USES OF CLASSIFICATION ERROR PROBABILITIES IN THE THREE-STEP APPROACH TO ESTIMATING COGNITIVE DIAGNOSIS MODELS

BY CHARLES JOSEPH IACONANGELO

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Dr. Jimmy de la Torre

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE DISSERTATION

Uses of Classification Error Probabilities in the Three-Step Approach to Estimating Cognitive Diagnosis Models

by Charles Joseph Iaconangelo

Dissertation Director: Dr. Jimmy de la Torre

Classification error probabilities (CEPs) are estimates of the amount of misclassification in the measurement model conditional on the true latent class memberships. CEPs can be used in several ways to improve the inferences drawn from cognitive diagnosis models (CDMs). To develop methodologies that facilitate the use of CDMs in practical research, this dissertation uses CEPs to accomplish three objectives: (1) to examine the conditional classification accuracy and generalizability of a cognitively diagnostic assessment; (2) to introduce correction weights that can improve a three-step approach for latent-class regression, which relate latent class memberships to predictor variables, and (3) to apply the same correction weights to select the best subset of predictor variables in the context of latent-class regression.

In the first study, an application of CEPs fills a gap in literature on CDM validity by serving as an index of classification accuracy conditional on the latent

class memberships. This index can also be extended to predict the classification accuracy of the assessment for a different population. Results show that the proposed index not only recovers the empirical values, but outperforms existing procedures based on the Monte Carlo approach.

In the second study, CEPs are used to improve the inferences in latent-class regression. Compared to the one-step procedure, which estimates the measurement model and regression parameters simultaneously, the three-step procedure is desirable from an applied researchers perspective because it simplifies latent-class regression by implementing the estimations involved in separate steps. However, it also leads to parameter estimation bias. This study uses CEP-derived weights to improve parameter estimation in various types of latent-class regression.

Finally, the third study extends the latent-class regression in the second study by incorporating a regularization procedure that permits variable selection. Results show that incorporating measurement error (as measured by CEP) in the variable selection process leads to a subset of nonredundant variables that more clearly shows the relationship between predictors and examinee classification. In addition, compared to the standard approach, using the CEP-based weights leads to fewer instances of estimation noncovergence.

With a general aim to address needs in conditional classification accuracy, correcting bias in parameter estimation, and high-dimension variable selection in the context of CDMs, this dissertation uses CEPs to accomplish three objectives: (1) to examine the conditional classification accuracy and generalizability of the assessment, (2) introduce correction weights for the three-step approach that result in improved parameter estimation, and (3) apply these correction weights to regularized latent-class regression to select variables.

Acknowledgements

I would like to acknowledge the time and effort invested in me by my advisor and mentor, Dr. Jimmy de la Torre. Thank you, Jimmy, for giving an undergrad liberal arts major a chance in a statistical PhD program. I never imagined that I would be doing half of what I'm doing now.

Likewise, I owe a huge debt of gratitude to my other mentor, Dr. Drew Gitomer, for all the subtle and not-so-subtle instruction I received. In spite of all the measurement work I did for him, I still feel like I got way more out of him than he ever got out of me.

My colleagues made the office an enjoyable and productive place to work - to Nate, Wenchao, Ragip, Lokman, Mehmet, Soo Lee, Eugene, Simon, and Scarlett - thank you for always being willing to discuss my research and offer your input.

Mom and Dad, I would never have done this without your help. As long as I can remember, you encouraged me to pursue my education.

Amelia, mi amor, thank you for supporting me through all of this. I won't forget it.

Table of Contents

Abstract ii

Acknowledgements iv

List of Tables ix

List of Figures x

1. Introduction 1

1.1. References 5

2. Conditional Classification Accuracy of Cognitive Diagnosis Assessments 7

2.1. Introduction 7

2.2. Cognitive Diagnosis Models 10

2.3. Conditional Classification Accuracy 13

2.3.1. Classification Accuracy for a Different Attribute Distribution 14

2.3.2. Off-Diagonal Entries of the Matrix of Classification Error Probabilities 15

2.4. Simulation Study 16

2.4.1. Design 16

2.4.1.1. Parametric Monte Carlo Approach 18

2.4.1.2. Classification Accuracy with a Different Attribute Distribution 19

2.4.2. Analysis 19
2.5. Results 20
2.5.1. Uniform Attribute Distribution 20
2.5.1.1. Comparing the Proposed Index and the Monte Carlo Approach 21
2.5.1.2. Latent Class $\boldsymbol{\alpha}_l = 1100$ 22
2.5.2. Higher-order Attribute Distribution 24
2.5.2.1. Comparing the Proposed Index and the Monte Carlo Approach 27
2.5.2.2. Latent Classes $\boldsymbol{\alpha}_l = 11000$ and $\boldsymbol{\alpha}_l = 00101$ 28
2.5.3. Classification Accuracy with a Different Attribute Distribution 31
2.6. Empirical Example 33
2.7. Discussion 34
2.8. References 36

3. Three-Step Estimation of Cognitive Diagnosis Models with Covariates 39
3.1. Introduction 39
3.2. Cognitive Diagnosis Models 42
3.2.1. The G-DINA Model 42
3.2.2. Latent Class Assignment 44
3.2.3. Matrix of Classification Error Probabilities 45
3.3. Modeling the Relationship between Covariates and Latent Classification 46
3.3.1. The One-Step Approach 46
3.3.2. The Uncorrected Three-Step Approach 48

3.3.3. The Three-Step Procedure with Latent-class Level Correction Weights . . . 49
3.3.3.1. Sample-Level Correction Weights . . . 49
3.3.3.2. Posterior-distribution Level Correction Weights . . . 50
3.3.3.3. Three-Step Approach with Attribute-Level Correction Weights . . . 51
3.4. Simulation Study to Evaluate the Performance of the Correction Weights . . . 53
3.4.1. Design . . . 54
3.4.2. Analysis . . . 56
3.4.3. Results . . . 58
3.4.3.1. Overall Bias and RMSE . . . 58
3.4.3.2. Bias and RMSE at Individual Parameter Level . . . 59
3.4.3.3. Separation of Likelihood . . . 63
3.4.3.4. Effective Sample Size . . . 65
3.5. Empirical Example . . . 66
3.6. Discussion . . . 69
3.7. References . . . 70

4. Variable Selection in the Three-Step Approach to Modeling Cognitive Diagnosis Models and Covariates: The Latent-Class Lasso 74
4.1. Introduction . . . 74
4.2. Cognitive Diagnosis Models . . . 76
4.3. Modeling the Relationship between Covariates and Latent Classification . . . 79
4.3.1. The One-Step Approach . . . 79
4.3.2. The Uncorrected Three-Step Approach . . . 80

4.3.3. The Three-Step Approach with Correction Weights 81
4.4. Variable Selection with the Lasso 82
4.4.1. The Latent-Class Lasso 84
4.5. Evaluating the Performance of the Correction Weights via Simulation Study 84
4.5.1. Design 85
4.5.2. Analysis 87
4.5.3. Results 89
4.5.3.1. Sparsity 89
4.5.3.2. Relevant Predictors Dropped 89
4.5.3.3. Correct Selection Rate 91
4.5.3.4. Overall ARMSE and ABIAS 92
4.5.3.5. ARMSE and ABIAS of the Relevant Predictors . 93
4.5.3.6. Individual Parameter Estimate 95
4.6. Empirical Example 97
4.7. Discussion 99
4.8. References 101

5. Conclusion 105

List of Tables

2.1. Uniform Attribute Distribution 25

2.2. Higher-Order Attribute Distribution 30

2.3. Predicting Classification Accuracy for a Different Population . . . 32

2.4. MCMI-III 34

3.1. Ten-Item Q-matrix 54

3.2. ABIAS - Attribute-Level Logistic Regression 60

3.3. ARMSE - Attribute-Level Logistic Regression 61

3.4. Bias - Attribute-Level Logistic Regression 62

3.5. RMSE - Attribute-Level Logistic Regression 63

3.6. Replications with Separated Likelihood 64

3.7. MCMI-III 68

4.1. Ten-Item Q-matrix 86

4.2. Sparsity 90

4.3. Proportion of Relevant Predictors Dropped 91

4.4. Correct Selection Rate 92

4.5. ARMSE and ABIAS 94

4.6. ARMSE and ABIAS of Relevant Predictors 95

4.7. Comparison of Parameter Estimates 97

4.8. MCMI-III 99

List of Figures

2.1. Latent class proportions under the higher-order attribute distribution 17
2.2. Mean difference of the proposed index under the uniform attribute distribution 21
2.3. RMSD of the proposed index under the uniform attribute distribution 22
2.4. Mean difference of both approaches under the uniform attribute distribution 23
2.5. RMSD of both approaches under the uniform attribute distribution 24
2.6. Mean difference of proposed index under the higher-order attribute distribution 26
2.7. RMSD of proposed index under the higher-order attribute distribution 27
2.8. Mean difference under the higher-order attribute distribution . . . 28
2.9. RMSD under the higher-order attribute distribution 29
3.1. Relationship between ABIAS and the PCV 66

Chapter 1
Introduction

Cogitive diagnosis models (CDMs) are a type of latent class model that offer several advantages over item response theory (IRT) models. In the typical application, IRT-based assessments are designed to measure a unidimensional or low-dimensional latent variable, which is usually interpreted as a general construct. The scores are subsequently used to rank examinees. By contrast, a cognitively diagnostic assessment (CDA) focuses on categorizing examinees according to mastery of specific attributes. In the educational measurement context, these would be academic skills. Alternatively, in a clinical psychology setting, these attributes could be construed as disorders. For diagnostic or formative assessment, these discrete (usually binary) attributes may be more relevant to the purpose of the assessment than the standard unidimensional construct from IRT (Junker & Sijtsma, 2001).

In spite of the potential advantages of implementing the CDM framework with assessments in formative classroom or clinical diagnostic settings, there remains the need to form validity arguments in favor of using the scores. Per the recommendations of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), if these test scores (or classifications) are to play a role in instructional or treatment decisions, then test users should qualify the decision inference (Kane, 2013). Qualifying the conclusions requires understanding the amount of measurement error

in the CDM classifications. In particular, researchers must be able to estimate the classification accuracy of the assessment if they are to implement a decision rule and a cost-benefit analysis. Because CDMs classify examinees based on fine-grained variation components, and because these classifications form the basis of inferences, the estimates of classification accuracy must be fine-grained as well. Specifically, the estimates of classification accuracy should be made at the latent-class level, which then allows for the end-user to study the validity of inferences made regarding the specific latent classes of interest. Furthermore, testing standards (American Educational Research Association et al., 2014) also demand an understanding of how the classification accuracy of an assessment generalizes to other populations of interest. This is a crucial part of the validity argument for CDMs.

A great deal of research has aimed to develop methodologies for CDAs, administering the assessment, selecting the correct model, validating the Q-matrix, and evaluating the accuracy and consistency in classifying examinees. However, there is a relatively paucity of literature on how to use these classifications in more exploratory research. In the field of educational measurement, research questions often focus on the relationship between student achievement and background variables. For example, this may occur in projects designed specifically to investigate student learning via CDMs, or it may occur in the context of secondary research that re-purposes already implemented assessments by relating the results to covariates of interest. The findings often have policy implications, which underscores the need to optimize the procedures available for researchers. To relate classifications to covariates, typically the most appropriate procedure is the one-step approach, which simultaneously estimates the measurement model (in this case a CDM) and the structural model (a regression model). In the literature on CDMs, this approach has been implemented in different ways. For one,

it has been used to evaluate differential item functioning (Park & Lee, 2014), where covariates can affect the probability of examinees answering a particular item correctly. For another, the one-step approach has also been implemented such that covariates affect the probability of examinees mastering a particular attribute, referred to as differential skill functioning. The latter is particularly important for exploratory researchers who are looking at the relationship between, for example, student learning and student or school covariates. In the clinical psychology setting, it allows researchers to investigate how well patient demographic information predicts the presence of personality disorders. However, for reasons of interpretability, the one-step may be less than ideal. For example, it is unclear if the attribute specification, Q-matrix specification, and examinee classification should depend on the covariates included in the model. The validity of inferences may be called into question. Therefore, although a one-step approach leads to best estimates (i.e., lowest bias) of the regression parameters when it is appropriate, other modeling approaches may be needed for researchers doing exploratory work that investigates the relationship between background variables and classifications.

Another practical constraint on modeling approaches is the simple fact that often times examinee item responses are not archived, or at least are not released to secondary researchers. These circumstances require the three-step approach, which separates the fitting of the CDM and the regression model. However, developments of the three-step approach from latent class models have not been applied to CDMs. Instead, the only options currently available for researchers relating CDM classifications to covariates involves treating the classifications as observed dependent variables. Ignoring measurement error in subsequent statistical analysis can have serious consequences, whether the error occurs in the independent variables (i.e., error-in-variables regression), or in the dependent variable (i.e.,

latent regression). The measurement error affects not only the regression parameter estimates (Bakk, Tekle, & Vermunt, 2013; Vermunt, 2010), but also the variable selection process (Bakk, Oberski, & Vermunt, 2013). The measurement error should be included in the procedure to ensure the best possible estimates and thus the most accurate inferences about the relationship between CDA-based classifications and background variables.

Objectives

The first issue to resolve, then, is how to quantify the measurement error in CDA. In these studies, measurement error is quantified via the matrix of classification error probabilities (CEPs). This is essentialy a $2^K \times 2^K$ contingency table that estimates the probability of being classified in latent class s given the true latent class is l. These estimates of measurement error are then used throughout all three studies to answer different research questions.

Study 1 interprets the CEPs directly to estimate the conditional classification accuracy via a proposed index of length 2^K. Furthermore, taking the weighted sum of the proposed index can estimate the overall classification accuracy of the assessment for any population of interest. Understanding the generalizability of the classification accuracy is another important part of the validity argument. The proposed index will be compared to a parametric Monte Carlo approach via simulation study. Overall, in study 1 the matrix of CEPs is used to draw conclusions about the overall accuracy of the assessment that are a key part of the validity argument for a CDA.

Studies 2 and 3 incorporate the matrix of CEPs into procedures for relating the classification to background variables. The CEPs are used to adjust three-step procedures such that the dependent variable is treated as latent rather than observed. This entails adjusting for measurement error by incorporating the CEPs

as weights in the objective function (i.e., the log-likelihood) of the regression model. Another type of weights related to the matrix of CEPs is proposed as well. In this study, the different correction weights and the standard approach to the three-step procedure will be used to estimate the multinomial logistic regression coefficients by regressing the assigned latent-classes onto the covariates. It also will evaluate regressing attribute classifications onto covariates. The aim of study 2 is to show that correction weights derived from the matrix of CEPs, and weights developed in this dissertation, can be used to adjust for measurement error and provide estimates of the relationship between covariates and classifications with less bias and lower RMSE.

Finally, Study 3 relies on the same correction weights and the same corrected three-step procedure developed in Study 2, but extends adjustments to include a variable selection process. The L_1 penalty is incorporated in the latent attribute-level regression log-likelihood along with the correction weights, and the regularized regression with cross validation is used to select variables. The study is designed to examine how adjustments for measurement error alter conclusions about the relationship between covariates and attribute mastery. In short, all three studies quantify measurement error via the matrix of CEPs and use that information to draw better conclusions about the assessment, thus improving the validity of research and decisions made involving CDA.

1.1 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2013). Relating latent class assignments to external variables: Standard errors for corrected inference. *Sociology*, *83*, 173-178.

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*, 272-311.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Park, Y., & Lee, Y. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Applied Psychological Measurement*, *38*, 376-390.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Chapter 2

Conditional Classification Accuracy of Cognitive Diagnosis Assessments

2.1 Introduction

Cognitively diagnostic assessments (CDAs) are often proposed as an alternative to item response theory (IRT)-based tests for diagnostic or formative purposes. IRT assessments typically score and rank examinees on a continuous unidimensional latent trait for what are often high-stakes purposes (Junker & Sijtsma, 2001). CDAs, on the other hand, measure multiple skills, referred to as attributes, which examinees are classified as either having mastered or not mastered (de la Torre & Lee, 2010). This makes them well-suited to low-stakes applications, such as formative assessments that are designed to directly support teaching and learning (DiBello & Stout, 2007). As the number of e-learning and intelligent tutoring systems proliferates (Newman, Bryant, Stokes, & Squeo, 2013), the corresponding methodology for integrating CDA as part of the curriculum has seen promising developments (e.g., Ye, Fellouris, Culpepper, & Douglas, 2016). Alternatively, CDAs have been shown to offer advantages over current practice in personalized clinical assessment, as demonstrated in de la Torre, van der Ark, and Rossi (2015), where the cognitive diagnosis model (CDM) framework was applied to the Millon Clinical Multiaxial Inventory III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009). Regardless of the application, however, the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) require an

in-depth understanding of the accuracy of inferences to justify implementation. Evaluating the accuracy of the classifications plays a crucial role in evaluating the claims made from a CDA (Kane, 2013). Indices that estimate the classification accuracy may allow test users to better determine the strength of the evidence - the classifications - for their proposed use. Test users also may be interested in making decisions regarding particular latent classes, and in that case, having estimates of the classification accuracy of latent classes would be of particular importance to the validity argument.

The salience of estimating score/classification accuracy has elicited a great deal of research in the field of test theory, and an exhaustive review will not be attempted here. In the CDM literature, one of the first contributions to this topic was a modification to the Kullback-Leibler matrix used in the cognitive diagnostic index that created attribute-level discrimination indices related to the correct classification of examinee mastery (Henson, Roussos, Douglas, & He, 2008). The index proposed by Cui, Gierl, and Chang (2012), $\hat{P}_a$, estimates the classification accuracy of latent classes marginalized to the test level. That is, the index estimates how accurately the assessment classified examinees overall, and for brevity will be referred to as the test-level classification accuracy. With a sufficiently large sample size, the sampling distribution of the index allows for the computation of standard errors as well, providing an estimate of the lower and upper bound on the accuracy (Cui et al., 2012). Unfortunately, because the index relies on the item response function, the adequacy of the CDM must be established. The impact of model misspecification was not investigated.

More recently, an index of classification accuracy of latent classes marginalized to the test level, $\hat{\tau}$, was developed by Wang, Song, Chen, Meng, and Ding (2015), who in the same paper proposed an accuracy index conditional on the attribute, $\hat{\tau}_k$. Corresponding standard errors were developed but not studied. The index $\hat{\tau}$,

like $\hat{P}_a$, provides an estimate of the proportion of correctly classified examines, yet unlike $\hat{P}_a$, $\hat{\tau}$ is computed from the examinee posterior distributions and requires much simpler calculations. Wang et al. (2015) compared the performance of $\hat{\tau}$ and $\hat{P}_a$ via simulation study, and the results indicated similar recovery rates of the empirical values (i.e., recovery of generated examinee classifications). It should be noted that the pairs of indices were investigated for the deterministic input, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model only. Furthermore, the $\hat{\tau}$ and $\hat{P}_a$ indices only estimate the accuracy of the CDA for the given sample, leaving the generalizability of the estimated accuracy unknown.

Although examinee classification accuracy of a diagnostic assessment is considered fundamental to the validity of its use, there are still gaps in the literature. Approaches for measuring the classification accuracy of the latent classes of CDA have not been developed, limiting study of the validity of inferences made about specific latent classes based on the assessment. Classification accuracy may be high overall, but low for some attribute patterns of interest. This work addresses a shortcoming of existing methodologies by proposing an index of the examinee classification accuracy that, in addition to being relatively straightforward to compute, estimates accuracy conditional on the latent class rather than marginalized to the test level. This can inform the practitioner of the effectiveness of the assessment in classifying specific latent classes of interest. Additionally, taking the weighted sum of the index over the latent classes returns an estimate of the test-level classification accuracy for any attribute distribution of interest. By generalizing the classification accuracy to other examinee populations, the index can provide a more in-depth look at the validity of the test across various situations (Kopriva, Thurlow, Perie, Lazarus, & Clark, 2016).

The rest of the manuscript is organized accordingly: The next section provides

background on CDMs, specifically the G-DINA model. Following that, the conditional classification accuracy index and some related techniques are presented. After that is the design and analysis of a simulation study that compares the proposed index to the parametric Monte Carlo approach and empirical values. An extension of the study examines how well the index can estimate the test-level classification accuracy for a different population. A brief example using real data is then provided. Finally, directions for future research are discussed.

2.2 Cognitive Diagnosis Models

The wide variety of CDMs in the literature can be organized according to how attributes are assumed to interact. For conjunctive models, of which the DINA model is the most well-known, only examinees that have mastered all attributes specified in the item are expected to answer correctly. On the other hand, disjunctive models, such as the deterministic, noisy "or" gate (DINO; Templin & Henson, 2006) model, expect any examinee that has mastered at least one of the item-attributes to answer correctly. Additive models, another class of CDMs, do not model attribute interactions; that is, mastery of each attribute does not affect the contribution of the others. Examples include the additive CDM (*A*-CDM; de la Torre, 2011), the logistic linear model (LLM; Maris, 1999), and the reduced reparametrized unified model (R-RUM; Roussos, Templin, & Henson, 2007). Simulation study results suggest that the additive nature of these models minimizes the negative impact of model misspecification, compared to the conjunctive and disjunctive models (Ma, Iaconangelo, & de la Torre, 2016).

The G-DINA Model

To eliminate model misfit as a factor in the following simulation study, and to ensure the results are not restricted to a particular class of CDM, a general, or saturated, model is used here. General models make no assumptions about how the attributes interact and thus subsume the constrained varieties. Several general models have been introduced in the literature - the general diagnostic model (GDM; von Davier, 2008), the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009), and the generalized DINA model (G-DINA; de la Torre, 2011). The G-DINA model is used throughout this study.

All of the aforementioned models specify via a Q-matrix (Tatsuoka, 1983) which of the K attributes are required by each of the J items. Summing row j of the Q-matrix yields K_j^*, the number of attributes required by item j. Letting $l = 1, \ldots, 2^K$ denote the latent classes, the examinee attribute vector is written $\boldsymbol{\alpha}_l = \{\alpha_{l1}, \ldots, \alpha_{lK}\}$, where the kth element of the vector is equal to one or zero, depending on whether the examinee has mastered or not mastered that attribute, respectively. To compute the item response function, denote the reduced attribute vector containing only the required attributes for item j by $\boldsymbol{\alpha}_{lj}^*$, where $l = 1, \ldots 2^{K_j^*}$. Furthermore, let $P(\boldsymbol{\alpha}_{lj}^*)$ be the probability that an examinee with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ answers item j correctly. The item response function of the G-DINA model is computed as,

$$P(\boldsymbol{\alpha}_{lj}^*) = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \phi_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}.$$

In this equation, ϕ_{j0} is the intercept for item j, ϕ_{jk} is the main effect due to α_k, and $\phi_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$; the interaction effect due to $\alpha_1, \cdots, \alpha_{K_j^*}$ is denoted by $\phi_{j12\ldots K_j^*}$.

Marginal maximum likelihood estimation of the item parameters, $\boldsymbol{\phi}$, requires the likelihood, written as

$$L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_l)^{X_{ij}}[1 - P_j(\boldsymbol{\alpha}_l)]^{1-X_{ij}}.$$

The likelihood is marginalized over the latent class proportions, $P(\boldsymbol{\alpha}_l)$, yielding

$$L(\boldsymbol{X}) = \prod_{i=1}^{N} \sum_{l=1}^{2^K} L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l),$$

the log of which is then optimized with respect to $\boldsymbol{\phi}$ via the Expectation-Maximization algorithm (Dempster, Laird, & Rubin, 1977). For more details on this, see de la Torre (2009b, 2011). The likelihood of $\boldsymbol{X}_i$ and latent class proportions are then used to compute the posterior distribution of examinee i, written as

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) \propto L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l).$$

The examinee posterior distribution is subsequently used to assign examinees to latent classes. Averaging over the examinee posterior distributions yields the estimated latent class proportions, $\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)/N = \hat{P}(\boldsymbol{\alpha}_l)$. This can be considered an estimate of the joint distribution of the attributes.

In this study, the examinee latent class assignment is done by the maximum a posteriori (MAP) classification method (Huebner & Wang, 2011), and the estimated attribute pattern of examinee i is denoted by $\hat{\boldsymbol{\alpha}}_i$. The possible latent class assignments are denoted by $\boldsymbol{\alpha}_s$, where $s = 1, \ldots, 2^K$. Note that $\boldsymbol{\alpha}_l$ is considered the possible true classification, whereas $\boldsymbol{\alpha}_s$ are the latent classes that may be realized according to the assignment rule.

2.3 Conditional Classification Accuracy

The posterior and latent class assignments are used to calculate the classification error probabilities, which are quantified by writing the estimated value conditional on the true value (Vermunt, 2010). This matrix of conditional classification error probabilities is calculated as

$$P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_{\boldsymbol{s}}]}{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)}, \tag{2.1}$$

where $I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_{\boldsymbol{s}}]$ is an indicator function equal to 1 when the estimated attribute pattern of examinee i is equal to latent class $\boldsymbol{\alpha}_{\boldsymbol{s}}$, and zero otherwise. Thus, $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})$ can be interpreted as the proportion of examinees with true latent class membership $\boldsymbol{\alpha}_l$ assigned to classification $\boldsymbol{\alpha}_s$. This is a $2^K \times 2^K$ contingency table of examinee latent classification proportions.

The proposed index, $\hat{\tau}_l$, can be computed directly from $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})$. The index $\hat{\tau}_l$ is the estimated probability of correctly classifying an examinee in latent class l. This is equivalent to the diagonal of $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})$, and is computed as,

$$\hat{\tau}_l = P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X}) I[s = l]$$

where $I[s = l]$ is the indicator function equal to 1 when the latent class assignment s is the same as the "true" latent class l. This is essentially the main diagonal of a contingency table that was computed using the posterior distributions and modal latent class assignments. Thus, $\hat{\tau}_l$ is a vector of length 2^K, where each element corresponds to the estimated proportion of examinees from each latent class that were correctly classified. Computation of the proposed index is straightforward, only requiring basic mathematical operations using matrices readily available from the estimation procedure.

Observe that $\hat{\tau}_l$ can be weighted according to the estimated latent class proportions and summed to compute $\hat{\tau}$, the index of classification accuracy marginalized to the test-level that was introduced by Wang et al. (2015):

$$\hat{\tau} = \sum_{l=1}^{2^K} \hat{\tau}_l \times \hat{P}(\boldsymbol{\alpha}_l). \tag{2.2}$$

This index estimates the empirical value of τ, which is the observed proportion of examinees classified in the same latent class as the generating data. The empirical value of τ can be computed as

$$\tau = \frac{\sum_{i=1}^{N} I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]}{N},$$

where $I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]$ evaluates whether the estimated attribute vector matched the generated values.

2.3.1 Classification Accuracy for a Different Attribute Distribution

Computing $\hat{\tau}$ in this manner allows for greater flexibility than the formula proposed in Wang et al. (2015). In Equation 2.2, $\hat{\tau}_l$ can be weighted according to any joint distribution, allowing the researcher or test developer to predict the classification accuracy of the assessment for examinees drawn from a different attribute distribution. For example, the observed sample may be drawn from a population with a uniform attribute distribution, whereas the test developer may be interested in the classification accuracy of the assessment for a sample drawn from a population with a higher-order attribute distribution. Given $\hat{\tau}_l$, the classification

accuracy of the assesment for a different sample, $\hat{\tau}^*$, can be estimated by,

$$\hat{\tau}^* = \sum_{l=1}^{2^K} \hat{\tau}_l \times P^*(\boldsymbol{\alpha}_l), \tag{2.3}$$

where $P^*(\boldsymbol{\alpha}_l)$ corresponds to the assumed proportion of examinees in latent class l. The values of $P^*(\boldsymbol{\alpha}_l)$ reweight $\hat{\tau}_l$ to reflect the latent class proportions of the sample of interest. Computing $\hat{\tau}^*$ is of particular significance because a key component of the validity argument is understanding how the classification accuracy generalizes across other examinee populations of interest (Kane, 2013; Pellegrino, DiBello, & Goldman, 2016).

2.3.2 Off-Diagonal Entries of the Matrix of Classification Error Probabilities

Although the index from the diagonal of the matrix of conditional classification error probabilities is the focus of this study, the off-diagonal entries can provide the researcher with an understanding of how examinees are misclassified. The rows and columns of the matrix correspond to $P(\boldsymbol{\alpha}_l)$ and $P(\boldsymbol{\alpha}_s)$, respectively. Each row entry of $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})$ estimates which latent class examinees were classified as $\boldsymbol{\alpha}_s$, conditional on their true classification, $\boldsymbol{\alpha}_l$ (note that each row of the matrix sums to one). For example, when $K = 2$, the four entries in row one of $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})$ are $P(\boldsymbol{\alpha}_s = 00|\boldsymbol{\alpha}_l = 00)$, $P(\boldsymbol{\alpha}_s = 10|\boldsymbol{\alpha}_l = 00)$, $P(\boldsymbol{\alpha}_s = 01|\boldsymbol{\alpha}_l = 00)$, and $P(\boldsymbol{\alpha}_s = 11|\boldsymbol{\alpha}_l = 00)$. The second entry in the row, $P(\boldsymbol{\alpha}_s = 10|\boldsymbol{\alpha}_l = 00)$, is the probability of being classified in latent class 10 when the examinee's true membership is 00. If additional, remedial instruction crucial for examinees classified in $\boldsymbol{\alpha}_l = 00$ were not provided for examinees classified in $\boldsymbol{\alpha}_l = 10$, then the value of $P(\boldsymbol{\alpha}_s = 10|\boldsymbol{\alpha}_l = 00)$ would inform the test-user of the probability that students would not be given important academic help. In other words, the value of

$P(\boldsymbol{\alpha}_s = 10|\boldsymbol{\alpha}_l = 00)$ could be used as part of a cost-benefit analysis when making decisions based on the assessment. More generally, these off-diagonal entries can provide a more comprehensive look at the classification rate for researchers exploring the outcome of the diagnostic assessment.

2.4 Simulation Study

A simulation study was designed to evaluate how well $\hat{\tau}_l$ approximates the empirical conditional classification accuracy across a variety of test conditions. For comparison, a parametric Monte Carlo approach to estimating $\hat{\tau}_l$ was included, with the details provided below.

2.4.1 Design

The factors manipulated in the simulation were sample size (N), test length (J), item quality (IT), and attribute structure. The $K = 5$ attributes followed either a higher-order structure or a uniform structure. The former was introduced by de la Torre and Douglas (2004) and relates the attributes to θ, a general ability to master the attributes. The probability of mastering attribute k for individual i can be written as

$$P_{ik} = \frac{\exp(\zeta_0\theta_i + \zeta_k)}{1 + \exp(\zeta_0\theta_i + \zeta_k)},$$

where ζ_0 is the slope and ζ_k is the intercept parameter of the k^{th} attribute, and θ_i is the ability of examinee i^{th} drawn from the standard normal distribution. The slope was fixed to 1, and the intercepts of the five attributes were set as 1, 0.5, 0, -0.5, and -1. The uniform attribute distribution can be derived from the higher-order structure by setting both the slope and intercepts equal to zero, $\zeta_0 = \zeta_1 = \ldots\zeta_K = 0$, thereby making the probability of mastering each attribute equal to 0.5. Note that independent attributes of varying difficulty can be generated

by setting the slope equal to zero and varying the intercept parameters. Figure 2.1 plots the proportion of examinees in each latent class under the higher-order attribute structure, with the horizontal line representing the uniform attribute distribution proportions, which were all equal to $1/2^K = 0.031$. The shape of the plot shows how the attributes were ordered from easiest to hardest to master, making less likely any attribute pattern with attribute k equal to 1 and attribute $k-1$ equal to 0.

Specifically, the six most-common latent classes, $\boldsymbol{\alpha}_l = 00000$, $\boldsymbol{\alpha}_l = 10000$, $\boldsymbol{\alpha}_l = 11000$, $\boldsymbol{\alpha}_l = 11100$, $\boldsymbol{\alpha}_l = 11110$, and $\boldsymbol{\alpha}_l = 11111$, have a perfect Guttman pattern (Guttman, 1950), reflecting the hierarchy of attribute difficulty. The proportion of examinees in each of these latent classes was above 0.07, much higher than the proportion under the uniform attribute structure. By contrast, the five least-common latent classes, $\boldsymbol{\alpha}_l = 00101$, $\boldsymbol{\alpha}_l = 00011$, $\boldsymbol{\alpha}_l = 01011$, $\boldsymbol{\alpha}_l = 00111$, and $\boldsymbol{\alpha}_l = 01111$, have not mastered the easiest attribute, $k = 1$, but have mastered the most difficult, $k = 5$. The proportion of examinees in each of these latent classes was less than 0.01, much lower than the proportion under the uniform structure.

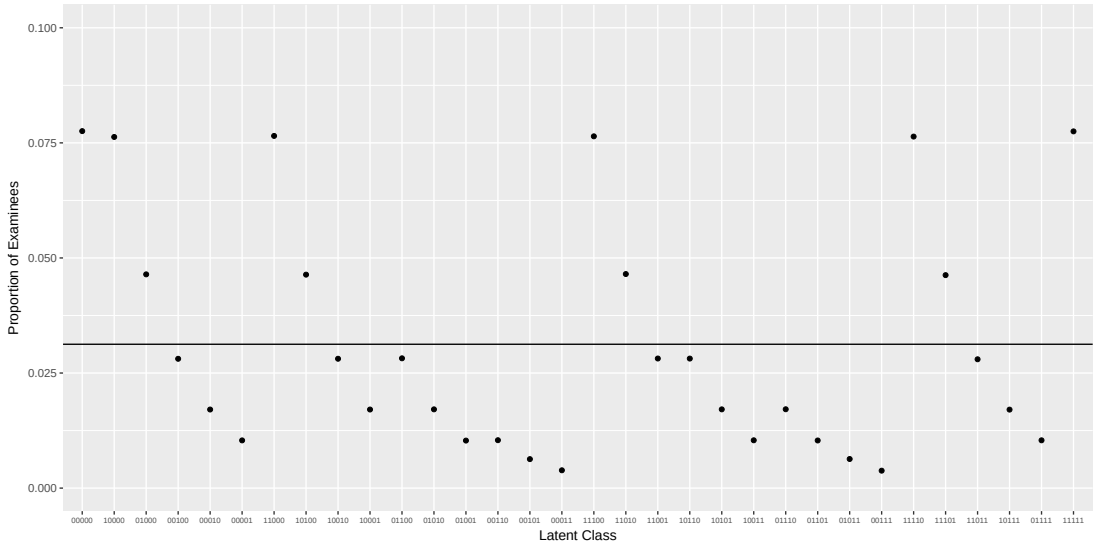

Figure 2.1: Latent class proportions under the higher-order attribute distribution

The sample sizes included were $N = 500$, 1000, 2000, and 5000. The test lengths were $J = 30$ and 60. For the G-DINA model used in this simulation study, item quality refers to the values of the guessing (g) and slip (s) parameters. High quality items had values $g = s = 0.1$, medium quality items $g = s = 0.2$, and low quality items $g = s = 0.3$. One hundred replications for each of the 48 conditions were generated. The various combinations of factors were designed to induce variability in the conditional classification accuracy, which allows for the investigation of the relationship between the indices and the empirical rates over a wide range of classification accuracies. For all replications, model and person parameters were estimated via the `GDINA` R Package (Ma & de la Torre, 2017).

2.4.1.1 Parametric Monte Carlo Approach

The only approach currently available for determining examinee classification accuracy conditional on the latent class is the parametric Monte Carlo approach. The first step of this method was to fit the G-DINA model to the sample, yielding estimates of the item parameters, $\hat{\boldsymbol{\phi}}$, and the latent class proportion parameters, $\hat{P}(\boldsymbol{\alpha}_l)$. The latter were used to draw a large sample of examinee attribute vectors, referred to as the resampled examinees. For the purposes of the simulation study, $100,000$ was determined to be sufficiently large. Item responses were generated based on the resampled attribute vectors and $\hat{\boldsymbol{\phi}}$. The G-DINA model, with item and latent class parameters fixed at $\hat{\boldsymbol{\phi}}$ and $\hat{P}(\boldsymbol{\alpha}_l)$, used the simulated item responses to classify the resampled examinees. The proportion of correct classification for each latent class was computed, returning the parametric Monte Carlo estimate of the conditional classification accuracy of the assessment, denoted by $\hat{\tau}_l^{mc}$. Because the parametric Monte Carlo approach employed resampling, this approach was substantially slower than computing the proposed index, particularly when the number of resampled examinees is set to be very large.

2.4.1.2 Classification Accuracy with a Different Attribute Distribution

An extension of the simulation was conducted to evaluate using the index $\hat{\tau}_l$ to compute $\hat{\tau}^*$. The estimated values were calculated for two scenarios. In the first scenario, the index $\hat{\tau}_l$ was computed based on the results obtained by fitting the G-DINA model to item responses from examinees drawn from a uniform attribute distribution. $P^*(\boldsymbol{\alpha}_l)$ was calculated by drawing a large sample ($N = 10,000,000$) from the higher-order distribution detailed above and calculating the latent class proportions of the sample. The index and latent class proportions were then used in Equation 2.3 to calculate the predicted classification accuracy of the assessment for a sample of examinees drawn from the higher-order distribution.

The second scenario entailed fitting a G-DINA model to item responses from examinees drawn from the higher-order attribute distribution (as detailed above), and calculating $\hat{\tau}_l$. Reflecting the uniform attribute distribution, $P^*(\boldsymbol{\alpha}_l)$ was a 2^K vector where each element was equal to $1/2^K$. In this case, $\hat{\tau}^*$ calculated via Equation 2.3 was equivalent to the unweighted average of τ_l.

2.4.2 Analysis

The estimates τ_l and τ_l^{mc} and the empirical values were compared across the conditions, with the empirical values of τ_l were computed as,

$$\tau_l = \frac{\sum_{r=1}^{Rep}\sum_{i=1}^{N} I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l]}{\sum_{i=1}^{N} I[\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l] \times Rep},$$

where Rep was the number of replications, and $I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l]$ evaluated whether the estimated examinee attribute pattern, $\hat{\boldsymbol{\alpha}}_i$, matched the generating value, $\boldsymbol{\alpha}_i$, conditional on the latent class l. The value $I[\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l]$ in the denominator is the number of examinees with true classification $\boldsymbol{\alpha}_l$.

The quality of the estimates was evaluated by computing the mean difference and root mean square difference (RMSD) of the parameter estimates across replications. The former was defined as,

$$\text{Mean Difference} = \frac{\sum_{r=1}^{Rep}(\hat{\tau}_l^r - \tau_l^r)}{Rep},$$

and the latter was defined as,

$$\text{RMSD} = \sqrt{\sum_{r=1}^{Rep}(\hat{\tau}_l^r - \tau_l^r)^2/Rep},$$

where $\hat{\tau}_l^r$ was the estimate of τ_l from replication r.

To evaluate the test-level classification accuracy for a different attribute structure, the empirical value of τ was computed as

$$\tau = \frac{\sum_{r=1}^{Rep}\sum_{i=1}^{N} I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]}{N \times Rep},$$

where $I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]$ evaluates whether the estimated attribute vector matched the generated values.

2.5 Results

2.5.1 Uniform Attribute Distribution

Figures 2.2 and 2.3 presents the mean difference and RMSD, respectively, of the proposed index under the uniform attribute structure. The boxplots show that the performance of the proposed index was distinguished by three groups of test conditions. The first and worst-performing group consisted of the test condition $N = 500$, $J = 30$, and low item quality, where the mean difference and RMSD exceeded 0.24 and 0.32, respectively. The second group consisted of two test

conditions, $N = 500$, $J = 60$, and low item quality, as well as $N = 1000$, $J = 30$, and low item quality. For this group, the range of mean difference values was 0.13 to 0.22, and the range of RMSD values was 0.17 to 0.27. The third group consisted of all other test conditions, under which the mean difference was 0.10 or less, and the RMSD was 0.16 or less. Overall, when test conditions were poor, the estimates of the conditional classification accuracy were also poor, on average 0.27 above the empirical value. By contrast, for reasonably favorable test conditions, the proposed index overestimated the empirical value by only 0.02, on average. The mean difference and RMSD appeared unrelated to the latent class, which was expected given an attribute structure where all examinees were equally likely to have each attribute.

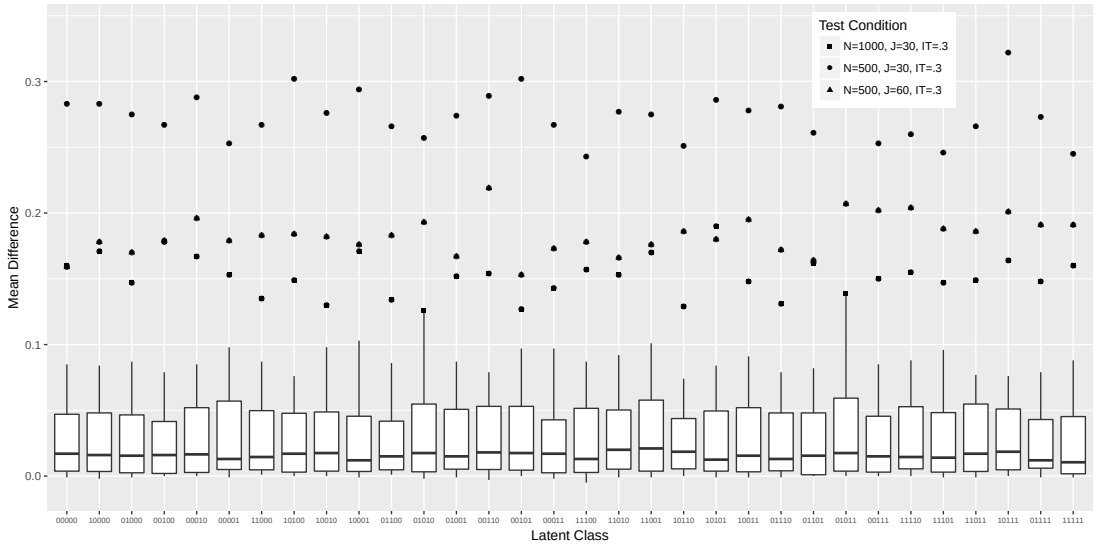

Figure 2.2: Mean difference of the proposed index under the uniform attribute distribution

2.5.1.1 Comparing the Proposed Index and the Monte Carlo Approach

The proposed index is compared to the Monte Carlo approach in Figures 2.4 and 2.5, which confirm that the mean difference and RMSD created approximately

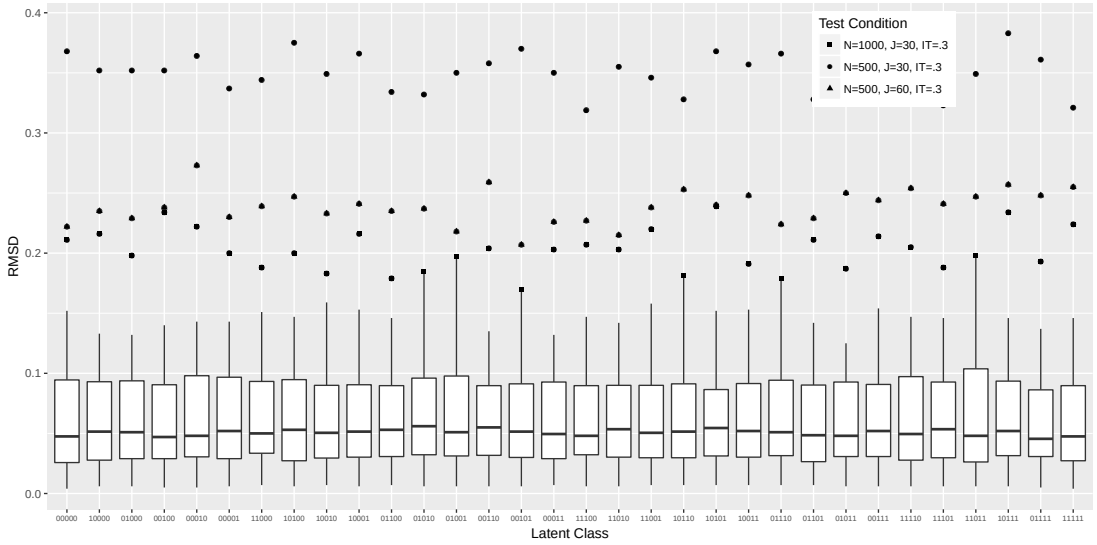

Figure 2.3: RMSD of the proposed index under the uniform attribute distribution

three clusters based on the test conditions. The plots illustrate not only the poor performance of both indices under unfavorable test conditions, but also the tendency of the proposed index to overestimate the empirical values across all conditions. The grouping of the plotted points around the 45 degree line illustrates the similarity of the estimates in terms of mean difference. The grouping of the plotted points above the line in Figure 2.5 indicates that the Monte Carlo approach led to slightly higher RMSD, suggesting that even with a large number of resampled attribute vectors, this approach led to more noisy estimates than the proposed index.

2.5.1.2 Latent Class $\boldsymbol{\alpha}_l = 1100$

To better understand the performance of the proposed index across all 24 conditions, Table 2.1 presents a more detailed look at the mean difference and RMSD of $\hat{\tau}_{11000}$ across all 24 conditions. As evidenced by the boxplots in Figures 2.2 and 2.3, the estimates of τ_{11000} were representative of the performance of all 2^K latent classes under the uniform attribute distribution. When item quality was high, the mean difference never exceeded 0.02, regardless of the approach. When

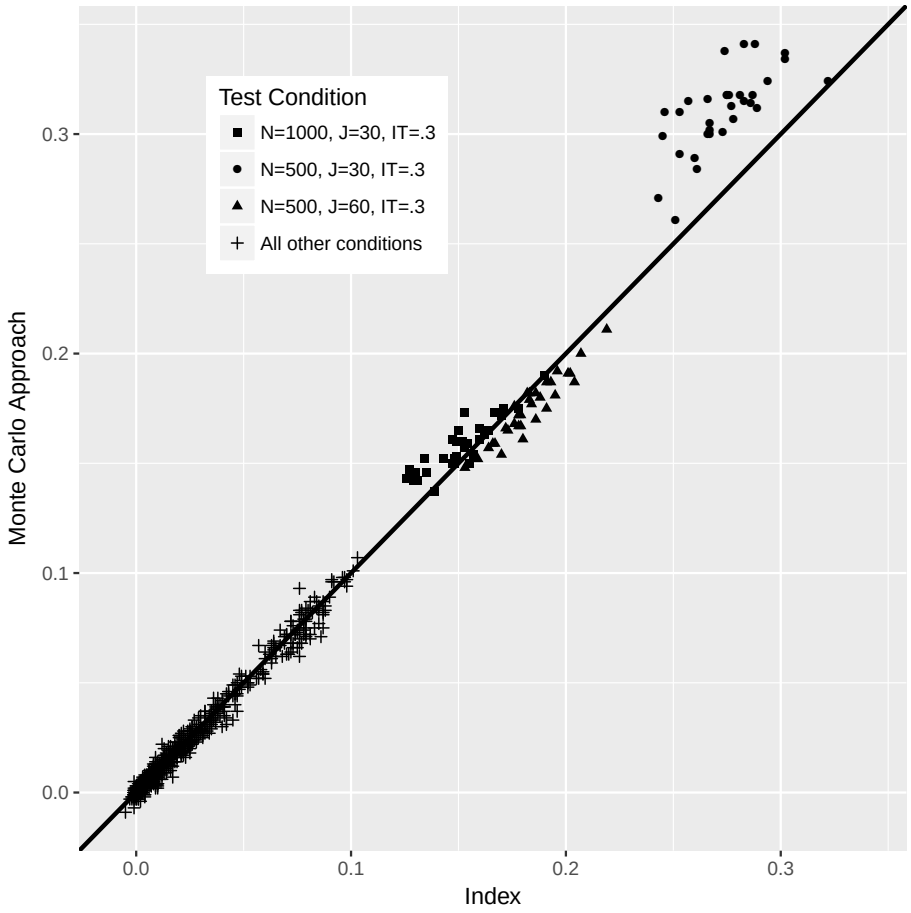

Figure 2.4: Mean difference of both approaches under the uniform attribute distribution

item quality was medium, the mean difference was 0.03 or less, with the exception of the $N = 500$ and $J = 30$ condition, where the mean difference increased to 0.07 and 0.08 for the index and Monte Carlo approaches, respectively. The mean difference under medium and high item quality indicated that well-estimated examinee posterior distributions led to very good recovery of the empirical values. When item quality was low, how well the estimates recovered the empirical values depended largely on the sample size. Specifically, the mean difference was inversely proportion to N, decreasing by approximately 50% each time the sample size doubled. When the sample size reached $N = 5000$, the mean difference did not exceed 0.03 even for tests with low item quality. Similarly, the RMSD of

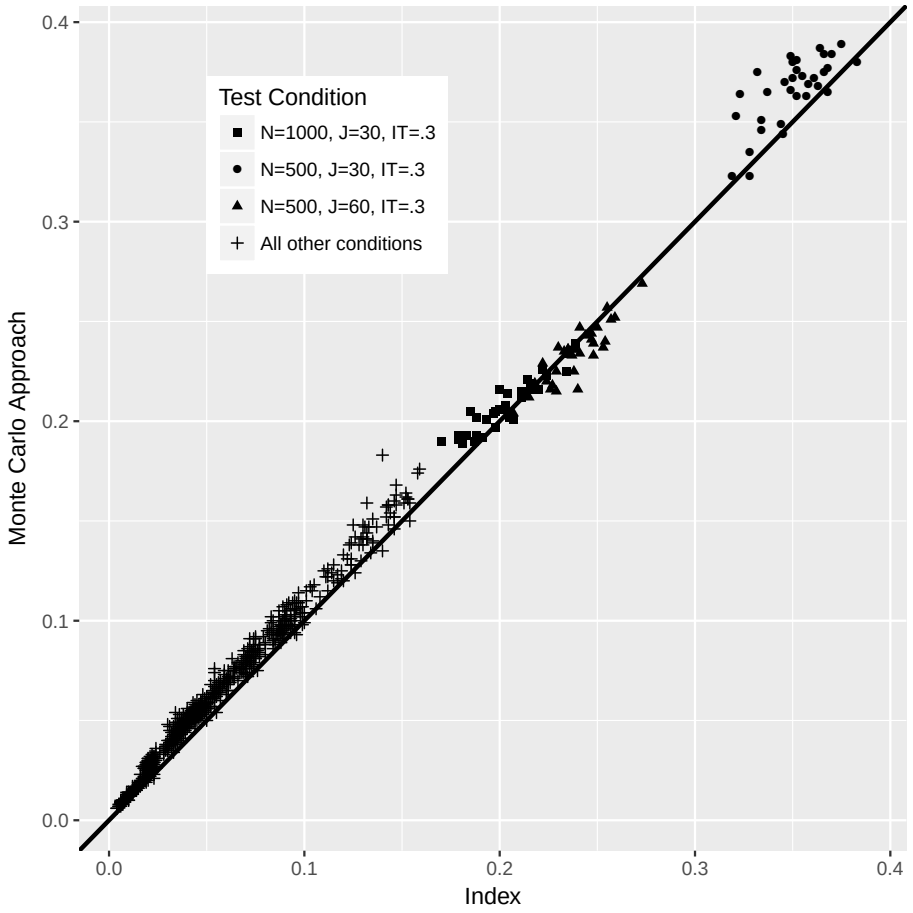

Figure 2.5: RMSD of both approaches under the uniform attribute distribution

$\hat{\tau}_{11000}$ was 0.34 when item quality was low and $N = 500$, but decreased to 0.05 when $N = 5000$. Overall, it appears that, even when the test conditions were poor, the proposed index can still recover the empirical classification accuracy provided the sample size was sufficiently large.

2.5.2 Higher-order Attribute Distribution

Under the higher-order attribute distribution, the relationship between the quality of the estimates and the sample size was particularly relevant. For both the mean difference and RMSD, the performance of the proposed index was not distinguished by three relatively distinct clusters of test conditions, as occurred under

Table 2.1: Uniform Attribute Distribution

			$\boldsymbol{\alpha}_l = 11000$			
			Mean Diff		RMSD	
IT	J	N	τ_l	τ_l^{mc}	τ_l	τ_l^{mc}
Low	30	500	0.27	0.30	0.34	0.35
		1000	0.14	0.15	0.19	0.19
		2000	0.08	0.08	0.10	0.10
		5000	0.03	0.03	0.05	0.06
	60	500	0.18	0.18	0.24	0.24
		1000	0.09	0.09	0.12	0.13
		2000	0.04	0.04	0.07	0.09
		5000	0.02	0.02	0.04	0.05
Medium	30	500	0.07	0.08	0.15	0.16
		1000	0.03	0.03	0.08	0.09
		2000	0.01	0.01	0.05	0.06
		5000	0.01	0.00	0.04	0.05
	60	500	0.03	0.02	0.09	0.10
		1000	0.01	0.01	0.06	0.07
		2000	0.01	0.01	0.04	0.05
		5000	0.00	0.00	0.02	0.03
High	30	500	0.02	0.02	0.07	0.08
		1000	0.01	0.01	0.05	0.05
		2000	0.00	0.00	0.03	0.04
		5000	0.00	0.01	0.02	0.03
	60	500	0.01	0.01	0.04	0.04
		1000	0.00	0.00	0.02	0.02
		2000	0.00	0.00	0.01	0.01
		5000	0.00	0.00	0.01	0.01

the uniform attribute distribution. Rather, the number of examinees in each latent class, in addition to the test length and item quality, heavily impacted the quality of the estimates. As illustrated in Figure 2.1, attribute patterns such as $\boldsymbol{\alpha}_l = 11000$ were relatively common, whereas attribute patterns such as $\boldsymbol{\alpha}_l = 00101$ were rare. Figure 2.6 contains the boxplots of the mean difference, by latent class, of the proposed index. From this figure, it can be concluded that the more common latent classes were relatively well-estimated, as evidenced by

the boxplot of $\boldsymbol{\alpha}_l = 11000$ having an interquartile range of 0.00 to 0.03. Referring back to Figure 2.2, under the uniform attribute structure, the boxplot of this same latent class had an interquartile range of 0.01 to 0.05. This shows that latent classes common under the higher-order structure were often estimated slightly better than the same latent class under the uniform structure. By contrast, the boxplot of a less-common latent class, $\boldsymbol{\alpha}_l = 00101$, had an interquartile range of 0.01 to 0.10 under the higher-order attribute structure and 0.00 to 0.05 under the uniform attribute structure.

The boxplots in Figure 2.7 evaluating performance by RMSD show a similar, though even more dramatic, pattern. The interquartile range of the RMSD values of $\boldsymbol{\alpha}_l = 11000$ was 0.03 to 0.10 under the uniform attribute distribution, whereas it decreased to a range of 0.02 to 0.06 under the higher-order attribute structure. For latent class $\boldsymbol{\alpha}_l = 00101$, however, it was 0.03 to 0.10 under the uniform and 0.08 to 0.20 under the higher-order. These patterns demonstrate the strength of the relationship between the number of examinees in a latent class and the quality of the estimates of conditional classification accuracy.

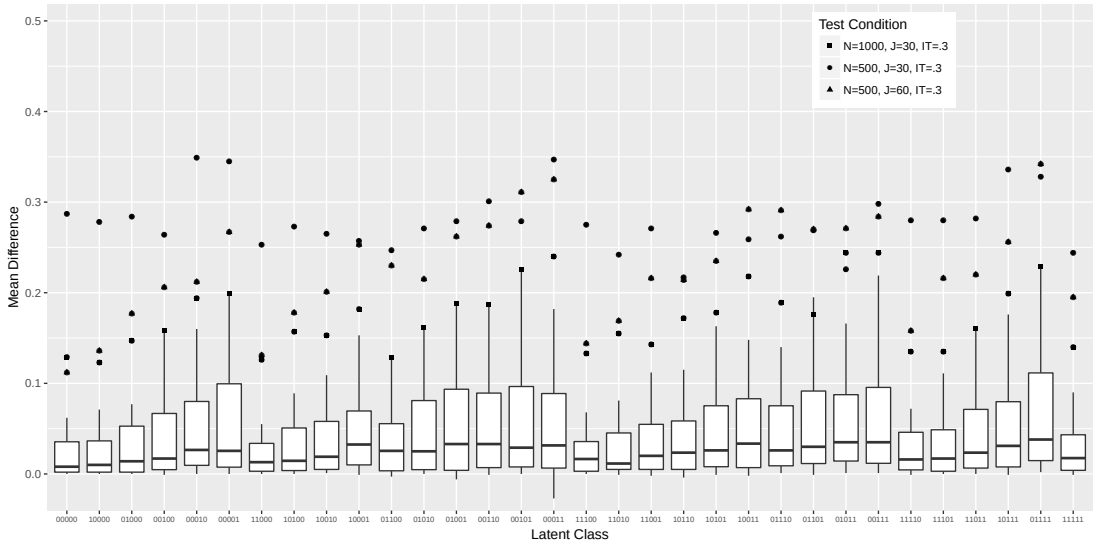

Figure 2.6: Mean difference of proposed index under the higher-order attribute distribution

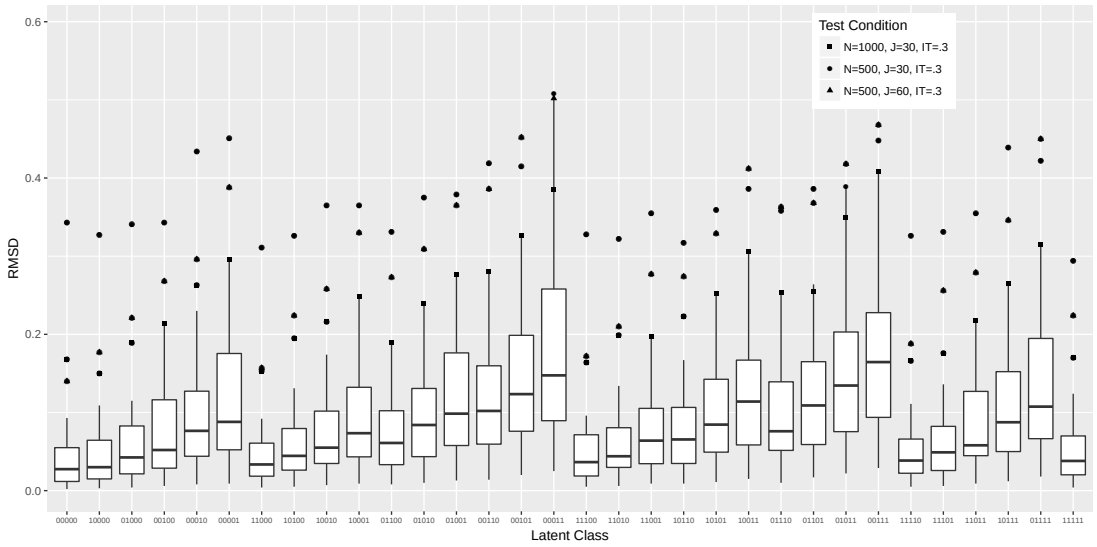

Figure 2.7: RMSD of proposed index under the higher-order attribute distribution

2.5.2.1 Comparing the Proposed Index and the Monte Carlo Approach

The proposed index is compared to the Monte Carlo approach in Figures 2.8 and 2.9. The two procedures tended to return similar estimates across favorable test conditions, although the pattern is less clear than in the case of the uniform attribute distribution because the similarity of the estimates also depended on the number of examinees in the latent class. Referring to Figure 2.8, the plotted mean difference reveals that discrepancies between the two approaches tended to occur when the mean difference was larger. For all latent classes under a higher-order attribute structure, when the mean difference of the index was less than 0.10, the maximum discrepancy between the proposed index and the parametric Monte Carlo approach was 0.03. This can be seen in the bunching of the scatterplot around the 45 degree line for x and y axis values less than 0.10. When the mean difference was 0.30, however, the maximum discrepancy was 0.18. Disparities between $\hat{\tau}_l$ and $\hat{\tau}_l^{mc}$ occurred only when both estimates were far from the empirical values. When investigating the conditional classification accuracy of real data,

a large discrepancy between the two estimates could be used as evidence that neither $\hat{\tau}_l$ nor $\hat{\tau}_l^{mc}$ is close to the true value.

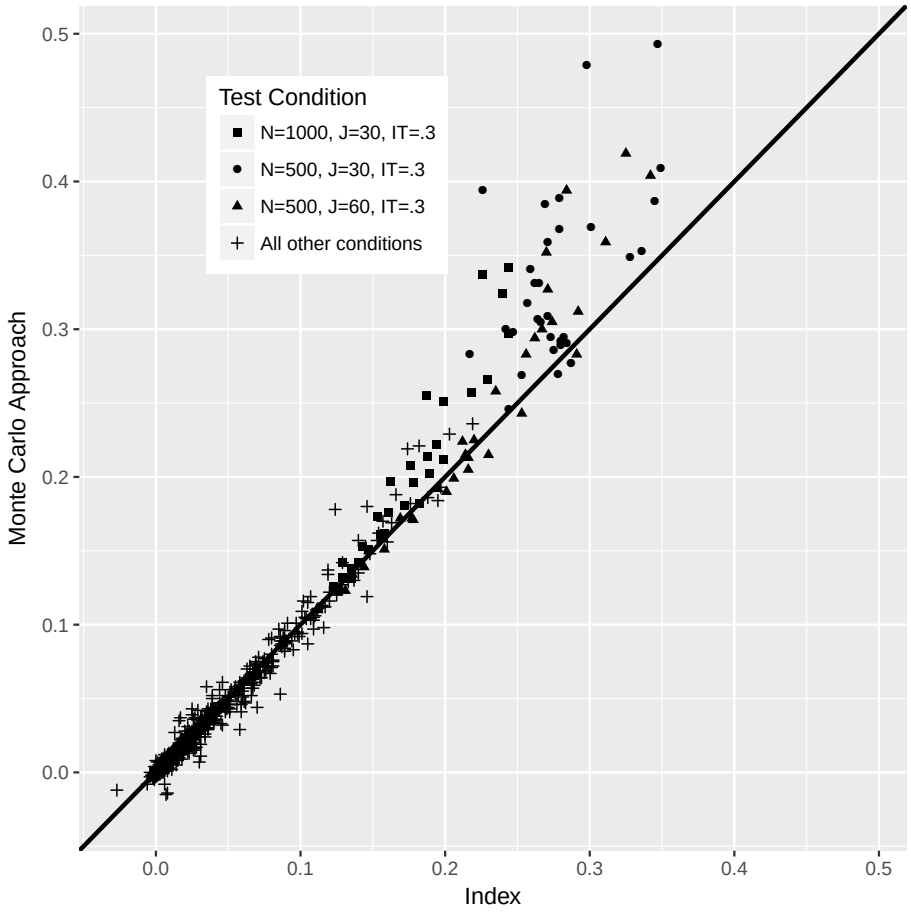

Figure 2.8: Mean difference under the higher-order attribute distribution

2.5.2.2 Latent Classes $\boldsymbol{\alpha}_l = 11000$ and $\boldsymbol{\alpha}_l = 00101$

The estimated classification accuracy of the two latent classes is presented in Table 2.2 to illustrate differences in quality across the 24 simulation study conditions. Excluding the three least-favorable test conditions, the largest mean difference of $\hat{\tau}_{11000}$ was 0.06, whereas the largest mean difference of $\hat{\tau}_{00101}$ was over three times larger, 0.19. The corresponding RMSD values were 0.07 and 0.28, respectively. The disparity in performance between the two latent classes was largest when the item quality was low, where the mean difference and RMSD for $\hat{\tau}_{00101}$ were

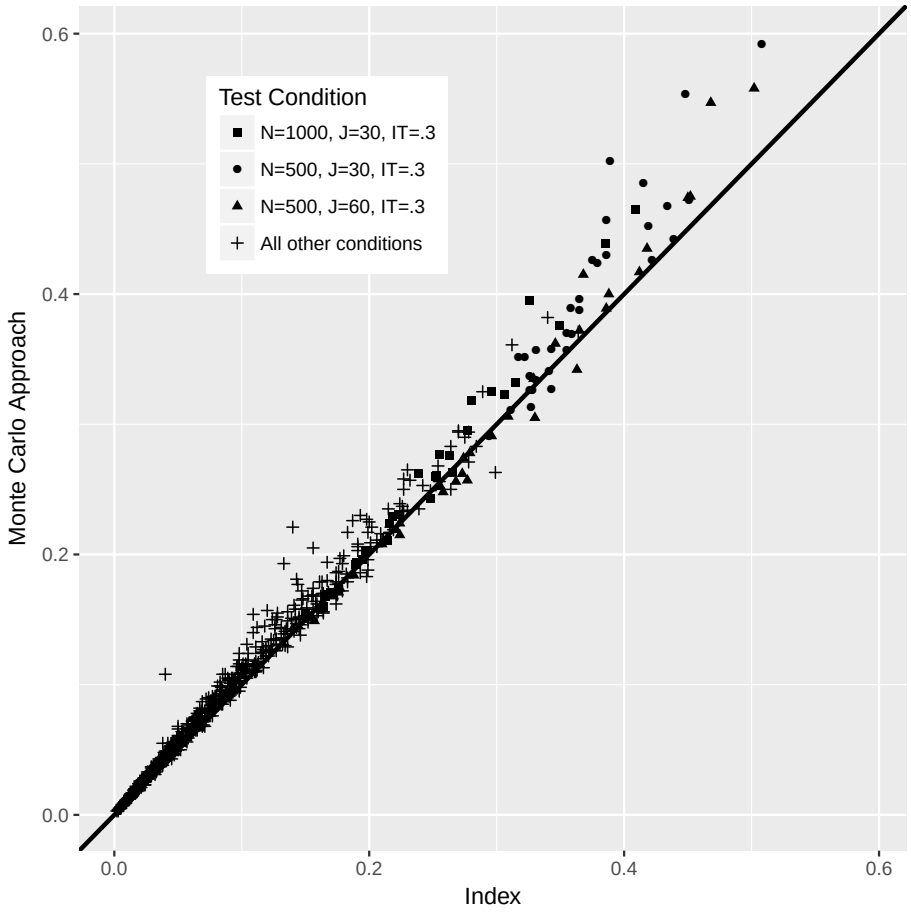

Figure 2.9: RMSD under the higher-order attribute distribution

frequently more than double that of $\hat{\tau}_{11000}$. Overall, the performance of the index followed the same pattern as under the uniform attribute distribution, except with worse overall recovery of empirical values for the less-common latent classes. Note that, here again, very favorable test conditions led to good recovery even for uncommon latent classes, with a maximum mean difference of $\hat{\tau}_{00101}$ of 0.04 when $J = 60$ and item quality was either medium or high. The maximum mean difference of $\hat{\tau}_{00101}$ when $N = 5000$ was also 0.04, indicating that large sample sizes compensated for shorter tests with lower quality items.

Additionally, Table 2.2 offers more insight into the discrepancies observed in Figures 2.8 and 2.9 between the proposed index and the parametric Monte Carlo approach. Specifically, differences in the quality of the estimates occured for

uncommon latent classes under the least-favorable test conditions. For example, for $\boldsymbol{\alpha}_l = 00101$ when $N = 500$, $J = 30$, and item quality was low, the proposed index and Monte Carlo approaches led to mean differences of 0.28 and 0.39, respectively. This discrepancy of 0.11 was much larger than for $\boldsymbol{\alpha}_l = 11000$, where the two approaches returned estimates within 0.01 of each other. Similar patterns were observed for the RMSD.

Table 2.2: Higher-Order Attribute Distribution

			$\boldsymbol{\alpha}_l = 11000$				$\boldsymbol{\alpha}_l = 00101$			
			Mean Diff		RMSD		Mean Diff		RMSD	
IT	J	N	τ_l	τ_l^{mc}	τ_l	τ_l^{mc}	τ_l	τ_l^{mc}	τ_l	τ_l^{mc}
Low	30	500	0.25	0.27	0.31	0.31	0.28	0.39	0.42	0.49
		1000	0.13	0.12	0.15	0.15	0.23	0.34	0.33	0.40
		2000	0.06	0.05	0.07	0.07	0.13	0.14	0.20	0.20
		5000	0.02	0.02	0.04	0.04	0.04	0.04	0.08	0.09
	60	500	0.13	0.12	0.16	0.15	0.31	0.36	0.45	0.48
		1000	0.05	0.05	0.07	0.08	0.19	0.19	0.28	0.28
		2000	0.03	0.03	0.05	0.05	0.10	0.09	0.17	0.16
		5000	0.01	0.01	0.03	0.03	0.03	0.04	0.08	0.10
Medium	30	500	0.06	0.05	0.09	0.10	0.10	0.08	0.29	0.33
		1000	0.03	0.03	0.06	0.07	0.08	0.07	0.20	0.23
		2000	0.02	0.02	0.04	0.04	0.02	0.02	0.13	0.14
		5000	0.01	0.01	0.02	0.02	0.00	0.01	0.08	0.09
	60	500	0.02	0.02	0.05	0.06	0.03	0.01	0.17	0.19
		1000	0.01	0.01	0.03	0.03	0.03	0.02	0.16	0.18
		2000	0.01	0.01	0.02	0.03	0.03	0.04	0.11	0.12
		5000	0.00	0.00	0.01	0.02	0.01	0.01	0.07	0.08
High	30	500	0.02	0.02	0.04	0.05	0.05	0.06	0.19	0.23
		1000	0.01	0.01	0.03	0.03	0.02	0.03	0.12	0.16
		2000	0.00	0.00	0.02	0.02	0.02	0.02	0.09	0.10
		5000	0.00	0.01	0.01	0.02	0.01	0.00	0.05	0.06
	60	500	0.00	0.00	0.01	0.02	0.00	0.00	0.05	0.07
		1000	0.00	0.00	0.01	0.01	0.00	0.00	0.04	0.04
		2000	0.00	0.00	0.01	0.01	0.00	0.00	0.03	0.03
		5000	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.03

2.5.3 Classification Accuracy with a Different Attribute Distribution

After estimating the classification accuracy of each latent class across each of the 48 conditions, the $\hat{\tau}_l$ computed in the simulation were subsequently weighted and summed to estimate τ^*, with the results presented in Table 2.3. The columns labeled "Predict Higher-order" contain: (1) τ, the empirical value when fitting the assessment to a sample drawn from a higher-order attribute distribution; (2) $\hat{\tau}^*$, calculated using τ_l from a uniform attribute distribution and $P^*(\boldsymbol{\alpha}_l)$ reflecting the higher-order distribution; and (3) $\hat{\tau}^*_{mc}$, which used $\hat{\tau}^{mc}_l$, the parametric Monte Carlo approach, rather than $\hat{\tau}_l$, the proposed index. The columns labeled "Predict Uniform" are the values from the reverse scenario, where τ is the empirical value when fitting the assessment to a sample drawn from a uniform distribution, and $\hat{\tau}^*$ is calculated using τ_l from a higher-order distribution and $P^*(\boldsymbol{\alpha}_l)$ reflecting the uniform attribute distribution.

Ignoring the three least-favorable test conditions, the proposed index returned values of $\hat{\tau}^*$ that differed from the empirical values by 0.06 or less, regardless of the attribute distribution. In 38 of the 48 total conditions, $\hat{\tau}^*$ predicted the classification accuracy within 0.03 of the empirical value. Performance of $\hat{\tau}^*$ was not consistently better or worse under either prediction scenario. For example, when $N = 2000$, $J = 30$, and item quality was low, predictions of τ^* for the higher-order distribution were closer to the empirical (0.02 difference) than predictions of τ^* for the uniform distribution (0.05 difference). However, when $N = 500$, $J = 30$, and item quality was medium, the results were reversed - predictions of τ^* for the uniform distribution were closer to the empirical (0.03 difference) than predictions for the higher-order (0.05 difference).

Table 2.3: Predicting Classification Accuracy for a Different Population

			Predict Higher-Order			Predict Uniform		
IT	J	N	τ	$\hat{\tau}^*$	$\hat{\tau}^*_{mc}$	τ	$\hat{\tau}^*$	$\hat{\tau}^*_{mc}$
Low	30	500	0.24	0.48	0.64	0.21	0.47	0.62
		1000	0.30	0.40	0.52	0.25	0.40	0.52
		2000	0.35	0.37	0.44	0.29	0.34	0.44
		5000	0.38	0.36	0.42	0.33	0.31	0.38
	60	500	0.49	0.63	0.70	0.44	0.62	0.73
		1000	0.55	0.59	0.64	0.51	0.57	0.65
		2000	0.57	0.58	0.62	0.53	0.54	0.60
		5000	0.59	0.57	0.61	0.55	0.53	0.58
Medium	30	500	0.62	0.67	0.72	0.60	0.63	0.71
		1000	0.65	0.66	0.70	0.62	0.62	0.68
		2000	0.67	0.66	0.69	0.64	0.61	0.66
		5000	0.68	0.65	0.68	0.64	0.61	0.65
	60	500	0.87	0.89	0.90	0.86	0.87	0.90
		1000	0.88	0.88	0.90	0.87	0.87	0.89
		2000	0.88	0.88	0.89	0.87	0.87	0.88
		5000	0.89	0.88	0.89	0.88	0.86	0.88
High	30	500	0.90	0.91	0.92	0.88	0.88	0.91
		1000	0.90	0.90	0.91	0.90	0.89	0.91
		2000	0.91	0.90	0.91	0.90	0.89	0.90
		5000	0.91	0.90	0.91	0.90	0.89	0.90
	60	500	0.99	0.99	0.99	0.99	0.98	0.99
		1000	0.99	0.99	0.99	0.99	0.99	0.99
		2000	0.99	0.99	0.99	0.99	0.99	0.99
		5000	0.99	0.99	0.99	0.99	0.99	0.99

Notably, the proposed index outperformed the parametric Monte Carlo approach under all but the most favorable test conditions, where both approaches returned similar estimates. Unlike the estimates of τ_l, the parametric Monte Carlo estimates of τ^* were not approximately the same as those from the proposed index, and were usually further from the empirical values. For example, under $N = 1000$, $J = 30$ and medium item quality, the empirical values of classification accuracy for the higher-order and uniform distributions were 0.65 and 0.62, respectively. The proposed index returned estimates of τ^* of 0.66 and 0.62,

whereas the Monte Carlo approach returned 0.70 and 0.68. Only when τ was greater than 0.90 did the discrepancies between approaches shrink to 0.01 or less.

2.6 Empirical Example

The utility of the proposed index for estimating the accuracy of classifications was demonstrated by its application to the Millon Clinical Multiaxial Inventory-III (MCMI-III), a personalized clinical assessment used to diagnose mental disorders (Millon et al., 2009). The dataset used here is from a Dutch-language version assessing 739 subjects that was fitted to the CDM framework in de la Torre et al. (2015) and Ma et al. (2016). It was determined that each of the thirty items measured one or more of three attributes: H = somatoform; SS = thought disorder; CC = major depression. Refer to Rossi, Elklit, and Simonsen (2010) for further details. Fitting the G-DINA model led to average estimates of guess and slipping parameters of 0.11 and 0.22, respectively, suggesting a test of medium to high item quality overall. Correlations among the attributes ranged from 0.73 to 0.84. After fitting the G-DINA model, the examinee posterior distributions were used to calculate $\hat{\tau}_l$ and $\hat{\tau}$. Because the true values are not available, the estimates were compared to the parametric Monte Carlo approach, $\hat{\tau}_l^{mc}$ and $\hat{\tau}^{mc}$.

Referring to Table 2.4, the majority of the examinees were classified as $\boldsymbol{\alpha}_l =$ 000 (47%) or $\boldsymbol{\alpha}_l = 111$ (27%), and the estimates of τ_l for these two largest latent classes were the same, 0.97 and 0.93, for both the proposed index and the parametric Monte Carlo approach. Some estimates of the smaller latent classes differed substantially, such as the 0.09 difference in the estimates of τ_{011}, to which 41 out of 739 patients (6%) were classified. Using the simulation study results as a guide, the test conditions, small latent-class size, and discrepancies between the two approaches suggested that the estimates of τ_{011} and τ_{110} may not be reliable, whereas the estimates of τ_{000} and τ_{111} were likely close to the empirical values.

At the test-level, the estimates of τ returned by both procedures differed only by 0.01, which was consistent with the results of the simulation study. Computing $\hat{\tau}_l$ required less than one-hundredth of a second of computing time, whereas the Monte Carlo approach with 100,000 resampled examinees required approximately 18 seconds. Overall, the results suggest that the CDA can classify examinees in the two most-common latent classes with a high degree of accuracy, which leads to high test-level accuracy. Classifying examinees in less-common latent classes, however, remains challenging.

Table 2.4: MCMI-III

	$\hat{P}(\boldsymbol{\alpha}_l)$	$\hat{\tau}_l$	$\hat{\tau}_l^{mc}$
$\boldsymbol{\alpha}_l = 000$	0.47	0.97	0.97
$\boldsymbol{\alpha}_l = 100$	0.05	0.39	0.40
$\boldsymbol{\alpha}_l = 010$	0.01	0.36	0.35
$\boldsymbol{\alpha}_l = 001$	0.06	0.59	0.65
$\boldsymbol{\alpha}_l = 110$	0.06	0.56	0.64
$\boldsymbol{\alpha}_l = 101$	0.02	0.55	0.51
$\boldsymbol{\alpha}_l = 011$	0.06	0.55	0.64
$\boldsymbol{\alpha}_l = 111$	0.27	0.93	0.93
		$\hat{\tau}$=0.84	$\hat{\tau}^{mc} = 0.85$

2.7 Discussion

CDAs offer the possiblity of assigning examinees to latent classes by measuring fine-grained components of variation that are of interest to the test-user. Before making any decision based on these classifications, it would be important to ask, how accurate the classifications are. Implementations of formative assessment may require important decisions to be made about particular latent classes, and if decisions related to educational instruction, or clinical diagnosis are to use

CDAs as evidence, estimates of the accuracy of the latent class, not just the overall accuracy of the assessment, should be available. This is fundamental to the interpretation of the test scores and the argument in favor of their intended use. This manuscript extends current methods to fill a gap in the literature by estimating the accuracy of latent classifications. Furthermore, because implementing a CDA will likely require a thorough understanding of how the accuracy of the test results generalize to other populations, this research proposes a relatively simple and accurate method for estimating the overall classification accuracy for any attribute distribution of interest.

In this manuscript, simulation studies investigated the recovery of the empirical values across test conditions and attribute distributions. For reasonably favorable test conditions, the proposed index overestimated the empirical values by 10% or less. The number of examinees in the latent class of interest heavily infuenced performance, which was particularly relevant under the higher-order attribute structure. The index returned estimates that were the same or better than the alternative, the parametric Monte Carlo approach, and only required basic matrix manipulation to compute. In addition, using the proposed index to predict classification accuracy for a different population was also computationally simple, producing values of $\hat{\tau}^*$ close to the empirical values under moderately favorable test conditions. The ease with which these indices are computed should encourage their routine use with CDAs, particularly when developing the validity argument.

The performance of the parametric Monte Carlo approach followed different patterns when estimating $\hat{\tau}^*$ compared to $\hat{\tau}_l$. This could be attributed to the fact that the parametric approach relies on $\hat{P}(\boldsymbol{\alpha}_l)$ and $\hat{\boldsymbol{\phi}}$, and although the former is changed to reflect the new population, the latter cannot be updated. Results indicated that, under less than ideal test conditions, the quality of $\hat{\boldsymbol{\phi}}$ depended on

the attribute distribution, which may explain why the parametric Monte Carlo approach performed worse than the proposed index when estimating τ^*. Note that this observation is about the noise in parameter estimates under suboptimal test conditions and does not contradict the findings of de la Torre and Lee (2010) regarding item parameter invariance of CDMs.

For short tests, low item quality, and/or uncommon latent classes, neither the proposed index nor the Monte Carlo approaches appear viable, returning values of $\hat{\tau}_l$ and $\hat{\tau}^*$ that were not close to the empirical values. It is still unclear why both approaches overestimated, but rarely underestimated, the empirical values. The overestimation suggests the need for a correction term in the computation of the index, which would shrink to zero as the sample size or quality of the posteriors improved. Additionally, the index should be evaluated in circumstances where the latent class structure is constrained (e.g., a hierarchical attribute structure). Extensions of this index for polytomous attributes would likewise be useful. Future simulation studies should include Q-matrix misspecification as a factor to evaluate how the accuracy and structure of the Q-matrix impacts performance. Finally, this research did not explore a way for the test user to quantify the amount of measurement error in the estimates of τ_l - the standard errors of $\hat{\tau}$, $\hat{\tau}_k$, $\hat{\tau}_l$, and $\hat{\tau}^*$ remain unstudied. A more robust validity argument for CDA will entail reporting these indices as well as their confidence intervals.

2.8 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19-38.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*, 115-127.

de la Torre, J., van der Ark, L., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 1-16.

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)*, *39*, 1-38.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285-291.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *The American soldier: Measurement and prediction* (Vol. 4). New York, NY: Wiley.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

Henson, Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*, 275-288.

Henson, Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

Huebner, A., & Wang, C. (2011). Comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407-419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Kopriva, R. J., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist*, *51*, 108-128.

Ma, W., & de la Torre, J. (2017). GDINA: The generalized DINA model framework [Computer software manual]. (R package version 1.2.0)

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*,

40, 200-217.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III Manual (4th ed.)*. Minneapolis, MN: Pearson Assessments.

Newman, A., Bryant, G., Stokes, P., & Squeo, T. (2013). *Learning to adapt: Understanding the adaptive learning supplier landscape*. Stamford, CT: Education Growth Advisors.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*, 59-81.

Rossi, G., Elklit, A., & Simonsen, E. (2010). Empirical evidence for a four factor framework of personality disorder organization: Multigroup confirmatory factor analysis of the Millon clinical multiaxial inventory - III personality disorder scales across Belgian and Danish data samples. *Journal of Personality Disorders*, *24*, 128-150.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*, 293-311.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Templin, J. L., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*, 457-476.

Ye, S., Fellouris, G., Culpepper, S., & Douglas, J. (2016). Sequential detection of learning in cognitive diagnosis. *British Journal of Mathematical and Statistical Psychology*, *69*, 139-158.

Chapter 3

Three-Step Estimation of Cognitive Diagnosis Models with Covariates

3.1 Introduction

The family of latent class models (LCMs), of which cognitive diagnosis models (CDMs) are part, offer a broad range of applications beyond classification that can be used to address a wide variety of research questions related to examinee performance. CDMs are a relatively new development, and the full suite of methodologies for applied researchers has yet to be developed. Specifically, techniques for modeling covariates along with CDMs has not been extensively discussed in the literature. For example, integrating structural models and measurement models has been studied extensively in the LCM context (for some examples, see Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Dayton & Macready, 1988, 2002; Huang & Bandeen-Roche, 2004). In the item response theory (IRT) context, modeling the relationship between scores and covariates via latent regression has been the subject of a great deal of research (for an overview, see Schofield, Junker, Taylor, & Black, 2015). However the details of incorporating covariates into CDMs has not received the same extensive treatment. The little work that has been done on the topic employed a one-step approach to incorporating covariates, which entails estimating the CDM and the regression model simultaneously. Ayers, Rabe-Hesketh, and Nugent (2013) implemented attribute-level logistic regression in both the DINA and higher-order

DINA models (de la Torre & Douglas, 2004), where the covariates affected the probability of the examinee mastering each attribute. This is referred to as differential skill functioning. Park and Lee (2014) extended this approach, constructing a logistic regression such that covariates influenced the probability of an examinee answering the item correctly. This can be conceived of as differential item functioning.

Compared to the one-step, the three-step approach offers the researcher additional flexibility in modeling the relationship between examinee classification and covariates. This approach treats classifications from the CDM as dependent variables and regresses them onto the covariates, which leads to downward bias in the parameter estimates. This is latent-class regression. Note that the techniques discussed here differ from errors-in-variables regression, where the measurement error is in the independent variables rather than the dependent variables (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006). In the IRT context, similar problems of poor parameter estimates in latent regression were addressed by developing plausible values, which are released for secondary analysis (Mislevy, 1991, 1993). In the LCM context, correction weights have been developed to adjust for bias in the latent regression parameters (Bakk, Tekle, & Vermunt, 2013; Bolck, Croon, & Hagenaars, 2004; Vermunt, 2010), permitting a wider variety of research questions to be studied. However, the solution mitigates much, but not all, of the downward bias, and the performance in terms of root mean square error (RMSE) has not been documented. Neither the corrected nor uncorrected three-step procedure has been explored in the CDM context. Unlike in LCA, in CDM the dependent variable in the third step could be either the latent class (i.e., attribute vector) or the attribute.

The shortcomings in the existing methodologies are problematic because secondary researchers likely will focus on the relationship of the covariates to the

latent classifications. In the educational measurement context, research questions frequently focus on the relationship between test scores and background variables, with the goal of determining which variables predict student achievement (see, for example, Abedi, Lord, & Hofstetter, 1998; Darling-Hammond, 2000; National Center for Education Statistics, 2011a). However, the one-step is not well-suited to the realities of secondary research, where item parameters and classifications have already been estimated and implemented. Furthermore, substantive experts may view as tautological interpretations of the relationship between covariates and classifications that were, in part, determined by the covariates (Bakk, Oberski, & Vermunt, 2013). There are circumstances under which this objection may be particularly compelling. For example , if the covariates were not collected at the same time as the item responses, the assumptions of the one-step model may be violated. The three-step approach may indeed lead to less biased estimates of the relationship between predictors and classification.

Adapted to the CDM framework, three-step methodologies for estimating CDMs would offer greater flexibility in modeling the relationship between covariates and fine-grained variation in examinee performance.

The rest of this paper is organized as follows: First, some relevant background on CDMs is covered. Next is a review of modeling the relationship between covariates and latent classes via the one-step and three-step procedures. After that is a section on correction weights in the three-step approach, including the proposed methodological advances. Following that is the simulation study to investigate the efficacy of the corrections in the CDM context, as well as a brief empirical example using real data. The paper will conclude with a discussion of the limitations and directions for future research.

3.2 Cognitive Diagnosis Models

The majority of CDM-related research has been done in the context of educational measurement, where the models are a relatively new development in a field dominated by IRT. The IRT framework has been built and refined to measure a uni- or low-dimensional construct on a continuum and rank examinees based on their performance (Junker & Sijtsma, 2001). The CDM framework is different in that it measures multidimensional skills referred to as attributes (de la Torre, 2011), on which examinees are classified as either having mastered or not mastered. Analysis can be done at either the attribute-level or the latent class-level; that is, the researcher may be interested in the mastery or non-mastery of particular attributes or the overall attribute-pattern classification.

A more detailed look at examinee skills could be used in low-stakes assessments to serve a diagnostic purpose, like, as an example from educational measurement, large-scale international assessments. The proliferation of e-learning platforms and intelligent tutoring systems marketed as formative assessment offer other potential applications. Alternatively, CDMs can be used in other contexts, like clinical psychology, where the multidimensional nature of the examinee classification provides advantages over other methods. For an example of this application, see de la Torre, van der Ark, and Rossi (2015) and the empirical example below.

3.2.1 The G-DINA Model

There are a multitude of CDMs in the literature that model the relationship among attributes in a variety of ways (see, for example, Rupp, Templin, & Henson, 2010), but a salient distinction is between reduced models and general models. The former makes particular assumptions about how the attributes interact when responding to an item, whereas the latter do not. There are several general

models - the general diagnostic model (GDM; von Davier, 2008), the log-linear CDM (LLM; Henson, Templin, & Willse, 2009), and the model used throughout this study, the generalized deterministic noisy "and" gate (G-DINA; de la Torre, 2011) model. A general model is used here to eliminate a potential source of model misfit and, because general models subsume reduced models, to bolster the generalizability of the findings.

The G-DINA model, like most CDMs, uses a Q-matrix (Tatsuoka, 1983) to specify the attributes used in each of the items. The number of attributes required for each particular item is denoted by K_j^*, where K refers to the number of attributes and the subscript j specifies the item. The number of required attributes can be calculated as $K_j^* = \sum_{k=1}^{K} q_{jk}$, where q_{jk} represents the k^{th} element of the j^{th} row of the Q-matrix. The examinee attribute vector can be written $\boldsymbol{\alpha}_l$=$\{\alpha_{l1}, \ldots, \alpha_{lK}\}$, where $l = 1, \ldots, 2^K$ denotes the latent classes, and $k = 1, \ldots, K$ the attributes. The kth element of the vector is 1 when the examinee has mastered the kth attribute, and is 0 when the examinee has not. Let $\boldsymbol{\alpha}_{lj}^*$ be the reduced attribute vector containing only the required attributes, where $l = 1, \ldots 2^{K_j^*}$. The probability of an examinee with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ answering item j correctly will be denoted by $P(\boldsymbol{\alpha}_{lj}^*)$. The G-DINA item response function is written as

$$P(\boldsymbol{\alpha}_{lj}^*) = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \phi_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk},$$

where ϕ_{j0} is the intercept for item j, ϕ_{jk} is the main effect due to α_k, $\phi_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\phi_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \cdots, \alpha_{K_j^*}$.

The parameter estimates can be obtained via marginalized maximum likelihood estimation. The likelihood is written

$$L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_l)^{X_{ij}}[1 - P_j(\boldsymbol{\alpha}_l)]^{1-X_{ij}}.$$

Marginalizing the likelihood is done by obtaining the weighted sum of the likelihood, where the latent class proportions, $P(\boldsymbol{\alpha}_l)$, are the weights. The marginalized likelihood is written as

$$L(\boldsymbol{X}) = \prod_{i=1}^{N} \sum_{l=1}^{2^K} L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l). \tag{3.1}$$

The log of Equation 4.2 can serve as the objective function and is optimized to estimate the item parameters. For more details on this, see de la Torre (2011). The posterior distribution of examinee i is a vector of length 2^K, written as

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) \propto L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l),$$

and normalized to sum to one.

3.2.2 Latent Class Assignment

The examinee posterior distributions were subsequently used in the second step of the three-step procedure to assign examinees to latent classes. In LCA, the examinee assignment may be either proportional, modal, or mean (Goodman, 2007). In this study, the vector of proportional assignment is equal to the estimated examinee posterior distribution, whereas modal and mean assignment correspond to the maximum a posteriori (MAP) and expected a posteriori (EAP) methods, respectively (Huebner & Wang, 2011). The latter was used to compute the marginalized attribute-level probabilities, $P(\alpha_k|\boldsymbol{Z}_i)$.

The estimated attribute pattern of examinee i is denoted by $\hat{\boldsymbol{\alpha}}_i$. In the following sections, $\boldsymbol{\alpha}_l$ refers to the potential true values of the attribute pattern, and $\boldsymbol{\alpha}_s$ refers to the possible value of the latent class assignment, where $s = 1, \ldots, 2^K$. That is, $\boldsymbol{\alpha}_s$ are the possible values of $\hat{\boldsymbol{\alpha}}_i$.

3.2.3 Matrix of Classification Error Probabilities

The examinee latent-class assignment and posterior distribution are used to calculate the matrix of conditional classification error probabilities (Vermunt, 2010), written as

$$P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_s]}{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)}, \tag{3.2}$$

where $I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_s]$ is an indicator function equal to 1 when the estimated attribute pattern of examinee i is equal to latent class $\boldsymbol{\alpha}_s$, and zero otherwise. Thus, $P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})$ can be interpreted as a $2^K \times 2^K$ contingency table containing the proportion of examinees with true latent class membership $\boldsymbol{\alpha}_l$ assigned to latent class $\boldsymbol{\alpha}_s$.

Note the relationship between the matrix of classification error probabilities and the pattern-level classification accuracy index (Wang, Song, Chen, Meng, & Ding, 2015),

$$\hat{\tau} = \sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X}) I[l = s] \times P(\boldsymbol{\alpha}_l)$$

where $I[l = s]$ is the indicator function equal to 1 when the latent class l is the same as the latent class s. The diagonal of the matrix of classification error probabilities, weighted by the proportion of examinees in each latent-class and summed, is equal to $\hat{\tau}$.

3.3 Modeling the Relationship between Covariates and Latent Classification

3.3.1 The One-Step Approach

The LCM literature offers two approaches for relating covariates to latent classification: the one-step and the three-step procedures (see, for example, Bakk, Tekle, & Vermunt, 2013; Bolck et al., 2004; Vermunt, 2010). In the following section, the observed examinee covariates are implemented at the latent-class level, where the covariates affect the probability of an examinee belonging to a latent category. The likelihood is written as

$$L(\boldsymbol{X}) = \prod_{i=1}^{N} \sum_{l=1}^{2^K} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) P(\boldsymbol{\alpha}_l|\boldsymbol{Z}_i),$$

where $P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)$ is the CDM relating the observed item responses to the latent classification, $\boldsymbol{Z}_i = \{Z_{i1}, \dots Z_{if}, \dots Z_{iF}\}$ is the covariate vector for examinee i, and $P(\boldsymbol{\alpha}_l|\boldsymbol{Z}_i)$ is the structural model relating the latent classifications to the covariates, here a multinomial logistic regression. The full-information maximimum likelihood (FIML) approach estimates both the CDM and multinomial logistic regression model parameters simultaneously by maximizing the following log-likelihood:

$$\log L_{FIML} = \sum_{i=1}^{N} \log \sum_{l=1}^{2^K} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) P(\boldsymbol{\alpha}_l|\boldsymbol{Z}_i).$$

A similar one-step procedure can be used to relate the covariates to the probability of an examinee having mastered an attribute. The relationship between the probability of mastering attributes $1 \dots K$ and belonging to latent class l can be written as

$$P(\boldsymbol{\alpha}_l) = \prod_{k=1}^{K} P(\alpha_k = 1)^{\alpha_{lk}} \times [1 - P(\alpha_k = 1)]^{1-\alpha_{lk}},$$

where α_{lk} is one when latent class l entails mastery of attribute k, and zero otherwise. Using this relationship, the log-likelihood can be written as

$$\log L_{FIML} = \sum_{i=1}^{N} \log \sum_{l=1}^{2^K} P(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) \prod_{k=1}^{K} P(\alpha_k = 1|\boldsymbol{Z}_i)^{\alpha_{lk}} \times [1 - P(\alpha_k = 1|\boldsymbol{Z}_i)]^{1-\alpha_{lk}},$$

and is optimized with respect to the CDM and the logistic regression model parameters, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$, respectively, to obtain the FIML estimates.

The one-step estimates the measurement model and the structural model simultaneously, producing unbiased estimates of the relationship between examinee classification and the covariates (Dayton & Macready, 1988). However, there are several circumstances under which the one-step procedure is undesirable or infeasible. Because the measurement model must be refitted when one or more covariates are dropped from the structural model, the estimated G-DINA parameters will change every time the latent class regression adds or drops a variable. This complicates the calibration of item parameters. Moreover, in educational measurement, the covariates tend to be collinear, and thus it may be desirable to employ dimension reduction techniques, like principal components analysis, to incorporate a large amount of covariate information without adding too many parameters to the objective function (National Center for Education Statistics, 2011b; von Davier, Sinharay, Oranje, & Beaton, 2006). Implementing covariates in this form via the one-step approach may improve estimation of the CDM parameters and/or may improve the classification accuracy, but estimating coefficients of the principal components prevents the researcher from studying the relationship between individual covariates and attribute mastery.

Additionally, in substantive research, it may be that the covariates should not contribute to the classification (Robitaille & Beaton, 2002). Alternatively, it may be that all the covariates (or the principal components of the covariates)

are needed to improve classification, but only the relationship among particular covariates and the classification is of interest (National Center for Education Statistics, 2008). Finally, the full item responses are often not released to secondary researchers. In these situations, a one-step procedure is impractical, if not impossible.

3.3.2 The Uncorrected Three-Step Approach

When the one-step procedure is not available or is inappropriate, the three-step procedure can be implemented. As described earlier, the first step is to estimate the CDM, the second step is to assign examinees to a latent class based on the posterior, and the third step is to estimate the relationship between the covariates and latent classification obtained in the second step. It should be noted that in the third step of the uncorrected approach, the classifications are no longer treated as latent, but rather as observed variables. Estimated classifications are regressed onto the covariates via a multinomial logistic regression, written as

$$P(\boldsymbol{\alpha}_l|\boldsymbol{Z}_i) = \frac{\exp(\beta_{l0} + \sum_{f=1}^{F} \beta_{lf} Z_{if})}{\sum_{l=1}^{2^K} \exp(\beta_{l0} + \sum_{f=1}^{F} \beta_{lf} Z_{if})}, \tag{3.3}$$

where the parameters of interest are the coefficients for each latent class, $\boldsymbol{\beta}^{MLR} = \{\beta_{l0}, \beta_{l1}, \ldots, \beta_{lF}\}$. They can be estimated by maximizing the objective function of the multinomial logistic regression model,

$$\log L = \sum_{i=1}^{N} \log P(\hat{\boldsymbol{\alpha}}_i|\boldsymbol{Z}_i), \tag{3.4}$$

where P refers to the probability function of Equation 3.3 evaluated at $\hat{\boldsymbol{\alpha}}_i$. That is, the latent class assignment of examinee i is used in place of the true latent class, $\boldsymbol{\alpha}_l$. In the uncorrected approach, the latent class assignment is treated as an

observed dependent variable in the multionomial logistic regression function. As will be demonstrated, this approach leads to biased estimates of $\boldsymbol{\beta}^{MLR}$, impacting the validity of the inferences made using this approach.

3.3.3 The Three-Step Procedure with Latent-class Level Correction Weights

3.3.3.1 Sample-Level Correction Weights

Equation 3.4 estimates the relationship between the estimated classifications and the covariates rather than the true classifications and the covariates. In contrast, the corrected three-step procedure incorporates the classification error probabilities to weight individual assignments in the objective function. These sample-level correction weights are denoted by SL_{il} and are calculated as

$$SL_{il} = P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l, \boldsymbol{X})I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_s]$$

which can be interpreted as using the column from the matrix of classification error probabilities that corresponds to the latent class assignment, s, of examinee i. The objective function of the multinomial logistic regression in the (corrected) third step can then be written as

$$\log L_{SL} = \sum_{i=1}^{N} \log \sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l|\boldsymbol{Z}_i)SL_{il}, \tag{3.5}$$

in effect computing a weighted average of the classifications, compared to Equation 3.4, which only uses the assigned classification. The sample-level corrected approach leads to reduction in the bias of the estimates of $\boldsymbol{\beta}^{MLR}$ in $P(\boldsymbol{\alpha}_l|\boldsymbol{Z}_i)$ (Vermunt, 2010).

To illustrate the computation of SL_{il}, in a hypothetical example with two

attributes, let the following be a matrix of classification error probabilties,

$$P(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_l) = \begin{bmatrix} .85 & .08 & .04 & .03 \\ .01 & .92 & .02 & .05 \\ .06 & .04 & .81 & .09 \\ .03 & .01 & .01 & .95 \end{bmatrix},$$

where the four columns correspond to latent classes assignments, $\boldsymbol{\alpha}_s = (0, 0)$, $\boldsymbol{\alpha}_s = (1, 0)$, $\boldsymbol{\alpha}_s = (0, 1)$, and $\boldsymbol{\alpha}_s = (1, 1)$, and the rows correspond to the true latent classes. If the examinee is assigned to class $\boldsymbol{\alpha}_s = (1, 0)$, then the second column would be used as the correction weights:

$$SL_{il} = \begin{bmatrix} .08 \\ .92 \\ .04 \\ .01 \end{bmatrix}.$$

These values would be incorporated into Equation 3.5 in estimating the multinomial logistic regression parameters, $\boldsymbol{\beta}^{MLR}$.

3.3.3.2 Posterior-distribution Level Correction Weights

The correction weights, SL_{il}, adjust for much, but not all, of the bias in the parameter estimates, as evidenced by the results of previous simulation studies in LCA (Bakk, Oberski, & Vermunt, 2013; Vermunt, 2010). The proposed correction weights were designed to reduce the discrepancy between one-step and corrected three-step parameter estimates. They are calculated as

$$PDL_{il} = \frac{P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)}{P(\boldsymbol{\alpha}_l)}, \tag{3.6}$$

and replace SL_{il} in the objective function of Equation 3.5. The main difference between SL_{il} and PDL_{il} is that the former uses the latent-class distribution from the entire sample, whereas the latter uses the posterior distribution from each individual examinee. That is, in SL_{il}, the same correction is applied to each examinee that has been assigned to the same latent class, whereas in the proposed weights, PDL_{il}, the correction is unique to each posterior distribution. Thus, SL_{il} and PDL_{il} will be referred to as sample-level and posterior-distribution level correction weights, respectively.

3.3.3.3 Three-Step Approach with Attribute-Level Correction Weights

In the CDM framework, the relationship between individual attributes and covariates also may be of interest, and correction weights were developed for a three-step procedure in this context as well. The probability of mastering each attribute is computed by aggregating the posterior probabilities $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$ into the marginal attribute-mastery probabilities, $P(\alpha_k = 1|\boldsymbol{X}_i)$. Note that $P(\alpha_k = 0|\boldsymbol{X}_i) = 1 - P(\alpha_k = 1|\boldsymbol{X}_i)$. The probability of mastering attribute k given the covariates can be modeled via logistic regression, written as

$$P(\alpha_k = 1|\boldsymbol{Z}_i) = \frac{\exp(\beta_{k0} + \sum_{f=1}^{F} \beta_{kf} Z_{if})}{1 + \exp(\beta_{k0} + \sum_{f=1}^{F} \beta_{kf} Z_{if})},$$

and $P(\alpha_k = 0|\boldsymbol{Z}_i) = 1 - P(\alpha_k = 1|\boldsymbol{Z}_i)$, where the parameters of interest are the coefficients for each attribute, $\boldsymbol{\beta}_{LR} = \{\beta_{k0}, \beta_{k1}, \ldots, \beta_{kF}\}$. The objective function for the uncorrected approach is

$$logL = \sum_{i=1}^{N} log[P(\hat{\alpha}_{ik}|\boldsymbol{Z}_i)],$$

where $\hat{\alpha}_{ik}$ is the estimated attribute k classification of examinee i, and is equal to 1 if the examinee is classified as having mastered the attribute, and 0 otherwise.

In the uncorrected attribute-level three-step procedure, the estimated attribute classification is treated as an observed dependent variable in a regression model.

A matrix of attribute-level classification error probabilities was computed in a manner similar to Equation 3.2, the classification error probabilities of the latent-classes. This yielded a 2×2 contingency table, denoted by $P(\alpha_q|\alpha_k, \boldsymbol{X})$, where α_q refers to the possible attribute classification values, and is equal to 0 or 1. Note that the α_k and α_q notation here is analogous to the $\boldsymbol{\alpha}_l$ and $\boldsymbol{\alpha}_s$ notation used in the latent-class three-step described above. The attribute-level matrix is calculated as

$$P(\alpha_q|\alpha_k, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\alpha_k|\boldsymbol{X}_i) I[\hat{\alpha}_{ik} = \alpha_q]}{\sum_{i=1}^{N} P(\alpha_k|\boldsymbol{X}_i)},$$

and can be interpreted as the proportion of examinees assigned attribute mastery α_q given true attribute mastery α_k. For example, the entries of column 1 are

$$P(\alpha_q = 0|\alpha_k = 0, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\alpha_k = 0|\boldsymbol{X}_i) I[\hat{\alpha}_{ik} = 0]}{\sum_{i=1}^{N} P(\alpha_k = 0|\boldsymbol{X}_i)},$$

and

$$P(\alpha_q = 0|\alpha_k = 1, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\alpha_k = 1|\boldsymbol{X}_i) I[\hat{\alpha}_{ik} = 0]}{\sum_{i=1}^{N} P(\alpha_k = 1|\boldsymbol{X}_i)},$$

where the former can be interpreted as the proportion of examinees correctly classified as not having mastered attribute k, and the latter as the proportion of examinees incorrectly classified as not having mastered attribute k.

The sample-level correction weights for examinee i use the column from the matrix of attribute-level classification error probabilities that corresponds to the examinee's estimated attribute classification, and were computed

$$SL_{ik} = P(\alpha_q|\alpha_k, \boldsymbol{X}) I[\hat{\alpha}_{ik} = \alpha_q]$$

The posterior-distribution level correction weights were calculated

$$PDL_{ik} = \frac{P(\alpha_k|\boldsymbol{X}_i)}{P(\alpha_k)},$$

where $P(\alpha_k)$ is the sample-level proportion of mastery of attribute k. The attribute-level logistic regression log-likelihood can be modified to find the estimates of the marginalized probability of mastering attribute k given the covariates, $P(\alpha_k|\boldsymbol{Z}_i)$, rather than the relationship between the attribute assignment and the covariates, $P(\hat{\alpha}_{ik}|\boldsymbol{Z}_i)$. The objective function of the corrected approach is written as

$$\log L = \sum_{i=1}^{N} \log \sum_{\alpha_k=0}^{1} P(\alpha_k|\boldsymbol{Z}_i) w_{ik},$$

where w_{ik} is equal to SL_{ik} or PDL_{ik}. Optimizing this objective function leads to corrected estimates of the parameters $\boldsymbol{\beta}_{LR}$ in $P(\alpha_k|\boldsymbol{Z}_i)$.

3.4 Simulation Study to Evaluate the Performance of the Correction Weights

A two-part simulation study was designed to investigate the ability of the sample-level correction weights, SL_{il} and SL_{ik}, and the posterior-distribution level correction weights, PDL_{il} and PDL_{ik}, to improve the estimates of the regression model parameters. The uncorrected three-step and the one-step procedures were included for comparison. Furthermore, the simulation study examined the extent to which the performance of the correction weights was related to the sample size and the degree of misclassification of the CDM.

3.4.1 Design

Factors manipulated were the sample size (N = 500, 1000, and 2000), the test length (J = 10 and 20), and item quality (High, Medium, and Low), operationalized as the values of the guessing and slip parameters ($g = s$ = .1, .2, and .3). The $K = 3$ attributes were generated under two conditions: an independent attribute structure, where correlations among the attributes were approximately zero, and a correlated attribute structure, where the correlations among the attributes ranged from .5 to .8, as suggested by Kunina-Habenicht, Rupp, and Wilhelm (2012) and Sinharay, Puhan, and Haberman (2011). The ten-item Q-matrix to be used in the study is presented in Table 4.1; it was doubled for the 20-item tests. Additionally, the number of examinee covariates tested were F = 3, 9, and 12, drawn from the multivariate standard normal distribution, $N(\mathbf{0}, \boldsymbol{I})$. In Part I of the study, the relationship between $\boldsymbol{\alpha}_l$ and $\boldsymbol{Z}_i$ assumed the form of a multinomial logistic regression, with latent-classes regressed onto covariates. For the three covariates ($F = 3$) condition, the matrix of true parameters used to generate the correlated

Table 3.1: Ten-Item Q-matrix

item	α_1	α_2	α_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	1	0
6	0	0	1
7	1	1	0
8	1	0	1
9	0	1	1
10	1	1	1

attribute structure was

$$\boldsymbol{\beta}_{corr}^{MLR} = \begin{bmatrix} -0.5 & 1 & 1 & 0.5 & 0.5 & 2 & 1.5 \\ 1 & -0.5 & 1 & 0.5 & 2 & 0.5 & 1.5 \\ 1 & 1 & -0.5 & 2 & 0.5 & 0.5 & 1.5 \end{bmatrix},$$

and the matrix of parameters used to generate the independent attribute structure was,

$$\boldsymbol{\beta}_{ind}^{MLR} = \begin{bmatrix} 1 & 0.5 & 0.5 & 1.5 & 1.5 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0.5 & 0.5 & 1 & 1 & 1.5 & 1.5 & 2 \end{bmatrix}.$$

$\boldsymbol{\beta}^{MLR}$ has $2^K - 1$ columns corresponding to the latent classes, with $\boldsymbol{\alpha}_l = (000)$ set as the reference group. The rows correspond to the number of covariates. When the number of covariates increased to nine and twelve, $\boldsymbol{\beta}^{MLR}$ was tripled and quadrupled, respectively. In Part II of the simulation study, the relationship between $\boldsymbol{\alpha}_k$ and $\boldsymbol{Z}_i$ assumed the form of a logistic regression. The same conditions were used, and another set of covariates were generated from the standard normal distribution and used with the generating parameters below in a logistic regression model. For the three covariate condition ($F = 3$), the matrix of parameters used to generate the correlated attribute structure was

$$\boldsymbol{\beta}_{corr}^{LR} = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix},$$

and the matrix of parameters used to generate the independent attribute structure was

$$\boldsymbol{\beta}_{ind}^{LR} = \begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}.$$

Here the three rows correspond to the number of covariates and the three columns to the number of attributes. When the number of covariates increased to nine and twelve, $\boldsymbol{\beta}_{corr}^{LR}$ was tripled and quadrupled, respectively. To simplify computations and analysis, and without loss of generality, the intercept β_0 was set equal to 0. In all 100 replications per condition, the estimates were computed by directly optimizing the log-likelihood via the L-BFGS-B (Nocedal & Wright, 2006) method in the `optim` package of R statistical computing software (R Core Team, 2016).

3.4.2 Analysis

Modifying the aforementioned factors varied the proportion of correctly classified attribute vectors (PCV), defined as,

$$PCV = \frac{\sum_{r=1}^{Rep}\sum_{i=1}^{N} I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]}{N \times Rep},$$

where *Rep* is the number of replications, and $I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]$ evaluates whether the estimated attribute vector matches the generated values. Across the tested factors, the PCV varied from approximately 42% to 98%. The classification accuracy of the G-DINA model, in turn, affected the accuracy of the parameter estimates of the regression model. The quality of the estimates was evaluated by computing the bias and root-mean-square error (RMSE) of the parameter estimates across

replications. The former is defined as,

$$bias = \frac{\sum_{r=1}^{Rep}(\hat{\beta}_{gf}^{(r)} - \beta_{gf})}{Rep},$$

where g represents either latent class l or attribute k, depending on which regression model is being estimated, $\hat{\beta}_{gf}^{(r)}$ is the estimate of β_{gf} from replication r, f is the covariate ($f = 1 \ldots F$), and Rep is the total number of replications; the latter is defined as,

$$RMSE = \sqrt{\sum_{r=1}^{Rep}(\hat{\beta}_{gf}^{(r)} - \beta_{gf})^2/Rep.}$$

To offer a look at the overall estimation across all parameters, the average absolute bias (ABIAS) and the average RMSE (ARMSE) were calculated as

$$ABIAS = \frac{\sum_{g=1}^{G}\sum_{f=1}^{F}\left|\sum_{r=1}^{Rep}(\hat{\beta}_{gf}^{(r)} - \beta_{gf})/Rep\right|}{F \times G}$$

and

$$ARMSE = \sqrt{\frac{\sum_{g=1}^{G}\sum_{f=1}^{F}\sum_{r=1}^{Rep}(\hat{\beta}_{gf}^{(r)} - \beta_{gf})^2}{Rep \times F \times G}},$$

respectively.

In addition to comparing the various approaches to model estimation, the study examined the relationship between the classification error and the performance of the correction procedures. Because the performance of the correction in the three-step procedure depends on separation among classes and the sample size (Vermunt, 2010), it was expected that the number of correctly classified examinees, referred to here as the effective sample size, N^*, would be closely related to the performance of the correction and thus could potentially inform research about the efficacy of the bias-corrected three-step procedure in real data

situations. $\hat{N}^*$ can be computed as,

$$\hat{N}^* = \sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)I[\boldsymbol{\alpha}_l = \hat{\boldsymbol{\alpha}}_i],$$

Note that $\hat{N}^*$ is also equal to $N \times \hat{\tau}$, where N is the sample size and $\hat{\tau}$ is the pattern-level marginal classification accuracy statistic of Wang et al. (2015).

3.4.3 Results

Although the estimation procedure of the attribute-level logistic regression in Part II of the simulation study is not subsumed by the latent-class level regression in Part I, the performance of the procedures across conditions were quite similar, showing similar patterns of bias and RMSE both overall and at the level of individual parameters. The two procedures also behaved similarly as the number of covariates increased. However, the quality of the latent-class regression parameter estimates were worse due to the extra parameters being estimated - $2^K \times F$ versus $K \times F$ in attribute-level regression. Because of this and the fact that the attribute structure did not have a substantial impact on the results, only the results of attribute-level logistic regression with three covariates and correlated attribute structure are presented in-depth here. The results in their entirety can be requested from the first author.

3.4.3.1 Overall Bias and RMSE

The ABIAS and ARMSE of each approach are presented in Tables 3.2 and 3.3, respectively. Consistent with previous research, no three-step procedure led to lower ABIAS or ARMSE than the one-step approach under any condition. Among the three-step approaches, the uncorrected three-step estimates tended to produce the most ABIAS and ARMSE, underscoring the need for correction weights.

Comparing the corrected three-step procedures across all conditions, the PDL_{ik} performed better than SL_{ik}^{PROP}, and as good or better than SL_{ik}^{MAP}. The PDL_{ik} weights led to greater improvement over SL_{ik}^{MAP} when item quality was low, or when item quality was medium and $J = 10$. For example, when $N = 1000$, $J = 20$, and item quality was low, PDL_{ik} outpeformed SL_{ik}^{MAP} 0.21 versus 0.25 in terms of ABIAS and 0.26 versus 0.31 in terms of ARMSE. When $N = 1000$, $J = 10$, and item quality was medium, PDL_{ik} returned ABIAS and ARMSE values of 0.18 and 0.23, respectively, whereas SL_{ik}^{MAP} returned corresponding values of 0.22 and 0.27.

The proposed correction weights not only performed as well or better than other three-step approaches, they led to the same results as the one-step when item quality was high, or when item quality was medium and $J = 20$. For example, when $N = 500$, $J = 10$ and item quality was high, both the one-step and PDL_{ik} led to ABIAS and ARMSE values of 0.18 and 0.23, respectively, whereas SL_{ik}^{MAP} returned values of 0.21 and 0.27. Similarly, when $N = 500$, $J = 20$, and item quality was medium, the one-step and PDL_{ik} again led to ABIAS and ARMSE values of 0.18 and 0.23, respectively, whereas SL_{ik}^{MAP} returned values of 0.20 and 0.26.

As expected, the discrepancy between the uncorrected and corrected approaches decreased when PCV approached 1.00. Specifically, when $N = 2000$, $J = 20$, and item quality was high, PCV was approximately 0.98, and the ABIAS of the uncorrected estimates was only .04 higher than the ABIAS of the PDL_{ik}-corrected approach.

3.4.3.2 Bias and RMSE at Individual Parameter Level

To illustrate how the different approaches performed at the level of the individual parameters, Tables 3.4 and 3.5 present the bias and RMSE in the estimates of

Table 3.2: ABIAS - Attribute-Level Logistic Regression

				Three-Step			
N	J	IT	One-Step	PDL_{ik}	SL_{ik}^{MAP}	SL_{ik}^{PROP}	Uncor
500	10	High	0.18	0.18	0.21	0.28	0.47
		Medium	0.24	0.26	0.32	0.60	0.93
		Low	0.47	0.77	0.80	0.77	1.32
	20	High	0.16	0.16	0.17	0.18	0.18
		Medium	0.18	0.18	0.20	0.30	0.55
		Low	0.27	0.35	0.41	0.61	1.08
1000	10	High	0.13	0.13	0.15	0.23	0.46
		Medium	0.16	0.18	0.22	0.58	0.92
		Low	0.29	0.44	0.50	1.06	1.28
	20	High	0.11	0.11	0.11	0.12	0.14
		Medium	0.12	0.12	0.14	0.27	0.54
		Low	0.18	0.21	0.25	0.69	1.04
2000	10	High	0.09	0.09	0.10	0.21	0.46
		Medium	0.11	0.13	0.16	0.59	0.91
		Low	0.18	0.25	0.31	1.32	1.25
	20	High	0.08	0.08	0.08	0.10	0.12
		Medium	0.09	0.09	0.10	0.26	0.54
		Low	0.13	0.15	0.18	0.73	1.02

Note. One-step: using the one-step approach; PDL_{ik}: using posterior-distribution level correction weights; SL_{ik}^{MAP}: using sample-level correction weights, with the latent class assignments performed using MAP; SL_{ik}^{PROP}: using sample level correction weights with the latent class assignment done using proportional assignment; Uncor: using the uncorrected three-step approach.

β_{13}, where the true value was equal to two. The bias in the estimates was again consistent with previous research - the one-step tended to overestimate parameter values, whereas the uncorrected three-step procedure led to downward bias, often severe, across all conditions. Both PDL_{ik} and SL_{ik}^{MAP} correction weights were able to adjust for attentuation across all conditions, however when item quality was low, the PDL_{ik} was better able to reduce bias and RMSE in parameter estimates compared to SL_{ik}^{MAP}. For example, when $N = 500$, $J = 20$, and item quality was low, the PDL_{ik} underestimated β_{13} by 0.19, whereas the SL_{ik}^{MAP}

Table 3.3: ARMSE - Attribute-Level Logistic Regression

				Three-Step			
N	J	IT	One-Step	PDL_{ik}	SL_{ik}^{MAP}	SL_{ik}^{PROP}	Uncor
500	10	High	0.23	0.23	0.27	0.37	0.52
		Medium	0.31	0.33	0.40	0.86	0.98
		Low	0.74	0.87	0.90	1.08	1.36
	20	High	0.20	0.20	0.21	0.23	0.22
		Medium	0.23	0.23	0.26	0.40	0.59
		Low	0.36	0.43	0.50	0.87	1.13
1000	10	High	0.16	0.16	0.18	0.29	0.49
		Medium	0.21	0.23	0.27	0.72	0.95
		Low	0.42	0.52	0.59	1.46	1.33
	20	High	0.14	0.14	0.14	0.16	0.18
		Medium	0.15	0.15	0.18	0.35	0.57
		Low	0.23	0.26	0.31	0.87	1.08
2000	10	High	0.11	0.11	0.13	0.26	0.49
		Medium	0.14	0.16	0.20	0.70	0.94
		Low	0.24	0.31	0.38	1.60	1.29
	20	High	0.10	0.10	0.10	0.12	0.15
		Medium	0.11	0.11	0.13	0.31	0.56
		Low	0.17	0.18	0.22	0.86	1.06

Note. One-step: using the one-step approach; PDL_{ik}: using posterior-distribution level correction weights; SL_{ik}^{MAP}: using sample-level correction weights, with the latent class assignments performed using MAP; SL_{ik}^{PROP}: using sample level correction weights with the latent class assignment done using proportional assignment; Uncor: using the uncorrected three-step approach

underestimated it by 0.28. The RMSE of the PDL_{ik} estimates was also lower, at 0.48, than the RMSE returned by SL_{ik}^{MAP}, 0.55. More generally, larger reductions in RMSE under low and medium item quality were seen using PDL_{ik} in place of SL_{ik}^{MAP}. It is worth noting that this pattern was consistent across all conditions tested, including the number of covariates; that is, the greater the bias and RMSE, the greater the improvement obtained by using PDL_{ik} rather than SL_{ik}^{MAP}.

Furthermore, when item quality was medium or high, the RMSE of the PDL_{ik} estimates was within 0.02 of the RMSE from the one-step approach. Under these

same conditions, the RMSE of the SL_{ik}^{MAP} estimates was within 0.08 of the one-step. The bias of PDL_{ik} and SL_{ik}^{MAP} estimates was within 0.02 of the one-step approach under eight and seven of the eighteen conditions, respectively. Finally, the performance of SL_{ik}^{PROP} was poor, sometimes dramatically over or under-estimating the parameter.

Table 3.4: Bias - Attribute-Level Logistic Regression

$\beta_{13} = 2$							
				Three-Step			
N	J	IT	One-Step	PDL_{ik}	SL_{ik}^{MAP}	SL_{ik}^{PROP}	Uncor
500	10	High	0.00	-0.04	-0.05	0.19	-0.59
		Medium	0.04	-0.15	-0.20	0.53	-1.11
		Low	0.27	-0.87	-0.90	-0.97	-1.57
	20	High	0.02	0.04	0.04	0.08	-0.11
		Medium	0.04	-0.05	-0.08	0.27	-0.65
		Low	0.19	-0.19	-0.28	0.66	-1.25
1000	10	High	0.01	0.02	-0.02	0.22	-0.53
		Medium	0.01	-0.05	-0.09	0.65	-1.09
		Low	0.13	-0.38	-0.43	1.10	-1.52
	20	High	-0.01	-0.01	-0.01	0.04	-0.14
		Medium	0.01	0.01	0.01	0.33	-0.63
		Low	0.05	-0.05	-0.09	0.92	-1.21
2000	10	High	0.01	0.01	0.01	0.25	-0.56
		Medium	0.00	0.02	-0.02	0.74	-1.07
		Low	0.01	-0.15	-0.25	1.49	-1.48
	20	High	0.02	0.02	0.02	0.07	-0.12
		Medium	0.01	0.01	0.02	0.30	-0.64
		Low	0.02	-0.03	-0.09	0.90	-1.22

Note. One-step: using the one-step approach; PDL_{ik}: using posterior-distribution level correction weights; SL_{ik}^{MAP}: using sample-level correction weights, with the latent class assignments performed using MAP; SL_{ik}^{PROP}: using sample level correction weights with the latent class assignment done using proportional assignment; Uncor: using the uncorrected three-step approach

Table 3.5: RMSE - Attribute-Level Logistic Regression

$\beta_{13} = 2$							
				Three-Step			
N	J	IT	One-Step	PDL_{ik}	SL_{ik}^{MAP}	SL_{ik}^{PROP}	Uncor
500	10	High	0.24	0.25	0.29	0.38	0.63
		Medium	0.32	0.34	0.40	0.82	1.12
		Low	0.92	0.99	1.03	1.17	1.57
	20	High	0.22	0.22	0.23	0.26	0.24
		Medium	0.25	0.26	0.28	0.42	0.68
		Low	0.45	0.48	0.55	1.16	1.27
1000	10	High	0.19	0.20	0.21	0.33	0.55
		Medium	0.26	0.24	0.32	0.82	1.09
		Low	0.47	0.56	0.65	1.56	1.52
	20	High	0.15	0.15	0.15	0.16	0.20
		Medium	0.17	0.17	0.20	0.41	0.65
		Low	0.25	0.26	0.34	1.09	1.21
2000	10	High	0.10	0.10	0.12	0.30	0.57
		Medium	0.16	0.16	0.22	0.83	1.08
		Low	0.26	0.32	0.44	1.73	1.48
	20	High	0.11	0.11	0.12	0.15	0.17
		Medium	0.11	0.11	0.13	0.34	0.65
		Low	0.18	0.19	0.23	0.98	1.22

Note. One-step: using the one-step approach; PDL_{ik}: using posterior-distribution level correction weights; SL_{ik}^{MAP}: using sample-level correction weights, with the latent class assignments performed using MAP; SL_{ik}^{PROP}: using sample level correction weights with the latent class assignment done using proportional assignment; Uncor: using the uncorrected three-step approach

3.4.3.3 Separation of Likelihood

One particular problem frequently encountered in the simulation study was separation of likelihood, defined as when the procedure returned values that approached infinity, or, in this context, were equal to the bounds set on the estimation method. This problem has been known since Albert and Anderson (1984), and has been addressed in numerous contexts (see; Heinze & Schemper,

Table 3.6: Replications with Separated Likelihood

			$F = 3$			$F = 9$			$F = 12$		
N	J	IT	One	PDL	SL	One	PDL	SL	One	PDL	SL
500	10	High	0	0	0	0	2	9	2	10	19
		Medium	1	1	1	4	15	17	7	49	61
		Low	1	1	2	16	17	37	19	47	85
	20	High	0	0	0	0	0	0	0	2	3
		Medium	0	0	0	0	5	5	1	12	23
		Low	0	0	0	4	11	17	9	30	58
1000	10	High	0	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	1	1	0	2	2
		Low	0	0	0	4	6	10	0	18	35
	20	High	0	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0	0
		Low	0	0	0	1	0	0	0	3	11
2000	10	High	0	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0	0
		Low	0	0	0	0	0	0	0	0	0
	20	High	0	0	0	0	0	0	0	0	0
		Medium	0	0	0	0	0	0	0	0	0
		Low	0	0	0	0	0	0	0	0	0

Note. One-step: using the one-step approach; PDL_{ik}: using posterior-distribution level correction weights; SL_{ik}^{MAP}: using sample-level correction weights, with the latent class assignments performed using MAP; SL_{ik}^{PROP}: using sample level correction weights with the latent class assignment done using proportional assignment; Uncor: using the uncorrected three-step approach

2002; Santos & Barrios, 2017), and, when it occurs, severely limits the practicability of a procedure. Replications with likelihood separation were considered as estimates that failed to converge and were eliminated from all computations. Table 3.6 compares the proposed correction weights, the next-best performing correction weight, SL_{ik}^{MAP}, and the one-step approach for attribute-level logistic regression. The table values indicate the number of replications, out of 100 total, where one or more estimate was equal to the bounds set on the estimation algorithm. When $F = 3$, separation of the likelihood occurred at most once out

of 100 replications for the PDL_{ik} and one-step approaches. When the number of covariates increased to $F = 9$ or 12, however, the number of replications with separated likelihood increased, particularly when $N = 500$ and item quality was medium or low. Compared to SL_{ik}^{MAP}, the proposed correction weights led to approximately half the instances of likelihood separation when $F = 9$ or 12. For example, when $N = 500$, $J = 20$, and item quality was medium, regressing attributes onto 12 covariates using SL_{ik}^{MAP} in a corrected three-step approach led to 23 occurrences of likelihood separation, compared to only 12 when using PDL_{ik}, and one when using the one-step approach. However, it should be noted that 12 independent covariates is relatively unlikely to occur in education and social science data. Some of the predictors would almost certainly be highly correlated, and it is unclear how that collinearity would affect the rate of nonconvergence. In cases where separation of likelihood is an issue, it can be resolved by using a prior distribution (Garre & Vermunt, 2006). Note that separation of likelihood occurred more frequently under latent-class regression, though the performance of the approaches followed the same patterns as described above. Likewise, the same conclusions can be drawn for both latent-class and attribute-level regression - overall, convergence failure appears to present an impediment to the scaling of the three-step procedure to more covariates. The proposed correction weights outperformed other three-step procedures in this regard, however the results still suggest the need for variable selection.

3.4.3.4 Effective Sample Size

The relationship between the PCV and the ABIAS of parameter estimates from both the PDL_{ik} and uncorrected approaches are plotted in Figure 3.1. As the linear trend in the plot would suggest, the correlation between the ABIAS of the uncorrected three-step procedure and the PCV was -0.99 when $F = 3$ and the

attribute structure was correlated. This relationship was strong across all conditions, with an average correlation of -0.96. As can be seen in Figure 3.1, the correction procedure undoes this linear relationship. Across all conditions, the correlation between the log-transformed ABIAS from PDL_{ik} and the PCV was high, at $= -0.76$. However, the log-transformed ABIAS was much more strongly correlated with the estimated effective sample size, with an average correlation across all conditions of -0.90. This strong relationship suggests that computing $\hat{N}^*$ could help a researcher predict the magnitude of the bias in parameter estimates when applying the corrected three-step procedure to real data.

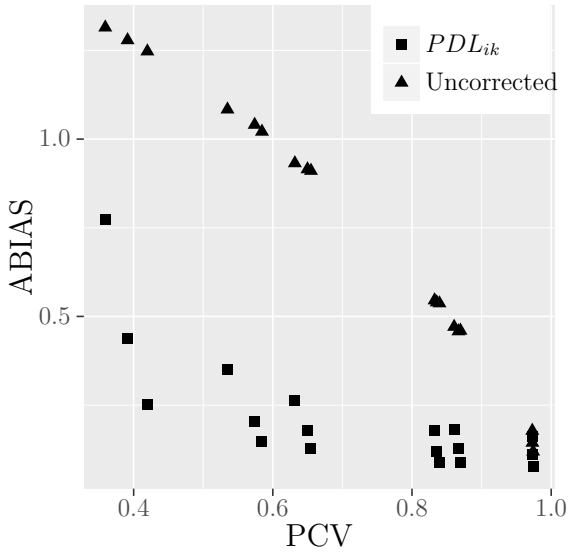

Figure 3.1: Relationship between ABIAS and the PCV

3.5 Empirical Example

Personality assessments administered by clinicians are intended to discern the true state of patients. The presence of particular disorders is observable through responses to items, making latent variable models prime candidates for modeling assessment data. Given the comorbidity of personality disorders, assessment items tend to measure more than one particular disorder, and researchers have

struggled with the measurement implications (Krueger & Eaton, 2010), particularly for traditional sum scores (Rossi, Elklit, & Simonsen, 2010). CDMs offer test developers a way to accommodate this cormorbidity when classifying patients by modeling attribute interactions. Through the incorporation of the comorbidity in the measurement model, CDMs may be able to better distinguish among the different psychological profiles presented by patients. Furthermore, researchers may want to understand the relationship between the patient covariates and the putative disorder profiles. This empirical example should illustrate how this relationship can be evaluated.

For this example of attribute-level regression, a Dutch-language version of the Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009) was used, with attribute and Q-matrix specification initially having been performed in de la Torre et al. (2015) and subsequently refined in Ma, Iaconangelo, and de la Torre (2016). The three attributes were: H = somatoform; SS = thought disorder; CC = major depression. Each of the 30 items measured 1, 2, or 3 of the attributes in 739 examinees. One dichotomous and one continuous covariate were used. The first was Setting, with 400 (54%) clinical patients, and 339 (46%) prisoners. The second covariate was Age, which ranged from 18 to 74 years. See Rossi et al. (2010) for more details on the dataset.

To show the improvements in parameter estimates from the corrected three-step procedures, a subset of less-discriminating items, referred to as subset B in Ma et al. (2016), was used in conjunction with 100 random samples of 300 examinees. The average attribute-level regression parameter estimates from the three-step procedures are presented in Table 4.8, along with the one-step parameter estimates obtained using the full dataset ($N = 739$, $J = 30$). Under the less favorable test condition ($N = 300$, item subset B), the correction weights, specifically the PDL_{ik} weights, returned estimates that were much closer to the

one-step estimates than the uncorrected values. For example, when regressing attribute H onto covariate Setting, the uncorrected three-step led to a parameter estimate of -1.47, whereas the PDL_{ik} returned a parameter estimate of 2.00, equal to the one-step estimate. The parameter estimates obtained when regressing attribute CC onto covariate Age suggest that the correction weights avoid overestimating non-significant covariates, though currently it is unclear how to implement a variable selection procedure in this context. No separation of likelihood occurred under any of the replications. However, when latent-class level regression models were fitted, over half of the replications with the less-favorable test conditions failed to converge. This was likely related to the highly uneven distribution of examinees across latent classes. Because of the problematic implementation, the results are not presented here. Overall, treating the one-step estimates of the attribute-level regression coefficients as the closest to the true values, the results corroborate the simulation study finding that the correction weights in general, and the PDL_{ik} weights specifically, can substantially improve parameter estimation when test conditions are poor.

Table 3.7: MCMI-III

			N=300, Item Subset B			
Attribute	Covariate	One-Step	PDL_{ik}	SL_{ik}^{MAP}	SL_{ik}^{PROP}	Uncor
H	Setting	-2.00	-2.00	-1.88	-2.13	-1.47
	Age	0.12	0.12	0.12	0.13	0.10
SS	Setting	-2.22	-2.19	-2.09	-2.27	-1.81
	Age	0.33	0.22	0.23	0.24	0.19
CC	Setting	-1.55	-1.57	-1.53	-1.60	-1.36
	Age	0.09	0.08	0.08	0.08	0.07

Note. One-step: using the one-step approach; PDL_{ik}: using posterior-distribution level correction weights; SL_{ik}^{MAP}: using sample-level correction weights, with the latent class assignments performed using MAP; SL_{ik}^{PROP}: using sample level correction weights with the latent class assignment done using proportional assignment; Uncor: using the uncorrected three-step approach

3.6 Discussion

The study extended existing LCM methodologies to the CDM framework and developed improved versions of the procedures. First, it applied the sample-level correction weights to a three-step procedure relating covariates to the latent classes using multinomial logistic regression. Second, it showed how these sample-level weights can be modified for attribute-level logistic regression. Furthermore, versions of the weights that used each examinee's posterior distribution were proposed for both the latent-class level and attribute-level regression. The simulation results showed that these proposed correction weights outperformed the best sample-level correction weights (SL_{ik}^{MAP}) in terms of bias and RMSE. The proposed correction weights also led to fewer instances of likelihood separation, improving the probability that a researcher would be able to study the relationship between classification and covariates. Given that the posterior-distribution level weights are no more difficult to compute than the versions already existing in the literature, they appear to be an unqualified improvement.

Not only did the proposed correction weights outperform the alternative three-step procedures, they often performed approximately as well as the one-step approach. In many or most of the tested conditions, a secondary researcher would have arrived at virtually the same parameter estimates using either the one-step or the PDL-corrected three-step approach. This affords the secondary researcher greater flexibility with modeling decisions while still obtaining parameter estimates that lead to valid interpretations about the relationship between the classification and the examinee covariates. The procedures presented here are also more straightforward than fitting a conditioning model, drawing plausible values, and regressing them onto covariates, although this could be investigated in future research. In terms of practical constraints on researchers, often the full item

responses are not released for secondary researchers - the three-step approaches presented here only require the examinee posterior distributions.

This work is not without shortcomings. Further research should focus on how to accommodate larger numbers of covariates in the regression model to address the separation of likelihood issue. Additionally, variable selection methods need to be adapted to the three-step estimation procedure, with particular attention paid to addressing the sort of collinear covariates seen in the social sciences. This will further improve the utility of the three-step approach.

3.7 References

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students NAEP math performance* (Tech. Rep. No. CSE Technical Report 478). CRESST/University of California, Los Angeles.

Albert, A., & Anderson, J. A. (1984). On the existence of maxmimum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1-10.

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification*, *30*, 195-224.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2013). Relating latent class assignments to external variables: Standard errors for corrected inference. *Sociology*, *83*, 173-178.

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*, 272-311.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-1386.

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

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. United Kingdom: Chapman and Hall.

Darling-Hammond, L. (2000). Teacher quality and student achievement: a review of state policy evidence. *Education Policy Analysis*, *8*, 1-44.

Dayton, C., & Macready, G. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173-178.

Dayton, C., & Macready, G. (2002). Use of categorical and continuous covariates in latent class analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (p. 213-233). Cambridge, UK: Cambridge University Press.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353.

de la Torre, J., van der Ark, L., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 1-16.

Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43-59.

Goodman, L. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, *37*, 1-22.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*, 2409-2419.

Henson, Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

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

Huebner, A., & Wang, C. (2011). Comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407-419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Krueger, R. F., & Eaton, N. R. (2010). Personality traits and the classification of mental disorders: Toward a more complete integration in dsm-5 and an empirical model of psychopathology. *Personality Disorders: Theory, Research, and Treatment*, 97-118.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59-81.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200-217.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III Manual (4th ed.)*. Minneapolis, MN: Pearson Assessments.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177-196.

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika*, *58*, 79-85.

National Center for Education Statistics. (2008). NAEP secondary analysis grant abstracts. Retrieved from `https://nces.ed.gov`

National Center for Education Statistics. (2011a). NAEP technical documentation: Achievement gaps. Retrieved from `https://nces.ed.gov`

National Center for Education Statistics. (2011b). NAEP technical documentation: NAEP population-structure models. Retrieved from `https://nces.ed.gov`

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. New York, NY: Springer.

Park, Y., & Lee, Y. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Applied Psychological Measurement*, *38*, 376-390.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Robitaille, D. F., & Beaton, A. E. (2002). *Secondary analysis of the TIMMS data*. USA: Springer.

Rossi, G., Elklit, A., & Simonsen, E. (2010). Empirical evidence for a four factor framework of personality disorder organization: Multigroup confirmatory factor analysis of the Millon clinical multiaxial inventory - III personality disorder scales across Belgian and Danish data samples. *Journal of Personality Disorders*, *24*, 128-150.

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Santos, K. C. P., & Barrios, E. B. (2017). Improving predictive accuracy of logistic regression model using ranked set samples. *Communications in Statistics - Simulation and Computation*, *46*, 78-90.

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, *80*, 727-747.

Sinharay, S., Puhan, G., & Haberman, S. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*, 29-40.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP):

Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26.* Amsterdam: Elsevier.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*, 457-476.

Chapter 4

Variable Selection in the Three-Step Approach to Modeling Cognitive Diagnosis Models and Covariates: The Latent-Class Lasso

4.1 Introduction

Cognitive diagnosis models (CDMs) are a relatively recent offshoot of latent class models (LCM), which date back to Lazarsfeld (1950) and have well-developed methodologies for not only classifying examinees, but for relating those classifications to covariates. The majority of the literature integrates the measurement and structural models in a simultaneous estimation procedure known as the one-step approach (for some examples, see Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Dayton & Macready, 1988, 2002; Huang & Bandeen-Roche, 2004). Additionally, a three-step approach was introduced, with correction weights developed to decrease bias in parameter estimates (Vermunt, 2010). Similar developments in the item response theory (IRT) context have been motivated by the National Assessment of Educational Progress (NAEP), which as part of its mission relates student achievement to student and school characteristics, and provides the necessary data for secondary researchers to study such relationships (for an overview, see Schofield, Junker, Taylor, & Black, 2015). In the literature, cognitively diagnostic assessments (CDAs) have been proposed for applications ranging from large-scale assessment, such as TIMMS (Lee, Park, & Taylan, 2011), to intelligent tutoring systems used in the classroom (Ye, Fellouris, Culpepper, & Douglas,

2016). In spite of the breadth of these potential applications, limited work has been done to develop methodologies that allow researchers to relate examinee classifications to covariates.

Ayers, Rabe-Hesketh, and Nugent (2013) developed a one-step approach that simultaneously estimated the measurement and structural models, in this case the deterministic input, noisy, "and" gate (DINA; Haertel, 1989) model, and the logistic regression model. The covariates affected the probability of the examinee mastering the attribute, known as differential skill functioning. Park and Lee (2014) extended the one-step such that the covariates affected the probability of the examinee answering the item correctly, known as differential item functioning. More recently, a three-step approach to estimating CDMs and covariates was proposed (Iaconangelo & de la Torre, 2016). This approach separates the fitting of the measurement model and the structural model, which is consistent with the way secondary research is often done.

There are a variety of methods designed to work with high-dimensional regression when many collinear covariates are available, as is often the case in the social sciences (National Center for Education Statistics, 2011). One popular example of these methods designed for shrinkage and variable selection is L_1 regularization, known as the least absolute shrinkage and selection operator (lasso), introduced in the regression context by Tibshirani (1996). However, other than the work done with the Bayesian lasso in Culpepper and Park (2017), there has been little to no investigation of the performance of the lasso in the latent-class regression context. In this research, the lasso will be implemented as a variable/model selection procedure. Correction weights developed in Iaconangelo and de la Torre (2016) that led to improvements in latent regression are incorporated with the L_1 regularization, and this research aims to determine if the use of correction weights also leads to more accurate variable selection. Specifically, in the case where a

large number of collinear variables are available, the correction weights may lead to greater sparsity - that is, more irrelevant predictors should be dropped from the model. For the researcher performing secondary data analysis, more accurate variable selection and more accurate structural parameter estimation should lead to better inferences about the relationship between the predictor variables and student attribute mastery.

The remainder of this manuscript is organized accordingly: First is a brief background on CDMs, followed by a review of the one-step and three-step approaches to modeling covariates and examinee classifications. After that is a summary of the lasso and cross-validation. The next section presents the latent-class lasso. Following that is a simulation study to compare the procedure to the standard lasso, as well as an application to real data. The paper concludes with a discussion of the uses and limitations of the work.

4.2 Cognitive Diagnosis Models

The IRT framework widely used in standardardized assessment typically features a unidimensional test that allows examinees to be ranked on a single latent trait. By contrast, the CDM framework measures multidimensional skills, referred to as attributes, on which examinees are classified as either having mastered or not mastered (de la Torre, 2011). The attributes (and thus the attribute vectors or latent classes) are typically determined before administering the assessment by content experts (DiBello & Stout, 2007), unlike traditional LCA, which determines the number of latent classes as part of the model-fitting process (McCutcheon, 1987; van der Ark, van der Palm, & Sijtsma, 2011). As a result, it is possible for a CDA to create latent classes that would not be statistically significant in cluster analysis methods but are nontheless of substantive interest. This detailed feedback on examinee skills could be used in classroom assessments via e-learning platforms

(Ye et al., 2016). Alternatively, CDMs can be used in other contexts, like patient reported outcomes, where the multidimensional nature of the patient classification could provide advantages over other methods. For applications to clinical psychology, see de la Torre, van der Ark, and Rossi (2015) and the empirical example in this manuscript.

The G-DINA Model

To maximize the generalizability of the research, a general CDM was used in the simulation study and real data example. General CDMs, as saturated models, subsume specific CDMs, which constrain parameters in a manner consistent with the way attributes are theorized to interact. There are three general CDMs in the literature: the general diagnostic model (GDM; von Davier, 2008), the log-linear CDM (LLM; Henson, Templin, & Willse, 2009), and the model used throughout this study, the Generalized DINA model (G-DINA; de la Torre, 2011).

The item-attribute relationship is specified by the Q-matrix (Tatsuoka, 1983). The number of attributes assessed in item j is denoted K_j^*, the row-sum of the Q-matrix. The examinee attribute vector can be written $\boldsymbol{\alpha}_l$=$\{\alpha_{l1}, \ldots, \alpha_{lK}\}$, where $l = 1, \ldots, 2^K$ denotes the latent classes, and $k = 1, \ldots, K$ the attributes. The kth element of the vector is 1 when the examinee has mastered the kth attribute, and is 0 when the examinee has not. Let $\boldsymbol{\alpha}_{lj}^*$ be the reduced attribute vector containing only the required attributes, where $l = 1, \ldots 2^{K_j^*}$. The probability of an examinee with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ answering item j correctly will be denoted by $P(\boldsymbol{\alpha}_{lj}^*)$. The G-DINA item response function is written as

$$P(\boldsymbol{\alpha}_{lj}^*) = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \phi_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk},$$

where ϕ_{j0} is the intercept for item j, ϕ_{jk} is the main effect due to α_k, $\phi_{jkk'}$ is the

interaction effect due to α_k and $\alpha_{k'}$, and $\phi_{j12...K_j^*}$ is the interaction effect due to $\alpha_1, \cdots, \alpha_{K_j^*}$.

The parameter estimates can be obtained via marginal maximum likelihood estimation. The likelihood is written as

$$L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_l)^{X_{ij}}[1 - P_j(\boldsymbol{\alpha}_l)]^{1-X_{ij}},$$

which is then marginalized across the latent class proportions, $P(\boldsymbol{\alpha}_l)$. The marginal likelihood function is written as

$$L(\boldsymbol{X}) = \prod_{i=1}^{N} \sum_{l=1}^{2^K} L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l),$$

the log of which is the objective function, optimized with respect to the item parameters. For more details on the estimation procedure, see de la Torre (2009b, 2011). The posterior distribution of examinee i is written as

$$P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i) \propto L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)P(\boldsymbol{\alpha}_l),$$

and normalized to sum to one. The examinee posterior distributions were subsequently used in the second step of the three-step procedure to assign the examinee attribute classification. Typically in the CDM literature, the maximum a posteriori (MAP) or expected a posteriori (EAP) classification method is used (Huebner & Wang, 2011).

This study also uses the EAP-based mean assignment to aggregate the posterior probabilities $P(\boldsymbol{\alpha}_l|\boldsymbol{X}_i)$ into the marginalized attribute-level probabilities, $P(\alpha_k|\boldsymbol{X}_i)$, where α_k is equal to 1 or 0, to indicate attribute mastery or non-mastery, respectively. Note that $P(\alpha_k = 0|\boldsymbol{X}_i) = 1 - P(\alpha_k = 1|\boldsymbol{X}_i)$.

In this work, α_k refers to the potential true values (0 or 1) of the attribute, and

the possible attribute assignments are drawn from the same sample space, and are denoted by α_q. The estimated examinee i classification on attribute k is written $\hat{\alpha}_{ik}$, and can be considered a realization of the possible attribute assignments.

4.3 Modeling the Relationship between Covariates and Latent Classification

4.3.1 The One-Step Approach

The advantage of using the one-step approach is that simultaneously estimating the measurement and structural models leads to unbiased estimates and the smallest standard errors (efficient estimator)(Bakk, Tekle, & Vermunt, 2013; Dayton & Macready, 1988; Iaconangelo & de la Torre, 2016; Vermunt, 2010). However, variable selection with a large number of collinear covariates requires re-fitting not only the regression model, but also the CDM, every time the predictors are modified. Furthermore, which predictors are included may even influence selection of a CDM and Q-matrix validation. This makes the procedure computationally intensive, particularly so when cross-validation is used with the lasso. Furthermore, in the context of secondary research, model selection, item parameter estimation, Q-matrix validation, and examinee attribute classification may have already been determined. However, in the one-step model, the item parameters and classifications will depend on the covariates in the model, complicating the validity of conclusions based on parameters/classifications that were not used in practice. That is, substantive researchers may see circular logic in interpreting the relationship between classifications influenced by covariates and the covariates themselves (Bakk, Oberski, & Vermunt, 2013). This circularity can be avoided with the three-step approach. In terms of practical restrictions on researchers, often the full item responses are not available for the one-step approach.

4.3.2 The Uncorrected Three-Step Approach

In the first step, the measurement model (here the G-DINA model) is fitted to the item responses. In the second step, examinees are assigned to a latent class, or, equivalently, assigned attribute mastery or non-mastery. In the third step, the relationship between covariates and latent classification is estimated. This research focuses on the relationship between the attribute classifications and the predictors rather than the latent classes and the predictors, because this approach can better accommodate large numbers of covariates and attributes - the number of parameters estimated is $K \times F$, rather than $2^K \times F$, where F is the number of examinee covariates. The estimated attribute classifications are regressed onto the covariates via a logistic regression model, written as

$$P(\alpha_k = 1|\boldsymbol{Z}_i) = \frac{\exp(\beta_{k0} + \sum_{f=1}^{F} \beta_{kf} Z_{if})}{1 + \exp(\beta_{k0} + \sum_{f=1}^{F} \beta_{kf} Z_{if})},$$

where $\boldsymbol{Z}_i = \{Z_{i1}, \ldots Z_{if}, \ldots Z_{iF}\}$ is the vector of F covariates for examinee i, and the parameters of interest are the attribute by covariate coefficients, $\boldsymbol{\beta} = \{\beta_{k0}, \beta_{k1}, \ldots, \beta_{kF}\}$. $P(\alpha_k = 1|\boldsymbol{Z}_i)$ can be interpreted as the marginalized probability of mastering attribute k given the covariates, $\boldsymbol{Z}$. Note that $P(\alpha_k = 0|\boldsymbol{Z}_i) = 1 - P(\alpha_k|\boldsymbol{Z}_i)$.

The uncorrected approach treats the estimated examinee attribute classification as an observed dependent variable in the logistic regression function, which can be written as

$$logL = \sum_{i=1}^{N} log[P(\hat{\alpha}_{ik}|\boldsymbol{Z}_i)].$$

As was demonstrated in Iaconangelo and de la Torre (2016), this leads to biased estimates of $\boldsymbol{\beta}$, impacting the validity of the inferences made using this approach.

4.3.3 The Three-Step Approach with Correction Weights

Iaconangelo and de la Torre (2016) developed sample-level and posterior-distribution level correction weights for attribute-level regression. A 2×2 matrix of attribute-classification error probabilities was computed

$$P(\alpha_q|\alpha_k, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\alpha_k|\boldsymbol{X}_i) I[\hat{\alpha}_{ik} = \alpha_q]}{\sum_{i=1}^{N} P(\alpha_k|\boldsymbol{X}_i)},$$

and can be interpreted as the proportion of examinees assigned attribute mastery α_q given true attribute mastery α_k. For example, the entries of column 2 are

$$P(\alpha_q = 1|\alpha_k = 0, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\alpha_k = 0|\boldsymbol{X}_i) I[\hat{\alpha}_{ik} = 1]}{\sum_{i=1}^{N} P(\alpha_k = 0|\boldsymbol{X}_i)},$$

and

$$P(\alpha_q = 1|\alpha_k = 1, \boldsymbol{X}) = \frac{\sum_{i=1}^{N} P(\alpha_k = 1|\boldsymbol{X}_i) I[\hat{\alpha}_{ik} = 1]}{\sum_{i=1}^{N} P(\alpha_k = 1|\boldsymbol{X}_i)},$$

where the former can be interpreted as the proportion of examinees incorrectly classified as having mastered attribute k, and the latter as the proportion of examinees correctly classified as having mastered attribute k.

The sample-level correction weights for examinee i with attribute-level latent class assignment α_q were equal to the column of the matrix $P(\alpha_q|\alpha_k, \boldsymbol{X})$ that corresponds to the latent class assignment, written

$$SL_{ik} = P(\alpha_q|\alpha_k, \boldsymbol{X}) I[\hat{\alpha}_{ik} = \alpha_q].$$

Using the examinee posterior distribution rather than the sample-level joint attribute distribution, the posterior-distribution level correction weights were calculated

$$PDL_{ik} = \frac{P(\alpha_k|\boldsymbol{X}_i)}{P(\alpha_k)},$$

where $P(\alpha_k)$ is the sample-level proportion of mastery of attribute k. The correction weights can be used in a modified attribute-level logistic regression log-likelihood, written

$$\log L = \sum_{i=1}^{N} \log \sum_{\alpha_k=0}^{1} P(\alpha_k|\boldsymbol{Z}_i) w_{ik}, \tag{4.1}$$

where w_{ik} is equal to SL_{ik} or PDL_{ik}. Equation 4.1 effectively averages the regression probabilities over the probabilities of latent class assignments rather than treating the estimated assignment as observed. This adjusts for the measurement error in the attribute assignment when estimating the regression parameters. Note that the SL_{ik} weights are the column of the matrix of classification error probabilities that corresponds to the latent class assignment. Optimizing this objective function led to reduced bias and RMSE in the estimates of the $\boldsymbol{\beta}$ parameters in $P(\alpha_k|\boldsymbol{Z}_i)$, the logistic regression model relating the attributes to the covariates (Iaconangelo & de la Torre, 2016).

4.4 Variable Selection with the Lasso

It is widely known that regularization is necessary for the model to generalize well to new data, and a popular choice is the lasso, which can be used for variable selection as well as shrinkage (Hastie, Tibshirani, & Friedman, 2009; Tibshirani, 1996). Specifically, it is a popular technique for shrinking the coefficients of poor predictors to zero, thereby removing irrelevant predictors. Shrinking estimates to zero creates a sparse statistical model, one having only a small number of relevant predictors, making it easier to estimate and interpret than a dense model (Hastie, Tibshirani, & Wainwright, 2015). The lasso can be implemented with a large number of predictors to identify a smaller subset that exhibit the strongest effects (Hastie et al., 2015). In the last step of the uncorrected three-step procedure, the L_1 penalty can be incorporated in the objective function of the attribute-level

logistic regression, written as

$$\log L_{lasso} = \sum_{i=1}^{N} \log\Big[P(\hat{\alpha}_{ik}|\boldsymbol{Z}_i) - \lambda \sum_{f=1}^{F} |\beta_{kf}|\Big], \tag{4.2}$$

where λ is referred to as the tuning parameter and controls the amount of shrinkage (Hastie et al., 2009). Increasing the size of the tuning parameter increases the L_1 penalty term and reduces the number of variables with nonzero coefficients in the model. Given the coefficient estimates and a λ value, a loss function can be used to assess the quality of the model. Although there are several loss functions in the lasso literature, they all measure the prediction error of the model for an independent test data from the same population. A procedure known as cross validation randomly divides the data into K-1 training datasets used to fit the model, and 1 test dataset used to estimate the performance via the loss function. Repeating the process over all K datasets ensures that all data is used as the test data, and this process returns K prediction errors that are averaged for each value of λ. A range of λ values is used and the lambda that corresponds to the minimal loss is selected as the optimal value of the tuning parameter. The cumulative process is referred to as K-fold cross validation. Thus, using a range of tuning parameters and comparing predictive accuracy of the various models allows the lasso to select variables. The researcher can then fit an unpenalized regression model to obtain the best estimate of the relationship between latent classification and the selected covariates. This process is known as the relaxed lasso (Meinshausen, 2007). The estimates resulting from the variable selection and subsequent corrected (unpenalized) regression model are referred to here as the final parameter estimates.

4.4.1 The Latent-Class Lasso

With the exception of the Bayesian lasso approach in Culpepper and Park (2017), the extensive literature on the lasso does not appear to specifically address using the lasso in the latent regression (or latent-class regression) context. This research modifies this well-known approach to address the fact that the dependent variable is a latent variable. Equation 4.2 uses the estimated examinee classification $\hat{\alpha}_{ik}$, effectively treating the dependent variable as observed, and is referred to here as the standard lasso. The proposed approach incorporates the L_1 penalty term in the objective function of the corrected three-step, Equation 4.1. This is referred as the latent-class lasso, and is written as

$$\log L_{LCL} = \sum_{i=1}^{N} \log \sum_{\alpha_k=0}^{1} P(\alpha_k|\boldsymbol{Z}_i) w_{ik} - \lambda \sum_{f=1}^{F} |\beta_{kf}|,$$

where w_{ik} can be either PDL_{ik} or SL_{ik}. This recognizes the uncertainty in the attribute assignment and finds a weighted average. In doing so, the correction weights account for the latent nature of the dependent variable, and in the process improve performance, in terms of variable selection and quality of the final parameter estimates.

4.5 Evaluating the Performance of the Correction Weights via Simulation Study

A simulation study was designed to compare the standard and latent-class lasso to variable selection in the third step of the three-step approach. The final parameter estimates from an unpenalized regression model represented the cumulative effect of incorporating or ignoring the correction weights in the procedure.

4.5.1 Design

In the studies, a sparse data condition is generated, consisting of 12 Gaussian covariates, $Z_1, Z_2, ...Z_{12}$, moderately correlated ($\rho = 0.5$) within blocks of four, but uncorrelated between blocks. The generating model has nonzero coefficients for three variables, one drawn from each block, and these coefficient values formed a matrix of

$$\boldsymbol{\beta} = \begin{bmatrix} 2.0 & 1.0 & 0.5 \\ 0.0 & 2.0 & 1.0 \\ 1.0 & 0.5 & 2.0 \end{bmatrix}.$$

Here the rows correspond to the number of covariates and the columns to the number of attributes. To simplify computations and analysis, and without loss of generality, the intercept β_0 was set equal to 0. These parameters were used to generate $K = 3$ attributes with correlations ranging from .5 to .8, as suggested by Kunina-Habenicht, Rupp, and Wilhelm (2012) and Sinharay, Puhan, and Haberman (2011). Factors manipulated were the sample size ($N = 500$, 1000, and 2000), the test length ($J = 10$ and 20), and item quality (High, Medium, and Low), operationalized as the values of the guessing and slip parameters ($g = s = .1$, .2, and .3). The ten-item Q-matrix to be used in the study is presented in Table 4.1. It was doubled for the $J = 20$ condition. Across the tested factors, the classification accuracy varied from a low of approximately 42%, when $N = 500$, $J = 10$, and item quality was low, to a high of approximately 98%, when $N = 2000$, $J = 10$, and item quality was high. 100 replications were generated for each condition.

In the first step, the G-DINA model was fitted to the item responses. The resulting examinee attribute classifications were then regressed onto the full 12 covariates. The standard and latent-class lasso were implemented, and the value of λ for each was determined by five-fold cross validation using predictive log-loss

Table 4.1: Ten-Item Q-matrix

item	α_1	α_2	α_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	1	0
6	0	0	1
7	1	1	0
8	1	0	1
9	0	1	1
10	1	1	1

as the criterion (Hastie et al., 2009; Wang & Gelman, 2015). The corrected approach used the nonzero coefficients from the latent-class lasso in an unpenalized (i.e., $\lambda = 0$) regression model with correction weights, returning the corrected final parameter estimates. The uncorrected approach used the nonzero coefficients from the standard lasso in an unpenalized regression model without correction weights, returning the uncorrected final parameter estimates. The corrected and uncorrected final parameter estimates were compared to evaluate the cumulative effect of treating the latent classifications as observed dependent variables.

Typical implementation of logistic regression involves the Newton-Raphson algorithm, and popular implementation of lasso, such as the `glmnet` R package (Friedman, Hastie, & Tibshirani, 2010; R Core Team, 2016), employs cyclical coordinate descent algorithms. In these cases, the implementation of the correction weights is not straightforward. However, the objective function including the correction weights can be optimized directly using the `lbfgs` package (Coppola, Stewart, & Okazaki, 2014) in R, which implements the Orthant-Wise Limited-memory Quasi-Newton algorithm (Andrew & Gao, 2007). This algorithm is based on the popular L-BFGS quasi-Newton method (Nocedal & Wright, 2006), but unlike L-BFGS, it can estimate L_1 regularized objective functions.

4.5.2 Analysis

The standard and latent-class lasso were evaluated according to how many of the irrelevant predictors were estimated as equal to zero, referred to as the sparsity and computed as,

$$\frac{\sum_{r=1}^{Rep}\sum_{k=1}^{K}\sum_{f=1}^{F} I[\hat{\beta}_{kf}^{(r)} = \beta_{kf}^{(r)} = 0]}{K \times F \times Rep},$$

where Rep is the number of replications, $\hat{\beta}_{kf}^{(r)}$ is the estimate of β_{kf} from replication r, f is the covariate ($f = 1 \dots F$), and $I[\hat{\beta}_{kf}^{(r)} = \beta_{kf}^{(r)} = 0]$ evaluates whether the coefficients that have a generating value of zero are estimated as zero. In addition to investigating how often the procedure correctly dropped irrelevant predictors, the simulation study computed the proportion of relevant predictors dropped by the variable selection procedure. This is computed as,

$$\frac{\sum_{r=1}^{Rep}\sum_{k=1}^{K}\sum_{f=1}^{F} I[\hat{\beta}_{kf}^{(r)} = 0, \beta_{kf}^{(r)} \neq 0]}{K \times F \times Rep},$$

where $I[\hat{\beta}_{kf}^{(r)} = 0, \beta_{kf}^{(r)} \neq 0]$ evaluates whether the coefficients with a non-zero generating value are estimated as zero.

The correct selection rate (CSR) is the proportion of replications that the procedure selected all of the relevant predictors and dropped all of the irrelevant predictors. That is, the CSR measures how often the perfect set of variables was selected. The number of correctly selected variables for replication r can be computed as

$$CS^{(r)} = \sum_{k=1}^{K}\sum_{f=1}^{F} I[\hat{\beta}_{kf}^{(r)} = \beta_{kf}^{(r)} = 0] + \sum_{k=1}^{K}\sum_{f=1}^{F} I[\hat{\beta}_{kf}^{(r)} \neq 0, \beta_{kf}^{(r)} \neq 0]],$$

where the first term is the number of correctly dropped predictors and the second

term is the number of correctly retained predictors, and

$$CSR = \sum_{r=1}^{Rep} I\big[CS^{(r)} = K \times F\big]/Rep,$$

where the indicator $I\big[CS^{(r)} = K \times F\big]$ is equal to one if the number of correctly selected predictors in the replication is equal to $K \times F$, the total number of predictors in the logistic regression model. Note that, unlike the sparsity and proportion of relevant predictors dropped, the CSR measures how often the procedure performs optimally.

In addition to variable selection, the significance of the proposed approach depends on whether it can deliver overall improvements in parameter estimation. Although Iaconangelo and de la Torre (2016) established that the corrected three-step procedure can provide better estimates of regression parameters, it did not investigate the performance of the correction weights when the classifications are regressed on a mix of relevant and irrelevant predictors. To study the impact of the variable selection procedure on the quality of all of the final parameter estimates, the average absolute bias (ABIAS) and the average root mean squared error (ARMSE) were employed. They are defined as

$$ABIAS = \frac{\sum_{g=1}^{K}\sum_{f=1}^{F}\left|\sum_{r=1}^{Rep}(\hat{\beta}_{kf}^{(r)} - \beta_{kf})/Rep\right|}{F \times K},$$

and

$$ARMSE = \sqrt{\frac{\sum_{k=1}^{K}\sum_{f=1}^{F}\sum_{r=1}^{Rep}(\hat{\beta}_{kf}^{(r)} - \beta_{kf})^2}{Rep \times F \times K}}. \tag{4.3}$$

Additionally, the ABIAS and ARMSE of the β_{kf} were used to evaluate the parameter estimates of the non-zero covariates. They are referred to here as the ABIAS or ARMSE of the relevant predictors, and denoted by $ABIAS_{rp}$ and $ARMSE_{rp}$, respectively.

4.5.3 Results

4.5.3.1 Sparsity

Overall, incorporating the correction weights into the L_1 regularization procedure led to substantially improved variable selection across all conditions. Referring to Table 4.2, the standard lasso dropped 76% to 89% of the irrelevant predictors, whereas the latent-class lasso using PDL_{ik} dropped 91% to 100%. As expected, more measurement error in the attribute classifications was associated with larger improvements in sparsity when using the latent-class lasso. Specifically, when the item quality was low, the latent-class lasso with PDL weights dropped between 91% and 100% of the irrelevant predictors, whereas the standard lasso dropped between 76% and 83%. Similar improvements in performance were observed when item quality was medium and $J = 10$, as exemplified by the 15% improvement in sparsity from using the latent-class lasso with a sample size of 2000. Less measurement error in the attribute classifications led to more modest improvements in the sparsity, as can be seen when item quality was high, $J = 20$, and $N = 1000$ or 2000, where the proposed approach with PDL weights led to 10% more sparsity. Note also that under both conditions the sample-level weights led to only 1% less sparsity. This typified the overall performance of the two correction weights. Across all tested conditions, the PDL_{ik} weights led to the same or slightly better sparsity than the SL_{ik} weights.

4.5.3.2 Relevant Predictors Dropped

Although the proposed approach led to dramatic improvements in sparsity, it also led to slightly higher percentage of relevant predictors dropped. This mainly occurred under the less-favorable test conditions. Table 4.3 shows that, under test conditions with medium and high item quality, the standard and latent-class

Table 4.2: Sparsity

N	J	IT	Stand	SL_{ik}	PDL_{ik}
500	10	High	0.82	0.95	0.96
		Medium	0.79	0.89	0.93
		Low	0.76	0.88	0.91
	20	High	0.86	0.95	0.98
		Medium	0.82	0.93	0.96
		Low	0.77	0.89	0.94
1000	10	High	0.85	0.98	0.99
		Medium	0.80	0.96	0.98
		Low	0.78	0.96	0.98
	20	High	0.88	0.97	0.98
		Medium	0.83	0.93	0.96
		Low	0.80	0.93	0.93
2000	10	High	0.87	0.99	0.99
		Medium	0.85	1.00	1.00
		Low	0.81	0.99	1.00
	20	High	0.89	0.98	0.99
		Medium	0.87	0.99	0.99
		Low	0.83	0.95	0.96

Note. Stand: Standard lasso approach; SL_{ik}: latent-class lasso approach, using sample-level correction weights; PDL_{ik}: latent-class lasso approach, using posterior-distribution level correction weights

lasso dropped a maximum of 3% and 5% of the relevant predictors, respectively. When $N = 500$, $J = 10$, and item quality was low, the sample-level and posterior distribution level correction weights led to 8% of the relevant predictors dropped, the highest under any condition. The standard lasso, in contrast, only dropped 5%. As the sample size increased to $N = 2000$, the latent-class lasso selected all relevant predictors. Only under the least favorable test conditions did the increase in sparsity from the latent-class lasso incur a cost in terms of relevant predictors dropped. The performances of the sample-level and posterior-distribution level correction weights were virtually indistinguishable.

Table 4.3: Proportion of Relevant Predictors Dropped

N	J	IT	Stand	SL_{ik}	PDL_{ik}
500	10	High	0.00	0.04	0.04
		Medium	0.03	0.05	0.05
		Low	0.05	0.08	0.08
	20	High	0.01	0.02	0.02
		Medium	0.03	0.04	0.04
		Low	0.04	0.06	0.06
1000	10	High	0.00	0.01	0.01
		Medium	0.02	0.03	0.04
		Low	0.03	0.04	0.05
	20	High	0.00	0.00	0.00
		Medium	0.00	0.01	0.02
		Low	0.00	0.01	0.00
2000	10	High	0.00	0.00	0.00
		Medium	0.00	0.00	0.00
		Low	0.00	0.00	0.00
	20	High	0.00	0.00	0.00
		Medium	0.00	0.00	0.00
		Low	0.00	0.00	0.00

Note. Stand: Standard lasso approach; SL_{ik}: latent-class lasso approach, using sample-level correction weights; PDL_{ik}: latent-class lasso approach, using posterior-distribution level correction weights

4.5.3.3 Correct Selection Rate

Referring to Table 4.4, the corrected procedures led to substantial improvements in the CSR, particularly when item quality was medium or low. In fact, when item quality as low, the standard lasso returned a CSR of no higher than 0.01, whereas the latent-class lasso with PDL correction weights returned CSR values ranging from 0.19, when $N = 1000$ and $J = 10$, to 0.80, when $N = 2000$ and $J = 20$. The corresponding values for the SL correction weights were 0.18 to 0.79. The relatively low CSR compared to the sparsity rates was due to the large number of predictors (27 irrelevant, 9 relevant). However, it is clear that the latent-class lasso

Table 4.4: Correct Selection Rate

N	J	IT	Stand	SL_{ik}	PDL_{ik}
500	10	High	0.04	0.65	0.68
		Medium	0.00	0.36	0.39
		Low	0.00	0.18	0.21
	20	High	0.04	0.60	0.62
		Medium	0.01	0.60	0.61
		Low	0.00	0.35	0.36
1000	10	High	0.21	0.88	0.90
		Medium	0.00	0.62	0.64
		Low	0.00	0.18	0.19
	20	High	0.32	0.94	0.95
		Medium	0.13	0.85	0.86
		Low	0.00	0.46	0.47
2000	10	High	0.63	0.96	0.96
		Medium	0.02	0.85	0.85
		Low	0.00	0.40	0.41
	20	High	0.88	0.96	0.97
		Medium	0.40	0.95	0.97
		Low	0.01	0.79	0.80

Note. Stand: Standard lasso approach; SL_{ik}: latent-class lasso approach, using sample-level correction weights; PDL_{ik}: latent-class lasso approach, using posterior-distribution level correction weights

leads to optimal variable selection with much greater frequency than the standard lasso. Comparing the performance via the CSR suggests that improvements in variable selection were greater when there was more measurement error in the CDM. This in accordance with the findings of Iaconangelo and de la Torre (2016), where simulation studies suggested that the amount of improvement in parameter estimates from the correction weights was related to the amount of measurement error in the CDM classifications.

4.5.3.4 Overall ARMSE and ABIAS

The selected variables from the standard and latent-class lasso were then used to create the final parameter estimates, which reflect the cumulative effect of the

procedures used. Referring to Table 4.5, the corrected approach led to lower ARMSE and ABIAS than the uncorrected approach, regardless of the test condition or specific correction weights. The difference was often dramatic, particularly when item quality was low. For example, when $N = 1000$, $J = 10$, and item quality was low, the uncorrected approach led to ARMSE and ABIAS of 0.42 and 0.18, whereas the PDL_{ik} corrected approach led to ARMSE and ABIAS of 0.10 and 0.24. As test conditions improved, the improvements in parameter estimates were more modest, particularly when item quality was high, where even the uncorrected approach returned low ABIAS and ARMSE. For example, when $N = 2000$, $J = 20$ and item quality was high, the difference in ARMSE and ABIAS between the uncorrected and PDL_{ik}-corrected approaches shrunk to 0.03 and 0.01.

4.5.3.5 ARMSE and ABIAS of the Relevant Predictors

Because of the greater sparsity of the latent-class lasso, the ARMSE and ABIAS of the final parameter estimates would likely be lower even if the estimates of the relevant predictors showed no improvement. It is important to directly compare how well the two approaches estimated the covariates that affected the classification. The ARMSE and ABIAS of the relevant predictors are presented in Table 4.6, and the results confirm that the corrected approach led to substantial improvements in the relevant predictors across all test conditions. The improvements were, like the overall ARMSE and ABIAS, more dramatic when item quality was low - for example, when $N = 1000$, $J = 20$, and item quality was low, the uncorrected approach led to more than twice the ARMSE compared to the PDL_{ik}-corrected approach (0.61 vs 0.30). Likewise, the ABIAS was more than 50% higher (0.48 vs 0.29). When item quality was high, the differences in ARMSE and ABIAS

Table 4.5: ARMSE and ABIAS

			ARMSE			ABIAS		
N	J	IT	Uncor	SL_{ik}	PDL_{ik}	Uncor	SL_{ik}	PDL_{ik}
500	10	High	0.17	0.14	0.14	0.09	0.06	0.06
		Medium	0.29	0.21	0.19	0.16	0.08	0.08
		Low	0.45	0.30	0.27	0.25	0.14	0.13
	20	High	0.15	0.11	0.11	0.06	0.05	0.05
		Medium	0.19	0.14	0.14	0.12	0.05	0.04
		Low	0.34	0.21	0.20	0.20	0.09	0.09
1000	10	High	0.15	0.09	0.09	0.09	0.04	0.04
		Medium	0.26	0.17	0.15	0.10	0.05	0.05
		Low	0.42	0.24	0.23	0.18	0.10	0.10
	20	High	0.11	0.07	0.07	0.05	0.03	0.03
		Medium	0.16	0.10	0.10	0.08	0.04	0.04
		Low	0.31	0.17	0.17	0.15	0.07	0.07
2000	10	High	0.13	0.06	0.07	0.06	0.03	0.03
		Medium	0.25	0.15	0.14	0.10	0.05	0.04
		Low	0.40	0.24	0.24	0.16	0.10	0.11
	20	High	0.08	0.05	0.05	0.03	0.02	0.02
		Medium	0.14	0.07	0.07	0.06	0.02	0.02
		Low	0.29	0.17	0.16	0.12	0.06	0.06

Note. Uncor: using the uncorrected three-step approach; SL_{ik}: using sample-level correction weights; PDL_{ik}: using posterior-distribution level correction weights

between the uncorrected and corrected approaches were smaller, though not inconsequential. For example, even under the most favorable test condition, the uncorrected approach led to an ARMSE and ABIAS of 0.14 and 0.08, while the PDL_{ik}-corrected approach returned 0.08 and 0.05. Again, the two corrected procedures performed very similarly, returning values of ABIAS that never differed more than 0.01. Likewise, the ARMSE of the estimates from the sample-level and posterior-distribution level weights differed more by than 0.01 only under three conditions, and never differed by more than 0.03.

Table 4.6: ARMSE and ABIAS of Relevant Predictors

			$ARMSE_{rp}$			$ABIAS_{rp}$		
N	J	IT	Uncor	SL_{ik}	PDL_{ik}	Uncor	SL_{ik}	PDL_{ik}
500	10	High	0.34	0.24	0.24	0.30	0.15	0.15
		Medium	0.55	0.27	0.27	0.45	0.23	0.23
		Low	0.90	0.54	0.51	0.76	0.51	0.49
	20	High	0.26	0.20	0.20	0.21	0.11	0.10
		Medium	0.33	0.19	0.19	0.26	0.14	0.14
		Low	0.66	0.35	0.36	0.54	0.34	0.33
1000	10	High	0.33	0.16	0.16	0.20	0.13	0.13
		Medium	0.51	0.25	0.24	0.40	0.28	0.24
		Low	0.84	0.45	0.44	0.68	0.41	0.40
	20	High	0.20	0.12	0.12	0.14	0.09	0.09
		Medium	0.28	0.14	0.14	0.22	0.11	0.11
		Low	0.61	0.31	0.30	0.48	0.30	0.29
2000	10	High	0.31	0.12	0.12	0.23	0.10	0.10
		Medium	0.48	0.23	0.22	0.38	0.18	0.18
		Low	0.79	0.48	0.48	0.63	0.42	0.42
	20	High	0.14	0.08	0.08	0.08	0.07	0.05
		Medium	0.27	0.10	0.10	0.21	0.11	0.10
		Low	0.57	0.32	0.34	0.45	0.31	0.28

Note. Uncor: using the uncorrected three-step approach; SL_{ik}: using sample-level correction weights; PDL_{ik}: using posterior-distribution level correction weights

4.5.3.6 Individual Parameter Estimate

To better illustrate the impact of the proposed approach at the level of the individual coefficients, the final parameter estimates of β_{11} are presented in Table 4.7, where the true value is equal to 2.00. Consistent with the analysis of the ARMSE and ABIAS, the corrected final parameter estimates were closer to the true values across all conditions, with particularly dramatic differences between the uncorrected and corrected approaches when the item quality was medium or low. Even when $N = 2000$ and $J = 20$, the uncorrected approach yielded estimates of 1.59 and 1.06 for medium and low item quality, respectively. For

comparison, the PDL-corrected approach returned estimates that were very close to the true value: 2.05 and 2.08.

The uncorrected approach consistently and often severely underestimated β_{11}, whereas the corrected approach tended to slightly overestimate β_{11}. For example, when $N = 1000$, $J = 10$, and item quality was medium, the estimate from the uncorrected approach was 1.19, and the estimate from the PDL-corrected approach was 2.08. When item quality was low or medium, the uncorrected approach led to estimates of β_{11} that were, at best, approximately 20% below the generating value. Only under the most favorable condition ($N = 2000$, $J = 20$, and high item quality, with a classification accuracy of 98%) was the measurement error sufficiently low for the downward bias to decrease below 5%. In contrast, the corrected approach with SL_{ik} and PDL_{ik} weights overestimated β_{11} under twelve and fourteen conditions, respectively, and never more than 5%. Note that the posterior-distribution level weights led to better estimates than the sample-level weights under thirteen out of eighteen conditions.

The quality of the uncorrected parameter estimates greatly depended on how favorable the test conditions were. When $N = 1000$, $J = 10$, and item quality was medium, the uncorrected estimate was 1.19, a severe underestimate; the PDL-corrected estimate was 2.08. Doubling the test length increased the uncorrected estimate to 1.61. The PDL-corrected estimate was virtually the same: 2.07. This highlights the way the correction weights can compensate for unfavorable test conditions, and is consistent with the findings of Iaconangelo and de la Torre (2016).

Table 4.7: Comparison of Parameter Estimates

$\beta_{11} = 2$					
N	J	IT	Uncor	SL_{ik}	PDL_{ik}
500	10	High	1.65	2.04	2.08
		Medium	1.15	2.00	2.10
		Low	0.61	1.46	1.53
	20	High	1.81	1.98	1.98
		Medium	1.58	1.98	2.04
		Low	0.96	1.96	1.95
1000	10	High	1.69	2.03	2.05
		Medium	1.19	2.15	2.08
		Low	0.67	1.82	1.93
	20	High	1.86	2.02	2.02
		Medium	1.61	2.07	2.07
		Low	1.03	2.11	2.09
2000	10	High	1.75	2.06	2.04
		Medium	1.21	2.15	2.07
		Low	0.72	2.13	2.08
	20	High	1.93	2.01	2.01
		Medium	1.59	2.05	2.05
		Low	1.06	2.14	2.08

Note. Uncor: using the uncorrected three-step approach; SL_{ik}: using sample-level correction weights; PDL_{ik}: using posterior-distribution level correction weights

4.6 Empirical Example

The effectiveness with which CDMs can provide information on fine-grained attributes recommends their application to clinical measurement instruments, such as the (Dutch-language version) Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009) used here in this example. As detailed in Rossi, Elklit, and Simonsen (2010), the dataset contains two predictors: Setting, indicating either a clinical patient (54%), or prisoner (46%); and Age, varying from 18 to 74. In de la Torre et al. (2015), the attributes were defined as the following disorders: H = somatoform; SS = thought disorder; and CC

= major depression. The Q-matrix specification for the thirty items was revised in Ma, Iaconangelo, and de la Torre (2016), resulting in a dataset of $N = 739$, and a subset of ten poorly discriminating items denoted as Subset B.

To show the improvements in variable selection from the latent-class lasso, 300 examinees and their responses to Subset B items were randomly sampled 100 times. The selected variables and the final parameter estimates obtained using both the less favorable test conditions ($N = 300$, Subset B) and the full dataset ($N = 729$, All Items) are presented in Table 4.8. Under the full dataset, both the standard and latent-class lasso dropped the Age variable in all three regression models. Under the less favorable test condition, however, the standard lasso did not drop the Age variable in any of the models, whereas the latent-class lasso did. This demonstrates the superior sparsity of the latent-class lasso. The final parameter estimates under the corrected approach were larger (in absolute value) than those from the uncorrected approach, which was consistent with the parameter estimates from the simulation study in Table 4.7. Under the less-favorable test condition, choice of lasso would alter the interpretation of the variables. Specifically, using the standard lasso, the researcher would conclude that the odds of 72 year old subject suffering from the thought disorder would be 1.87 times that of a subject of average age (36 years). In contrast, applying the latent-class lasso would lead to the conclusion that the odds were equal. Similarly, the corrected approach would lead the researcher to conclude that the odds of a clinical patient having the thought disorder were approximately eleven times that of a prisoner, compared to about six when using the uncorrected approach. When the test conditions were more favorable, the discrepancy was reduced, due to both approaches dropping the Age variable. However, even when the full dataset was used, the corrected approach estimated the odds of a clinical patient having the disorder as approximately twelve times that of a prisoner, whereas the uncorrected approach

estimated the odds at around seven.

Table 4.8: MCMI-III

		$N = 729$, All Items			$N = 300$, Item Subset B		
Attribute	Covariate	Uncor	SL_{ik}	PDL_{ik}	Uncor	SL_{ik}	PDL_{ik}
H	Setting	-1.79	-2.11	-2.17	-1.47	-1.99	-2.07
	Age	-	-	-	0.10	-	-
SS	Setting	-1.97	-2.45	-2.49	-1.81	-2.31	-2.37
	Age	-	-	-	0.19	-	-
CC	Setting	-1.58	-1.66	-1.70	-1.36	-1.61	-1.60
	Age	-	-	-	0.07	-	-

Note. Uncor: using the uncorrected three-step approach; SL_{ik}: using sample-level correction weights; PDL_{ik}: using posterior-distribution level correction weights

4.7 Discussion

The latent-class lasso adjusts for measurement error in examinee classification when relating attribute mastery to background variables. In the simulation studies and empirical example, the proposed procedure outperformed the alternative in terms of model sparsity, as well as bias and RMSE of the final parameter estimates. Both corrected approaches evaluated via simulation study produced much better quality parameter estimates than the standard approach. The posterior-distribution level correction weights tended to perform slightly better than the sample-level weights. The enhanced sparsity makes for easier interpretation of the model, and the higher-quality parameter estimates can lead to more valid conclusions about the relationship between attribute mastery and background variables.

Although the research presented here is related to the latent regression work done in the context of IRT and large-scale assessment, it is not solely aimed at developing analogs for CDA. Rather, it is also designed for other common

applications of CDA in the literature, such as implementing diagnostic assessment in the curriculum and using the results to guide student instruction. In the same vein, CDA can be used in a clinical setting, to help guide the work of practicioners - the empirical example presented here is an example of this. The classifications from these small-scale assessments can then serve as a starting point for an exploratory analysis that relates the classifications to examinee covariates. It is assumed that the three-step approach will be used, because the secondary researcher has to work with the test as it was implemented. Specifically, the three-step approach accommodates the fact that the item parameter estimation, model selection, Q-matrix validation, and attribute clasification has already been completed. The latent-class lasso presented here is an attempt to supply the secondary researcher with a methodological tool for these circumstances.

The techniques presented here are alterations of well-known methods, which should make them relatively accessible to researchers. Although the latent-class lasso was developed here in the CDM context, there is not reason to suppose that this approach cannot be applied to latent class models more generally. In fact, this approach could be used to implement the lasso with any latent variable model. The adjustment for measurement error can be computed for the estimation of the posterior distribution, and this correction can then be incorporated into subsequent regression models. This method offers additional flexibility for the researcher, who can apply this corrected three-step approach in a variety of ways. For example, the latent-class (or latent variable) lasso could be used to select variables, and then those variables could be modeled using a one-step approach. This could reduce computational time while still yielding the gold-standard parameter estimates. The work here is limited, however, in that the methods cannot relate one latent-variable to another, with no predictor. Along the same lines, this work

does not accommodate research questions that ask how the attribute classifications predict distal outcomes. This is a relatively common line of research and it would be valuable to develop procedures for it in the CDA context, like has been done for IRT assessments (Schofield et al., 2015). IRT methodologies are well-developed for large-scale assessment settings, such as NAEP. There is extensive literature on how to account for the influence of covariates and prepare data for use by secondary researchers while minimizing the risk of statistical errors such as, to name just one example, omitted variable bias. It is still unclear to what extent these methodologies and practices should extend to the CDM framework and a more diffuse, small-scale implementation of assessments. Because CDMs promise to classify examinees based on fine-grained components of variation, it seems natural that secondary researchers will seek to relate these classifications to not only examinee covariates, but school and district covariates as well. Similar research questions could be asked in the context of clinical psychology. The latent-class lasso presented here is an attempt to supply the secondary researcher with a methodological tool for these circumstances. However, because secondary research could potentially impact policy, a more thoroughly developed and coherent framework of statistical analysis is needed to ensure the validity of the conclusions.

4.8 References

Andrew, G., & Gao, J. (2007, June). *Scalable training of l1-regularized log-linear models.* Paper presented at the 24th international conference on Machine learning, New York, NY.

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification*, *30*, 195-224.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2013). Relating latent class assignments to external variables: Standard errors for corrected inference. *Sociology*, *83*, 173-178.

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*, 272-311.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-1386.

Coppola, A., Stewart, B., & Okazaki, N. (2014). lbfgs: Limited-memory bfgs optimization [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=lbfgs` (R package version 1.2.1)

Culpepper, S. A., & Park, T. (2017). Bayesian estimation of multivariate latent regression models: Gauss versus Laplace. *Journal of Educational and Behavioral Statistics*.

Dayton, C., & Macready, G. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173-178.

Dayton, C., & Macready, G. (2002). Use of categorical and continuous covariates in latent class analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (p. 213-233). Cambridge, UK: Cambridge University Press.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

de la Torre, J., van der Ark, L., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 1-16.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285-291.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22. Retrieved from `http://www.jstatsoft.org/v33/i01/`

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of statistical learning*. New York, NY: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. United Kingdom: Chapman and Hall.

Henson, Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

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

Huebner, A., & Wang, C. (2011). Comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407-419.

Iaconangelo, C. J., & de la Torre, J. (2016, July). *Three-step estimation of cognitive diagnosis models with covariates.* Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59-81.

Lazarsfeld, P. (1950). The logical and mathematical foundation of latent structure analysis. In S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld, S. Star, & J. Clausen (Eds.), *Measurement and prediction* (p. 362-412). Princeton, NJ: Princeton University Press.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144-177.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200-217.

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park, CA: SAGE publications.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, *52*, 374-393.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III Manual (4th ed.).* Minneapolis, MN: Pearson Assessments.

National Center for Education Statistics. (2011). NAEP technical documentation: Achievement gaps. Retrieved from `https://nces.ed.gov`

Nocedal, J., & Wright, S. (2006). *Numerical optimization.* New York, NY: Springer.

Park, Y., & Lee, Y. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Applied Psychological Measurement*, *38*, 376-390.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rossi, G., Elklit, A., & Simonsen, E. (2010). Empirical evidence for a four factor framework of personality disorder organization: Multigroup confirmatory factor analysis of the Millon clinical multiaxial inventory - III personality disorder scales across Belgian and Danish data samples. *Journal of Personality Disorders*, *24*, 128-150.

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika*, *80*, 727-747.

Sinharay, S., Puhan, G., & Haberman, S. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*, 29-40.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *68*, 267-288.

van der Ark, L., van der Palm, D., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*, 380-392.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307.

Wang, W., & Gelman, A. (2015). Difficulty of selecting among multilevel models using predictive accuracy. *Statistics and Its Interface*, *8*, 153-160.

Ye, S., Fellouris, G., Culpepper, S., & Douglas, J. (2016). Sequential detection of learning in cognitive diagnosis. *British Journal of Mathematical and Statistical Psychology*, *69*, 139-158.

Chapter 5
Conclusion

The work presented here attempts to begin to address the various ways measurement error affects the validity of CDM-related research. A variety of methodologies for estimating and correcting for classification error have been presented. Taken together, the work presented here makes several contributions to the CDM literature and develops methodologies that should help lay the groundwork for implementations of CDA in a variety of real-world applications.

The proposed index $\hat{\tau}_l$ returned estimates of conditional classification accuracy that, with the exception of the worst conditions, were within 10% of the true value. This index promises researchers an evaluation of not only the overall, or test-level, classification accuracy, but of how well examinees in particular latent classes of interest can be expected to be classified. When evaluating the overall accuracy for a sample from another population, the simulation study results indicated that the $\hat{\tau}^*$ index that the empirical rates were well-recovered under all but the least favorable test conditions.

A three-step method was developed that secondary researchers, in particular, may find appealing. Under all but the most favorable conditions with almost perfect classification accuracy, the use of the estimated examinee latent class assignments as the dependent variable in the regression led to poor parameter estimates. The sample-level correction weights, adapted from the latent class analysis literature, substantially reduced the bias and RMSE of the regression model parameter estimates. Furthermore, the posterior-distribution level weights

developed in this work improved the quality of the estimates more so than the sample-level weights, often returning estimates that were approximately as good as the one-step approach (the gold standard) in terms of bias and RMSE. As the amount of error in the classification increased, so did the improvement attributed to the correction weights. The correction weights were applied to an L_1 regularized regression to create the latent-class lasso. This modification of a popular, widely used approach to model selection led to large improvements not only in the variables selected, but in the final parameter estimates. That is, the standard lasso and uncorrected three-step approach resulted in a lack of model sparsity and severely attenuated final parameter estimates of the relevant predictors. In comparison, the latent-class lasso and corrected three-step approach often selected the exact set of correct predictors and the estimated coefficients were much less biased and had smaller RMSE.

By addressing the effects of measurement error in the CDM framework, this research addressed issues central to the validity argument for CDAs. Note that although the proposed methodologies focus on the CDM framework, there is no reason to suppose that the methods cannot apply to latent class analysis more generally. All three studies were connected by the common thread of the matrix of CEPs, which is an estimate of the measurement error. Study 1 uses that matrix directly, as an estimate of the classification accuracy. Studies 2 and 3 use it indirectly, as a way of accounting for the measurement error, making the multinomial logistic regression a latent-class logistic regression.

Likewise, all three studies illustrate how the conclusions drawn from the CDA are altered by measurement error, and how those effects, in terms of variable selection, bias, and RMSE, can be corrected by accounting for measurement error in the procedures. An index is proposed to tackle the problem of estimating classification accuracy, both for the given sample and hypothetical samples of

interest, which is crucial for the validity argument. This is particularly important for implementing a cost-benefit analysis that allows the test user to based decisions on the CDA classification. For clinical psychologists, the decision to treat or not treat for a particular collection of disorders may depend on the classification, and so it is important to have an estimate of the accuracy of this particular patients classification, without resorting to using the estimated marginal (test-level) classification. Similarly, for an educator, the decision to administer remedial instructional may hinge on the latent class assignment, and an estimate of how likely that assignment is to be correct is crucial to making that decision. This could play a role in deciding whether the assessment was appropriately constructed for use across a broad number of possible attribute distributions, as might be the case for those in charge of administering exams in large urban school zones.

A corrected three-step procedure aims to provide tools for secondary researchers that allow for connecting student achievement to covariates by proposing techniques to accounting for error in the classifications. By implementing the proposed approaches, the practitioner implicitly acknowledges that treating latent classifications as observed variables measured without error can be potentially very misleading, thus harming the validity of interpretations made from estimated coefficients. Because the conclusions of secondary research often influence policy, the validity of those conclusions is of particular importance. That is, it is of the utmost importance that the coefficient estimates be as accurate and precise as possible. Note that better coefficient estimates can only be obtained by using item responses, or, at least, item parameters. However, it is often the case that only the scores or classifications are released to secondary researchers. The three-step procedure in Study 2 only requires examinee posterior distributions.

A common model selection procedure is likewise adapted to address measurement error. Specifically, it shows the discrepancy between models selected and

parameters estimated both ignoring the latent nature of the variable, and adjusting for the measurement error. This observation, that measurement error must be accounted for, is not new. But it has not been addressed within the CDM context. In fact, the CDM literature contains a relative paucity of methodologies designed to relate classifications to covariates. This is surprising given how often researchers investigate the relationship between student achievement and student, school, district, and community variables. Study 3 creates the latent-class lasso, a modification of a well-known technique. In spite of its popularity, the literature on the lasso lacked any way of incorporating measurement error associated with the dependent variable into the model.

Looking forward, there are many directions for future research. For the clinical diagnostic setting, practicioners and treatment developers may be interested in a way to accommodate patient covariates in a longitudinal framework, to evaluate patient improvement over the course of treatment. These methods could be developed for educational settings as well, to evaluate the relationship between the learning progression and student variables. More broadly, researchers may aim to relate two classifications to each other, which would require adjustments for measurement error in both the independent and dependent variables.

More fundamentally, the constructions of the matrix of CEPs may be altered to accommodate polytomous CDMs. Measuring classification error may require modifications given this item format. This adaptation is particularly relevant given the use of Likert data in clinical applications as well as the use of partial credit items in educational settings.