**330.6**
Large DIF	**305.2**	**-647.3**	**286**	**-606.8**	**-320.2**	**505.7**	**312.3**	**-501.2**	**319.2**	**-561.2**	**-303**	**544.9**
Unequal classes												
Small lambda												
Small DIF	**381.1**	**-103.9**	**409.8**	**-134.6**	**-411.3**	**115.2**	**418.3**	**-129.7**	**400.9**	**-129.1**	**-427.3**	**137.3**
Large DIF	**439**	**-197.9**	**444.6**	**-205.1**	**-502.5**	**219.5**	**473.5**	**-202.6**	**471.3**	**-206.6**	**-482.9**	**227.3**
Large lambda												
Small DIF	**395.7**	**-140.1**	**433.5**	**-148.2**	**-412**	**106.5**	**405.1**	**-113.7**	**414.7**	**-121.3**	**-413.1**	**119.7**
Large DIF	**443.5**	**-235.4**	**470**	**-279.4**	**-500.1**	**179.3**	**535.5**	**-192.3**	**506.2**	**-234.5**	**-510.4**	**246.4**

Note: Standardized bias with absolute value greater than 40 is denoted with light shading; standardized bias with absolute value greater than 100 is denoted with dark shading and bold text.

Table 9.
Baseline endorsement probabilities under the impact-only model for all data-generating models.

	DIF Items						Non-DIF Items	
	Y3		Y4		Y8		Y5	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class	Class 2
Intercept-only DIF								
Equal classes								
Small lambda								
Small DIF	**372.40**	**-359.70**	-65.15	66.70	26.43	-26.06	40.11	-37.18
Large DIF	**168.30**	**-161.40**	-85.35	85.41	10.52	-16.14	**108.80**	**-109.80**
Large lambda								
Small DIF	**442.70**	**-429.60**	49.16	-42.59	**101.80**	**-105.50**	8.33	3.84
Large DIF	**721.50**	**-726.90**	**184.70**	**-187.00**	**239.20**	**-256.80**	12.05	-16.59
Unequal classes								
Small lambda								
Small DIF	**337.50**	**-246.20**	-85.16	34.75	13.28	-15.01	68.73	-3.68
Large DIF	4.66	39.42	9.14	62.24	45.81	38.76	**411.30**	**-131.90**
Large lambda								
Small DIF	**371.70**	**-366.20**	-13.78	**-108.60**	66.56	**-141.20**	6.93	-4.18
Large DIF	**616.50**	**-666.30**	51.51	**-295.80**	**149.80**	**-333.50**	23.57	2.40
Loading-only DIF								
Equal classes								
Small lambda								
Small DIF	**-122.90**	**-189.30**	50.09	16.18	1.97	-22.89	-13.04	-21.70
Large DIF	**-176.70**	**-361.00**	**122.80**	-17.49	33.64	-65.98	-14.93	-36.85
Large lambda								
Small DIF	**-104.10**	**-204.40**	69.47	-5.19	31.32	-36.55	-0.84	-10.48
Large DIF	**-115.90**	**-390.70**	**181.50**	-49.80	**109.90**	-99.58	-7.74	-17.04
Unequal classes								
Small lambda								
Small DIF	**-121.30**	**-152.30**	33.44	-16.35	-2.99	-36.55	-11.84	-10.17
Large DIF	**-183.20**	**-290.40**	83.35	-44.84	0.29	-67.73	-0.57	-13.26
Large lambda								
Small DIF	**-134.20**	**-157.90**	50.24	-28.72	17.13	-48.67	-10.33	0.12
Large DIF	**-179.60**	**-383.30**	**133.10**	**-126.10**	59.46	**-139.50**	-4.05	-8.71
Intercept-and-loading DIF								
Equal classes								
Small lambda								
Small DIF	**219.70**	**-470.80**	-71.76	39.43	-16.66	-50.00	62.61	-2.71
Large DIF	**182.10**	**-319.20**	**-104.60**	63.44	-6.86	-32.31	**105.50**	-54.37
Large lambda								
Small DIF	**212.50**	**-607.00**	-20.35	**-112.90**	23.51	**-183.80**	6.97	2.07
Large DIF	**407.00**	**-868.60**	30.33	**-287.30**	**103.20**	**-355.50**	29.87	-0.33
Unequal classes								
Small lambda								
Small DIF	**210.30**	**-170.80**	-69.43	77.35	4.50	30.90	**136.90**	9.20
Large DIF	-7.01	-3.26	-18.76	80.88	14.60	36.48	**526.30**	**-116.10**
Large lambda								
Small DIF	**212.00**	**-549.00**	-37.81	**-204.60**	16.08	**-249.80**	26.76	4.77
Large DIF	**399.50**	**-832.80**	-30.43	**-317.70**	57.72	**-436.00**	39.49	16.05

Note: Standardized bias with absolute value greater than 40 is denoted with light shading; standardized bias with absolute value greater than 100 is denoted with dark shading and bold text.

Table 10.
Baseline endorsement probabilities under the intercept DIF model for all data-generating models.

	DIF Items						Non-DIF Items	
	Y3		Y4		Y8		Y5	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Intercept-only DIF								
Equal classes								
Small lambda								
Small DIF								
Large DIF								
Large lambda								
Small DIF								
Large DIF								
Unequal classes								
Small lambda								
Small DIF								
Large DIF								
Large lambda								
Small DIF								
Large DIF								
Loading-only DIF								
Equal classes								
Small lambda								
Small DIF	**-123.50**	**-154.00**	39.95	19.52	-7.46	-12.41	-12.51	-24.59
Large DIF	**-145.90**	**-206.80**	74.63	-9.06	0.09	-47.10	-17.35	-52.71
Large lambda								
Small DIF	**-114.60**	**-180.20**	58.91	1.09	23.52	-28.56	-0.70	-11.65
Large DIF	**-142.00**	**-344.40**	**152.60**	-31.25	91.71	-85.72	-8.20	-21.30
Unequal classes								
Small lambda								
Small DIF	**-282.80**	-18.96	35.04	-32.87	-29.33	-27.07	-22.25	-38.03
Large DIF	**-821.70**	-12.66	31.06	-79.42	-71.69	-59.38	-27.30	**-106.00**
Large lambda								
Small DIF	**-270.60**	-45.38	51.29	-21.41	-0.63	-28.85	-14.72	-9.51
Large DIF	**-541.10**	**-104.40**	96.70	**-108.60**	-4.73	-89.45	-22.71	-52.82
Intercept-and-loading DIF								
Equal classes								
Small lambda								
Small DIF	**-158.30**	**-181.90**	-51.14	-64.71	-62.93	-74.94	-18.99	-25.21
Large DIF	**-172.30**	**-231.10**	-86.71	**-139.60**	-84.07	**-123.20**	-30.77	-29.42
Large lambda								
Small DIF	**-186.40**	**-235.10**	-72.18	-92.87	-80.32	-93.60	-20.46	-15.55
Large DIF	**-249.40**	**-345.70**	**-142.30**	**-202.90**	**-134.30**	**-204.20**	-17.49	-22.00
Unequal classes								
Small lambda								
Small DIF	**-298.10**	-50.81	-46.85	-59.69	-72.10	-41.47	-20.66	-32.44
Large DIF	**-378.90**	-92.58	**-104.10**	-92.86	**-133.70**	-76.79	-47.25	-59.80
Large lambda								
Small DIF	**-329.30**	-99.84	-51.05	-55.72	-87.59	-62.59	-10.22	-15.45
Large DIF	**-496.50**	**-161.30**	**-170.60**	**-116.50**	**-179.10**	**-122.90**	-19.00	-21.23

Note: Standardized bias with absolute value greater than 40 is denoted with light shading; standardized bias with absolute value greater than 100 is denoted with dark shading and bold text.

Table 11.
Adjusted Rand Index (ARI) statistics comparing true and estimated class membership under all data-generating and fitted models.

	Adjusted Rand Index (ARI)		
	Fitted Model		
	Impact-only	Uniform DIF	Nonuniform DIF
Intercept-only DIF			
Equal classes			
Small lambda			
Small DIF	0.632	0.716	0.714
Large DIF	0.351	0.697	0.694
Large lambda			
Small DIF	0.901	0.926	0.925
Large DIF	0.866	0.91	0.909
Unequal classes			
Small lambda			
Small DIF	0.635	0.746	0.741
Large DIF	0.07	0.726	0.721
Large lambda			
Small DIF	0.909	0.932	0.932
Large DIF	0.872	0.92	0.919
Loading-only DIF			
Equal classes			
Small lambda			
Small DIF	0.703	0.702	0.743
Large DIF	0.651	0.638	0.783
Large lambda			
Small DIF	0.919	0.919	0.929
Large DIF	0.901	0.9	0.937
Unequal classes			
Small lambda			
Small DIF	0.738	0.743	0.773
Large DIF	0.689	0.702	0.818
Large lambda			
Small DIF	0.934	0.932	0.941
Large DIF	0.914	0.903	0.944
Intercept-and-loading DIF			
Equal classes			
Small lambda			
Small DIF	0.639	0.704	0.73
Large DIF	0.451	0.681	0.726
Large lambda			
Small DIF	0.898	0.917	0.925
Large DIF	0.87	0.905	0.919
Unequal classes			
Small lambda			
Small DIF	0.431	0.746	0.762
Large DIF	0.076	0.718	0.761
Large lambda			
Small DIF	0.907	0.93	0.935
Large DIF	0.865	0.911	0.927

Note: Shading denotes particularly severe values of ARI, with darker shading indicating lower concordance between true and estimated class membership.

Table 12.
Percent of replications with improper solutions, sensitivity, and specificity in the model-based and posthoc testing procedures.

	Model-based testing procedure			Posthoc testing procedure		
	Proportion Improper Solutions in Itemwise Models	**Sensitivity**	**Specificity**	**Proportion Improper Solutions in Itemwise Models**	**Sensitivity**	**Specificity**
Intercept-only DIF						
Equal classes						
Small lambda						
Small DIF	0.00	0.98	0.97	0.00	0.65	0.97
Large DIF	0.01	1.00	0.98	0.18	0.81	0.97
Large lambda						
Small DIF	0.00	0.99	0.97	0.00	0.92	0.99
Large DIF	0.00	1.00	0.97	0.00	1.00	0.99
Unequal classes						
Small lambda						
Small DIF	0.00	0.95	0.98	0.28	0.85	1.00
Large DIF	0.06	0.97	0.98	0.29	0.67	0.93
Large lambda						
Small DIF	0.00	0.98	0.97	0.03	0.91	0.99
Large DIF	0.01	1.00	0.97	0.09	0.99	0.99
Loading-only DIF						
Equal classes						
Small lambda						
Small DIF	0.00	0.67	0.97	0.00	0.61	0.99
Large DIF	0.00	0.98	0.98	0.00	0.99	0.99
Large lambda						
Small DIF	0.00	0.71	0.97	0.00	0.68	1.00
Large DIF	0.00	0.99	0.97	0.00	0.99	0.99
Unequal classes						
Small lambda						
Small DIF	0.01	0.48	0.98	0.18	0.40	1.00
Large DIF	0.01	0.96	0.98	0.31	0.92	0.99
Large lambda						
Small DIF	0.00	0.54	0.97	0.03	0.51	0.99
Large DIF	0.00	0.98	0.96	0.02	0.98	0.99
Intercept and loading DIF						
Equal classes						
Small lambda						
Small DIF	0.00	0.84	0.98	0.01	0.74	0.99
Large DIF	0.00	0.97	0.98	0.16	0.81	0.99
Large lambda						
Small DIF	0.00	0.88	0.97	0.00	0.83	0.99
Large DIF	0.00	0.99	0.97	0.00	0.99	1.00
Unequal classes						
Small lambda						
Small DIF	0.01	0.77	0.98	0.18	0.60	0.98
Large DIF	0.02	0.90	0.98	0.22	0.53	0.94
Large lambda						
Small DIF	0.00	0.80	0.97	0.02	0.76	0.99
Large DIF	0.00	0.96	0.97	0.09	0.90	0.99

Note: Shading denotes particularly poor performance, with darker shading indicating higher rates of improper solutions, as well as lower sensitivity and specificity.

Table 13.
Alcohol Use Disorder (AUD) criteria used in the current study.

Item	Criterion	Notes
1	Role impairment	
2	Used in dangerous situations	
3	Legal problems	Dropped from DSM-5
4	Drinking despite problems with family and friends	
5	Uncontrolled drinking	
6	Unsuccessful quit attempts	
7	Spent a lot of time drinking	
8	Gave up activities for drinking	
9	Continued use despite health or psychological problems	
10	Tolerance	
11	Withdrawal symptoms	
12	Craving	New to DSM-5

Table 14.

Class enumeration statistics for unconditional models in both samples.

		Sample 1							Sample 2						
K	#Param	LL	BIC	AIC	LMR	LMR p.val	BLRT	BLRT p.val	LL	BIC	AIC	LMR	LMR p.val	BLRT	BLRT p.val
1	12	-1748.4	3569.43	3520.8	--	--	--	--	-1754.6	3581.33	3533.1	--	--	--	--
2	25	-1504.8	**3160.88**	3059.57	481.112	**0**	487.227	0	-1530.8	**3212.13**	3111.67	441.787	0	447.434	0
3	38	-1466.7	3163.28	3009.3	75.315	0.0629	76.272	0	-1505.9	3240.6	3087.89	49.152	**0.0145**	49.78	0
4	51	-1452.2	3212.98	3006.32	28.615	0.2814	28.979	**0**	-1487.6	3282.08	**3077.13**	36.296	0.2015	36.76	**0**
5	64	-1438.8	3264.84	**3005.51**	26.477	1.00E-04	26.814	0.1224	-1476.6	3338.4	3081.21	21.641	0.362	21.917	0.3333

Table 15.
Itemwise model-based DIF test results in both samples.

	Sample 1				**Sample 2**			
	Age	Gender	White	Time	Age	Gender	White	Time
Item 1	●	●						●
Item 2					●			
Item 4	●							
Item 5				●	●	●		●
Item 6								
Item 7		●						
Item 9				●				●
Item 10	●				●			
Item 11	●							
Item 12	●							

Table 16.

Class membership and item parameters in the final model.

	Sample 1									
			Covariate effects							
	Intercept		Gender		Age		White		Visit	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Class membership										
P(High-symptom)	-0.465	0.481	0.097	0.290	0.268	0.094	0.648	0.288	-0.484	0.262
High-symptom class										
Item 1	-0.283	0.350	1.132	0.353	0.256	0.118				
Item 2	0.071	0.234								
Item 4	1.636	0.321			0.222	0.201				
Item 5	-4.307	0.867							-0.620	0.254
Item 6	0.715	0.234								
Item 7	1.036	0.356	1.274	0.403						
Item 9	-3.833	0.657							-2.812	0.458
Item 10	-0.987	0.248			0.242	0.089				
Item 11	1.867	0.306								
Item 12	2.994	0.528			-0.497	0.216				
Low-symptom class										
Item 1	2.979	0.366	1.132	0.353	0.256	0.118				
Item 2	2.802	0.311								
Item 4	5.653	0.841			-0.427	0.081				
Item 5	-0.422	0.396							-0.620	0.254
Item 6	3.153	0.366								
Item 7	5.319	0.702	1.274	0.403						
Item 9	-0.682	0.576							-2.812	0.548
Item 10	1.352	0.205			0.242	0.089				
Item 11	4.521	0.710								
Item 12	5.132	0.591			-0.497	0.216				

	Sample 2									
			Covariate effects							
	Intercept		Gender		Age		White		Visit	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Class membership										
P(High-symptom)	0.796	0.621	0.740	0.266	0.294	0.098	0.681	0.310	-0.758	0.273
High-symptom class										
Item 1	-0.382	0.245								
Item 2	-1.417	0.659			0.285	0.132				
Item 4	1.410	0.277								
Item 5	-5.750	1.116	-0.644	0.267					-0.801	0.255
Item 6	0.518	0.214								
Item 7	0.202	0.247								
Item 9	-3.505	0.662							-2.360	0.402
Item 10	-2.673	0.959			0.366	0.187				
Item 11	1.896	0.311								
Item 12	3.073	0.492								
Low-symptom class										
Item 1	2.173	0.219								
Item 2	2.014	0.693			0.285	0.132				
Item 4	15.000									
Item 5	-1.134	0.436	-0.644	0.267					-0.801	0.255
Item 6	2.323	0.236								
Item 7	4.046	0.512								
Item 9	-0.299	0.576							-2.360	0.402
Item 10	1.812	0.602			-0.116	0.108				
Item 11	4.039	0.471								
Item 12	5.004	0.785								

Figure 1.
Summary of all models fitted in Chapter 2.

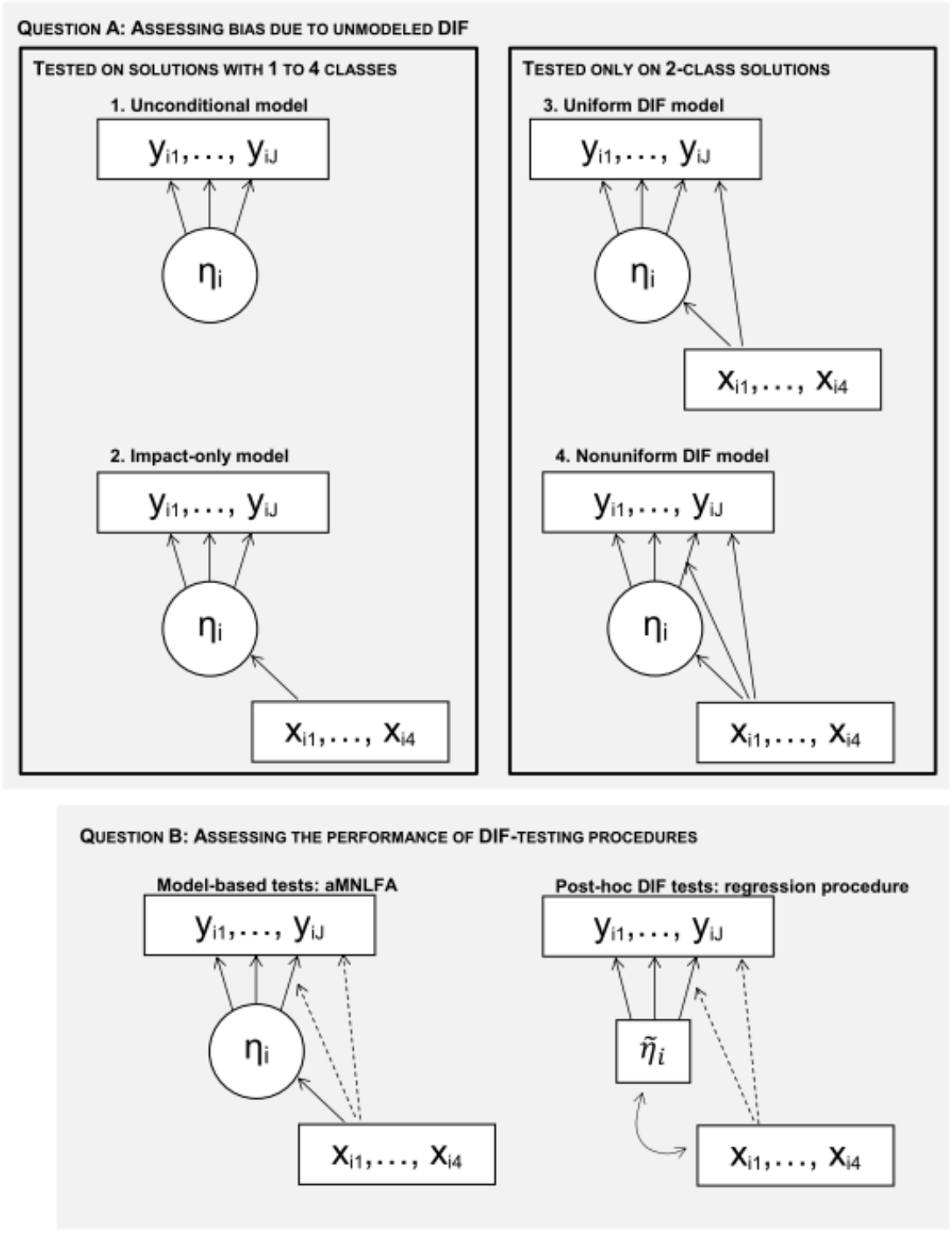

Figure 2. Logit parameter estimates $\hat{\delta}_{81k}$ in classes 1 and 2 across all models.

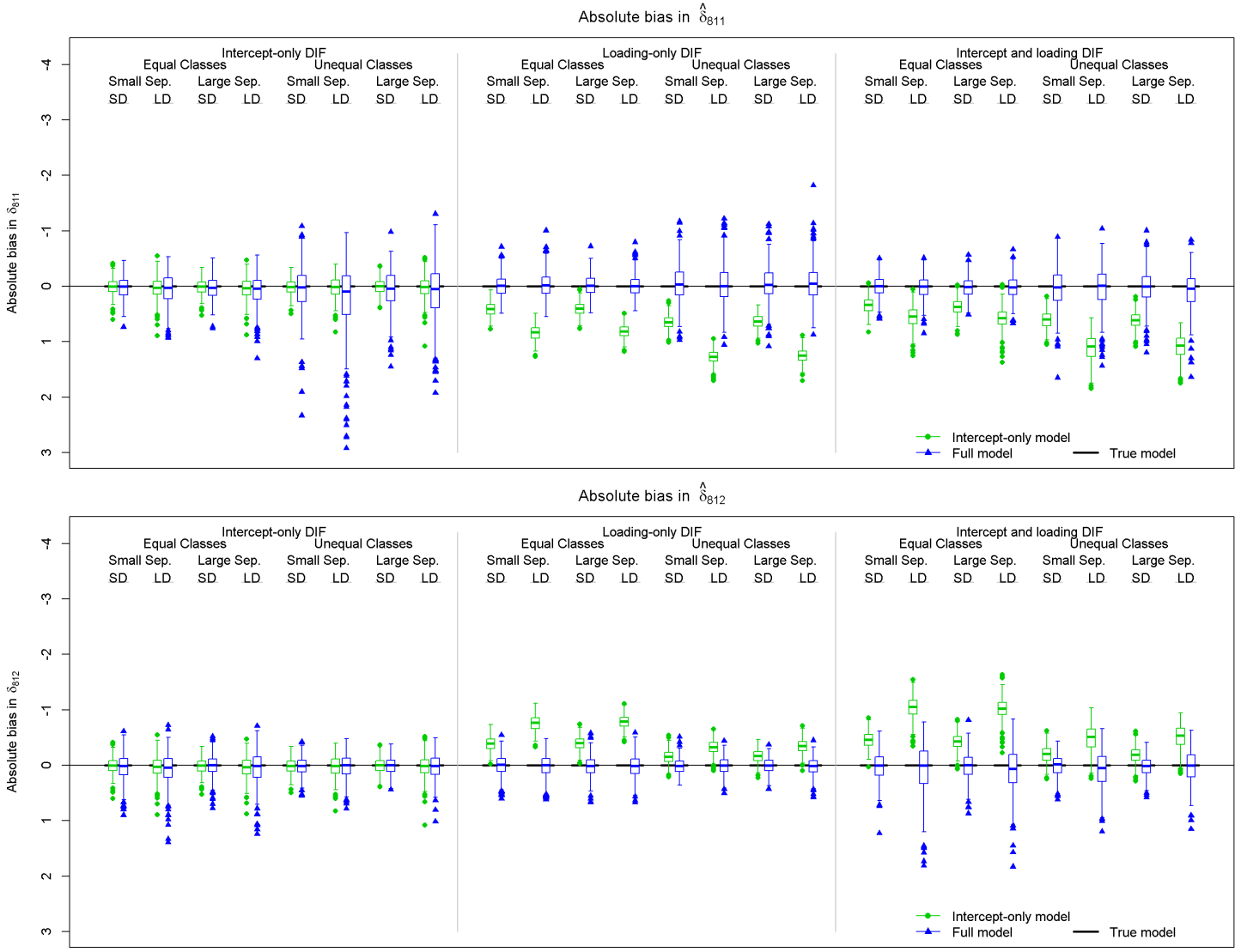

Figure 3.

Average baseline endorsement probabilities $\hat{\mu}_{0jk}$ in large DIF conditions.

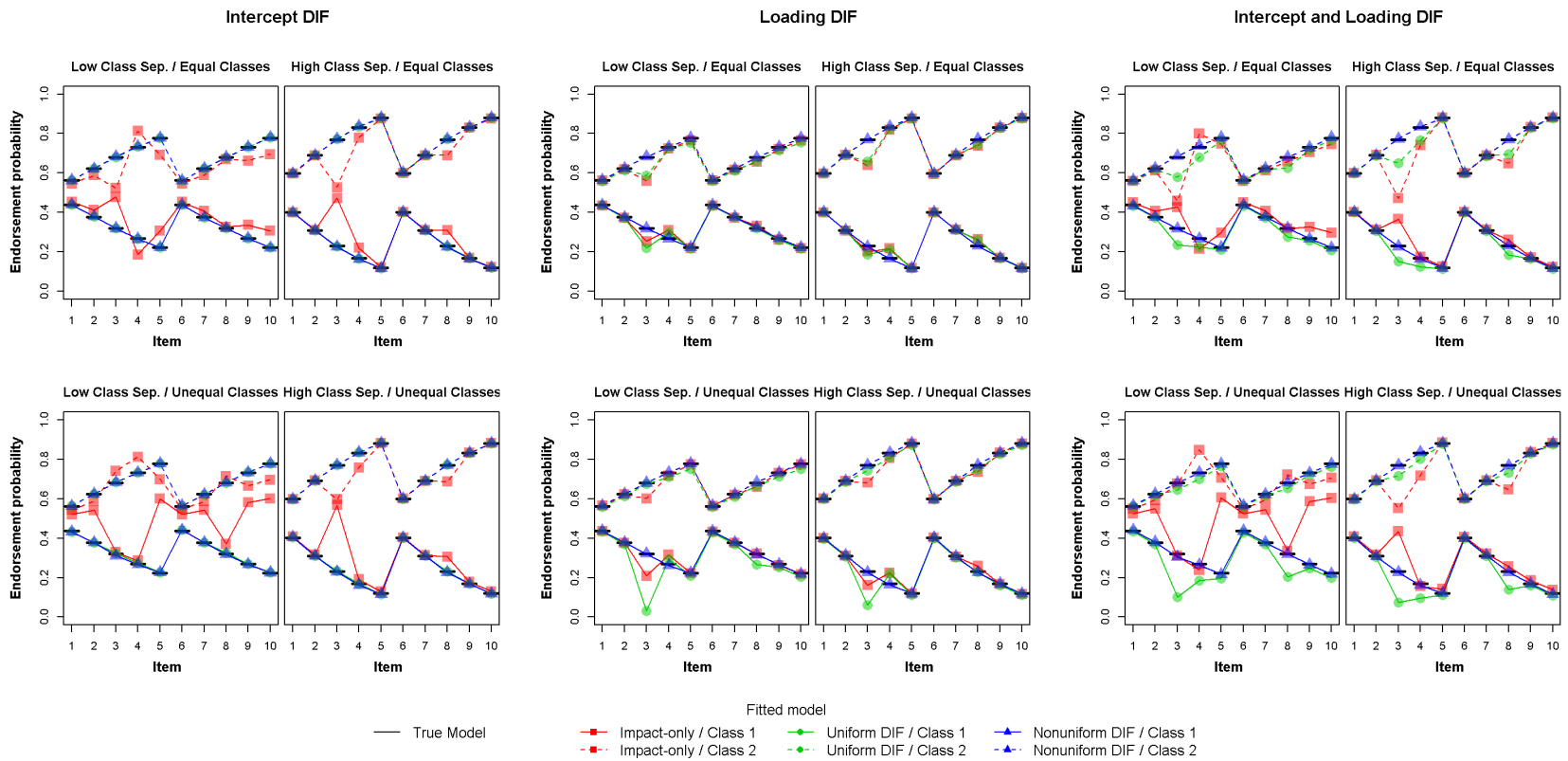

Figure 4.

Average estimates of individual predicted probabilities $\hat{\mu}_{ijk}$ in large DIF conditions.

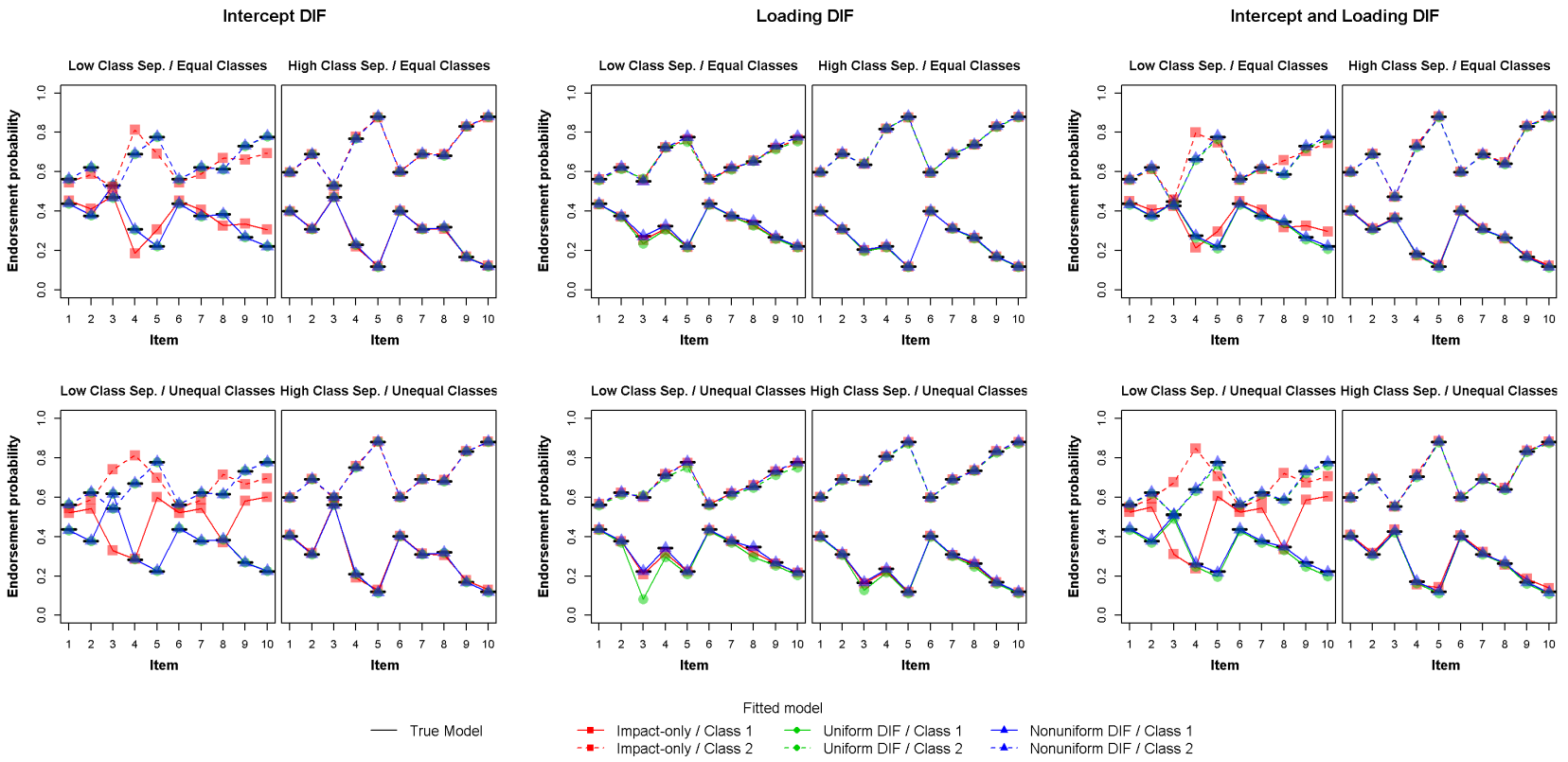

Figure 5.

Distribution of randomly-selected individual predicted probabilities $\hat{\mu}_{i3}$ in large DIF conditions.

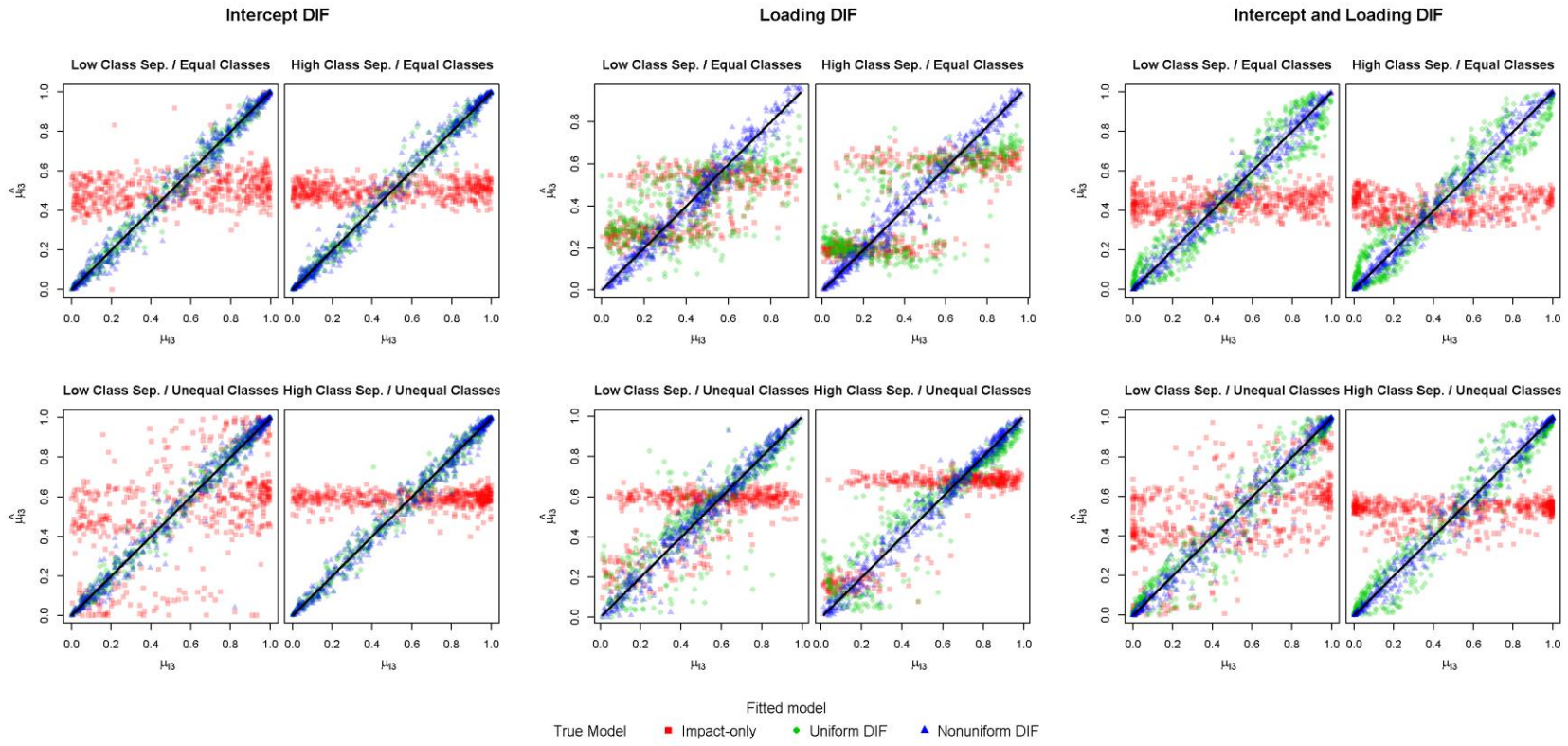

Figure 6: Estimates of prevalence of membership to class 1 across all models.

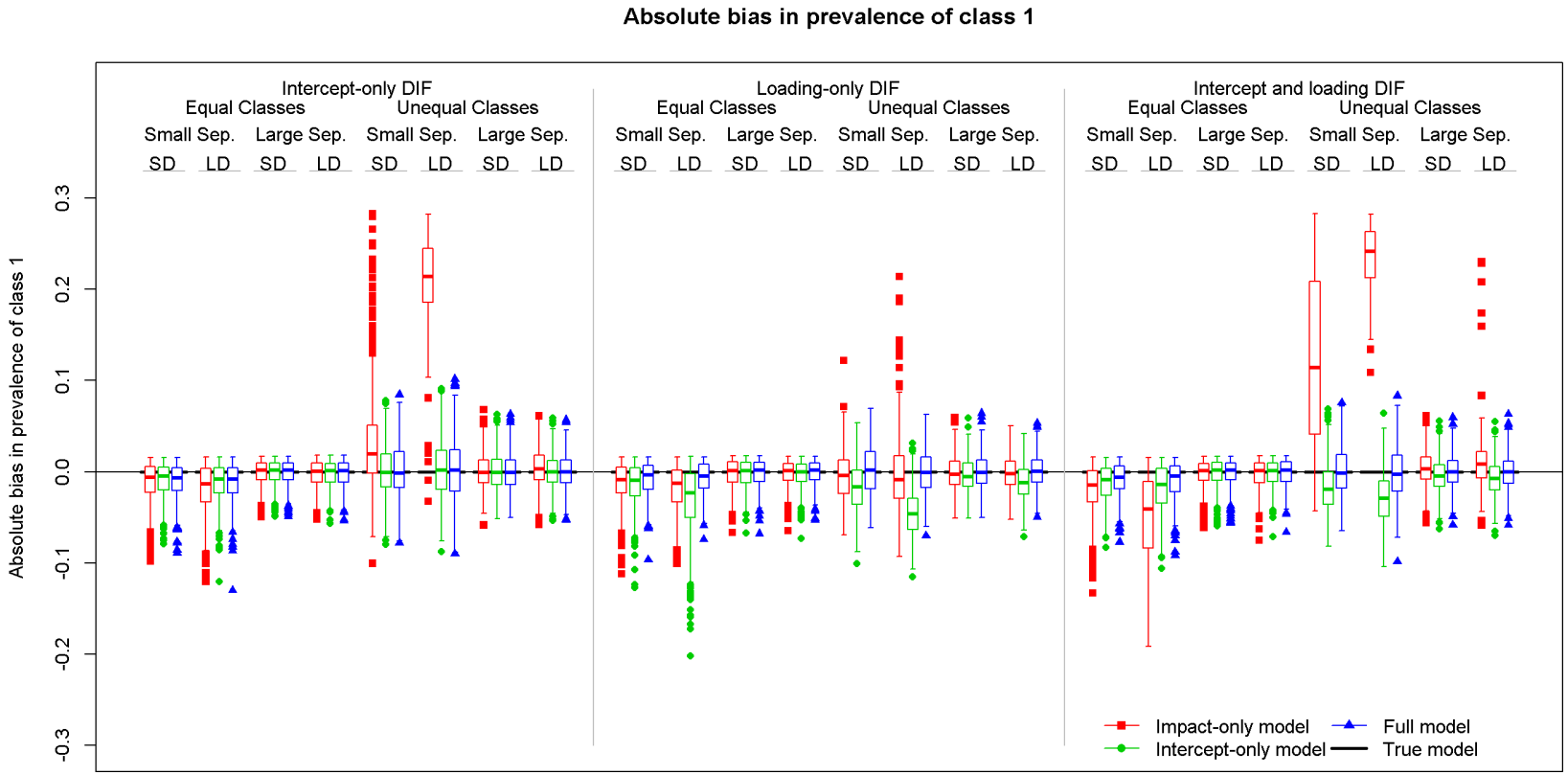

Figure 7. Model-implied endorsement probabilities for both the impact-only and full models.

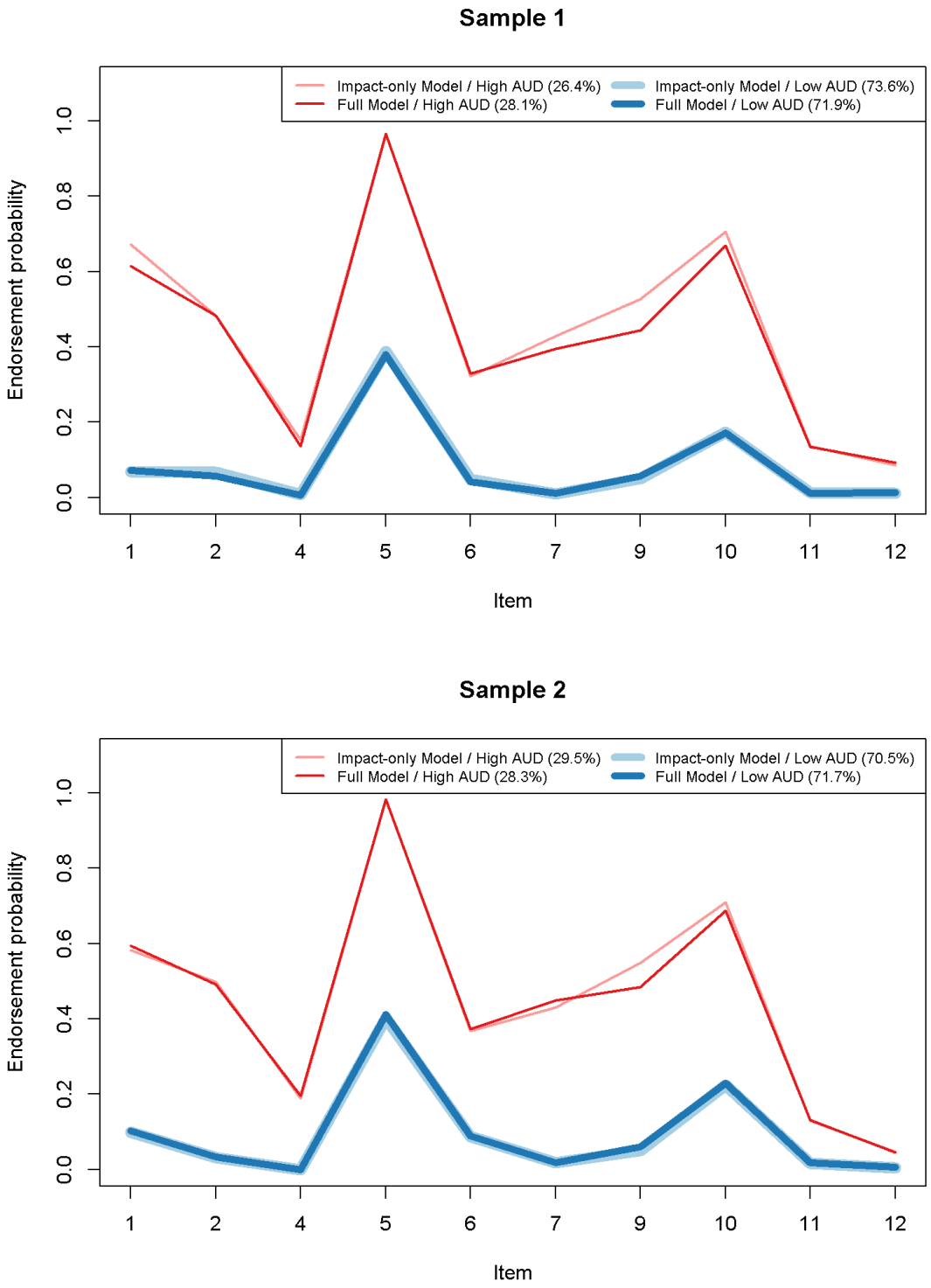

Figure 8. Model-implied endorsement probabilities for male and female subjects under the full model.

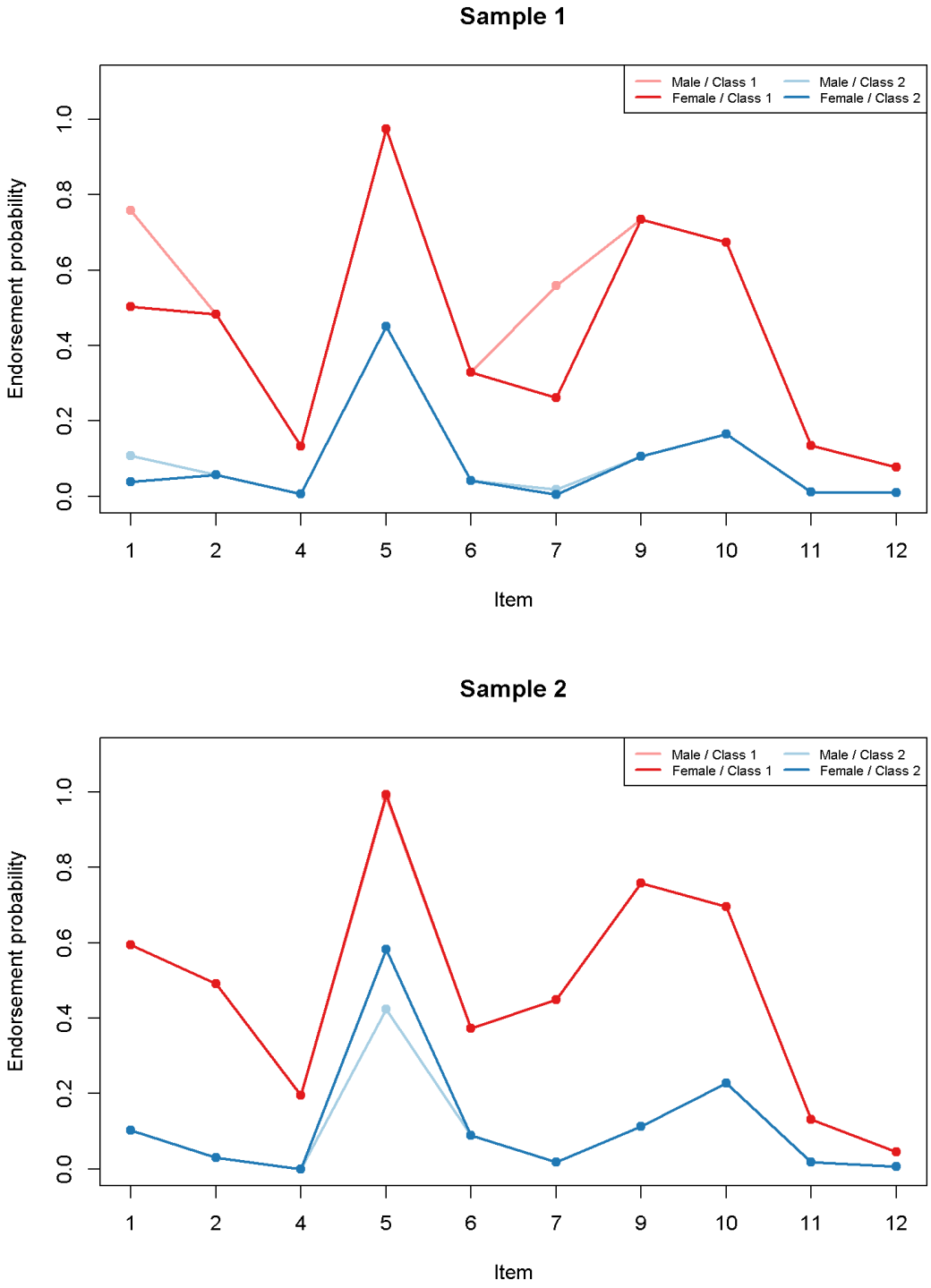

Figure 9. Model-implied endorsement probabilities for subjects of different ages under the full model.

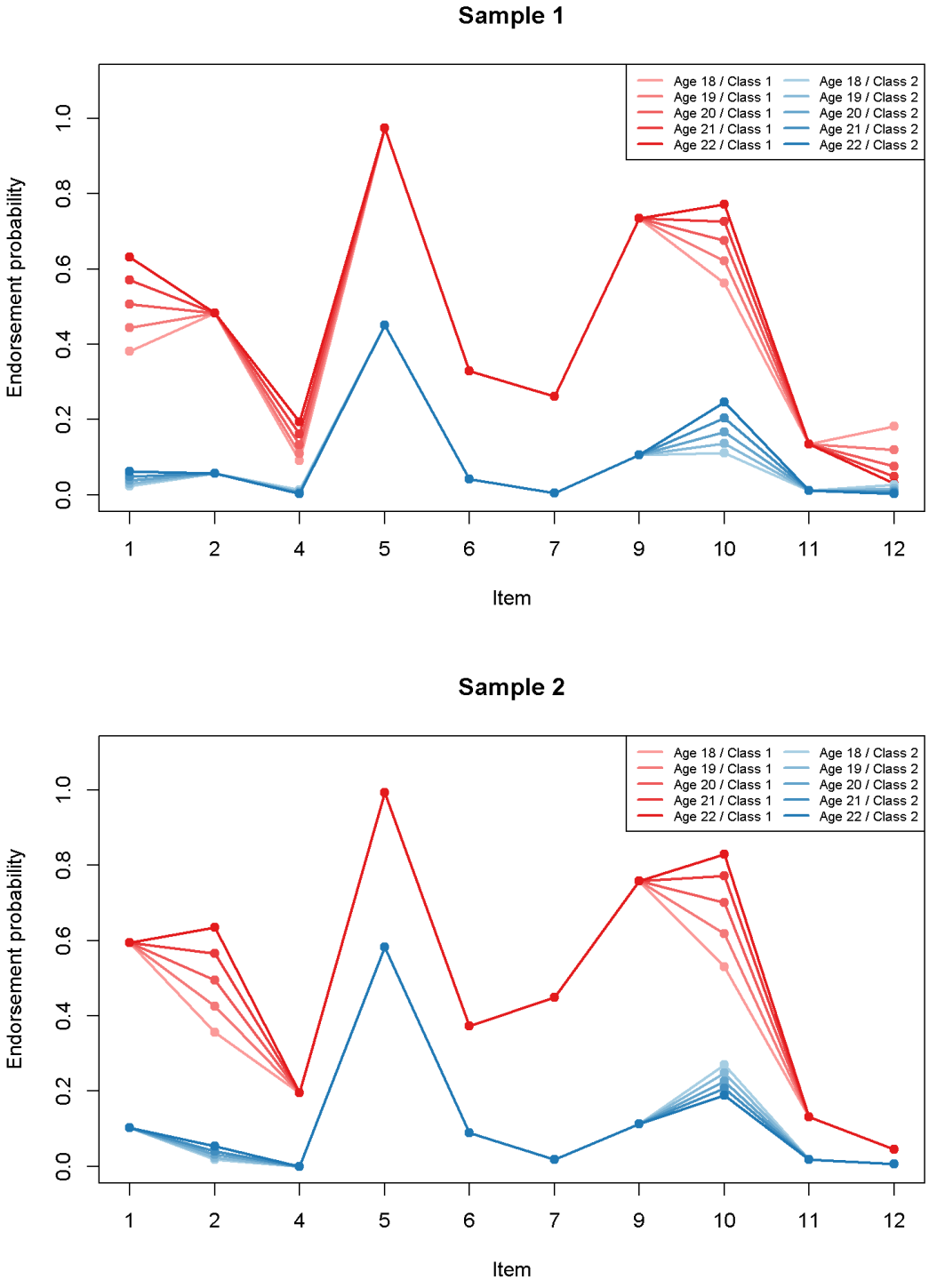

Figure 10. Model-implied endorsement probabilities for subjects seen at visit 1 and visit 2 under the full model.