A Model That Combines Diagnostic Classification Assessment with Mixture Item Response Theory Models

by

Hye-Jeong Choi

(Under the direction of Allan S. Cohen and Jonathan L. Templin)

Abstract

The purpose of this dissertation is to present a new psychometric model that combines a Mixture Rasch model with a diagnostic model. We refer to this model as a diagnostic classification mixture Rasch model (DCMixRM). The motivation for the development of the DCMixRM is twofold. First, the DCMixRM is designed to provide rigorous explanation as to factors that are potentially causing the latent classes to form. In doing so, this model uses attribute mastery states as covariates. Second, the DCMixRM is also designed to connect assessment to instruction by furnishing diagnostic information along with a general ability level.

This model consists of two components: measurement and structural components. The measurement component includes specification of item responses through simultaneously taking into account three sets of latent variables, such as a general ability, latent class membership, and mastery profiles of attributes. In the structural component, characteristics of three latent variables are specified, including distributions of ability, latent class, and mastery profile. Further, in this model, we specify the relationship among these variables, particulary the association between latent class and mastery profile.

The DCMixRM has several advantages: it provides a way to detect heterogeneity in the population; it yields more accurate classification of latent classes; it provides a rigorous explanation about features of latent classes; it allows us to examine incompleteness of the Q-matrix; and it allows us to make inferences on a global ability as well as on mastery profiles formed over the set of attributes.

A series of simulation studies were conducted to evaluate the quality of estimation process for the DCMixRM in terms of convergence and recovery of model parameters. For the simulation study, two sets of tests were designed: 30 items involving 3 attributes (A3I30), 20 items involving 4 attributes (A4I20). Under each condition, sample size, similarity of ability means across latent classes, and strength of relationship between latent class and mastery profile were manipulated. Although for some conditions, convergence appeared problematic, results showed that the model parameters were well recovered enough to lead appropriate inferences on the model parameters.

We also applied the model to two empirical data sets, including an international reading and a statewide mathematics tests to give an illustration of how the model can be used. Further research directions were discussed as well.

Index words: Latent covariate, Local dependence, Multidimensionality, Latent class model (LCM), Mixture Rasch Model (MixRM), Log-linear Cognitive Diagnosis Model (LCDM), Diagnostic Classification Mixture Rasch Model (DCMixRM)

A Model That Combines Diagnostic Classification Assessment with Mixture Item Response Theory Models

by

Hye-Jeong Choi

B.A., Seoul National University, Seoul, Korea, 1992

M.A., Seoul National University, Seoul, Korea, 2001

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2010

A Model That Combines Diagnostic Classification Assessment with Mixture Item Response Theory Models

by

Hye-Jeong Choi

Approved:

Major Professors: Allan S. Cohen
Jonathan L. Templin

Committee: Deborah L. Bandalos
Robert A. Henson
Karen Samuelsen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

Dedication

This is dedicated to my late father and my mother.

Acknowledgments

Unlike in a deterministic world, in a probabilistic world, we talk a lot about uncertainty and randomness which instill fears within our lives. However, I believe in Karma; there are always reasons people connect to others in relationships. In particular, in order to be a parent of, a teacher for, or a friend with another, a very special Karma is required from the eternal previous lifetimes. Writing a dissertation can be a lonely and isolating experience, yet due to existence of my special people from the Karma, for me it became a form of catharsis leading to understanding myself, and I was able to see hopes out of fears I had encountered. Thus it is time for me to acknowledge several individuals among whom I owe deepest gratitude.

I could not find a single word that I can convey my gratefulness to my advisor, Dr. Allan S. Cohen. I am heartily thankful to him. Not only has he supported me with all the resources during my graduate studies at UGA, but also he has demonstrated his faith in me even when I truly doubted myself. My deepest appreciation goes to my co-advisor, Dr. Jonathan L. Templin. Without his infectious enthusiasm and unlimited zeal about diagnostic models and his persistent help, I could not have completed the simulation studies. Because of them I could see the potential even through my despair.

I owe a special debt to Drs. Bandalos, Olejnik, and Kim. They have provided me with a tremendous graduate education. They have taught me how to think about problems as a researcher; they have shown me how to do research; and they have encouraged and inspired me to become an independent scholar. I have been privileged to be their student.

Drs. Samuelsen and Henson deserve a special thanks as my dissertation committee members and for their valuable suggestions and comments over writing the dissertation. I have enjoyed their questions because they have always sharpened my mind about the model.

Being a graduate student has brought me the frustration, pain, dissatisfaction, excitement, meditation, and rich rewards. My graduate studies, however, would not have been the same without the social and academic challenges and diversions provided by all my friends in Athens. In particular, my enormous debt of gratitude can hardly be repaid to my friends Laine Bradshaw and Gregory McClure who not only proof-read multiple versions of all the chapters of this dissertation, but also provided many stylistic suggestions and substantive challenges to help me improve my presentation and clarify my arguments.

In probability theory, conditional probability is one of the most important concepts. However, I have learned that love of my family has never been conditional on anything. They always love me whatever I am. Despite valuing becoming an independent scholar at this moment, I still rely heavily on their love and I am just grateful for their unconditional love. Without their love, I could not survive in this long journey.

Table of Contents

Page

Chapter

List of Figures

List of Tables

Chapter 1

INTRODUCTION

> Assessments should be valued for their utility in facilitating constructive adaptations of educational programs and for service in offering *executable* advice to both students and teachers. Testing and learning should be integral events, guiding the growth of competence. (Glaser, 1990, p. 480)

This chapter outlines the purpose and scope of the current study. We begin with a brief description of mixture IRT models focusing on the way these models handle population heterogeneity. Next, we describe the topic of this study, followed by a description of the purpose and significance of the study. Finally, we present an overview of subsequent chapters.

1.1 Statement of Problem

Item response theory (IRT) models have been used to provide estimates of examinee proficiency and of the item parameters that are included in the model. The basic idea of IRT models is that associations among item responses can be accounted for by a latent trait or examinees' ability. This is premised on two related assumptions: local independence and unidimensionality. The local independence assumption states that the observed items are independent of each other conditional on the value of the latent trait (Lazarsfeld & Henry, 1968). This definition is a mathematical way of stating that the latent variable explains why the observed items are related to one another. On the other hand, the unidimensionality assumption refers to the assumption that items on the test measure a single dimension, ability. That is, a single underlying latent variable can account for the association between items.

Although these are considered two separate assumptions, the unidimensionality assumption is not an additional assumption because the local dependence follows automatically from unidimensionality (Lord, 1980). We detail these two assumptions later in Chapter 2. When these assumptions hold, IRT models have desire features which would include item characteristics are not group-dependent, ability estimates are not test-dependent, and a model provides a measure of precision for each ability level(Hambleton, Swaminathan, & Rogers, 1991). Because of these features, it can be used for solving issues in testing such as equating tests or constructing adaptive tests.

It is frequently the case, however, that local dependencies exist among item responses, even after controlling for examinees' ability. This has concerned psychometric researchers because this local dependence is a serious violation of an IRT assumption. Research has shown that there are at least two different sources of variation that can result in these dependencies (Steinberg, Thissen, & Wainer, 2000). First, not all individuals solve test items in the same way because individuals' cognitive patterns can differ from one another. Second, a test consists of several sets of items, each set measuring related but distinct components designed to assess a common latent trait. Both of these sources of variation cause local dependence, resulting in secondary dimension(s) in the data. For the purposes of this dissertation, we consider the former source of variability as leading to heterogeneity in the examinee population and the second source of variability as leading to multidimensionality in the test items.

Rost (1990) developed the mixture Rasch model (MixRM) to address the local dependence caused by the first source, which is the population heterogeneity. He combined a Rasch model with a latent class model (LCM) to detect the existence of two or more population subtypes or latent classes in the data. This is done while also obtaining model-based item and ability estimates within each latent class. This idea has been extended to more general IRT models (von Davier & Rost, 2006). Interest in the mixture IRT (MixIRT) models has grown as researchers have become aware that finding latent subgroups in the data may not be that uncommon. Further, such models have been shown to provide a useful framework for detecting and explaining the differential behavior of members of these latent groups. To illustrate how to employ MixRM, Rost (1990) fit the model to a physics test data set and found two latent groups of examinees who responded differently to a set of physics test items. Members in these two latent groups were characterized as practice- or theory-oriented groups because one group was better at practice-related items, and the other group was better at theory-based items. Similarly, in spacial visualization tasks Mislevy, Wingersky, Irvine, and Dann (1991) found two latent groups that differed in their use of problem solving strategies for answering test items. One latent group tended to use a rotational strategy, and the other latent group tended to use an analytical strategy to solve the same mathematics problems.

As shown in both studies above, after identifying latent groups using a MixIRT model, the next step is to explain the difference(s) between the groups. That is, one needs to determine what might be the sources of variation that made these different latent classes form. Unfortunately, the resulting solution from fitting MixIRT models does not by itself necessarily lead directly to an explanation about why latent classes response differently to the same items. Most current applications of MixIRT models, in fact, rely on post-hoc analyses to identify characteristics of examinees in different latent classes (e. g. Maij-de Meij, Kelderman, & van der Flier, 2008; Cohen & Bolt, 2005; Hong & Min, 2007). One way to improve some of the explanatory power of such models is to include covariate(s) to help explain latent class membership.

Smit, Kelderman, and van der Flier (1999, 2000, 2003) have shown that inclusion of external collateral information, (i.e., as covariate(s)) in the MixIRT model has the potential to be useful for providing information that can be helpful in describing characteristics of latent classes to the extent that latent class membership is related to that information. They also argue that this use of covariates improves the accuracy of detection of latent classes and latent class membership, even when sample sizes or differences between latent classes are small.

To date, typical covariates include manifest variables such as age, gender, and ethnicity. Rost (1990) and Mislevy et al. (1991) note, however, that such manifest variables might not be as informative as would be needed for describing the differences in response patterns among the latent groups. More useful variables for this purpose would be those that can provide a more direct explanation as to why these differences in responding may have occurred. For example, it would be useful if a model could be developed to indicate why some individuals were more likely to use a rotational strategy in solving mathematical problems while others tended to use an analytical strategy.

In this regard, Cohen, Gregg, and Deng (2005) found that differences in mastery of specific mathematical skills were associated with membership in latent classes detected in a statewide mathematical testing program. From this result, they concluded that the prior knowledge may have resulted in members of different latent groups taking different paths to arrive at the same answers to test questions. In other words, differences in mastery state of various aspects of mathematics required to correctly answer the items on the test may be one reason why individuals in different latent groups display different response patterns.

This indicates that multidimensionality in a set of test items, which is the second source of variation that causes local dependence, can be related to different response patterns among latent classes, or population heterogeneity. What would be useful is to be able to determine if these skills or components of knowledge can be used to help explain latent class membership and differences among latent classes. That is, if a test is made up of items to cover several different, albeit related skills, and if examinees differ in mastery status on these skills, then the dissimilarity in examinees' prior mastery status on these skills could affect their response strategies. For example, some examinees may have had more prior knowledge of geometry than others, and thus it may be more effective for them to use a rotational strategy than an analytical strategy to solve certain mathematics items.

One challenge in using individuals' mastery statuses on particular aspects of ability as covariates for understanding differences among members of distinct latent groups, however, is that these states are themselves latent variables which we cannot observe these states directly. In order to use the mastery statuses, or more precisely, the latent mastery states, as exploratory information, it is necessary to first detect mastery statuses so that they can be employed as covariates for helping to explain membership in the latent groups.

To find a solution to this issue, in this study we proposes a model which is designed to extract information about individuals' mastery states as well as a general ability and then also use that information as covariates to explain latent class membership in a MixIRT model. In doing so, we combine a MixRM with a diagnostic classification model (DCM) to identify individual mastery states on the individual skills needed for answering the questions on the test. We focus only on the MixRM although other MixIRT models could be used.

As is described in detain in Chapter 2, DCMs can be considered restricted latent class models. These models provide a pattern of mastery and non-mastery for each examinee, thereby providing substantial diagnostic information that can be used to better understand examinee performance. Unlike traditional IRT models, these models do not require the unidimensionality assumption. In fact, they explicitly specify multidimensionality of test items, and thus allow one to handle multidimensionality in item responses, if necessary.

Hereafter we refer to this new model as a diagnostic classification mixture Rasch model (DCMixRM). This model retains the benefits of the LCM, the Rasch model, and a diagnostic classification model without having some of the drawbacks of each. Specifically, the model identifies individuals' mastery states in skill knowledge and simultaneously uses this information for characterizing latent classes detected by the MixRM model. That is, DCMixRM permits classifying individuals' mastery states in a multidimensional space as well as detecting population heterogeneity. Furthermore, it uses different mastery states to describe response pattern heterogeneity, even when examinees have the same level of general ability.

In summary, the basic idea of the model proposed in this study is that if or when two sources of local dependence exist, including them jointly in the model can resolve the local dependence issue and render a description of their relationship.

1.2 The Purpose of the Study

The main purpose of this dissertation is to propose a new model which combines a MixRM and a diagnostic assessment model to directly relate latent class membership to mastery status and to show how this model can be carried out through a maximum likelihood estimator using a standard structural equation modeling software. A unique component addressed in the model developed in this study is the focus on the impact of the mastery state on latent class membership. In so doing, the model takes advantage of two sources of the secondary dimensionality in item response data: heterogeneity and multidimensionality.

The model, the DCMixRM, offers three advantages: (1) it furnishes rich information about examinees' performance including information about a general ability, the mastery states of the examinee on each skill (also known as an attribute), and membership in latent classes, if the population is composed of several latent classes; (2) it improves the power to detect latent classes by using latent covariate(s) extracted directly from item responses; and (3) it provides a means of using examinees' mastery states to account for why latent classes may respond differently to items.

An important concern in the application of any new models is the availability of software to estimate model parameters. In this study, commercially available standard software, M*plus* Version 5.21 (Muthén & Muthén, 1998-2007), is used to implement the DCMixRM.

1.3 Overview of Chapters

The remainder of this dissertation is organized as follows: In Chapter 2, relevant psychometric models are reviewed including the Rasch Model, the latent class model (LCM), and the log-linear cognitive diagnosis model (LCDM). The review focuses on strengths and weaknesses of each model. At the end of Chapter 2, the DCMixRM is introduced by showing how a Rasch model, an LCM and an LCDM contribute to and can be combined to form the DCMixRM. The primary challenge in this model is that three person-related latent variables (ability, mastery profile, and class membership) are embedded in one model, all of which need to be simultaneously estimated from the data. For such a complex model, it is necessary to verify whether the model and estimation procedures behavior appropriately in estimating relevant parameters. Hence, in Chapter 3 we present a series of simulation studies conducted to evaluate the estimation process with respect to the recovery of model parameters under varying conditions. In Chapter 4 we demonstrates how to apply this new model to large-scale data sets from reading and mathematics tests. Finally, in Chapter 5, we conclude the study by presenting discussion of results and directions for future research.

Chapter 2

THEORETICAL FRAMEWORK

This chapter discusses how to integrate three latent variable models to formulate the latent space involved in the DCMixRM proposed in this dissertation. We begin by reviewing several relevant latent variable models, including latent class models (LCMs), mixture Rasch models (MixRMs), and diagnostic classification models (DCMs). Among these models, DCMs have been recently developed and these models intend to classify individuals' mastery states on the skills needed for answering the questions on the test. We focus on Log-linear Cognitive Diagnosis Models (LCDMs) because they provide the most general framework for DCMs. Advantages and limitations of each model are described, followed by a discussion of how the advantages of each model provide the bases for building the diagnostic classification mixture Rasch model (DCMixRM) proposed in this study. In the last section of this chapter, the characteristics of the DCMixRM are described in detail.

2.1 Latent Class Models (LCMs)

Latent Class Models (LCMs). Lazarsfeld and Henry (1968) delineated that latent structure models describe "the probability relation between the set of observed indicators and the inferred position of the units involved in an empirical study" (p. 3), and "particular latent structure *model* is a specification of the nature of the latent space, and of how the *item probabilities* vary within the space" (p.16).

One of latent structure models is the latent class models (LCMs). These models include the latent space consisting of a finite number of points called latent classes. The relation between the latent space and the observable indicators is defined through the *axiom of local independence*. Lazarsfeld and Henry (1968) defined the local independence as:

> *Within a latent class, α, responses to different items are independent. The within class probability of any pattern of response to any set of items is the product of the appropriate marginal probabilities.* (p. 22)

In other words, the underlying latent variable explains why the observed indicators are related to one another. This is because if such a latent variable exists, after controlling for the latent variable, the dependencies among the indicators should vanish. The local independence assumption, the discrete nature of the latent variables, and the use of probability are the most important elements of LCMs.

To outline LCMs in a mathematical form, suppose that (1) a sample of N individuals is drawn from a population, (2) the population is composed of a mixture of G latent classes that are mutually exclusive and exhaustive, (3) the proportion of individuals in each latent class, $\pi_1, \pi_2, ..., \pi_G$, is unknown, and (4) the class membership of each individual is also unknown. Let $Y_{ij} \in \{0, 1\}$ be a random variable indicating a response of individual $i \in \{1, 2, ..., N\}$ to item $j \in \{1, 2, ..., T\}$, and $g \in \{1, 2, ..., G\}$ denote the latent class membership of the individual. Under the assumption of local independence, the marginal probability of the correct response, $P(Y_{ij} = 1)$, can be written as

$$P(Y_{ij} = 1) = \sum_{g=1}^{G} \pi_g P(Y_{ij} = 1|G = g), \tag{2.1}$$

where π_g denotes the latent class probability and $P(Y_{ij} = 1|G = g)$ is the conditional probability that individual i in class g correctly answers item j.

As shown in Equation 2.1, the probability of obtaining response Y_{ij} is a weighted average of the class-specific probabilities. The latent class probability π_g and the conditional probability $P(Y_{ij} = 1|G = g)$, therefore, are the two main components of LCMs. The latent class probability π_g describes the distribution of latent classes; the number of classes and the relative sizes of these classes. Since the conditional probability $P(Y_{ij=1}|G = g)$ indicates the probability that individuals in class g correctly respond to item j, it can be used to describe the characteristics of latent classes. This use of the conditional probability can be considered an analogue of a factor loading in factor analysis (McCutcheon, 1987).

Since both the latent class probability and conditional probability are probabilities, they should satisfy properties of a probability: (1) both should be non-negative, (2) the sum of the latent class probabilities over all latent classes should equal one ($\sum_{g=1}^{G} \pi_g = 1$), and (3) within each class the conditional probability over all categories for each item sum to 1, for example, for binary responses, $P(Y_{ij} = 1|G = g) + P(Y_{ij} = 0|G = g) = 1$ (Dayton & Macready, 2007).

The main purpose of LCMs is to define a set of latent classes within which items are locally independent. Once the parameters in LCMs are estimated, it is possible to classify each individual into the appropriate latent class. The modal rule is the most commonly used for assigning an individual to a latent class (McCutcheon, 1987). In this way, a person is classified into the likely class with the highest a posteriori probability of the membership by utilizing the Bayes rule:

$$P(G = g|\mathbf{Y} = y) = \frac{P(G = g)P(\mathbf{Y} = y|G = g)}{P(\mathbf{Y} = y)}. \tag{2.2}$$

LCMs have been used for detecting latent heterogeneity in the population in several research areas: medical research (e.g., Laumann, Paik, & Rosen, 1999; Bucholz et al., 1996), economics (e.g., Jedidi, Jagpal, & DeSarbo, 1997; Eckstein & Wolpin, 1999; Deb & Trivedi, 2002; Thacher & Morey, 2003), psychometric research (e.g., Thomas & Horton, 1997; Uebersax, 1999), and educational research (e.g., Aitkin, Bennett, & Hesketh, 1981; Brown, Askew, Baker, Denvir, & Millett, 1998). More general overviews of LCMs, including identification issues, can be found in Lazarsfeld and Henry (1968), Goodman (1974), and Hagenaars and McCutcheon (2002).

Limitations of the LCMs. The LCMs allow for differences in response probabilities across latent classes, and yet, the conditional probability is the same for all members in the same class. Stated differently, all individuals within a latent class have the same response probability for a given item (e.g., identical probability of a correct answer to an item). Since this assumption is too strong for some cases, it is difficult to realize in practice, and may even result in "spurious" latent classes being detected in order to reconcile the data structure to the latent class model (Uebersax, 1999).

2.2 Mixture Rasch Models (MixRMs)

When introducing latent structure models, Lazarsfeld and Henry distinguished latent class models from latent trait models depending upon whether the latent space is continuous or discrete. Whereas LCMs assume a discrete latent variable to account for the relationship among item responses, as latent trait models, item response theory (IRT) models assume a continuous latent space as cause of association among indicators. However, MixRMs combine both discrete and continuous latent variables as the latent space in the models. In MixRMs, a continuous latent variable is modeled by a Rasch model, and a discrete latent variable is modeled by an LCM. Below, we begin with the Rasch model, and then describe how it can be combined with an LCM. We also extend how a covariate or covariates can be incorporated with MixRMs.

The Rasch Model. Rasch (1960/80) emphasized the importance of the use of probability in describing the response behavior of an examinee on a psychological or educational assessment. He specified the probability of correct answer through relationship between item and ability as

> *every person has a certain probability of solving correctly each problem of a given kind and his probability (P) is - independently of the answers to the preceding problems - given by formula 2.3 where ξ is a characteristic of the person and δ of the problem*
>
> $$P = \frac{\xi}{\xi + \delta}. \tag{2.3}$$

Even though his choice of the model was rather simple, this model has quite important features: it assumes statistical independence between examinees and items (this is the local independence assumption); it separates item and person parameters; and it yields a scale for latent ability. This model is equivalent to the one parameter logistic IRT model as

$$P(Y_{ij} = 1|\theta_i) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}, \tag{2.4}$$

where Y_{ij} indicates response of individual i to item j; θ_i is an ability parameter of individual i; b_j is a difficulty parameter of item j; and $P(Y_{ij} = 1|\theta_i)$ denotes the probability of a correct answer to item j given ability level θ (Baker & Kim, 2004). This probability is a function of examinee ability and the item difficulty and is the same as Equation 2.3. When examinee ability equals item difficulty, the probability of an examinee answering the item correctly is .5. For each ability, performance increases as difficulty decreases or ability increases. That is, items and individuals are strictly monotonically ordered.

As with LCMs, the local independence assumption is necessary in the Rasch model for linking the observed item responses and the latent variable. In other words, it is assumed that responses to any items are uncorrelated for a given ability level. Furthermore, the latent space in the Rasch model consists of only one continuous variable, examinees' ability, and this is also referred to as *unidimensionality*.

As with LCM, local independence or unidimensionality may be too strong, and it may be violated in practice (Dorans & Kingston, 1985). That is, responses may be determined by more than a single latent variable. When this is the case, the Rasch model does not accurately describe the generation of item responses. Several alternative models have been developed to relax this assumption, including the multidimensional IRT model (Reckase, 1997), the Hybrid model (Yamamoto, 1987), and the Saltus model (Wilson, 1989).

It is also possible that the population under consideration may not be homogeneous with respect to the latent variable, contrary to assumption in the Rasch Model (Rost, 1990). Instead, the population may be composed of several latent classes, and further it may be that individuals in different latent classes may possess different response propensities to an item, which may, in turn, result in different probabilities of a correct response. One way to model such data is to use a MixRM.

Mixture Rasch Models (MixRMs). As discussed in the previous chapter, "heterogeneity" in a population has been a concern among researchers in educational and psychological measurement. The question is whether or not the latent ability "turns out to be scalable traits for *all* individuals when analyzed with a latent trait model accounting for differential scalability in latent subpopulations" (Rost, Carstensen, & von Davier, 1997, p. 324). Rost (1990) suggested the mixture Rasch models (MixRMs) where the Rasch model is consolidated with an LCM to overcome "the deficiencies of both approaches and retain their positive features" (p. 271).

The main ideas of MixRMs are: (1) the observed item responses arise from a number of subpopulations, or latent classes; (2) the latent classes are assumed to be mutually exclusive and exhaustive; and (3) the Rasch model holds in each latent class, and yet different sets of item difficulties may hold in different classes. Then, for a MixRM, the probability of the correct response can be written as

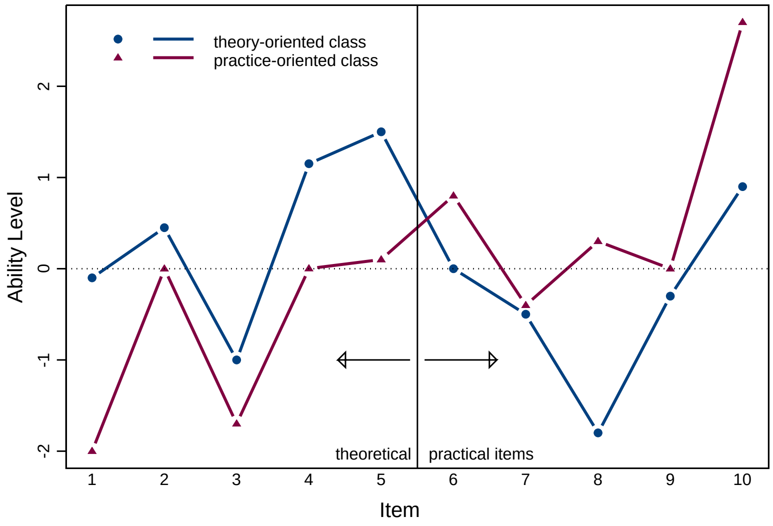
$$P(Y_{ij} = 1|\theta_i) = \sum_{g=1}^{G} \pi_g P(Y_{ij} = 1|\theta_i, g) = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_i - b_{jg})}{1 + \exp(\theta_i - b_{jg})}, \tag{2.5}$$

where θ_i denotes the ability of person i as in the Rasch model, π_g is latent class probability as in an LCM, and $P(Y_{ij} = 1|\theta_i, g)$ denotes the conditional probability that person i gives a correct response to item j given the class membership g and ability level θ; b_{jg} has a critical meaning here such that it denotes item difficulty of item j for class g, indicating that each class has its own item difficulty. It should be noted that unlike items, individuals have one ability parameter given their latent class membership, and therefore examinees do not have subscripts to indicate their class membership.

This model allows not only heterogeneity across latent classes but also variability among persons' ability within a latent class. The class membership can be understood as reflecting *qualitative* differences in response patterns across latent classes while ability can be understood as reflecting *quantitative* differences in ability level. If item difficulties are on the same scale, the item difficulty patterns among latent classes can also help to characterize the latent classes. Furthermore, the MixRM (and MixIRT models in general) may be viewed as one solution for handling the possibility of over-extraction of the number of classes that may arise with LCMs, caused by local dependence (as noted in Chapter 1). As a result, the MixRM would be expected to recover "true" latent classes more accurately than the conventional LCMs.

Due to such flexibility and the potential for finding meaningful latent classes, interest in MixRMs has been growing in several research areas, including marketing (e.g., Kaiser & Keller, 2001), medical research (e.g., Bonnefon, Eid, Vautier, & Jmel, 2008), and longitudinal data analysis (e.g., Meiser, Stern, & Langeheine, 1998; Feddag, 2008). This is particularly the case in educational and psychological assessment, where applications of MixRMs include studies on different problem solving strategies (e.g., Mislevy & Verhelst, 1990; Mislevy et al., 1991; Rijkes & Kelderman, 2006; Rost & von Davier, 1993), differential item functioning (DIF) (e.g., Samuelsen, 2005; Schultz-Larsen, Kreiner, & Lomholt, 2007; Van Nijlen & Janssen, 2008), test speededness (e.g., Bolt, Cohen, & Wollack, 2002; Meyer, 2008; Mroch, Bolt, & Wollack, 2005), faking or desirability tendency in personality questionnaires (e.g., Maij-de Meij et al., 2008), subtypes in personality (e.g., Hong & Min, 2007; Meiser & Machunsky, 2008; Rost et al., 1997), different mastery types of skills (e.g., Bolt, 1999; Rost, 1990), and measurement invariance (e.g., Eid & Rauber, 2000). von Davier and Carstensen (2007) provide a more extensive description of applications of MixRMs.

As mentioned briefly in Chapter 1, in general, analyses with the MixRM are conducted in sequence. First, one fits candidate models with different numbers of latent classes. Next, one selects the best fitting model among the candidates from the first step. Since these models are not nested within one another, the model selection decision is made by comparing information indices such as Akaike's information criterion (AIC), Bayesian Information Criterion (BIC), or Deviance information criterion (DIC) (see, for example, Li, Cohen, Kim, & Cho, 2009). The last step is to provide an interpretation of the parameters for the chosen model. In



Note. The figure was modified and reprinted with permission from Rost (1990, p. 277).

Figure 2.1: Rost's Illustration of a Mixture Rasch Model Application to a Physics Test

this regard, the class-specific item difficulties may be useful for examining the heterogeneity among the populations defined by the different latent classes. In addition, the item difficulty patterns of each class can help one infer what may have caused the classes to form. This is because, in part, the class-specific item difficulties may reveal distinctive profiles for each class in response propensities (i.e., which items are more or less difficult for the members of the different classes). On the basis of the profile of the class-specific item difficulties, the researcher might be able to make some inferences about the second dimension or dimension(s) causing the classes to form.

To illustrate how the MixRMs can work for test items, Rost (1990) fit a physics test with three different MixRMs, each with a different number of latent classes. Taking into account interpretability along with fit statistics, he concluded that there existed two distinct latent classes: a practice-oriented latent class and a theory-oriented latent class. This interpretation was drawn by scrutinizing the item response pattern of each latent class. As shown in Figure 2.1, for the theory-oriented class, the first five items turned out to be easier, whereas for the practice-oriented latent class, the last five items were easier. An important point illustrated by Rost (1990) is that determining how many latent classes remain is not simply a matter of statistical fit indices, but it is also important to have a substantive rationale for which model fits better as it is this rationale that helps to delineate what causes the latent classes to form in the population (see, for example, Bolt et al., 2002; Mislevy et al., 1991).

Note that depending on research design and research questions, a MixRM analysis can be conducted as either exploratory or confirmatory. The exploratory analysis can be carried on to attempt to understand the latent structure in the population and the nature of the qualitative differences between classes. On the other hand, the confirmatory analysis can be used to test hypotheses about different interpretations. In both types of analyses, a test is conducted to determine which model fits better (see, for example, Li, Cohen, & Bottge, 2007). If a researcher conducts an exploratory analysis, ad hoc analyses can be used to interpret the resulting latent classes. For instance, after finding three classes in the data for a depression scale using a MixRM (i.e., loss-of-libido, hopelessness, and genuine depression class), in order to characterize latent classes, Hong and Min (2007) fit a multinomial logistic model to the data set in which posterior membership was regressed on gender. They concluded that females tend to belong to a genuinely depressed class more than males do. As in exploratory factor analysis, in the absence of a theoretical justification for the interpretation, however, the exploratory MixRM analyses alone provide no guarantee that any particular interpretation is accurate.

Limitations of MixRMs. The MixRMs have the potential for developing a scale for each latent class, and yet this comes with possible disadvantages. First, the number of parameters in MixRMs increases exponentially as the number of latent classes increases because the model requires estimating ability and membership parameters for each examinee and each item parameters for each class. As a result, to obtain reliable estimates, larger samples are required as the number of latent classes increases (Li et al., 2007). Second, as pointed out previously, MixRMs themselves do not provide an explicit rationale as to what causes the latent classes to form. In most cases, in order to provide an accurate explanation for how the latent classes differ, researchers must either have a well-grounded theoretical rationale in advance or they need to conduct appropriate ad hoc analyses to relate the class membership to external (and typically manifest) variables. Care needs to be taken in either case as the results may be confounded by classification, measurement, and other possible types of errors.

To address both issues, Smit et al. (1999) have suggested using additional information in the form of covariates to help improve the detection of latent classes in MixRMs. As Smit et al. (1999) note, the effectiveness of the covariate is based on the strength of the relation with latent class membership. In the next section, we detail their idea of the inclusion of covariates in MixRMs.

Mixture Rasch Models (MixRMs) with Covariate(s). One of the most challenging tasks in utilizing a MixRM is to determine what may have caused the heterogeneity in the population. For example, Mislevy et al. (1991) encountered two groups of people who employed different cognitive strategies to solve the same spatial visualization problems. Even after two latent groups were detected, however, it was still necessary to determine who chose the rotation strategy and who chose the analytic strategy. Further, it was necessary to describe why some individuals had difficulty with problems dealing with *length*, and yet were able to solve problems about *degree of rotation* and vice versa.

The same type of question is present with the use of LCMs. To address this issue with LCMs, Dayton and Macready (1988) proposed concomitant-variable (or covariate) latent class models in which covariate(s) are included to predict class membership. This type of models is also known as a latent class regression model because the model forms a logistic regression model. In the case of two classes with one covariate, the probability of a person being in the first class conditioning on a covariate would be formulated as

$$\pi_{1|x} = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{2.6}$$

where β_0 is an intercept, x denotes the covariate, and β_1 is a logistic regression coefficient for x, which indicates the magnitude of impact of x on a person belonging to the first class compared with the second class. To generalize Equation 2.6 to multiple classes with multiple covariates, let $\mathbf{X} = (x_1, x_2, ..., x_p)$ denote realization of p predictors. Then, the probability that person i with $\mathbf{X}$ falls into class g can be represented by

$$\pi_{g|\mathbf{X}} = \frac{\exp(\beta_{g0} + \boldsymbol{\beta}_g' X)}{1 + \sum_{g=1}^{G-1} \exp(\beta_{g0} + \boldsymbol{\beta}_g' X)}, \tag{2.7}$$

where β_{g0} is an intercept of class g, $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, ..., \beta_{gp})$ are logistic regression coefficients of a set of covariates $\mathbf{X} = (x_1, x_2, ..., x_p)$ for class g, and the last class, or G is the arbitrary chosen reference class. The β 's indicate the influence of log-odds that an individual falls into class g compared with being in class G. Covariates used in this way have impact only on the latent class membership, but not on the conditional probability of an item response. The item response is still a function of item and person parameters within class. We revisit to distinguish this equation from Equation 2.18 later.

Latent class regression models have several advantages: (1) they allow for testing hypotheses about the relationships between latent class membership and covariates in the model while taking estimation errors into consideration; (2) since the models take a logistic regression form, they are flexible in the sense that covariates can be either continuous or categorical, and interactions between covariates can be easily modeled via products of their respective terms; and (3) the models can be applied to small data sets because the covariates may provide additional information and provide more degrees of freedom for estimation. See Chung (2003) for detailed descriptions regarding estimation procedures for LCMs with covariates models.

Smit et al. (1999, 2000, 2003) presented the same idea applied to MixRMs. They suggest using collateral variable(s) to predict membership in classes by simply substituting π_g in

Equation 2.5 with $\pi_{g|\mathbf{X}}$ in Equation 2.7 but keeping the conditional probability as follows

$$P(Y_{ij} = 1|\theta_i, \mathbf{X}) = \sum_{g=1}^{G} \overbrace{\frac{\exp(\beta_{g0} + \boldsymbol{\beta}'_g X)}{1 + \sum_{g=1}^{G-1} \exp(\beta_{g0} + \boldsymbol{\beta}'_g X)}}^{\pi_{g|\mathbf{X}}} \overbrace{\frac{\exp(\theta_i - b_{jg})}{1 + \exp(\theta_i - b_{jg})}}^{Rasch} \tag{2.8}$$

where β_{g0} denotes an intercept for class g, and $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, ..., \beta_{gp})$ is a vector of logistic regression coefficients of covariates, and θ and b_{jg} are ability and a class-specific item difficulty, respectively. They argued that by including collateral variable(s), one may possibly obtain more stable identification of latent classes and reduced standard errors of estimation. Cho, Cohen, and Kim (2006) noted that the use of covariates, in fact, improved recovery of both item difficulty and group membership, and this use may reduce the attenuation in estimation of the relationship between covariates and class membership which can occur in ad hoc analyses noted previously. To some extent, the use of covariates also has the potential to aid in detecting latent classes when the sample sizes or differences among latent classes are small (Cho, Cohen, Kim, & Bottge, 2007; Li et al., 2007).

To date, most covariates considered in the MixRM have been manifest variables such as race, gender, English language learner status, etc. Such covariates are useful to the extent they are associated with the factor(s) causing the latent classes to form. Even though they may be proxies for the underlying causes, unfortunately most manifest variables are not themselves causally related to why the latent classes form. As a result, these manifest covariates actually may tell us little about what have caused the classes to form.

For example, Samuelsen (2005) included ethnicity and gender as covariates of a MixRM in DIF analyses of an English language test in order to investigate whether such variables were related to the underlying cause of the DIF. She, however, did not find a sufficiently strong association between latent class and either gender or ethnicity to adequately explain the DIF present in an English language proficiency test. Rather, by investigating patterns of item difficulties between two classes, she suggested that the type of instruction or the cognitive style of the students were more plausible explanations for the causes of the differences observed in item functioning between the classes. As another example, Lee, Fradd, and

Sutman (1995) found that students with adequate knowledge were more adept at applying appropriate problem solving strategies specific to the tasks than were their peers who had less knowledge in science class. They concluded content knowledge had something to do with the use of particular strategies for solving problems on a science test.

These studies suggest that examinees' knowledge states may result in different ways of thinking or different ways of approaching problems, and these differences may result in the formation of latent classes. From here, we may conclude that to understand the formation of latent classes, it may be instructive to better understand the structure of examinees' knowledge relative to the use of particular strategies for answering test items. One challenge is that, as with ability in IRT, knowledge state is a latent construct, and adding this kind of variable to the MixRM requires adding a second set of latent variables. Fortunately, it is possible to take advantage of recent advances in the development of DCMs for describing examinees' knowledge states. In this dissertation, we make use of this to facilitate the development of the DCMixRM. Before detailing DCMs, in the next section, we connect local dependence and multidimensionality.

2.3 Local dependence and Multidimensionality

The assumption of local independence is pivotal in both LCMs and IRT models. This assumption implies that the latent variables imposed by the model should fully account for the associations between observed item responses. Mathematically speaking, item responses can be related to each other only though the imposed latent space. Therefore, in a fixed latent space, the joint distribution of item responses is equal to the product of the marginal item distributions. This can be written as

$$P(Y_{i1} = 1, Y_{i2} = 1, ..., Y_{iT} = 1|\Theta_i) = \prod_{j=1}^{T} P(Y_{ij} = 1|\Theta_i), \tag{2.9}$$

where Y_{ij} are the item responses, Θ_i are parameters in the latent space, and $P(Y_{ij} = 1|\Theta_i)$ are the conditional probabilities of a correct answer given the parameters.

Local dependencies, however, are often observed in practice. These dependencies can arise from several sources including fatigue, examinee knowledge or ability in specific areas, speededness, item format, passage dependence, item chaining (i.e., relation of item responses through a particular stimulus), raters, or item content etc. (Yen, 1993). Some dependencies are irrelevant to the latent variable of interest, while others are necessarily related to a dimension of the latent space being measured.

Ignoring local dependence among items may result in overestimating the precision of the measurements and biased estimates, and therefore it also may lead to misinterpretation of the latent space (Yen, 1993; Wainer & Thissen, 1996). This is particularly the case when the local dependence suggests multidimensionality in test items. Hence, it is useful to consider the distinction between violation of local independence due to idiosyncratic features of test format and the departures from unidimensionality (Rosenbaum, 1988). The former can be considered nuisance variation, but the latter may be substantially significant in detecting the latent space which needs to be accounted for in modeling item responses.

While discussing the construct validity of IRT models, Steinberg et al. (2000) present a distinction between two types of sources which can cause the multidimensionality, or local dependence in test items: differences in examinees' cognitive patterns (between group), and testing formats (within group). Previously, we also mentioned that examinees' cognitive difference may result in heterogeneity in population and testing formats can result in multidimensionality.

To copy with local dependence, testlet models (e.g., Bradlow, Wainer, & Wang, 1999), hybrid models (e.g., Yamamoto, 1987), rater effect models (e.g., Patz, Junker, Johnson, & Mariano, 2002), and method effect models (e.g., Tomás & Oliver, 1999) have been developed. Each model focuses on specific source to cause the local dependence, or a secondary dimension among items, but typically all of these models consider this secondary dimension a nuisance dimension or noise, which is unrelated to the latent variable measured by test items. Bifactor models, however, provide an alternative for resolving the local dependence and accounting for the a set of secondary dimension, or multidimensionality, among item responses. This is done by assuming that the latent space is made up of multiple domain specific factors along with a global factor (Holzinger & Swineford, 1937). The DCMixRM is similar to bifactor models in the sense that it assumes that there are several domain or group factors as well as a general factor.

In the next section, we describe DCMs in greater detail. We show how DCMs can be used to model discrete latent variables in multidimensional space. Moreover, we present how this multidimensional space can be used to account for heterogeneity in population.

2.4 Diagnostic Classification Models (DCMs)

IRT models have been successfully applied in many areas because by using these models one can establish a common metric for expressing estimates of item and ability parameters (Baker & Kim, 2004). One concern with the use of IRT models, however, is that the ability measures usually provide only a coarse description of the latent variable that explains examinees' item responses. IRT models, particularly unidimensional IRT models, provide a linear explanation of the latent construct but may not lead to a more direct understanding of the possible factors involved. In other words, IRT models are useful for ranking, comparing examinees, or even for predicting who will do well in future, but IRT scales do not generally provide sufficient diagnostic information for helping teachers intervene with students.

At the heart of this criticism is a recognition of the need for alternative assessment methods from which one may draw more detailed diagnostic information about examinees, and from which one may directly connect that information to examinees' specific instructional needs. Considered that diagnostic information or feedback about examinees' weaknesses and strengths plays a fundamentally important role of assessment in improving learning, this agrement is valid. Emphasizing the instructional uses of tests, Linn (1986) argued that

> a test that reliably rank orders students in terms of global test scores provides a teacher with relatively little information about the nature of a student's weak-

> nesses, errors, or gaps. For example, the knowledge that a student scores, say, in the 10th percentile on a standardized arithmetic test suggests the student has a general weakness in the area of arithmetic relative to his or her peers. However, such a score does not, by itself, indicate the source of the problem or what should be done to improve the student's level of achievement; that is, it lacks diagnostic information. (p. 1158)

Nichols, Chipman, and Brennan (1995) articulated the same concern:

> The assessment must respond to questions like these: What is the appropriate level of detail to represent performance for the purpose of diagnosis? What statistical approaches are most useful for making inferences about the procedural knowledge underlying performance? How do you take into account the learning that is, after all, the goal of the tutor? (p. 3)

The DCM family of models provides one response to this need. These models are intended to yield diagnostic information specific to the mastery state of individual knowledge components needed for correctly answering test items. Several terms have been used to refer to these components, including attributes (e.g., Tatsuoka, 1983), skills (e.g., Francis et al., 2006), components (e.g., Kruidenier, 2002), and sub-skills (e.g., Moseley, 2004). *Attribute* is usually taken as the generic term in DCMs to characterize these specific components in test items. In this dissertation, the terms *attribute* and *mastery profile* are used to indicate a knowledge component and pattern of mastery status on these components, respectively. Attributes defined for DCMs are intended to be used for providing diagnostic information in a form that can be used for instructional decisions (Nichols et al., 1995). In this section, we describe some variations of DCMs, but we focus more on describing the log-linear cognitive diagnosis model (LCDM).

Diagnosis Classification Models (DCMs). The primary purpose of DCMs is to provide a mathematical model that can be used to describe individuals' knowledge states by referring to the presence or absence of the set of attributes needed for correctly answering test items. The mastery status of an examinee on each of these attributes is assumed to provide useful information on the examinee's strengths and weaknesses. Rupp, Templin, and Henson (2010) offer the following definition of diagnostic classification models:

> Diagnostic classification models are probabilistic confirmatory multidimensional latent variable models with a complex loading structure. They are suitable for modeling categorical response variables and contain categorical latent variables that generate latent classes, which are used to classify respondents. Thus, they enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes at a comparatively fine grain size. (p. 108)

This definition points out several important features of DCMs: the models are confirmatory; more than one attribute can be involved in solving an item; latent variables are discrete and multidimensional; and the models yield classification of examinees in terms of knowledge states.

This definition covers the following models, most of which have been developed over the last two decades: the Rule Space Model (Tatsuoka, 1983); the Restricted Latent Class Model (Haertel, 1989); the Reparameterized Unified (Fusion) Model (RUM, DiBello, Stout, & Roussos, 1995); the Noisy Inputs, Deterministic And Gate Model (NIDA, Maris, 1999); the Deterministic Inputs, Noisy And Gate Model (DINA, Junker & Sijtsma, 2001); the Bayes Nets Model (Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002); the general diagnostic model (GDM, von Davier, 2005); the Deterministic Input, Noisy Or Gate Model (DINO, Templin & Henson, 2006); and the Log-linear Cognitive Diagnosis Model (LCDM) (Henson, Templin, & Willse, 2009). See Fu (2005), Junker (1999), and Templin (2004) for a more comprehensive review.

As detailed below, these models differ in how they conceptualize relationships among attributes, but as defined, they have one thing in common: all the models classify examinees based on mastery states on a set of attributes. The usual way of classifying examinees is to use mastery profiles. Typically, these are represented in vectors consisting of 1's to indicate mastery and 0's to indicate non-mastery. For instance, if an examinee is classified as having mastery profile (1,1,0,0), this is taken to mean the individual is thought to have mastered the first two attributes but not the last two attributes.

Specifying mastery status in this way requires understanding knowledge structures in specific and detailed forms as the basis for classification. All DCMs explicitly include this knowledge structure in the model. This is done through an item-by-attribute incidence matrix, commonly referred to as a Q-matrix (Tatsuoka, 1990). The Q-matrix is designed to represent the relationship that is assumed between items and attributes. It indicates which attributes are required for successful performance on specific items. The entries in the Q-matrix can be viewed as loading indicators that specify a factor structure in a confirmatory factor analysis (Rupp & Templin, 2008).

In order to understand how a Q-matrix works with DCMs, let $\mathbf{Q}$ be the $K \times T$ attribute-by-item incidence matrix in which items are typically organized in rows and attributes in columns, q_{jk} be the entry that lies in the j_{th} row and the k_{th} column of a matrix:

$$q_{jk} = \begin{cases} 1 & \text{if item } j \text{ involves attribute } k \\ 0 & \text{otherwise} \end{cases}$$

and $\boldsymbol{\alpha}_i = (\alpha_1, \alpha_2, ..., \alpha_K)$ be a vector of an examinee's mastery profile:

$$\alpha_i = \begin{cases} 1 & \text{if person } i \text{ mastered attribute } k \\ 0 & \text{otherwise.} \end{cases}$$

Combining a Q-matrix and an examinee mastery profile produces a predicted, deterministic response matrix. Table 2.1 depicts a Q-matrix with five items and four attributes along with the predicted responses of the person who is presumed to have mastered first two attributes but not the last two items, (1,1,0,0). According to this Q-matrix, the first item requires mastery of the first attribute, the second item requires the first two attributes, the third requires the fourth attribute, the fourth requires the second and third attributes, and the last item requires the third and fourth attributes. As shown in the table, in general, the

Table 2.1: Hypothetical Q-matrix of Five Items with Four Attributes

Item	α_1	α_2	α_3	α_4	Response [a]
1	1	0	0	0	1
2	1	1	0	0	1
3	0	0	0	1	0
4	0	1	1	0	0/1[b]
5	0	0	1	1	0

[a] Predicted responses of mastery profile (1,1,0,0).

[b] It depends on whether the posited model requires intersection or union among attributes for answering correctly to an item.

entries are binary (i.e., indicating the absence or presence of an attribute), but they could be ordinal or even continuous (i.e., indicating the degree to which an attribute is present) (see Karelitz, 2004; Templin, 2004; von Davier, 2005).

Ideally, given this Q-matrix, a person, who has the mastery profile of (1,1,0,0), should respond correctly to Items 1 and 2, but not to Items 3 and 5 as indicated in the last column of the table. This can be done through a simple matrix multiplication as follows

$$\mathbf{Q} \times \boldsymbol{\alpha}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0/1 \\ 0 \end{bmatrix}.$$

As indicated in the table, however, it is not clear whether or not the person can answer Item 4 correctly because this person has mastered only one out of two required attributes. In addition, there are some sources that may cause such a response pattern to vary in real data. These include alternate strategy selection, incompleteness of the Q-matrix, low positivity in an attribute, and slip (i.e., random error) (DiBello et al., 1995). When formulating the probabilistic model, each model handles this variation slightly differently. We elaborate this below in the discussion of compensatory and conjunctive models.

All DCMs rely heavily on the Q-matrix because it maps item responses onto those underlying attributes that can be used to give a diagnosis of individuals' knowledge status with respect to mastery or non-mastery status. As a result, the Q-matrix is an essential component in DCMs. Along these lines Junker (1999) asserted:

> The Q-matrix is thus essentially an accounting device that describes the 'experimental design' of tasks or responses in terms of underlying attributes that the responses are intended to be sensitive to. Although it has gained prominence in recent years as a tool for task analysis in the work of Tatsuoka (e.g. Tatsuoka, 1990, 1995), it or something like it, would be present in any well specified model of task performance in terms of underlying student attributes or task features. (p. 13)

Tatsuoka (1990) emphasized a systematic or automatic interpretation of mastery state using the Q-matrix by describing the purposes of the Q-matrix is: (1) to make it easier to construct a set of items relevant for diagnosing weaknesses or strengths of knowledge states in terms of the attributes of interest, and (2) to extract a set of binary patterns describing performance over the given items, each pattern being produced by a systematic application of a combination of the set of attributes.

There are a number of ways to develop a Q-matrix. In educational testing, Q-matrices are typically constructed based on learning or cognition theory, experts' judgement, empirical research, or examinees' think-aloud protocols. (Buck & Tatsuoka, 1998). Also, various methods can be integrated for constructing a Q-matrix. Gierl, Tan, and Wang (2005), for instance, used both examinees' responses and experts' judgement by employing exploratory and confirmatory factor analytic methods to analyze mathematics and critical reading items on the SAT.

To date, DCMs have been being applied to several testing programs including the Science Education for Public Understanding Project as a prototypical example of a DCM-grounded classroom-formative assessment (Stout, 2007), and the PSAT/NMSQT (Preliminary SAT/National Merit Scholarship Qualifying Test) as an example of a summative assessment (DiBello & Crone, 2001, 2002; Wang & Gierl, 2007).

Thus far, we have described common features of DCMs, but as mentioned before, each model provides unique ways of linking observable responses on test items to the latent attributes. Broadly speaking, all DCMs can be classified into two categories, compensatory or conjunctive, based on the way the model conceptualizes the compensatory relationship among attributes. In a compensatory model, mastery of a subset of all the required attributes can compensate for the lack of mastery of the remaining attributes. In a conjunctive model, all attributes defined in the Q-matrix have to be mastered in order to respond correctly on an item. The DINO model is an example of a compensatory model and the DINA model is an example of a conjunctive model (see Junker & Sijtsma, 2001 and Templin, Henson, & Douglas, in press, for more details of the distinction). As described in the next section, the LCDM provides a general framework for DCMs and can include both compensatory and conjunctive models.

Log-linear Cognitive Diagnosis Model (LCDM). To specify relationship between item and attributes, the LCDM utilizes one of categorical data analysis methods for contingency tables, a log-linear model framework. A very simple log-linear model can be employed to analyze a contingency table in which two categorical data present. Assuming one of categorical data in the table is a latent variable rather than manifest one, the LCDM can be formulated as exactly the same way as the regular log-linear model. Let $\boldsymbol{\alpha}_i = (\alpha_1, \alpha_2, ..., \alpha_K)$ be a vector representing mastery states of examinee i on attributes, $\mathbf{q}_j = (q_{j1}, q_{j2}, ..., q_{jK})$ is a vector denoting attributes required by item j, and K indicates the total number of attributes. For the LCDM, Henson et al. (2009) adopted a log-linear model to define the probability of a correct response as

$$\log\left[\frac{P(Y_{ij}=1|\boldsymbol{\alpha}_i,\mathbf{q}_j)}{1-P(Y_{ij}=1|\boldsymbol{\alpha}_i,\mathbf{q}_j)}\right]=\lambda_{0j}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j), \tag{2.10}$$

or equivalently

$$P(Y_{ij}=1|\boldsymbol{\alpha}_i,\mathbf{q}_j)=\frac{\exp[\lambda_{0j}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)]}{1+\exp[\lambda_{0j}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)]} \tag{2.11}$$

where i and j indicate person and item, respectively; $\boldsymbol{\lambda}_j$ represents a vector of logistic regression coefficients of attributes indicating the effects of mastery of attributes on a correct response; $\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)$ is a set of linear combinations of $\boldsymbol{\alpha}_i$ and $\mathbf{q}_j$ which connects a Q-matrix into the probability; and λ_{0j} is an intercept.

To better understand the model, let's take a close look at each parameter. λ_j's, as indicated in Equation 2.10, are influences of mastery attributes on a correct response compared with non-mastery of any of the attributes. All parameters are given in the odds ratio scale, and therefore, λ_{0j} is the odds ratio of correct response from those who have not mastered any of the attributes. $\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)$ on the right side of Equation 2.10 designates the effect structure of attributes in the model as follows

$$\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)=\sum_{k=1}^{K}\lambda_{jk}(\alpha_k q_{jk})+\sum_{k=1}^{K}\sum_{u>k}\lambda_{jku}(\alpha_k\alpha_u q_{jk}q_{ju})+\ldots. \tag{2.12}$$

where λ_{jk}'s in the first part of the right side of Equation 2.12 represent the main effects of each attribute on item j, and λ_{jku}'s in the second part of the equation indicate the two-way interaction effects of the combinations of attributes k and u on items j. For instance, Item 4 in Table 2.1 requires two attributes, $q_{42}=1$ and $q_{43}=1$. The probability of the correct response on that item takes the form

$$P(Y_{i4}=1|\boldsymbol{\alpha}_i,\mathbf{q}_4)=\frac{\exp(\lambda_0+\lambda_2\alpha_2+\lambda_3\alpha_3+\lambda_{23}\alpha_2\alpha_3)}{1+\exp(\lambda_0+\lambda_2\alpha_2+\lambda_3\alpha_3+\lambda_{23}\alpha_2\alpha_3)}. \tag{2.13}$$

Therefore, for the examinee in Table 2.1 who has mastered the first and second attributes, or mastery pattern (1,1,0,0), the probability of getting a correct answer for this item is determined by only quantities of λ_0 and λ_2. This is because the examinee has not mastered the third attribute, $\alpha_3 = 0$, and hence neither λ_3 nor λ_{23} has an impact on the probability of getting a correct answer. Depending on how many attributes are included in the test, the LCDM may include main effects for each attribute, two-way interactions, three-way interactions, and so forth.

As with other DCMs, the LCDM is a confirmatory latent class model because it imposes restrictions on the parameters for classifying latent classes. It does this by setting some conditional probabilities to be equal across latent classes. Therefore, an important part of the LCDM specification concerns imposing appropriate restrictions on model parameters. Each restriction relates to identity, design, and monotonicity on probability. First, as with other logit models for categorical data, a baseline is required for model identification. In this model, the nonmastery class, consisting of those who have not mastered any of the attributes, is arbitrarily taken as the baseline. Second, as a factor loading design matrix in a confirmatory factor analysis, a Q-matrix specifies the structure of $\boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)$. For instance, for Item 4 $\boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)$ forms $\lambda_2\alpha_2 + \lambda_3\alpha_3 + \lambda_{23}\alpha_2\alpha_3$ as shown in Equation 2.13. This is because the Q-matrix specifies that this item requires Attributes 2 and 3 for a correct answer. For Item 2, however, it should be $\lambda_1\alpha_1 + \lambda_2\alpha_2 + \lambda_{12}\alpha_1\alpha_2$, as Attributes 1 and 2 are required by the item. Finally, sets of constraints are included to ensure the monotonicity of the model, namely, the probability increases as examinees master more attributes. In doing so, all main effects must be positive, and interaction effects are constrained to be greater than the negative values of all related main effects: for Item 4, $\lambda_2 \geq 0$, $\lambda_3 \geq 0$ and $\lambda_{23} \geq$ -minimum (λ_2, λ_3) (i.e., negative value of the smaller of λ_2 or λ_3) to ensure $\lambda_2 + \lambda_{23} \geq 0$ and $\lambda_3 + \lambda_{23} \geq 0$.

Because of the use of log-linear models with latent variables, the LCDM has several advantages over other DCMs. First, the LCDM provides a general framework for DCMs in the sense that one may define any DCM by adding simple constraints for relevant parameters of either the main effects or the interaction effects (i.e., λ's).

Second, the LCDM allows one to specify the compensatory relationship at the item-level in one model. That is, some items can be specified as compensatory and others as conjunctive.

This can be accomplished by manipulating the interaction terms for individual item in the model because the interaction effects indicate whether or not there is a conjunctive relation among attributes. As the interaction effect is positive, there is an additional effect above and beyond each main effect on the probability of the right answer as in a conjunctive model. However, when the interaction effect is less than zero, there is no effect of mastery of multiple attributes and the model becomes compensatory. For instance, if one has reason to believe that the first two attributes in Table 2.1 have a compensatory relationship, but the second and third attributes do not have such relationship, one can fit Item 2 to a compensatory model by setting $\lambda_2 \geq 0$, $\lambda_3 \geq 0$ and $\lambda_{12} = 0$, yet fit Item 3 to a conjunctive model by constraining $\lambda_2 = 0$, $\lambda_3 = 0$ and $\lambda_{23} \geq 0$.

Third, it is easy to interpret coefficients in the LCDM. For example, λ_1 indicates the main effect of possessing Attribute 1 on the probability of getting a correct answer. Likewise, λ_{12} represents an interaction effect for a correct answer if an examinee has mastered both Attributes 1 and 2.

Fourth, this model enables one to test hypotheses about the Q-matrix or about a posited relationship among attributes by examining the significance level of λ's in the model. The magnitudes of the λ's can also be used to help prioritize attributes.

Limitations of DCMs. The Q-matrix is a fundamental part of all DCMs including the LCDM. Hence, verification of a Q-matrix is essential, and misspecification and incompleteness of the Q-matrix can result in inaccurate estimates. Rupp and Templin (2008) have noted that misspecification of a Q-matrix has an impact on both parameter estimation and accuracy of classification. Incompleteness of Q-matrix refers to the fact that an item may require one or more attributes in the Q-matrix in addition to the ones that have been specified (DiBello et al., 1995; Hartz, 2002). DiBello et al. (1995) further caution that examinees may employ different response strategies than those assumed in the Q-matrix. Incompleteness due to either of these poses a threat to the validity of inferences drawn from a DCM.

In addition, unlike IRT models, DCMs are defined by discrete latent spaces, and therefore it is not possible to obtain a continuous scale regarding general ability from DCMs. In order to fully support instruction, assessment should provide two types of information: attainment assessments which deal with whether a student has attained the desired instructional goals; diagnostic assessments which are designed to provide specific information on individual learning deficiencies and misunderstandings in order to regulate learning processes (Nitko, 1995).

Several models have been proposed to attempt to overcome these issues, including the Rule Space model (Tatsuoka, 1983), the RUM (DiBello et al., 1995), and the Higher Order DINA model (de la Torre & Douglas, 2004). Currently, none of these models is capable of handling incompleteness in the Q-matrix and at the same time of providing a continuous scale on a general latent variable.

In this dissertation, we propose a diagnostic classification mixture Rasch model (DCMixRM) as an alternative way for resolving these issues. In so doing, a MixRM and an LCDM are combined into the DCMixRM. As previously noted, a MixRM is itself a combination of a Rasch model and an LCM. In a MixRM, an LCM and a Rasch model capture different aspects of examinees' responses: the Rasch model describes quantitative differences in examinees' ability within a latent class, and hence it enables one to obtain quantitative information regarding a continuous latent ability; the LCM captures qualitative differences among the latent classes. In the mean while, the LCDM attempts to provide even finer, more detailed analyses by specifying the attributes needed to answer each item on the test and incorporating that information into the mastery profile. The DCMixRM includes these aspects of each component model and, as a result, can provide richer information to characterize examinees' responses. In the next section we describe the model and estimation procedures.

2.5 Diagnostic Classification MixRM (DCMixRM)

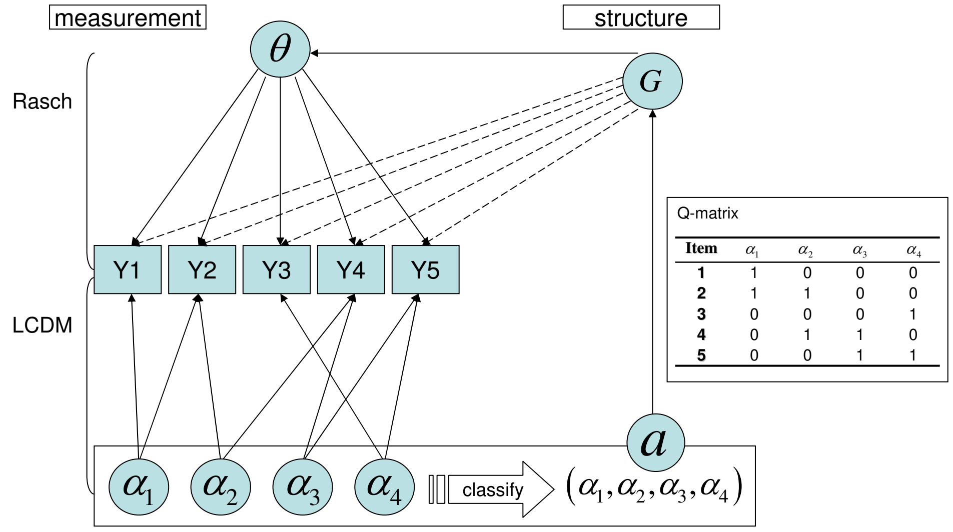
The motivation for the development of the DCMixRM was twofold. First, the DCMixRM was designed to provide explanatory information regarding factors that were potentially causing the latent classes to form. In order to accomplish it, this model includes mastery states of attributes as covariates. Second, the DCMixRM was designed to rigorously connect assessments to instruction by furnishing diagnostic information along with summative information.

In doing so, the DCMixRM incorporates the LCM, the IRT model, and the LCDM. Rasch models are the simplest among IRT models, and in general, item difficulties are easier to estimate than other item parameters. Hence, here we limit the IRT model to a Rasch model, and yet its extension to other IRT models such as a 2-parameter, 3-parameter, or polytomous IRT models is straightforward. In the rest of this chapter, we detail assumptions, specifications, and advantages of the model.

2.5.1 General Framework of DCMixRM

As indicated previously, the purposes of the DCMixRM are (1) to obtain an estimate of examinee's general ability, θ, (2) to classify an examinee into one of the mastery profiles pre-specified by the Q-matrix, $\boldsymbol{\alpha}$, (3) to detect heterogeneity in response patterns in the population, g, and (4) to provide a plausible explanation as to why one latent class may respond differently from others using mastery profiles. We begin with introducing the following assumptions of the model.

Local Independence. As with other latent structure models, local independence is assumed for this model. However, item responses are locally independent conditioning *jointly* on examinees' ability, class membership, and mastery profile. In other words, we redefine the local independence assumption by extending a latent space to three types of latent variables. This results in relaxing the strong local independence of an LCM and a Rasch model.



Item	α_1	α_2	α_3	α_4
1	1	0	0	0
2	1	1	0	0
3	0	0	0	1
4	0	1	1	0
5	0	0	1	1

Note. The Q-matrix is provided in the rectangle at the right side of the figure.

Figure 2.2: A Schematic Representation of the Diagnostic Classification Mixture Rasch Model (DCMixRM)

Exclusivity and Exhaustivity. Both latent classes and mastery profiles are mutually exclusive and exhaustive. Mutual exclusivity means that the intersection of any latent classes or mastery profiles is null; that is, there is no overlapping among latent classes or mastery profiles. Exhaustivity means that the union of all the latent classes covers entire classes and the union of all the mastery profiles covers the entire profiles.

Monotonicity. The probability of a correct response to an item increases as ability increases or as an examinee masters more attributes.

Compensatory. A general ability and attribute mastery in the model are compensatory such that having higher ability may compensate for lack of one or more requisite attribute(s), and vice versa.

Figure 2.2 illustrates how the DCMixRM can be understood given the Q-matrix in Table 2.1. As portrayed in this figure, the DCMixRM consists of four elements: multivariate outcomes (item responses, Y's); a continuous latent variable (ability, θ); a discrete latent variable (latent classes, G); and a vector of discrete latent variables (attributes, α's). These four elements constitute a measurement model and a structural model connected by regression components. Central to this model is the specification of hierarchical conditional relationships: the measurement model specifies the probability of the observed responses conditional on the three latent variables such as ability, mastery profile, and membership, whereas in the structural model these latent variables themselves are regressed on one another.

To be more specific, in the measurement component, observed item responses are jointly regressed on θ and α's through the Rasch model and the LCDM. As shown in the bottom of the figure, it is possible for one item to include more than one attribute. In addition, the dotted lines from G to each item indicate each latent class may have different sets of item difficulty parameters as is the case in the MixRM. Next, in the structural model, ability is regressed on class membership, and class membership is regressed on mastery profile to explain latent class as covariates. As shown in Figure 2.2, the classified mastery profile is the variable of interest in the structural model while the mastery of an individual attribute is the variable of interest in the measurement model.

The most challenging task here is that of the four elements in the model, three variables are latent and have to be inferred from the data. The question then becomes how we estimate these three latent variables from a set of item responses. This requires a set of constraints in specifying the model. In the next section, the model constraints and estimation procedures are described.

2.5.2 Model Specification

The DCMixRM is made up of two components: the measurement component and the structural component. The measurement component contains the probabilistic statement for item responses conditioning on item and person parameters; the structural component describes the relationship among latent person variables such as ability, class membership, and mastery profile. Two pieces of information are required to fully specify the DCMixRM: (1) the conditional probability of a correct response given the latent space in the measurement model, and (2) the joint distribution of ability, mastery profile, and class membership of examinees in the structural component.

To begin, we combine the Rasch model with the LCDM in the measurement model. Recall Y_{ij} and θ_i denotes the response of person i to item j and person i's ability, respectively. With the Rasch model, assuming local independence with one continuous latent variable, the probability of a correct response to item j can be written as

$$P(Y_{ij} = 1|\theta_i) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}, \tag{2.14}$$

where b_j is the difficulty parameter of the item as in Equation 2.4. Likewise, in the LCDM the probability of a correct response on item j can be written as

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \mathbf{q}_j) = \frac{\exp[\lambda_{0j} + \boldsymbol{\lambda}_j' \mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)]}{1 + \exp[\lambda_{0j} + \boldsymbol{\lambda}_j' \mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)]}, \tag{2.15}$$

where $\boldsymbol{\lambda}_j$ denotes a vector of coefficients of mastery effects on the correct response to the item, λ_{0j} is an intercept, and $\mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)$ is a set of linear combinations of $\boldsymbol{\alpha}_i$ and $\mathbf{q}_j$ in which $\boldsymbol{\alpha}_i$ represents mastery state, and $\mathbf{q}_j$ is a vector of attributes for the item in a Q-matrix as in Equation 2.11.

Not only are both the Rasch Model and LCDM probabilistic models, but they also rely on the local independence assumption: item responses are independent of each other conditional on the latent variable(s). If item responses are assumed to depend both on the general ability and the mastery profile, and if ability and attribute mastery are assumed to be compensatory, then these two models can be combined and revised as

$$P(Y_{ij} = 1|\theta_i, \boldsymbol{\alpha}_i, \mathbf{q}_j) = \frac{\exp[\theta_i - b_{j0} + \boldsymbol{\lambda}_j' \mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)]}{1 + \exp[\theta_i - b_{j0} + \boldsymbol{\lambda}_j' \mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)]}, \tag{2.16}$$

where for identification purposes, the intercept of LCDM (λ_0) is absorbed into item difficulty and becomes b_{j0}. Further, in case where there are more than two latent classes in the

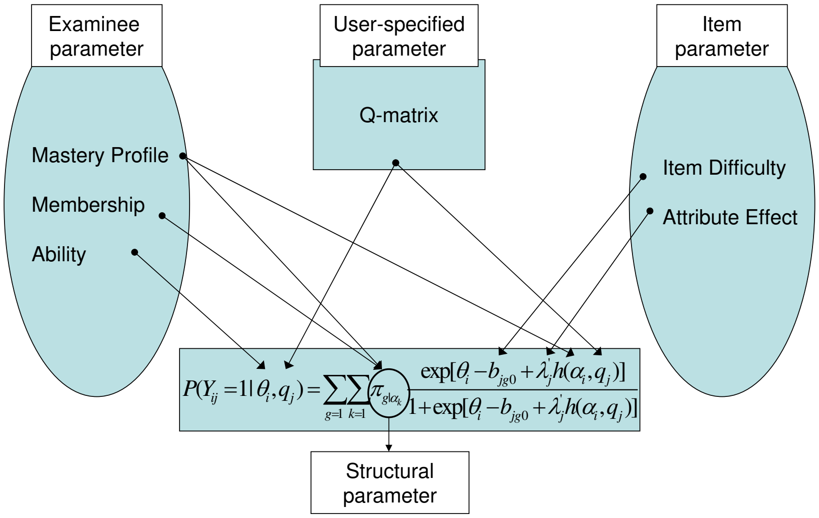

Examinee parameter
User-specified parameter
Item parameter
Q-matrix
Mastery Profile
Membership
Ability
Item Difficulty
Attribute Effect
$P(Y_{ij}=1|\theta_i,q_j)=\sum_{g=1}\sum_{k=1}\pi_{g|\alpha_k}\frac{\exp[\theta_i-b_{jg0}+\lambda_j' h(\alpha_i,q_j)]}{1+\exp[\theta_i-b_{jg0}+\lambda_j' h(\alpha_i,q_j)]}$
Structural parameter

Figure 2.3: Configuration of Connections among Components of DCMixRM

population, and each class may have its own item difficulty parameters, then the probability can be reformulated as

$$P(Y_{ij}=1|\theta_i,g,\boldsymbol{\alpha}_i,\mathbf{q}_j)=\frac{\exp[\theta_i-b_{jg0}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)]}{1+\exp[\theta_i-b_{jg0}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)]} \tag{2.17}$$

where b_{jg0} indicates item difficulty for class g reflecting that each class is allowed to have its own set of item difficulty parameters.

Next, Figure 2.3 sketches how each part is connected in the model. In particular, this figure points out that in the structural component of the model, the latent class is regressed on the mastery profile. Here, both the latent class and the mastery profile are discrete variables, and hence a multinomial logit is a natural choice as a link function to specify the relationship between these two variables (Agresti, 2002). Recall $\boldsymbol{\alpha}_k=(\alpha_1,\ldots,\alpha_k)$ is a vector of mastery profile and K is the number of non-redundant profiles. The mastery profiles are assumed to be mutually exclusive, and hence each individual belongs to only one of the profiles. For instance, if there are four attributes in a Q-matrix, there exist 16 mastery profiles (i.e.,

$2^4 = 16$). If an examinee possesses all four attributes, the mastery profile for this person can be expressed by $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1, 1, 1, 1)$. However, when a dummy coding system is used to denote each mastery profile, the mastery profile of this individual can be represented as $\boldsymbol{\alpha}$=$(0_1, 0_2, \ldots, 1_{16})$. This indicates that this examinee is classified into the $16th$ profile, and in this dummy coding system, this is also equivalent to $\boldsymbol{\alpha}$=$(0_1, 0_2, \ldots, 0_{15})$, and as a result those who master all attributes serve as the baseline in the model. That is, in the case of four attributes included in the Q-matrix, 15 profiles are non-redundant profiles (i.e., the total number of profiles minus one). Now, let $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, \ldots, \beta_{gk})$ be another vector of logistic regression coefficients relating the mastery profiles to latent classes. The probability that person i belongs to class g conditional on the mastery profile $\boldsymbol{\alpha}_k$ can be written

$$\pi_{g|\boldsymbol{\alpha}_k} = \frac{\exp(\beta_{g0} + \boldsymbol{\beta}'_g \boldsymbol{\alpha}_k)}{1 + \sum_{g=1}^{G-1} \exp(\beta_{g0} + \boldsymbol{\beta}'_g \boldsymbol{\alpha}_k)}, \tag{2.18}$$

where β_{g0} is an intercept. As in a regular multinomial logit model, the exponentiated coefficients (β's) indicate the magnitude of influences of the mastery profiles on the log-odds that an individual belongs to class g compared with the last latent class G. That is, these coefficients may be interpreted as estimated odds ratios. It should be noted that β_{gk}'s are set to zero because the last profile is a reference group for the model.

To illustrate how these parameters in the model can be interpreted, we take an example from Lanza, Collins, Lemmon, and Schafer (2007). Using an LCM with covariates, the authors investigated adolescent drinking behavior in the United States. Table 2.2 (taken from Lanza et al., 2007) summarizes the effects of covariates on adolescent drinking patterns. School skipping and grades were included as covariates, and five classes were detected in the study. “Non drinkers” and “No skip” were selected (arbitrarily) as baselines. One plausible interpretation based on this table is that when conditioning for grades, adolescents who skipped school were five times (i.e., $e^{1.6} \doteq 5$) more likely to become “Heavy drinkers” than those who did not skip school at all.

It is noteworthy that in a MixRM covariate(s) only affect latent class membership (as is shown in Equation 2.8), but in a DCMixRM the mastery status appears twice. It first

Table 2.2: An Illustration of Covariate Effect on Latent Class Membership

	Skipped School		Grades	
Class	β	Odds Ratio	β	Odds Ratio
Nondrinkers [a]	.0	1.0	.0	1.0
Experimenters	.4	1.5	-.2	.8
Drinkers	.7	2.0	-.4	.7
Bingers	.9	2.5	-.3	.7
Heavy drinkers	1.6	5.0	-.5	.6

Note. The table was reprinted with permission from Lanza et al. (2007, p. 688).
a. Reference group.

appears in the measurement component as all attributes relevant to an item have influence on the conditional probability (see Equation 2.17). Mastery profiles appears again in the structure component where they are used to explain latent class membership (as shown in Figures 2.2, 2.3 and Equation 2.18).

Advantages of the DCMixRM. The DCMixRM has several advantages over the MixRM or the LCDM alone. First, it allows us to handle incompleteness of the Q-matrix. As discussed in the previous chapter, the Q-matrix is pivotal in DCMs. If relevant attributes are incorrectly omitted in the Q-matrix, this incompleteness of the Q-matrix can be a threat to the validity of inferences drawn from a DCM. By introducing θ, the DCMixRM can potentially overcome some of this issue. That is, all necessary but unspecified attribute(s) in the Q-matrix can be absorbed into θ. Second, the model provides a way to detect heterogeneity in the population even at the same ability level. This is done through the LCM component of the model. Third, the model has potential to yield more accurate classification of latent class membership and reduce standard errors in classification because it incorporates the mastery profile as a covariate. Fourth, it allows for testing hypotheses regarding relationships between latent classes and mastery profiles, and as a result, it is possible to directly provide a rigorous explanation about features of latent classes. Fifth, the model furnishes two sets of inferences about examinees' performances at different grain sizes: inference on a global ability level and inferences on mastery profiles that yield a finer level of information about sources of examinees' weaknesses or strengths in knowledge states. As mentioned earlier, estimating the three different latent variables is a challenging task. Thus, in the next chapter, we present a series of simulation studies in order to evaluate recovery of parameters, and also analyses of two sets of empirical data to illustrate how to implement and how to use the DCMixRM.

Chapter 3

SIMULATION STUDY

In Chapter 2, a review was presented of the following models: the LCM, Rasch Model, MixRM, and LCDM. We also presented the development of the DCMixRM as a combination of these models. The primary purposes of the DCMixRM are to extend the latent space in order to better account for the associations among item responses and to support inferences about the secondary dimension(s) that may have caused the population heterogeneity. In the latter regard, this was done by inclusion of a set of latent covariates, specifically, mastery profiles. Although this model has the potential for disclosing richer information about examinees' learning processes than is available using the usual MixIRT models, it is first necessary to determine that the estimation process yields accurate and stable parameter estimates. This is because, with complicated models, even when parameters are statistically identifiable and substantively sensible, it does not imply that the model parameters can be reliably or accurately estimated. Hence, the simulation study described below was conducted in order to evaluate the behavior of the model under varying practical testing conditions. In addition, two empirical data analyses were conducted on reading comprehension and mathematics tests to illustrate how the model can be implemented and used in practice.

3.1 Research Design

To be realistic and informative simulation studies depend on the representativeness of the conditions modeled. However, it is challenging to construct conditions that include all factors in a simulation study because the real world is too complicated to represent. It is essential to balance practicality with fidelity to the real world one is attempting to represent (Bandalos, 2006).

In the current simulation study, two test conditions were considered: one test condition included three attributes with 30 items (A3I30), and the other had four attributes with 20 items (A4I20). The first simulation had fewer attributes but more items and the latter had more attributes but fewer items. The first condition was chosen based upon the characteristics of the tests for State accountability programs; the second condition was more similar to tests in international testing programs. In both simulations, the population was assumed to be composed of two latent classes. Throughout the simulation study, the following questions were considered in evaluating the parameter estimation for the DCMixRM:

1. How well does the estimation converge to yield estimates with acceptable standard errors?
2. How well are the item parameters of the model recovered?
3. How accurately are the examinee parameters estimated?
4. How well are the structural parameters of the model recovered?

As is the case with both the LCM and the LCDM, mis-specification of either or both the number of latent classes or the Q-matrix could be important estimation issues. In this study, however, we assumed that the "correct" number of latent classes and the "correct" specification of the Q-matrix were realized, and therefore it was assumed that we fit the "correct" model in terms of the number of latent classes and the specification of the Q-matrix.

3.1.1 Design of the Simulation Study

For each test condition (A3I30 or A4I20), the factors manipulated as independent variables in the current simulation study were (1) three levels of sample size (2,000, 5,000, and 10,000 examinees), (2) two levels of strength of association between class membership and mastery profile (moderate and strong relation), and (3) two levels of mean ability differences between latent classes (equal and unequal ability means). As all three factors were crossed, the simulation study had 12 conditions for the A3I30 and the A4I20 conditions, and each condition had 100 replications.

Sample Size. In this study, examinees were cross-classified by latent class membership and mastery profile. For the A3I30 condition, this resulted in a 2-by-8 cross-table (16 cells) and for the A4I20 condition, a 2-by-16 cross-table (32 cells). In order to have enough examinees for every cell, relatively large sample sizes were chosen: 2,000, 5,000, and 10,000.

Strength of Relationship of Latent Class and Mastery Profile. Two levels of the strength of association between latent class and mastery profile were manipulated: moderate and strong association. Since the relationship between latent class and mastery profile was of primary interest in the structural component of this model, the impact of the strength of association on the relevant parameter estimates was considered important. Cramér's V was used to determine the size of the strength because Cramer's V measures the strength of association between two polytomous categorical variables in contingency tables regardless of table size. It is defined as

$$V = \sqrt{\frac{\chi^2}{N(q-1)}} \tag{3.1}$$

where N is the sample size and q is the number of columns or rows, whichever is smaller (Cramér, 1946, p. 443). Cramér's V is bound by 0 and 1; the closer V is to 0, the smaller the association between the categorical variables; the closer V is to 1, the stronger the association between variables. Following convention, the strength of the relationship is interpreted as follows: values between $0 \sim .30$ indicate a weak association; values between $.31 \sim .60$ indicate a moderate association; and values $> .60$ indicate a strong association.

In the A3I30 condition, the relationship was manipulated as follows: those who belonged to Class 1 tend to lack Attribute 3; those who belonged to Class 2 tended to have mastered Attribute 3. For the A4I20 condition, those who had mastered both Attributes 3 and 4 tended to belong to Class 2, and those who had not mastered either Attribute 3 or 4 tended to belong to Class 1. The resulting proportions of examinees within each cell of each condition are shown in Table 3.1.

Item Difficulty across Latent Classes. Table 3.2 represents item difficulties for each condition. In the A30I30 condition, item difficulties for Class 1 were sampled from a uniform distribution between -3.0 and 2.8. By subtracting 1 from difficulties of 10 items (33%) that involved Attribute 3, we generated item difficulties for the second latent class. Difficulties of the remaining 20 items were invariant between latent classes. In the A4I20 condition, item difficulties for Class 1 were sampled from a uniform -2.7 to 2.7, and then difficulties of 8 items (40%) that required both Attributes 3 and 4 were altered for Class 2. These conditions resulted that item difficulty means differed by .33 and .40 between two latent classes in the A3I30 and A4I20 conditions, respectively. These differences can be considered effect sizes of the latent classes. According to Cohen's guideline, d=.2, .5, and .8 are small, medium, and large effects for two groups. Following this guideline, we considered that the item mean differences are small for both conditions.

Q-matrix and Effect Size of Attributes. Table 3.3 displays the Q-matrix for this study. It had low complexity in that there was an average of 1.5 and 1.6 attributes per item under the A3I30 and A4I20 conditions, respectively. The main effects of all attributes were set to .25 and the interaction effects of two attributes were set to .05. In other words, as examinees mastered more attributes, the probability of getting a correct response increases, and as examinees mastered two attributes, the probability of getting a right answer increases above the two main effects.

Ability of Latent Classes. Ability parameters were randomly sampled from normal distributions: for the equal ability means condition, N(0,1) was used to generate ability parameter; for the different ability means condition, N(0,1) and N(1,1) were used to generate ability parameter for the first class and the second class, respectively. In this way, the two simulated latent classes differed in mean ability, but their variances were the same.

Table 3.1: True Values of Examinees Proportions in Latent Classes and Mastery Profiles

		Moderate[a]		Strong[b]		
	Profile	C1	C2	C1	C2	Marginal
A3I30	1 (000)	73	27	88	12	14.2
	2 (001)	27	73	12	88	10.5
	3 (010)	73	27	88	12	10.5
	4 (011)	27	73	12	88	11.6
	5 (100)	73	27	88	12	10.5
	6 (101)	27	73	12	88	11.6
	7 (110)	73	27	88	12	11.6
	8 (111)	27	73	12	88	19.2
	Marginal	49	51	48	52	100.0
		Moderate[c]		Strong[d]		
	Profile	C1	C2	C1	C2	Marginal
A4I20	1 (0000)	73	27	88	12	0.9
	2 (0001)	73	27	88	12	4.0
	3 (0010)	73	27	88	12	4.0
	4 (0011)	27	73	12	88	17.9
	5 (0100)	73	27	88	12	4.0
	6 (0101)	73	27	88	12	6.6
	7 (0110)	73	27	88	12	6.6
	8 (0111)	27	73	12	88	10.9
	9 (1000)	73	27	88	12	4.0
	10 (1001)	73	27	88	12	6.6
	11 (1010)	73	27	88	12	6.6
	12 (1011)	27	73	12	88	10.9
	13 (1100)	73	27	88	12	6.6
	14 (1101)	73	27	88	12	4.0
	15 (1110)	73	27	88	12	4.0
	16 (1111)	27	73	12	88	2.4
	Marginal	54	46	56	44	100.0

[a] Cramér's V=.461; [b] Cramér's V=.761; [c] Cramér's V=.458; [d] Cramér's V=.758.

Table 3.2: Item Difficulty Parameters under Each Condition

A3I30[a]									A4I20[b]					
Item	C1	C2	Item	C1	C2	Item	C1	C2	Item	C1	C2	Item	C1	C2
1	-3.0	-3.0	11	**-1.0**	**-2.0**	21	1.0	1.0	1	-2.7	-2.7	11	.0	.0
2	-2.8	-2.8	12	**-.8**	**-1.8**	22	1.2	1.2	2	-2.4	-2.4	12	.3	.3
3	-2.6	-2.6	13	**-.6**	**-1.6**	23	1.4	1.4	3	-2.1	-2.1	13	**.6**	**-.4**
4	-2.4	-2.4	14	**-.4**	**-1.4**	24	1.6	1.6	4	-1.8	-1.8	14	**.9**	**-.1**
5	-2.2	-2.2	15	**-.2**	**-1.2**	25	1.8	1.8	5	-1.5	-1.5	15	1.2	1.2
6	-2.0	-2.0	16	.0	.0	26	**2.0**	**1.0**	6	-1.2	-1.2	16	1.5	1.5
7	-1.8	-1.8	17	.2	.2	27	**2.2**	**1.2**	7	**-.9**	**-1.9**	17	**1.8**	**.8**
8	-1.6	-1.6	18	.4	.4	28	**2.4**	**1.4**	8	**-.6**	**-1.6**	18	**2.1**	**1.1**
9	-1.4	-1.4	19	.6	.6	29	**2.6**	**1.6**	9	-.3	-.3	19	**2.4**	**1.4**
10	-1.2	-1.2	20	.8	.8	30	**2.8**	**1.8**	10	.0	.0	20	**2.7**	**1.7**

[a] 3-attribute with 30-item condition
[b] 4-attribute with 20-item condition

Item Characteristic Curves. Figure 3.1 illustrates item characteristics of Items 15 and 26 for the A3I30 condition. For Item 15, four curves were required to describe the probability of a getting correct answer for two latent classes because it required only one attribute. For each class, we needed two curves: one for those who had mastered the required attribute and another for those who had not. For Item 26, however, six curves were required because, as the Q-matrix indicates, the item required two attributes. So, three curves were needed to describe the probability patterns for each class: a curve for those who had mastered both of attributes, a curve for those who had mastered either of the attributes, and a curve for those who had mastered none of the attributes.

In general, all the curves had similar shapes. There was no intersection among curves. Since it was based on the Rasch model, the probability of a correct response increases as ability increases. For Item 15, the probabilities of Class 2 were always higher than those of

Table 3.3: The Q-matrices under the A3I30 and A4I20 Conditions

A3I30												A4I20									
Item	α_1	α_2	α_3	Item	α_1	α_2	α_3	Item	α_1	α_2	α_3	Item	α_1	α_2	α_3	α_4	Item	α_1	α_2	α_3	α_4
1	1	0	0	11	0	0	1	21	1	0	1	1	1	0	0	0	11	1	0	1	0
2	1	0	0	12	0	0	1	22	1	0	1	2	1	0	0	0	12	1	0	1	0
3	1	0	0	13	0	0	1	23	1	0	1	3	0	1	0	0	13	1	0	0	1
4	1	0	0	14	0	0	1	24	1	0	1	4	0	1	0	0	14	1	0	0	1
5	1	0	0	15	0	0	1	25	1	0	1	5	0	0	1	0	15	0	1	1	0
6	0	1	0	16	1	1	0	26	0	1	1	6	0	0	1	0	16	0	1	1	0
7	0	1	0	17	1	1	0	27	0	1	1	7	0	0	0	1	17	0	1	0	1
8	0	1	0	18	1	1	0	28	0	1	1	8	0	0	0	1	18	0	1	0	1
9	0	1	0	19	1	1	0	29	0	1	1	9	1	1	0	0	19	0	0	1	1
10	0	1	0	20	1	1	0	30	0	1	1	10	1	1	0	0	20	0	0	1	1

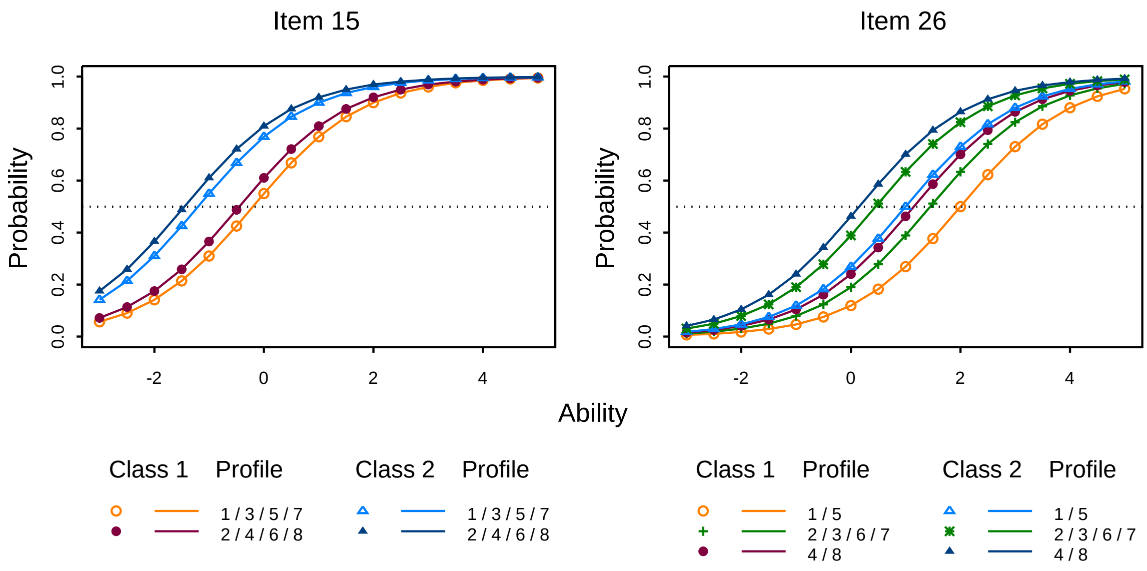


Figure 3.1: Item Characteristic Curves of Two Items under the A3I30 Condition

Class 1. Such patterns were similar for Item 26, but the probability of a correct answer for those who had mastered both attributes and belonged to Class 1 was very close to that for Class 2 for those who had not mastered any of them.

Data Generation. First, distributions of mastery profiles were generated and then joint distributions with latent classes were generated. Using these distributions, each simulee was assigned to a mastery profile and a latent class. Given class membership, a value of ability for each simulee was drawn from a normal distribution using the distributions previously described. Finally, item responses for each condition were simulated using examinees' parameters and conditional probability function. Codes written in SAS software, version 9.1 of the SAS system for Windows, were used to generate data sets.

3.2 Model Specifications

As discussed in Chapter 2, each latent structural model has its own way to identify the model which includes placing a series of constraints. In order to implement the DCMixRM via M*plus*, several constraints were required. First, factor loadings for all items on θ were set equal to 1. This is how to specify a Rasch model component in a DCMixRM. Further, to anchor the metrics between latent classes, the mean and variance of the first latent class were set to be zero and one, respectively. The mean of the second class was set free to be estimated for the unequal means conditions, but for the equal mean conditions, it was fixed to zero as for the first class. This choice was purely arbitrary and determined for convenience.

Second, characteristics of mastery profiles were established by setting constraints on item thresholds for each mastery profile. The constraints were done to ensure monotonicity. That is, all the main effects of the attributes were set to be positive, $\lambda_k \geq 0$, and the interaction effects were constrained to be greater than the negative values of the smaller of the main effects (i.e., $\lambda_{ku} \geq -\lambda_k$ and $\lambda_{ku} \geq -\lambda_u$). These constraints are the same as those required by the LCDM.

Finally, the structural component of the DCMixRM employs a multinomial logit model where mastery profiles relate to latent classes. Multinomial logit models require reference groups. We chose the second latent class and all mastery profiles as reference groups, or baselines. This choice was made purely arbitrarily.

The M*plus* Version 5.21 (Muthén & Muthén, 1998-2007) was used to estimate the model with these constraints. The choice of M*plus* was made for two main reasons. First, with M*plus* a marginal maximum likelihood estimator (MLE) is employed with various algorithms for searching for parameter estimates to maximize the likelihood. The MLE has desired properties as an estimator such as consistency, efficiency, and asymptotic normality. These are very useful and important properties for statistical inference. In particular, for finite mixture models which the DCMixRM belongs to, by imposing proper constraints, the MLE yields consistent estimates (McLachlan & Peel, 2000). McLachlan and Peel also argued that "the lack of identifiability is not of concern in the normal course of events in the fitting of mixture models by maximum likelihood, say, via the EM algorithm" (p. 27). Further, in addition to the Expectation Maximization (EM) algorithm, several algorithms are utilized to optimize MLE in M*plus*, including Quasi-Newton, Fisher scoring, and Newton-Raphson. Also, for continuous latent variables, numerical integration is carried out with or without adaptive quadrature in combination with rectangular integration, Gauss-Hermite integration, or Monte Carlo integration. As a result of this flexibility, computation time can be substantially reduced compared to other approaches such as Markov Chain Monte Carlo estimation (MCMC).

Second, one of the main barriers limiting the use of DCMs in practice is that accessible computer programs for the models are not readily available (de la Torre, 2009; Templin et al., in press). To this end, Templin et al. have demonstrated that it is possible to fit several DCMs via M*plus* by imposing a series of constraints since DCMs can be viewed as restricted LCMs. In this study, we extend this implementation to the DCMixRM.

3.3 Evaluation Criteria

The main question of the simulation study was whether the estimation procedure yielded reliable parameter estimates for the DCMixRM. In answering this question, three criteria were employed: convergence rate, recovery of parameters, and accuracy of classification. We describe each criterion in this section.

Convergence Rate. In order to make inferences on parameters from the statistical analyses, it is necessary that the model converges properly. This is particulary the case with the MLE because the MLE is an estimator that estimates the values of the parameters that maximize the likelihood of the observed data via an iterative process. When the algorithm is unable to arrive at values which meet prescribed criteria, the resulting estimates cannot be trusted for making inferences.

In M*plus*, there are four criteria used to monitor convergence through the observed-data log likelihood: the change in both the absolute and relative log likelihood, the change in any class count, the observed-data log likelihood derivation criterion, and the capability of computing the standard errors (information matrix positive definite) (Muthén & Muthén, 1998-2004, p. 35). In this study, we calculated convergence rates by ratios of the number of converged data sets to the number of generated data sets.

Recovery Evaluation. The recovery of item difficulties (b's), attribute effects (λ's), coefficients of mastery profile effect on class membership (β's) and ability (θ's) were of interest. The bias, relative bias, root mean square error (RMSE) of each parameter estimate, and Pearson correlation between parameter and estimate values were employed to evaluate the quality of parameter estimation for those parameters as each gives slightly different information about the quality of the estimation. By definition, the Pearson correlation reflects the degree of linear relationship between the parameters and estimates. The bias provides information on the amount of difference between generating values and realized estimates. Relative bias gives the same information but after taking into account the size of the parameter. Values of zero for bias and relative bias measures indicate the estimation results in unbiased parameter estimates; the signs on these indices provide information about under- or over-estimation, respectively. Additionally, since the RMSE is a function of both the bias and variance of estimates, it can help inform about the efficiency of the estimates; the smaller, the better. Each is defined as follows

$$\text{Bias } (\widehat{\Omega}) = E(\widehat{\Omega}) - \Omega \tag{3.2}$$

$$\text{Relative Bias } (\widehat{\Omega}) = \frac{E(\widehat{\Omega}) - \Omega}{\Omega} \tag{3.3}$$

$$\text{RMSE } (\widehat{\Omega}) = \sqrt{E(\widehat{\Omega} - \Omega)^2}, \tag{3.4}$$

where $E(\bullet)$ indicates the expected values, and Ω and $\widehat{\Omega}$ denote parameters and estimates, respectively.

In calculating the bias of item difficulties, different scales between the generating values and the realized estimates might be problematic. To take these scale differences into account, parameter estimates from each replication were placed onto the parameter scale through the mean and sigma equating method which is both simple and preferred (Kolen & Brennan, 2004). The equating was carried out as follows

$$b_j^* = b_j - (\bar{b}_j - \bar{b}_p), \tag{3.5}$$

where b_j^* denotes the equated difficulty for the simulated item, b_j is the difficulty estimate of that item, and $\bar{b}_j$ and $\bar{b}_p$ denote the means of the item difficulty estimates for the simulated data set and parameters, respectively.

Accuracy of Classification. The DCMixRM includes two discrete latent variables, latent class membership and mastery profile. To assess degree of classification accuracy with respect to these variables, we investigated group and individual levels of classification accuracy. In terms of group level, we calculated marginal and joint proportions; for the individual level, we computed hit rates and κ measures. The κ is an indicator of the degree of agreement between two categorical variables against that which might be expected by chance. A value of 1 implies perfect agreement, and a value of 0 implies no agreement. It is defined as

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+j}}{1 - \sum \pi_{i+}\pi_{+j}}, \tag{3.6}$$

where π_{ii} is the observed agreement, and $\pi_{i+}\pi_{+j}$ is the chance agreement (Cohen, 1960).

3.4 Results

The results of this study are presented in three parts: in the first part, convergence rates are presented for both the A3I30 and A4I20 conditions. In the remaining two parts, the results for the A3I30 and the A4I20 conditions are presented separately. For each condition, the quality of estimation of the model is discussed in terms of item, examinee, and relationship parameter estimates. The results presented here are only from replications that converged. We use three acronyms to refer to each condition in this section: N indicates the sample

Table 3.4: Convergence Rate across Different Conditions

	A3I30				A4I20			
Sample	DM	DT	SM	ST	DM	DT	SM	ST
N02	29	32	81	84	24	34	53	47
N05	29	42	79	81	32	33	53	52
N10	34	38	83	90	48	46	61	59

Note. In the second row, the first letter refers ability mean condition and the second letter refers strength of relationship between latent class and master profile. For instance, DT refers a condition where two classes have different ability means and have strong relationship with mastery profile. Also, in first column N02 indicates the sample size 2000.

size; D/S indicates whether two classes have different (D) or same (S) means in ability; M/T indicates whether the association between latent class and mastery profile is moderate (M) or strong (S). For example, N02DM indicates the condition that has 2000 examinees, different or unequal ability means across latent classes, and moderate relationship between latent class and mastery profile.

3.4.1 Convergence Rate

Table 3.4 and Figure 3.2 summarize the results for convergence rates across the various conditions. In the figure, results for the A3I30 condition are presented in the left panel and those under the A4I20 condition are presented in the right panel. Overall, similar patterns for convergence rates were observed under both conditions. First, the effect of the mean condition on the convergence rate was noticeable. The equal mean conditions, in which the two latent classes had the same means in ability, showed better convergence rates than the unequal mean conditions. Second, in the equal mean ability condition, the larger the sample,

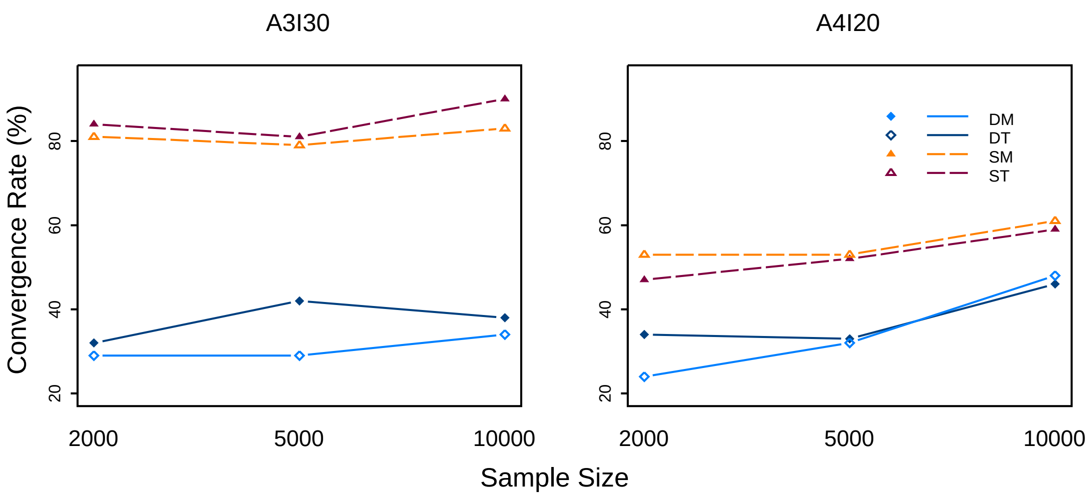


Figure 3.2: Convergence Rate across Different Conditions

the better the convergence rate. Also, the stronger the relationship between mastery profile and latent class, the better the convergence rate.

As clearly shown in Figure 3.2, convergence for the A3I30 conditions was much better than for the A4I20 conditions. The A3I30 condition achieved about 80% convergence rate, even for the sample size of 2,000. As expected, the highest convergence rate occurred for the equal means × strong relationship condition with the largest sample size (e.g. STN10). In contrast, under the A4I20 conditions, even for a sample size of 10,000 examinees and a strong relationship between latent class and mastery profile, a number of replications failed to converge.

The A4I20 condition appeared to be sensitive to sample size, but in general, the equality of mean ability and the degree of relationship had more impact than sample size on the convergence rate, and hence these two factors were of primary interest in investigating quality of recovery of parameter estimates. For this reason, in next sections we report results for some cases, after we collapsed the sample size conditions.

Table 3.5: Correlation between Item Parameters and Estimates (A3I30)

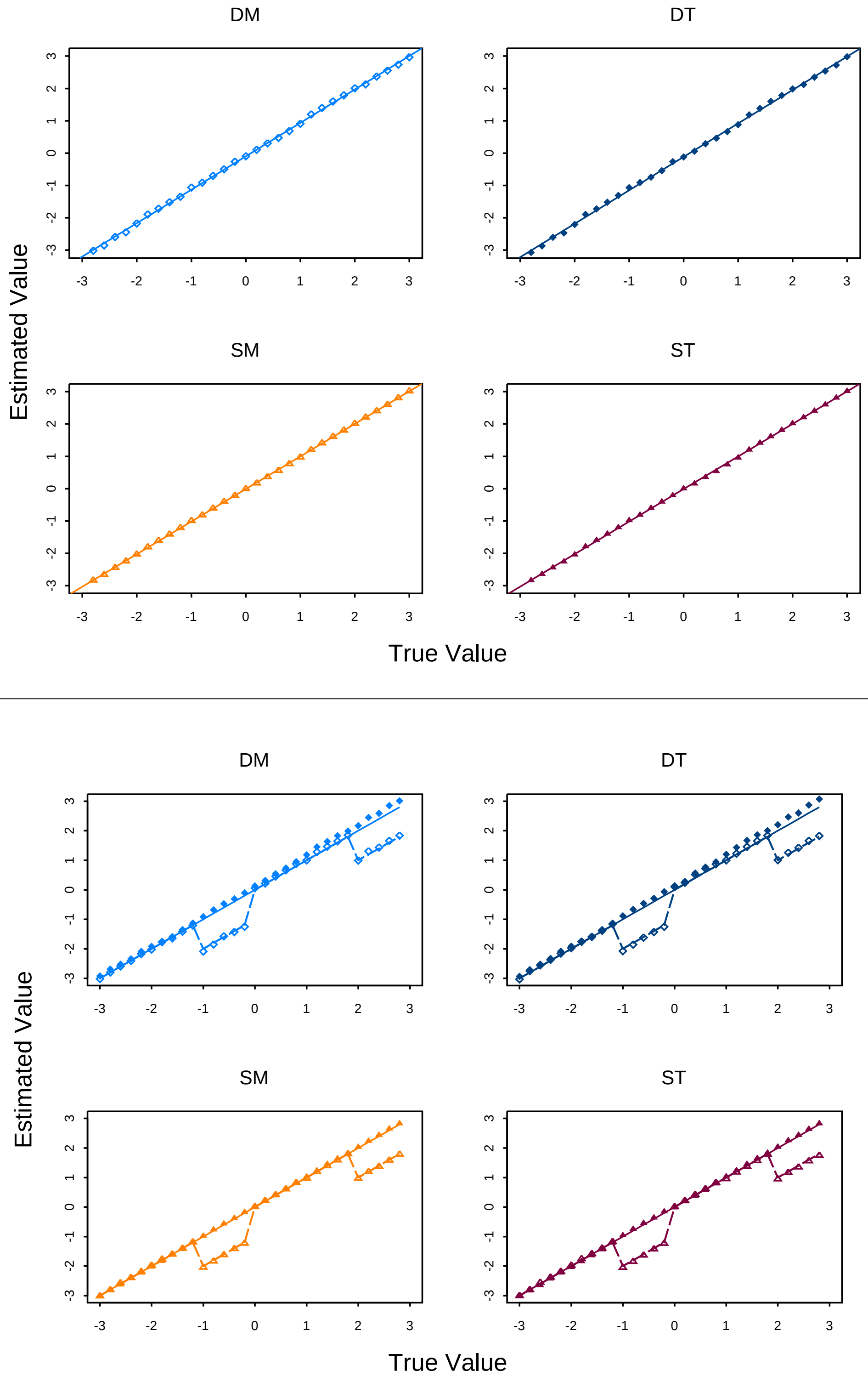
	DM			DT			SM			ST		
	N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
Difficulty	.999	.999	1.000	.998	.999	.999	1.000	1.000	1.000	1.000	1.000	1.000
Attribute	.879	.894	.929	.863	.961	.929	.993	.989	.992	.991	.990	.989

3.4.2 Results under the 3-Attribute with 30-Item Condition (A3I30)

Recovery in Item Parameter. The recovery of item parameters was evaluated by correlation, bias, relative bias, and RMSE statistics. The summary of results across conditions is presented in Tables 3.5 and 3.6 and Figures 3.3 to 3.5.

First, as shown in Table 3.5, the correlations suggest the item difficulty and attribute effect parameters were well recovered across all conditions. The correlations between attribute effect parameters and estimates were slightly lower than those for item difficulties. This was particulary the case for the small sample size in the unequal mean conditions (i.e., DTN02).

Plots of the true and estimated values are presented in Figure 3.3 to help describe the recovery accuracy of the item difficulty parameters with sample size conditions collapsed. In the figure, the solid lines indicate perfect recovery, and the dots indicate the estimated values. The plots in the upper panel are the overall results under all conditions; those in the lower panel show results by latent class: the straight lines represent item difficulties for Class 1 and the crooked lines for Class 2. The dots appear to be well aligned with the perfect recovery line, suggesting that the true values were successfully recovered and in addition, the relationship between the true and estimated values was likely linear. However, estimates deviated from the true values for the very difficult items for Class 1 under the unequal mean conditions. It should be noted that label switching is an important concern for mixture



Note. The lines drawn are the lines Y(Estimated)=X(True).

Figure 3.3: Comparison of True and Estimated Values of Item Difficulty (A3I30)

Table 3.6: Bias and RMSE of Item Parameter Estimates (A3I30)

	Bias				Relative Bias				RMSE			
	DM	DT	SM	ST	DM	DT	SM	ST	DM	DT	SM	ST
Item Difficulty												
N02	-.005	-.008	-.006	-.009	.053	.045	.000	-.007	.298	.296	.131	.125
N05	-.006	-.009	-.006	-.009	.037	.024	-.003	-.010	.226	.218	.095	.089
N10	-.007	-.010	-.006	-.009	.015	.017	-.003	-.012	.175	.179	.083	.072
Average	-.006	-.009	-.006	-.009	.035	.028	-.002	-.010	.233	.231	.103	.095
Attribute Effect												
N02	.150	.145	.018	.017	.197	.202	.044	.057	.407	.407	.096	.088
N05	.107	.097	.014	.012	.153	.163	.031	.043	.327	.309	.092	.074
N10	.061	.057	.013	.010	.074	.125	.026	.041	.232	.238	.097	.068
Average	.106	.100	.015	.013	.175	.182	.038	.050	.322	.318	.095	.077

distribution modeling (Chung, 2003), and yet for this particular model, this result indicates that label switching did not take place during estimation.

Next, Figure 3.4 and Table 3.6 summarize overall bias, relative bias and RMSE results. Figures 3.5 and 3.6 display recovery at the item level. As indicated in Equations 3.3 and 3.4, relative bias takes into account magnitude of parameter values in calculating bias of an estimator, and RMSE assesses the quality of an estimator in terms of its variation and biasedness, or the efficiency and bias of the estimator. Each statistic informs quality of estimates from a slightly different perspective. Figure 3.4 clearly suggests that item difficulty parameters were well recovered even though the estimates were slightly negatively biased; bias values ranged from -.005 to -.010. The relative bias and RMSEs for item difficulty were smaller in the equal mean conditions, and they also decreased as sample size increased. Figure 3.5 shows bias and RMSEs estimates at each difficulty level. From this figure, we observed the salient phenomenon that the directions of bias were changed at the point of item difficulty

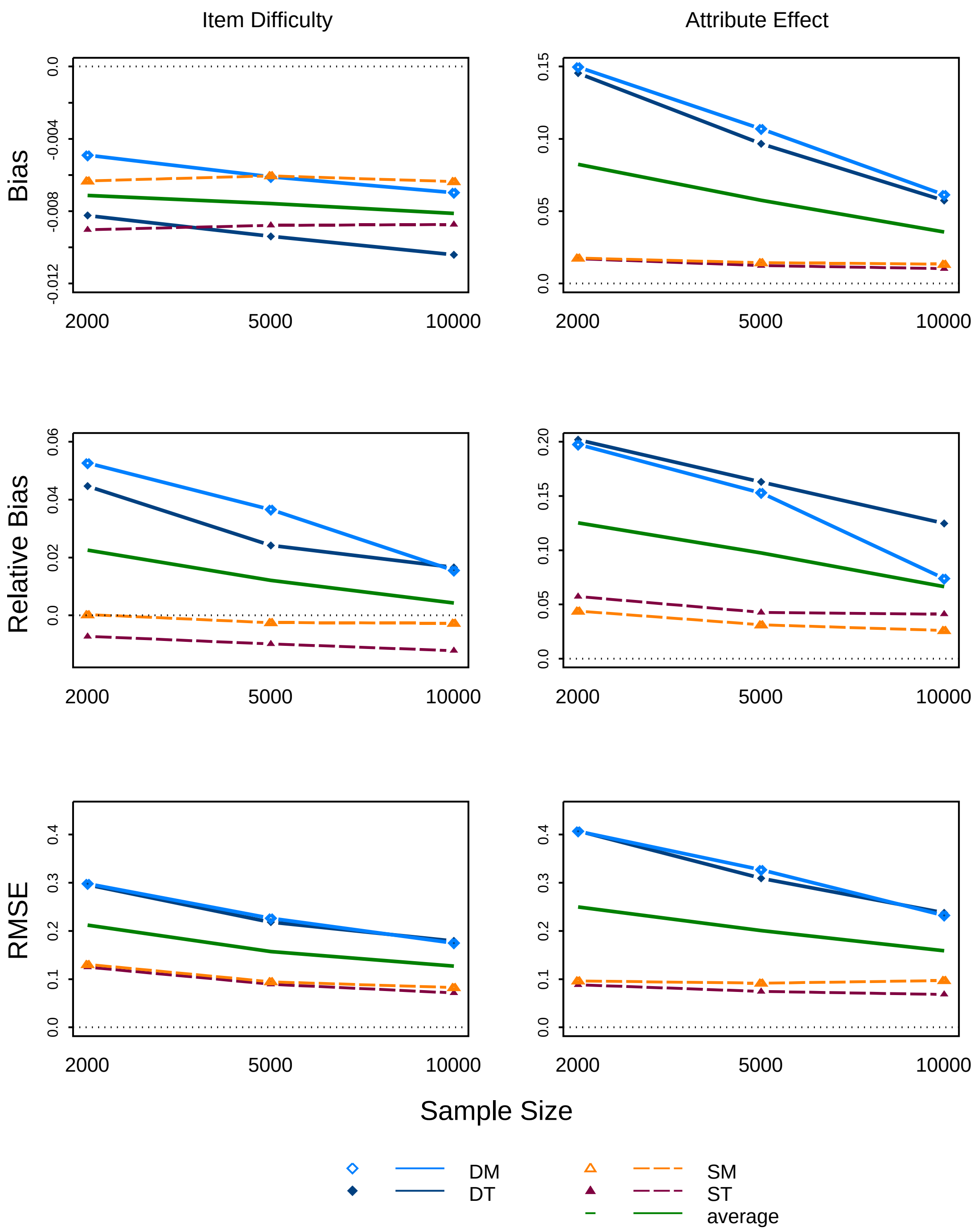


Figure 3.4: Overall Bias and RMSE of Item Parameter Estimates (A3I30)

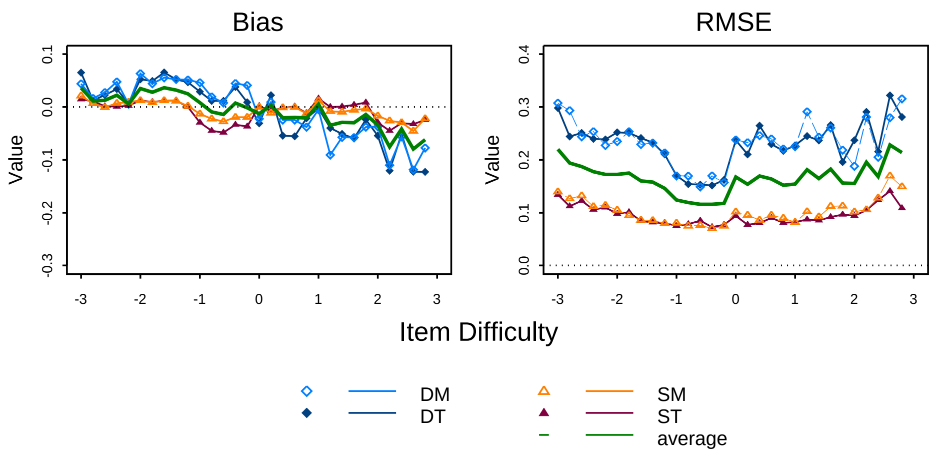
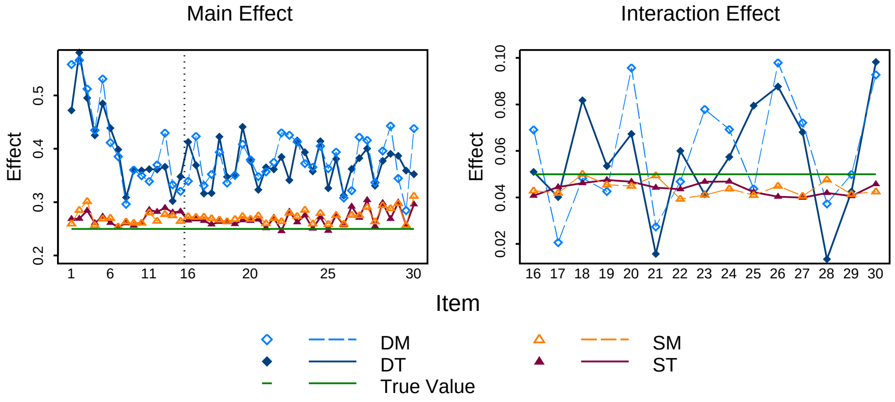


Figure 3.5: Bias and RMSE of Item Difficulty of Each Item (A3I30)



Note. The solid lines refer to the true values.

Figure 3.6: Bias in Attribute Effect Estimates of Each Item (A3I30)

of zero. That is, for easy items, difficulties were overestimated; for difficult items, difficulties were underestimated. Figure 3.4 does not show any impact of mean condition on overall bias in item difficulty parameter estimates, but Figure 3.5 reveals that under the equal mean conditions, item difficulty was better recovered, and at the the point of item difficulty of zero the estimators yielded the least bias. These results are consistent with previous research on item difficulty.

On the other hand, the attribute effect estimates were positively biased, ranging from .012 to .150. This indicates that the attribute effect parameters were overestimated (see Table 3.6). For instance, in the unequal means condition of DTN02, the relative bias approached .20 (i.e., 20%). The results show a pattern in bias that for the equal mean conditions, attribute effect estimates were less biased than under the unequal mean conditions, and further, for the unequal mean conditions, as sample size increased, bias decreased. Similar results were also observed with RMSEs and relative bias. Results at the item level, however, revealed that the estimates themselves deviated from their true or generating values. This appeared to be more pronounced in the unequal mean conditions and can be seen clearly in Figure 3.6. This was particulary the case with items that involved only one attribute such as Items 1 to 15.

Recovery of Examinee Parameters. The quality of the estimation of the examinee parameters was evaluated with correlations between ability parameters and estimates, the hit rate and Cohen's κ for classification of latent class membership and of mastery profile. Table 3.7 shows correlations for ability. Under the unequal mean conditions, the correlations were higher than those under the equal mean ability conditions. This pattern is the opposite of what was observed for item parameter estimates. Figure 3.7 shows the recovery of ability parameters for the 2,000 examinees sample size (i.e., the smallest sample size condition). The four plots in the upper panel show the relationship between ability parameters and estimates; the four plots in the lower panel show the residuals. Although the relationship between the parameters and estimates appears to be linear in the upper panel, the residuals

Table 3.7: Correlation between Ability Parameters and Estimates (A3I30)

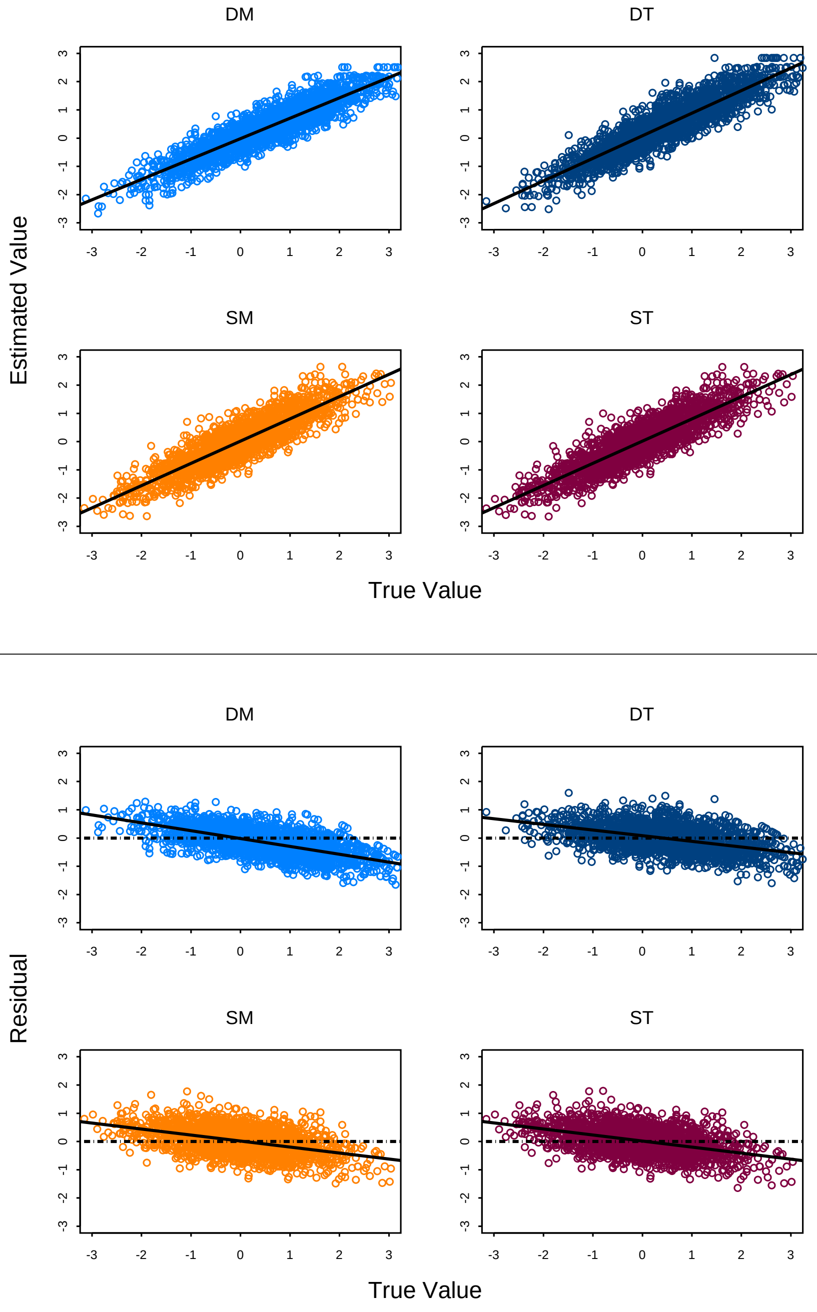
	DM			DT			SM			ST		
	N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
θ	.901	.906	.906	.902	.904	.905	.883	.883	.882	.881	.880	.880

in the lower panel suggest that the ability estimates were overestimated for the lower ability group and underestimated for the higher ability group. This can be understood as the scale shrinkage phenomenon in IRT. See (Lord, 1975) for further discussion.

The hit rate and Cohen's κ for the class membership and mastery attribute are shown in Tables 3.8 and 3.9, respectively. For class membership, the larger the sample size, the better the hit rate. Under the unequal mean conditions, the hit rates were higher. Further, under the strong relationship conditions (i.e., the strong association between latent class and mastery profile), the highest hit rate of .806 was observed. Also, κ measures suggest latent class membership was well detected.

In terms of mastery attributes, however, the hit rates ranged from .547 to .647, suggesting the model did not perform well in identifying the mastery attributes. Consistently, the resulting κ's suggest that the observed agreement may be due to chance (see Table 3.9). Note that in the current study, the marginal probability of mastering each attribute was used for classifying the mastery attribute of each individual. This differs from classification based on posterior probability and in this way, we were able to take into account the uncertainty in classification.

Recovery in Structural Component Parameter. One of the main purposes of the current study is to use mastery profiles to help explain latent class membership. This is accomplished in the form of the relationship parameters in the structural component of the



Note. The sample size is 2000 and the solid line is the fit line.
At the lower panel, the dot line is y=0.

Figure 3.7: Scatter Plot of Ability Parameter (θ) Estimates (A3I30)

Table 3.8: Hit Rate of Membership in Latent Classes (A3I30)

	DM			DT			SM			ST		
	N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
Hit Rate	.784	.793	.795	.796	.802	.806	.704	.710	.712	.716	.721	.723
κ	.568	.585	.590	.592	.603	.610	.407	.419	.423	.431	.441	.446

Table 3.9: Hit Rate of Classification in Attribute Mastery States (A3I30)

		DM			DT			SM			ST		
		N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
Hit Rate	α_1	.548	.549	.551	.550	.550	.550	.556	.557	.555	.554	.555	.554
	α_2	.548	.547	.549	.549	.551	.550	.558	.561	.559	.560	.561	.561
	α_3	.584	.588	.595	.618	.623	.631	.605	.606	.606	.642	.646	.647
κ	α_1	.091	.094	.097	.095	.096	.097	.111	.113	.108	.107	.109	.107
	α_2	.093	.093	.095	.094	.100	.097	.115	.121	.116	.119	.121	.121
	α_3	.168	.177	.189	.236	.247	.263	.212	.213	.212	.285	.293	.294

Table 3.10: Recovery of Ability (θ) Distribution Parameter Estimates (A3I30)

Parameter[a]		Mean				Standard Deviation			
		DM	DT	SM	ST	DM	DT	SM	ST
C1	N(0,1)	-.201	-.200	-.209	-.209	.746	.748	.854	.853
C2	N(1,1)	1.179	1.203	.202	.203	.745	.747	.881	.885

[a] Parameters include the mean and variance of ability for each class. For the same mean conditions, the means and variances are the same across two latent classes.

model. The structural component of the model also includes other important features of the population such as distributions of ability, marginal proportions of the latent classes and mastery profiles, and relationships between attributes. The results for these components are given in Tables 3.10, 3.11, and 3.12. First, Table 3.10 contains means and standard deviations for ability of each class. As can be seen in this table, under all conditions, estimated mean abilities approached the theoretically expected values. This suggests the ability distribution of the population was reasonably well captured. The estimated standard deviations of ability, however, were smaller than the true values. This was particularly the case for the unequal mean conditions. The reduction in standard deviations of ability is an indicator of scale shrinkage (as discussed earlier in this section of the recovery of examinee parameter estimation).

Second, the resulting marginal proportions of latent class and mastery profile in the DCMixRM are reported in Table 3.11. The estimated marginal proportions for each mastery profile were close to the true distributions and the discrepancies appeared to be negligible: the calculated marginal proportions differed from the true distribution by less than 6%.

Third, the results of the tetrachoric correlations between attributes are given in Table 3.12. The correlations for the population were around .170. Under all conditions, correlations appeared to be underestimated even though under the equal mean conditions, the correlation

Table 3.11: Recovery of Marginal Proportion of Latent Classes and Mastery Profiles (A3I30)

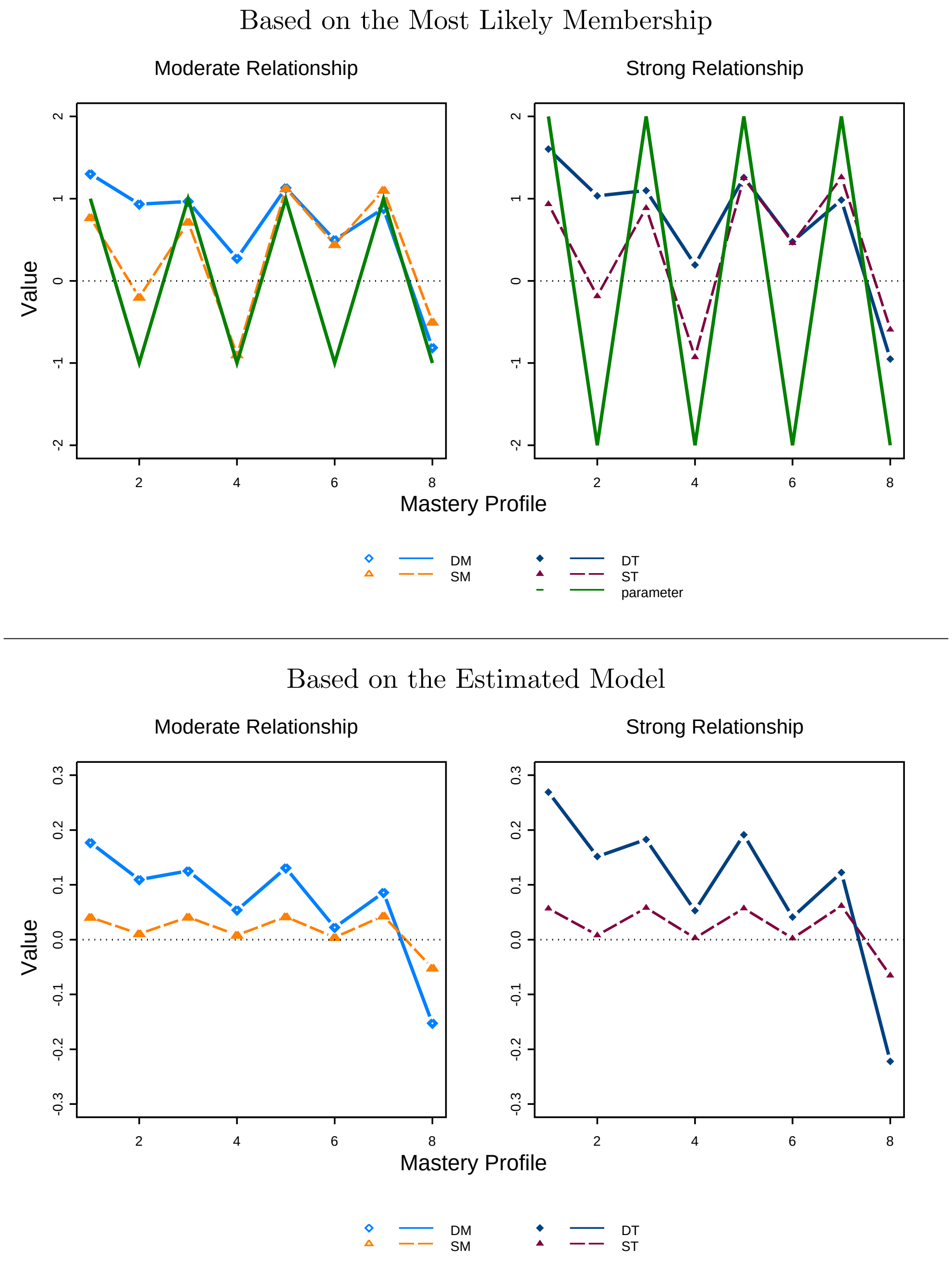
Profile	True[a]	DM	DT	SM	ST
1 (000)	14.23	12.82	12.42	13.92	13.45
2 (001)	10.54	10.83	11.42	11.06	11.16
3 (010)	10.54	11.07	11.27	12.39	12.12
4 (011)	11.65	11.34	10.95	9.90	10.33
5 (100)	10.54	11.15	10.29	8.93	9.03
6 (101)	11.65	12.96	12.85	14.80	14.70
7 (110)	11.65	14.31	14.81	15.56	15.71
8 (111)	19.21	15.51	16.00	13.45	13.50
Class 1	48.00	47.84	47.04	48.38	48.16
Class 2	52.00	52.16	52.96	51.62	51.84

[a] True values of marginal proportion of each profile and latent class.

Table 3.12: Recovery of Tetrachoric Correlation between Attributes (A3I30)

	True[a]		DM		DT		SM		ST	
	α_2	α_3	α_2	α_3	α_2	α_3	α_2	α_3	α_2	α_3
α_1	.170	.169	.081	.078	.122	.086	.126	.039	.131	.010
α_2		.169		-.061		-.053		-.176		-.176

[a] True values of correlation.


Based on the Most Likely Membership

Based on the Estimated Model

Note. The solid line without any marks refers to the true values. Owing to the scale alteration, the solid line for the true value is dropped at the lower panel.

Figure 3.8: Comparison of True and Estimated Values of Relationship Parameters between Latent Class and Mastery Profile (A3I30)

Table 3.13: Conditional Proportion of the Latent Class (A3I30)

	Moderate Relationship						Strong Relationship					
	True[a]		DM		SM		True[a]		DT		ST	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
1 (000)	73.11	26.89	60.78	39.22	56.40	43.60	88.08	11.92	64.47	35.53	58.31	41.69
2 (001)	26.89	73.11	52.77	47.23	33.36	66.64	11.92	88.08	52.58	47.42	31.79	68.21
3 (010)	73.11	26.89	53.23	46.77	55.16	44.84	88.08	11.92	53.97	46.03	57.06	42.94
4 (011)	26.89	73.11	37.82	62.18	20.15	79.85	11.92	88.08	33.04	66.96	18.33	81.67
5 (100)	73.11	26.89	56.54	43.46	64.59	35.41	88.08	11.92	57.10	42.90	65.47	34.53
6 (101)	26.89	73.11	42.45	57.55	48.49	51.51	11.92	88.08	39.18	60.82	46.81	53.19
7 (110)	73.11	26.89	51.23	48.77	64.32	35.68	88.08	11.92	51.71	48.29	65.94	34.06
8 (111)	26.89	73.11	32.25	67.75	37.65	62.35	11.92	88.08	29.47	70.53	35.62	64.38

[a] True values for each condition.

between α_1 and α_2 seemed to be well recovered. This was particularly the case for the correlation between α_2 and α_3.

In addition, Tables 3.13 and 3.14, and Figure 3.8 display the recovery in the relationship parameters. M*plus* provides two kinds of class probability estimates, or mixing proportion estimates: posterior probability-based and estimated model-based mixing proportions. These may be consistent with each other, but they also can yield quite different results.

The class proportions conditioning on mastery profiles are given in Table 3.13. In spite of the poor recovery of the relationship parameters, under the equal mean conditions, the conditional distribution patterns were correctly scratched. However, their relationships were underestimated as presented in In Table 3.14. The upper part of the table gives classification results based on the posterior probability of class membership, and the lower part gives the model-based classification results. This order is the same for Figure 3.8. According to these results, it appears that one cannot recover relationship parameters using the model-based

Table 3.14: Recovery of Relationship Parameters between Latent Class and Mastery Profile (A3I30)

	Based on the Posterior Probability									
	Moderate Relationship					Strong Relationship				
Profile	True	DM		SM		True	DT		ST	
1 (000)	1	1.300	(0.579)	.767	(0.293)	2	1.603	(0.660)	.932	(0.266)
2 (001)	-1	.932	(0.957)	-.200	(0.315)	-2	1.035	(0.824)	-.186	(0.314)
3 (010)	1	.965	(0.736)	.713	(0.288)	2	1.098	(0.847)	.884	(0.259)
4 (011)	-1	.274	(1.002)	-.900	(0.409)	-2	.194	(0.930)	-.928	(0.346)
5 (100)	1	1.130	(1.020)	1.119	(0.329)	2	1.256	(0.943)	1.250	(0.292)
6 (101)	-1	.498	(0.875)	.440	(0.259)	-2	.477	(0.676)	.460	(0.212)
7 (110)	1	.881	(1.116)	1.100	(0.279)	2	.984	(0.834)	1.258	(0.255)
8 (111)	-1	-.813	(0.624)	-.503	(0.188)	-2	-.951	(0.528)	-.592	(0.183)
	Based on the Estimated Model									
	Moderate Relationship					Strong Relationship				
Profile	True	DM		SM		True	DT		ST	
1 (000)	1	.177	(.110)	.040	(.042)	2	.269	(.120)	.057	(.044)
2 (001)	-1	.109	(.217)	.011	(.044)	-2	.152	(.182)	.008	(.036)
3 (010)	1	.125	(.156)	.041	(.041)	2	.183	(.153)	.058	(.037)
4 (011)	-1	.054	(.260)	.008	(.046)	-2	.053	(.225)	.003	(.028)
5 (100)	1	.131	(.223)	.041	(.045)	2	.191	(.194)	.057	(.040)
6 (101)	-1	.022	(.206)	.004	(.036)	-2	.041	(.171)	.002	(.026)
7 (110)	1	.086	(.262)	.043	(.034)	2	.122	(.189)	.062	(.031)
8 (111)	-1	-.152	(.176)	-.052	(.054)	-2	-.222	(.154)	-.066	(.048)

Note. The values in parentheses are standard deviations.

Table 3.15: Correlation between Item Parameters and Estimates under the A4I20

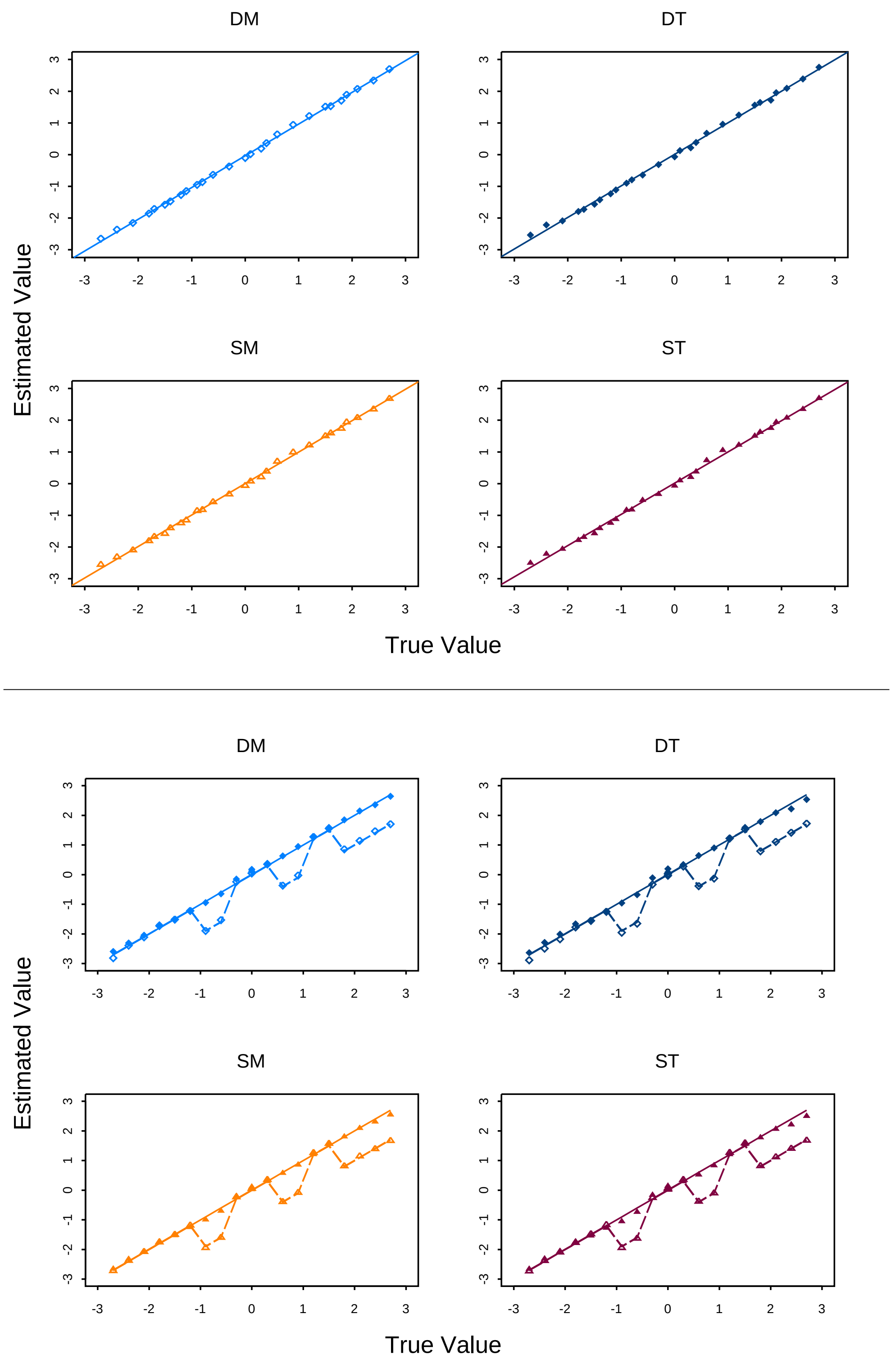
	DM			DT			SM			ST		
	N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
Difficulty	.999	.999	.999	.998	.998	.999	.999	.999	.999	.999	.999	.999
Attribute	.903	.942	.977	.898	.942	.959	.959	.947	.944	.946	.927	.930

mixing proportions, but it might be possible to obtain a general pattern of relationship using the posterior probability-based estimates. Examining results using posterior probability of membership suggests that the general pattern of relationship between latent class and mastery profile was fairly reproduced, but they were underestimated. In addition, the resulting estimates were similar in patterns and values although for the moderate relationship conditions, the parameters were better recovered. For the equal mean conditions, the estimation yielded smaller standard deviation and better recovery than for the unequal mean condition.

3.4.3 Results under the 4-Attribute with 20-Item Condition (A4I20)

Recovery in Item Parameter. The results for item parameter recovery are presented in Tables 3.15 and 3.16 and Figures 3.9 to 3.11. As in the A3I30 condition, the item parameters were recovered well with the DCMixRM. To be specific, as can be seen in Table 3.15, under the A4I20 condition, correlations between attributes were better recovered than those under the A3I30 condition, even for the unequal mean and small sample size conditions.

Plots of the true and estimated values are presented in Figure 3.9 to help describe the recovery accuracy of the item difficulty parameters with sample size conditions collapsed. In the figure, the solid lines indicate perfect recovery, and the dots indicate the estimated values. The plots in the upper panel are the overall results under all conditions; those in the lower panel show results by each class: the straight lines represent item difficulties for



Note. The lines drawn are the lines Y(Estimated)=X(True).

Figure 3.9: Comparison of True and Estimated Values of Item Difficulty (A4I20)

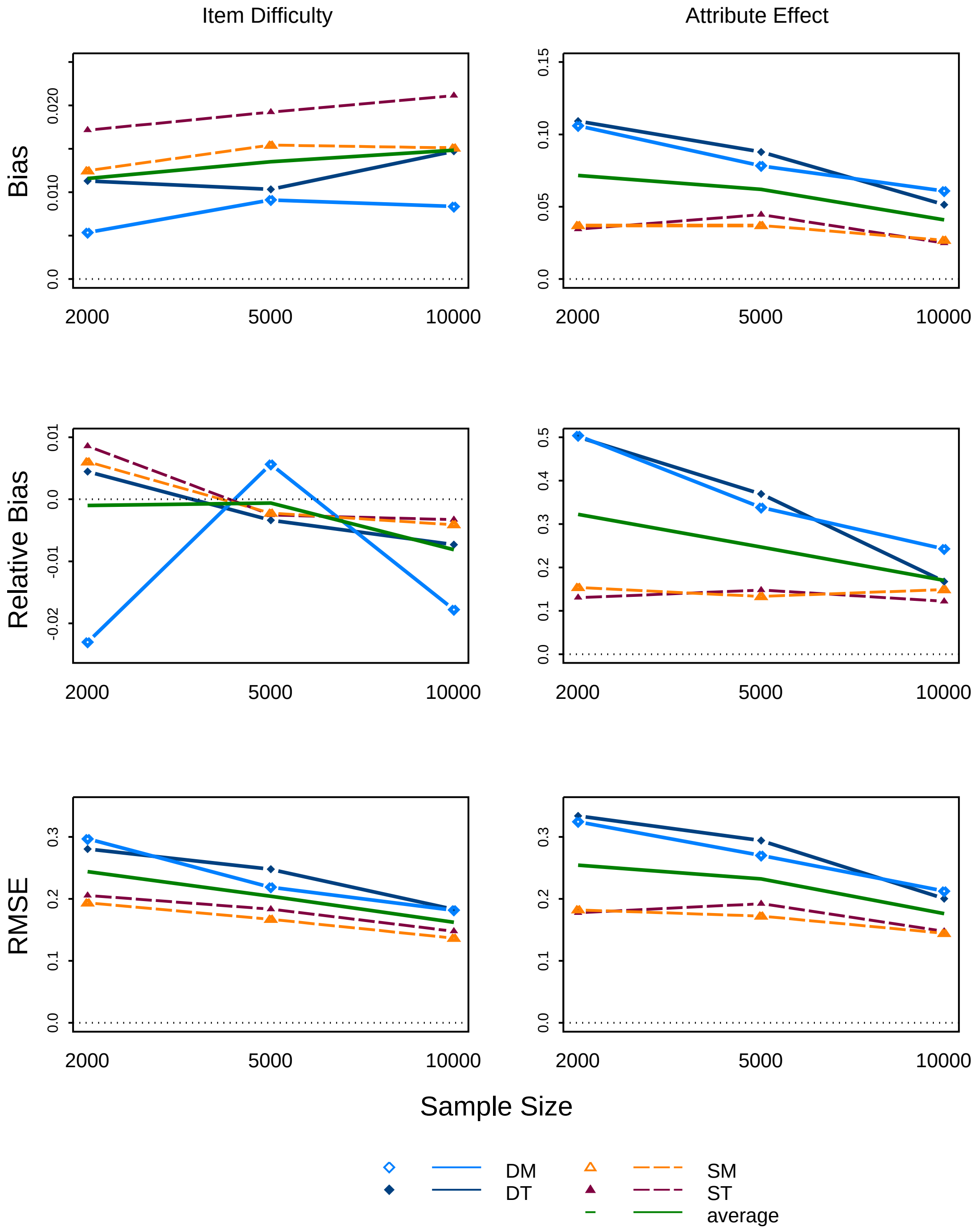


Figure 3.10: Overall Bias and RMSE of Item Parameter Estimates (A4I20)

Table 3.16: Bias and RMSE of Item Parameter Estimates (A4I20)

	Bias				Relative Bias				RMSE			
	DM	DT	SM	ST	DM	DT	SM	ST	DM	DT	SM	ST
Item Difficulty												
N02	.005	.011	.012	.017	-.023	.004	.006	.009	.297	.280	.194	.206
N05	.009	.010	.015	.019	.006	-.003	-.002	-.003	.219	.248	.167	.184
N10	.008	.015	.015	.021	-.018	-.007	-.004	-.003	.182	.183	.137	.148
Average	.008	.012	.014	.019	-.012	-.002	.000	.001	.232	.237	.166	.179
Attribute Effect												
N02	.106	.109	.037	.034	.504	.501	.154	.131	.324	.334	.182	.178
N05	.078	.088	.037	.045	.338	.369	.133	.148	.270	.294	.172	.192
N10	.061	.051	.027	.025	.243	.168	.149	.122	.213	.200	.144	.148
Average	.082	.083	.034	.035	.362	.346	.145	.134	.269	.276	.166	.173

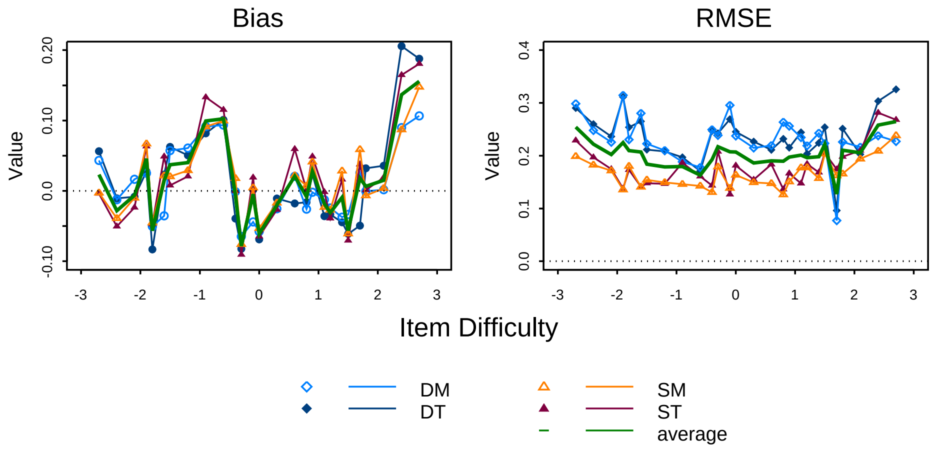
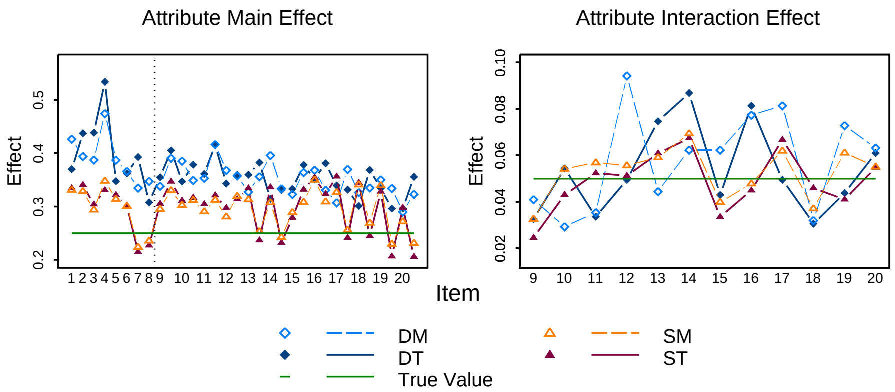


Figure 3.11: Bias and RMSE of Item Difficulty for Each Item (A4I20)



Note. The solid lines refer to the true values.

Figure 3.12: Bias in Attribute Effect Estimates of Each Item (A4I20)

Class 1 and the crooked lines those for Class 2. All the plots in this figure indicate that item difficulty parameters were well recovered because the dots appear to agree with the solid lines. In addition, as indicated by the plots in the lower panel, the results here imply that labeling switching on the latent classes did not occur (see discussion in the section on the A3I30 condition).

The recovery was also evaluated by examining bias, relative bias, and RMSE. As can be seen in Table 3.16, under all conditions, the overall bias in item difficulty estimates were slightly positive, and yet the magnitudes of the bias were small enough to be essentially negligible; the largest value was .021 under the strong relationship with equal mean conditions. Figure 3.11 further shows bias and RMSEs at each level of item difficulty. Unlike the A3I30 condition where a systematic pattern in bias was observed, no systematic pattern was exhibited except that the estimator yielded the most bias in extremely difficulty items.

For the attribute effect parameter, the estimation process seemed to have some difficulty in recovering the true values, particularly for the unequal mean conditions (see Figure 3.12). Attribute effects were substantial positively biased; this was more evident in the relative bias indices (see Table 3.16). The bias decreased, however, as sample size increased or ability means were equal across latent classes. To illustrate this, under the DWN02 condition, the magnitude of bias was .109, but under the STN10 condition, it was .025 which was much smaller. This pattern was also observed for the RMSEs for both item difficulty and attribute effect parameter estimates.

Recovery in Examinee Parameter. Tables 3.17 to 3.19 and Figure 3.13 contain results about the correlations between ability parameters and estimates and about the hit rates with respect to latent class membership and mastery profile. It can be seen that under the unequal mean conditions, the correlations were slightly higher than those under the equal mean conditions, and yet all these correlations were lower than those under the A3I30 condition (see Tables 3.17 and 3.7 for comparison)

Second, Figure 3.13 illustrates how well the ability parameters were recovered with the sample size of 2,000 (i.e., the smallest sample size conditions). As explained in the section on the A3I30 condition, the four plots in the upper panel display linear relationships between parameters and estimates and four in the lower panel display the residuals. The patterns were similar to those with the A3I30 condition: the relationship between the ability parameters and estimates appears to be linear in plots at the upper panel, and yet the residuals in the lower panel suggest that the ability estimates for the lower ability group were overestimated while the ability estimates for the higher ability group were underestimated. This is consistent with previous research that discussed scale shrinkage for ability (Lord, 1975). Furthermore, for the unequal mean conditions, the figure suggests that an upper limit, or ceiling effect, seems to appear on the ability estimates.

Finally, Tables 3.18 and 3.19 show the hit rate and Cohen's κ for class membership and attribute mastery status. The same pattern as with the A3I30 conditions can be seen in these tables. The larger the sample size, the better the hit rate for class membership. Also, the unequal mean conditions were better for detecting class membership, with κ indices supporting the finding that under the unequal mean conditions, latent class membership was

Table 3.17: Correlation between Ability Parameters and Estimates (A4I20)

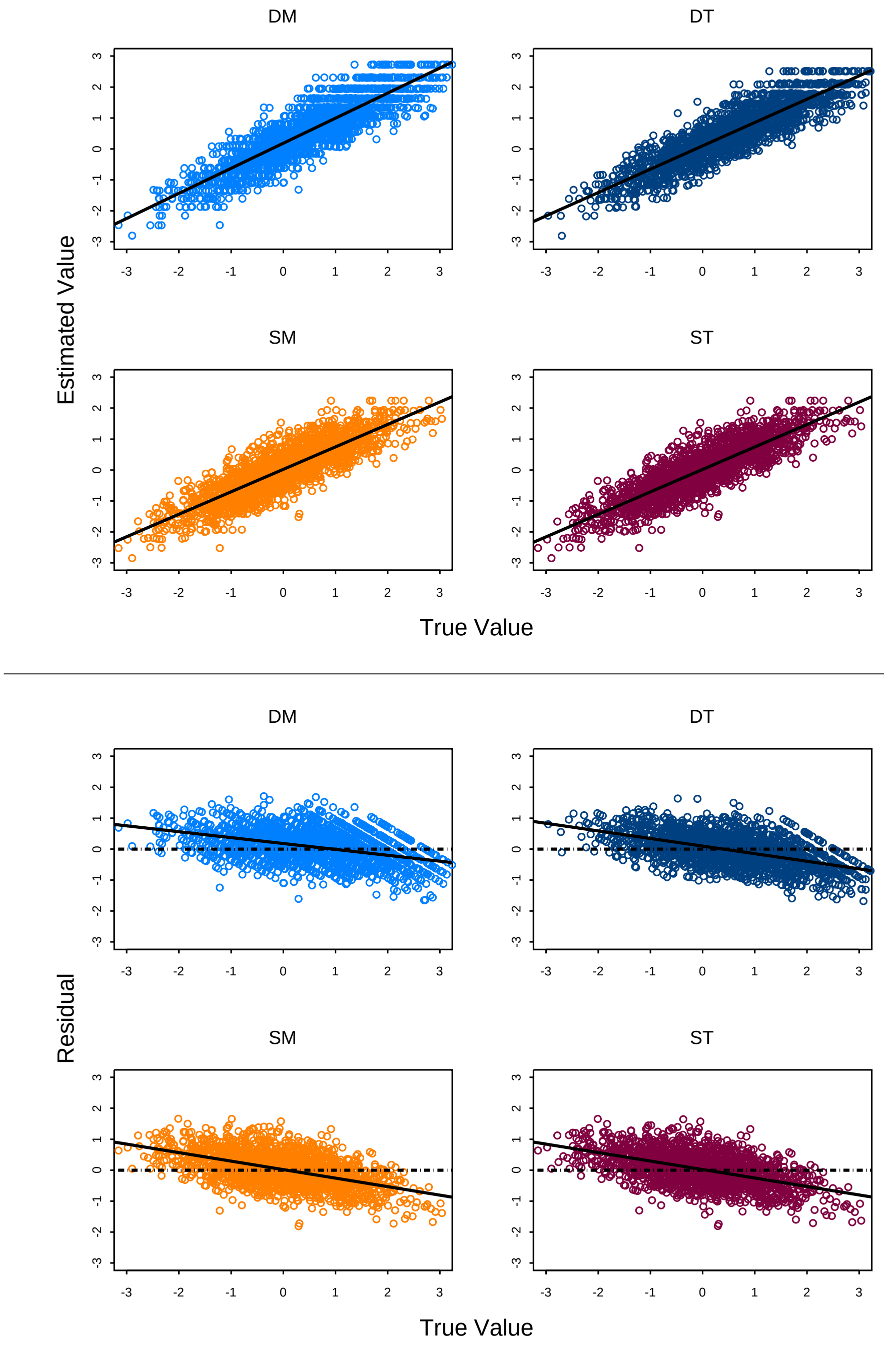
	DM			DT			SM			ST		
	N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
θ	.878	.881	.881	.880	.879	.881	.857	.857	.856	.858	.858	.858

Table 3.18: Hit Rate and Kappa of Class Membership in Latent Class (A4I20)

	DM			DT			SM			ST		
	N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
Hit rate	.761	.763	.766	.755	.760	.764	.652	.660	.664	.643	.653	.659
κ	.519	.523	.528	.504	.513	.520	.303	.317	.326	.283	.300	.313

well detected. In terms of attribute mastery state, however, the hit rates were essentially the same regardless of the conditions, ranging from .514 to .534 (also see κ's in Table 3.19), and were even smaller than those in the A3I30 condition.

Recovery of the Structural Component Parameter. In this section, we present results about the structural component parameters. These include ability distributions, marginal proportions of the latent classes and mastery profiles, and the relationships between mastery profile and latent class membership. The summary results are presented Tables 3.20, 3.21, and 3.22. Across all conditions (see Table 3.20), the estimated ability means approached the theoretically expected values, but the estimated standard deviations were much smaller than the expected values. This suggests that ability means were reasonably captured but underestimated the scale parameters, or standard deviations, resulting in scale shrinkage.



Note. The sample size is 2000, and the solid line is the fit line.
At the lower panel, the dot line is y=0.

Figure 3.13: Scatter Plot of Ability Parameter (θ) Estimates (A4I20)

Table 3.19: Hit Rate and Kappa of Classification in Mastery State of Attributes (A4I20)

		DM			DT			SM			ST		
		N02	N05	N10	N02	N05	N10	N02	N05	N10	N02	N05	N10
Hit	α_1	.519	.523	.530	.523	.522	.533	.527	.526	.527	.523	.524	.526
rate	α_2	.521	.525	.532	.520	.522	.534	.526	.528	.529	.524	.524	.525
	α_4	.517	.520	.525	.515	.514	.523	.527	.526	.531	.525	.526	.532
	α_3	.430	.439	.458	.422	.420	.435	.440	.442	.450	.419	.422	.427
κ	α_1	.047	.053	.068	.056	.053	.074	.056	.056	.058	.050	.052	.055
	α_2	.048	.053	.069	.049	.049	.072	.051	.056	.057	.047	.050	.051
	α_3	.027	.032	.043	.022	.021	.035	.056	.055	.061	.052	.052	.062
	α_3	-.013	-.015	-.021	-.047	-.043	-.068	-.008	-.010	-.012	-.036	-.039	-.044

Table 3.20: Recovery of Ability (θ) Distribution Parameter Estimates (A4I20)

Parameter[a]		Mean				Standard Deviation			
		DM	DT	SM	ST	DM	DT	SM	ST
C1	N(0,1)	-.242	-.273	-.248	-.252	.695	.683	.790	.793
C2	N(1,1)	1.150	1.100	.229	.224	.687	.684	.846	.843

[a] Parameters include the mean and variance of ability for each class. For the same mean conditions, the means and variances are the same across two latent classes.

Table 3.21: Recovery of Marginal Proportion of the Latent Class and Mastery Profile (A4I20)

Profile	True[a]	DM	DT	SM	ST
1 (0000)	0.89	3.96	3.84	5.97	5.75
2 (0001)	4.00	6.20	6.07	5.57	5.85
3 (0010)	4.00	4.91	4.52	5.75	5.48
4 (0011)	17.93	6.12	5.78	5.81	5.83
5 (0100)	4.00	4.84	4.97	5.18	5.17
6 (0101)	6.60	6.30	6.08	6.48	6.57
7 (0110)	6.60	6.09	6.40	5.59	5.49
8 (0111)	10.88	5.15	5.04	5.81	6.03
9 (1000)	4.00	4.75	4.70	4.42	4.46
10 (1001)	6.60	6.82	7.33	7.20	7.43
11 (1010)	6.60	7.21	6.76	6.88	6.46
12 (1011)	10.88	7.25	7.14	7.75	8.02
13 (1100)	6.60	6.45	6.62	5.76	5.62
14 (1101)	4.00	7.41	7.73	8.15	8.06
15 (1110)	4.00	9.30	9.59	6.94	6.91
16 (1111)	2.43	7.25	7.45	6.77	6.89
Class 1	56.01	54.44	56.47	52.15	53.05
Class 2	43.99	45.56	43.53	47.85	46.95

[a] True values of marginal proportion of each profile and latent class.

Table 3.22: Recovery of Tetrachoric Correlation between Attributes (A4I20)

	True[a]			DM			DT			SM			ST		
	α_2	α_3	α_4	α_2	α_3	α_4	α_2	α_3	α_4	α_2	α_3	α_4	α_2	α_3	α_4
α_1	.213	.303	.305	.046	.074	-.001	.040	.060	.001	.026	.048	.084	.017	.043	.100
α_2		.304	.304		-.005	.060		.034	.079		.041	.018		.045	.021
α_3			.142			.123			.139			.113			.125

a True values of correlation.

Table 3.23: Conditional Proportion of the Latent Class (A4I20)

	Moderate Relationship						Strong Relationship					
	True[a]		DM		SM		True[a]		DT		ST	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
1 (0000)	73.11	26.89	66.81	33.19	61.41	38.59	88.08	11.92	64.21	35.79	59.75	40.25
2 (0001)	73.11	26.89	53.78	46.22	37.88	62.12	88.08	11.92	50.67	49.33	38.31	61.69
3 (0010)	73.11	26.89	58.37	41.63	66.99	33.01	88.08	11.92	56.93	43.07	67.49	32.51
4 (0011)	26.89	73.11	45.83	54.17	36.23	63.77	11.92	88.08	48.08	51.92	40.34	59.66
5 (0100)	73.11	26.89	61.16	38.84	63.86	36.14	88.08	11.92	55.90	44.10	61.46	38.54
6 (0101)	73.11	26.89	44.10	55.90	31.33	68.67	88.08	11.92	40.39	59.61	29.87	70.13
7 (0110)	73.11	26.89	54.52	45.48	70.19	29.81	88.08	11.92	52.70	47.30	69.67	30.33
8 (0111)	26.89	73.11	37.40	62.60	34.25	65.75	11.92	88.08	39.24	60.76	35.64	64.36
9 (1000)	73.11	26.89	46.62	53.38	57.14	42.86	88.08	11.92	42.07	57.93	53.65	46.35
10 (1001)	73.11	26.89	45.27	54.73	35.04	64.96	88.08	11.92	39.61	60.39	33.66	66.34
11 (1010)	73.11	26.89	48.79	51.21	64.21	35.79	88.08	11.92	45.62	54.38	63.57	36.43
12 (1011)	26.89	73.11	39.69	60.31	36.07	63.93	11.92	88.08	40.35	59.65	39.29	60.71
13 (1100)	73.11	26.89	46.71	53.29	60.35	39.65	88.08	11.92	44.70	55.30	56.62	43.38
14 (1101)	73.11	26.89	30.32	69.68	29.03	70.97	88.08	11.92	30.08	69.92	26.09	73.91
15 (1110)	73.11	26.89	46.46	53.54	68.65	31.35	88.08	11.92	43.75	56.25	65.99	34.01
16 (1111)	26.89	73.11	22.36	77.64	30.01	69.99	11.92	88.08	22.85	77.15	30.66	69.34

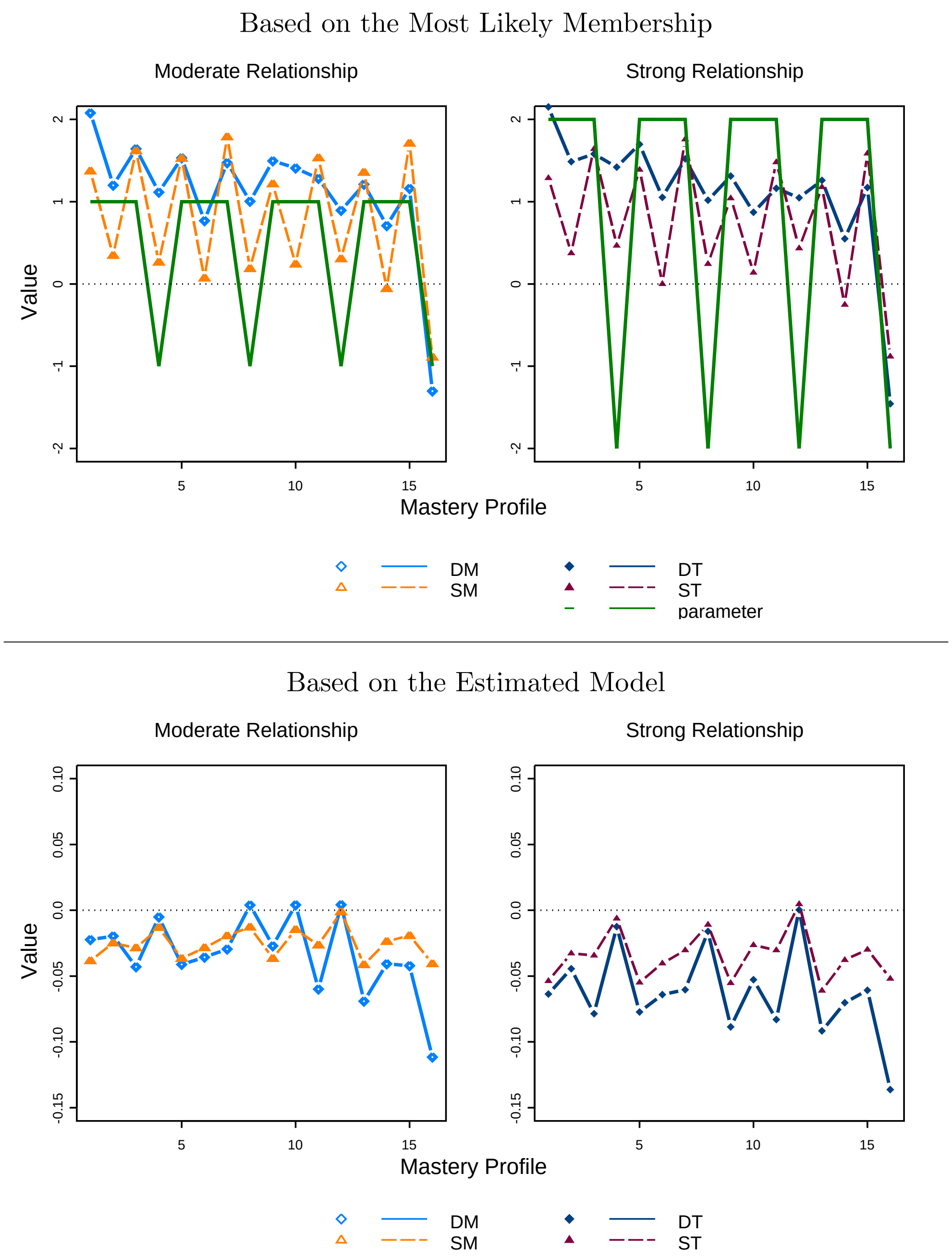
[a] True values for each condition.

Table 3.24: Recovery of Relationship Parameters between Latent Class and Mastery Profile (A4I20)

	Based on the Posterior Probability									
	Moderate Relationship					Strong Relationship				
Profile	True	DM		SM		True	DT		ST	
1 (0000)	1	2.078	(0.757)	1.371	(0.806)	2	2.152	(1.589)	1.288	(0.806)
2 (0001)	1	1.200	(2.014)	.345	(0.739)	2	1.487	(1.626)	.373	(0.859)
3 (0010)	1	1.643	(0.854)	1.618	(0.837)	2	1.579	(0.754)	1.640	(0.811)
4 (0011)	-1	1.115	(0.863)	.262	(0.764)	-2	1.421	(1.632)	.466	(0.701)
5 (0100)	1	1.532	(1.947)	1.526	(0.834)	2	1.698	(1.601)	1.388	(0.882)
6 (0101)	1	.767	(2.006)	.070	(0.672)	2	1.052	(1.490)	-.001	(0.705)
7 (0110)	1	1.473	(0.872)	1.786	(0.668)	2	1.520	(1.528)	1.758	(0.714)
8 (0111)	-1	1.004	(1.758)	.185	(0.585)	-2	1.017	(1.422)	.241	(0.569)
9 (1000)	1	1.496	(1.744)	1.215	(0.841)	2	1.311	(2.511)	1.042	(0.789)
10 (1001)	1	1.409	(1.788)	.239	(0.630)	2	.870	(1.811)	.136	(0.747)
11 (1010)	1	1.282	(1.063)	1.532	(0.787)	2	1.163	(0.923)	1.480	(0.679)
12 (1011)	-1	.893	(0.968)	.303	(0.571)	-2	1.047	(1.692)	.433	(0.557)
13 (1100)	1	1.210	(1.039)	1.356	(0.747)	2	1.258	(1.732)	1.176	(0.744)
14 (1101)	1	.709	(1.855)	-.057	(0.525)	2	.551	(1.382)	-.255	(0.590)
15 (1110)	1	1.160	(1.001)	1.708	(0.627)	2	1.172	(1.486)	1.585	(0.652)
16 (1111)	-1	-1.303	(0.643)	-.890	(0.490)	-2	-1.455	(1.314)	-.882	(0.497)

	Based on the Estimated Model									
	Moderate Relationship					Strong Relationship				
Profile	True	DM		SM		True	DT		ST	
1 (0000)	1	-.022	(.096)	-.038	(.082)	2	-.064	(.087)	-.054	(.077)
2 (0001)	1	-.020	(.133)	-.025	(.069)	2	-.044	(.138)	-.033	(.078)
3 (0010)	1	-.043	(.122)	-.029	(.081)	2	-.079	(.121)	-.034	(.069)
4 (0011)	-1	-.005	(.120)	-.013	(.076)	-2	-.012	(.141)	-.006	(.062)
5 (0100)	1	-.041	(.141)	-.037	(.082)	2	-.077	(.125)	-.055	(.071)
6 (0101)	1	-.036	(.150)	-.029	(.068)	2	-.064	(.130)	-.040	(.063)
7 (0110)	1	-.030	(.135)	-.019	(.063)	2	-.060	(.138)	-.030	(.057)
8 (0111)	-1	.004	(.158)	-.013	(.052)	-2	-.016	(.106)	-.011	(.050)
9 (1000)	1	-.027	(.189)	-.037	(.074)	2	-.089	(.125)	-.055	(.066)
10 (1001)	1	.004	(.146)	-.015	(.069)	2	-.053	(.126)	-.026	(.062)
11 (1010)	1	-.060	(.164)	-.026	(.072)	2	-.083	(.145)	-.030	(.057)
12 (1011)	-1	.004	(.119)	-.001	(.063)	-2	.000	(.139)	.005	(.047)
13 (1100)	1	-.069	(.100)	-.041	(.071)	2	-.092	(.160)	-.061	(.060)
14 (1101)	1	-.041	(.134)	-.024	(.054)	2	-.070	(.118)	-.038	(.055)
15 (1110)	1	-.042	(.125)	-.019	(.048)	2	-.061	(.134)	-.030	(.050)
16 (1111)	-1	-.112	(.081)	-.041	(.058)	-2	-.136	(.096)	-.052	(.064)

Note. The values in parentheses are standard deviations.



Based on the Most Likely Membership

Based on the Estimated Model

Note. The solid line without any marks refers to the true values. Owing to the scale alteration, the solid line for the true value is dropped at the lower panel.

Figure 3.14: Comparison of True and Estimated Values of Relationship Parameters between Latent Class and Mastery Profile (A4I20)

Next, Table 3.21 contains the estimates of the marginal proportions for latent classes and mastery profiles. The estimation process yielded comparable results for the latent classes but did not do as well for the mastery profiles. This is also seen in the tetrachoric correlations between attributes as presented in Table 3.22. Although in general, the tetrachoric correlations between α_3 and α_4 were recovered well, the estimation process failed to capture the relationships between attributes and, in fact, underestimated them.

Finally, Tables 3.24 and 3.23 and Figure 3.14 present results for the recovery of the relationship parameters. The upper parts of Table 3.24 and Figure 3.14 present the classification results based on the posterior probability of class membership; the lower part presents results based on the posited model. Several observations can be made from the results. First, the results show that regardless of the strength of relationship conditions, the estimation process yielded similar sets of parameter estimates. Second, as with the A3I30 condition, using the model-based mixing proportions the relationship parameters were not recovered well; the classification based on the posterior probability provided better estimates for relationship parameters. Third, under the moderate relationship conditions, all relationship parameters were overestimated; however, under the strong relationship conditions, parameter estimates tended toward moderate values; the parameters having high values were underestimated whereas the parameters having low values were overestimated.

The class proportions conditioning on mastery profiles are given in Table 3.23. In spite of the poor recovery of the relationship parameters, under the moderate relationship and equal mean conditions, the patterns of conditional proportions for the 12 profiles were close to truth, and under unequal mean conditions, for nine of 12 mastery profiles the conditional proportions were close to true values. Despite that the general picture of the association was well captured, however, the strengths of association were underestimated.

3.5 Summary and Conclusions

The item and ability parameters of the DCMixRM appeared to be well recovered. Even so, item difficulty tended to be slightly underestimated and attribute effect slightly overestimated. Under the A3I30 condition, at the point of item difficulty of zero, the direction of bias was altered: underestimation occurred at high difficulty items and overestimation occurred at low difficulty items. These phenomena can be referred to as scale shrinkage. This is consistent with the previous research on item difficulty estimation in IRT. Scale shrinkage also occurred with ability parameter estimation. For attribute effect estimates, bias tended to be larger for single-attribute items compared to bias in two-attribute items.

For the structural component of the DCMixRM, the distribution of latent classes was recovered well. More interestingly, the latent classes appeared to be well detected with the use of the mastery profiles. It also appears that the use of the mastery profiles (i.e., the use of the attributes as covariates on the structural parameter) reduced or possibly prevented label switching. As indicated earlier, label switching occurs because the likelihood is not sufficiently informative to the estimation process (McLachlan & Peel, 2000). Yet, the inclusion of covariates, such as the mastery profiles, may be sufficiently informative to become inequality constraints on the parameter space to prevent the label switching (see McLachlan & Peel, 2000). This can also be interpreted in such a way that, as Smit et al. (1999, 2000, 2003) suggest, the inclusion of mastery profiles as covariates may be sufficiently informative to detect latent classes.

Under the A3I30 condition, the relationship between the latent classes and mastery profiles were fairly recovered. As a result, the model appears to be supported for providing information about characteristics of latent classes in the form of mastery profiles. The short test length condition (A4I20) appeared to cause some difficulties in the estimation process in terms of detecting the relationship parameters. For both cases (i.e., A3I30 and A4I20), the classification based on the posterior probability worked better in recovering the association between the latent classes and mastery profiles. It should be noted, however, that the classification based on the posterior probability does not take into account uncertainty in classification. Also, the correlations between attributes appeared to be attenuated. These results may have been due to the small magnitudes of the correlations that were used in generating data sets.

For the unequal mean conditions, the convergence rates were not as good as for the equal mean conditions. To better understand this, it is useful to note that there are at least two types of non-convergence: (1) estimation stops without converging because the maximum number of iterations was reached (hard stopping rule) or (2) estimation failed to converge because it was difficult to optimize the function being fit (soft stopping rule). For the current study, examination of the optimization history of the analyses suggested that all of the non-convergence results occurred due to difficulties in optimizing the parameter space (soft stopping rule). The largest number of iterations that was run was 120 for the data set for the condition of 2,000 examinees with moderate relationship between latent class and profile under the different mean condition (DMN02). This suggests that the non-convergence occurred due to the complexity of the parameter space. It is possible that the MLE used here may have had difficulty in finding maximization points over the relatively complex parameter space of the DCMixRM.

To deal with this complexity, other estimators might be considered to improve the estimation for the DCMixRM; Bayesian estimation via MCMC is one possibility. Limiting the parameter space may also be a useful alternative for improving the optimization process for parameter estimation. For instance, fixing certain parameters that are not of interest or imposing constraints that reflect a conjecture or theoretical framework may help improve the process.

In this study, the magnitudes of attributes and interaction effects were set to be the same across all attributes. Furthermore, small values were chosen for these as parameters. These two factors may have caused problems in recovering attribute effects. One problem faced here is that no testing program yet exists that combines measurement of both a general proficiency and mastery state on attributes. As a result, deciding what the appropriate magnitudes should be for both item difficulties and attribute effects was difficult. In this study, item difficulties were generated from a uniform distribution with mean of zero and a bound of -3 and 3, and ability was generated from a normal distribution with a mean of zero and a variance of 1. Given these restrictions, there was little room to set the magnitudes of attribute effects in formulating probability of getting a correct answer. Since the DCMixRM is a probabilistic model with three sets of parameters, such as θ (ability), b (item difficulty), and λ (attribute effect), the influence on probability of getting a correct answer is bound by relative magnitude to one another. Further study is needed to investigate reasonable values for attribute effects and how these values might affect on parameter recovery and convergence for the DCMixRM.

Chapter 4

ILLUSTRATIVE ANALYSES: TWO EMPIRICAL STUDIES

Two empirical data analyses are presented in this section to illustrate the use of the DCMixRM. The intent of this presentation is to illustrate how to fit the model and how to interpret results for large-scale assessments. Data from a reading test and data from a mathematics test were used for this purpose. The reading test was from an international assessment program, The Progress in International Reading Literacy Study (PIRLS) 2006, and the mathematics test was administered in a Midwestern state as a part of state's accountability program.

The two analyses are presented separately. In each section, the description of the data set is presented first, followed by descriptive statistics and the model selection process used with the DCMixRM. Then the results are provided for estimates of item, examinee, and structural parameters.

4.1 DCMixRM with Reading Comprehension: PIRLS 2006

Overview of The Progress in International Reading Literacy Study. The Progress in International Reading Literacy Study (PIRLS) is an international comparative trend study of reading literacy. It is coordinated by the International Association for the Evaluation of Educational Achievement (IEA) (Mullis, Martin, Kennedy, & Foy, 2007). The assessment is taken by fourth-grade students (primarily nine- and ten-year-old) every five years beginning in 2001. The target grade was chosen because it is an important point at which students begin to transition from learning *how to read* to *how to read to learn.* PIRLS provides information about reading literacy that complements two other international assessment programs.

These are the IEA's Trends in International Mathematics and Science Study (TIMSS), which assesses achievement at fourth and eighth grades, and the Programme for International Student Assessment (PISA), which assesses reading literacy of 15-year-old students.

The Reading Development Group (RDG), the Questionnaire Development Group (QDG), and the National Research Coordinators (NRCs) were involved in developing the PIRLS framework. Reading literacy was defined as

> the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment. (Mullis, Kennedy, Martin, & Sainsbury, 2006, p. 3)

Based on this definition, the three groups focused on three aspects of students' reading literacy in constructing the PIRLS assessment program: purposes for reading, processes of comprehension, and reading behaviors and attitudes. Of these, the reading purposes and comprehension processes provide the foundation for the written assessment of reading literacy. The PIRLS focuses on two purposes for reading: reading for literary experience and reading for acquiring and using information. Four processes were identified for the comprehension process:

- Focus on and retrieve explicitly stated information (Retrieval),
- Make straightforward inferences (Inference),
- Interpret and integrate ideas and information (Integration), and
- Examine and evaluate content, language and textual (Evaluation).

These elements were used as the blueprint for writing items, and hence they were used here to build the Q-matrix which is described in the following section.

Data and Q-matrix. PIRLS 2006 consisted of 10 passages, five literary and five informational. Each passage was accompanied by approximately 12 questions, 126 items in total.

As is done for the National Assessment of Educational Progress (NAEP), PIRLS implements a rotated booklet design. In this design, each student is administered one booklet consisting of two passages. 14,109 examinees who took test items in Booklet 10 in 2006 were sampled. Booklet 10 was chosen because it had the most multiple-choice items and because it included one information-related and one literary-related passage. Table 4.1 gives a summary of distributions of gender and nationality of the examinees in the sample.

PIRLS was developed to measure overall proficiency level in reading comprehension. That is, the test is scored to provide an ability score rather than a diagnostic estimate of examinees' strengths or weaknesses on specific attributes. Hence, a Q-matrix was not developed or used by the test developers. For purposes of this example, therefore, the Q-matrix was constructed using the four reading comprehension processes used for constructing the test. Table 4.2 presents the Q-matrix along with descriptive statistics. As shown in the table, each item requires only a single attribute. This is described as a simple structure Q-matrix. The descriptive statistics of items indicate many students had difficulty in solving Items 3, 4 and 13, as less than half of the students answered these items correctly. Two sample items are given in Table 4.3. Both items require inference but with different contexts. These items were not included in the analysis described in this example, but they do illustrate how items on the PIRLS are constructed.

Model Selection and Descriptive Statistics. Three DCMixRMs were fit to the PIRLS 2006 data. These consisted of DCMixRMs for one- to three-classes. This was done to determine the number of latent classes that best fit the data. Fit indices used were used to guide the selection of the best fitting model. These included AIC, BIC, and entropy of three models. As shown in Table 4.4, although entropy indicated a one-class DCMixRM was the best, AIC and BIC suggested that a two-class DCMixRM was the best fitting model. There is no research on model fit indices appropriate for a DCMixRM, however, previous research on MixIRT models (Li et al., 2009) suggested that BIC and AIC functioned well for this purpose. Based on the AIC and BIC, therefore, we decided to use the two-class DCMixRM

Table 4.1: Demographic Compositions of Examinees in PIRLS 2006

	Frequency	Percent		Frequency	Percent
Gender[a]					
Female	6,883	48.78	Male	7,218	51.16
Country					
Austria	339	2.40	Poland	318	2.25
Bulgaria	255	1.81	Qatar	440	3.12
Taipei	297	2.11	Romania	275	1.95
Denmark	266	1.89	Russia	320	2.27
France	285	2.02	Singapore	424	3.01
Georgia	307	2.18	Slovakia	357	2.53
Germany	525	3.72	Slovenia	352	2.49
Hong Kong	309	2.19	South Africa	875	6.20
Hungary	243	1.72	Spain	259	1.84
Iceland	237	1.68	Sweden	283	2.01
Indonesia	302	2.14	Trinidad & Tobago	267	1.89
Iran	366	2.59	Macedonia	262	1.86
Israel	262	1.86	United States	344	2.44
Italy	241	1.71	England	267	1.89
Kuwait	254	1.80	Scotland	248	1.76
Latvia	284	2.01	Belgium Flemish	309	2.19
Lithuania	313	2.22	Belgium French	299	2.12
Luxembourg	332	2.35	Canada Ontario	265	1.88
Moldova	266	1.89	Canada Quebec	252	1.79
Morocco	213	1.51	Canada Alberta	271	1.92
Netherlands	285	2.02	Canada British Columbia	280	1.98
New Zealand	420	2.98	Canada Nova Scotia	294	2.08
Norway	247	1.75	**Total**	14,109	100.00

[a] missing=8

Table 4.2: Q-matrix along with Mean and SD of Items in PIRLS 2006

Informational passage							Literary passage						
Item	α_1	α_2	α_3	α_4	Mean	SD	Item	α_1	α_2	α_3	α_4	Mean	SD
1	0	0	0	1	.87	.33	8	0	0	0	1	.73	.44
2	0	0	0	1	.50	.50	9	1	0	0	0	.73	.45
3	1	0	0	0	.46	.50	10	1	0	0	0	.80	.40
4	0	1	0	0	.45	.50	11	0	1	0	0	.67	.47
5	0	0	1	0	.76	.43	12	0	0	1	0	.54	.50
6	0	0	1	0	.64	.48	13	0	0	1	0	.62	.49
7	0	0	1	0	.73	.44	Total	3	2	5	3	8.50	3.10

Note. α_1 Retrieval; α_2 Inference; α_3 Integration; and α_4 Evaluation.
Means are proportions of examinees who answered correctly to the item.

Table 4.3: Two Released Items that Requires α_2 (Inference) from PIRLS 2006

Passage	Item
Literary	Why was the clay eventually taken out of the bin? ⓐ All the other lumps of clay were used. * ⓑ It was on top of the other lumps of clay. ⓒ The boy chose that lump because he especially liked it. ⓓ The teacher told the boy to use that lump
Informative	Why does the article tell you that ‘a mug of boiling water thrown in the air would freeze before it hit the ice’? ⓐ to tell you how hot the water is in Antarctica ⓑ to show you what they drink in Antarctica ⓒ to tell you about scientists’ jobs in Antarctica ⓓ to show you how cold it is in Antarctica *

Note. The items were reprinted with permission from Mullis et al. (2006).
* Answer key.

Table 4.4: Goodness-of-Fit Statistics of DCMixRMs for PIRLS 2006

	Class 1	Class 2	Class 3
Number[a]	41	71	101
AIC	200315	**199395**	199467
BIC	200625	**199931**	200230
Entropy	**.56**	.50	.41

[a] Number of parameters that were estimated in the model.

solution. Recall that since four attributes were involved in the test, it was possible to have 16 mastery profiles within each class.

Item Parameter Estimates. Recall that to compare the class-specific item difficulties, it was necessary first to place the item difficulties for each latent class on the same scale. As done with item difficulties in the simulation study in the previous chapter, it was done by the mean and sigma method (see Equation 3.5). Since class 1 was used as baseline, all item difficulties were on the scale of Class 1. Class-specific item difficulty and attribute effect parameter estimates are presented in Table 4.5. The results are also visually displayed in Figure 4.1. Difficulties for Items 1 and 6 were virtually the same across latent classes. Items 7, 9, and 10 were easier for Class 1 whereas Items 2, 3 and 12 were easier for Class 2. Figure 4.2 depicts two sets of item characteristic curves for two items for the same passage. This figure demonstrates that although Items 10 and 12 belonged to the same passage, they functioned differently in each latent class. Specifically, Item 10 was easier for examinees who had not mastered the attribute required for the item, if they were in Class 1. The same examinees tended to have difficulty in solving Item 12. This pattern does not appear to be related to passage types as would be modeled by testlet or method effect models.

Attribute effects for Items 2, 3, 4, and 13 were not significant at the .05 level and was zero for Item 8. Among the items that had significant and relatively large attribute effects,

Table 4.5: Item Difficulty and Attribute Effect of PIRLS 2006

	Item Difficulty		Attribute Effect			
Item	Class 1	Class 2	λ_1	λ_2	λ_3	λ_4
1(0001)	-2.731	-2.867				.173*
2(0001)	1.777	.851				4.280
3(1000)	1.951	.536	3.555			
4(0100)	2.005	1.678		4.341		
5(0010)	-1.807	-1.492			.656*	
6(0010)	-.737	-.848			.946*	
7(0010)	-1.852	-.517			1.162*	
8(0001)	-1.642	-1.901				.000*
9(1000)	-1.820	-.905	.577*			
10(1000)	-2.766	-1.251	.585*			
11(0100)	-1.395	-1.066		.197*		
12(0010)	.081	-.853			.986	
13(0010)	-.467	-.766			1.263	

Note. λ_1 Retrieval; λ_2 Inference; λ_3 Integration; λ_4 Evaluation.
* p <.05 for attribute effects.

Items 5, 6, 7, and 12 required α_3 (integration) and Items 9 and 10 required α_1 (retrieval). These results suggest this test may not be a good one for measuring mastery states of either α_2 (inference) or α_4 (evaluation). For each of these attributes, only two items were included in this test measuring these attributes. It is likely that there was not a sufficient number of items to estimate mastery state on these attributes.

Ability Distribution and Latent Classes. The ability distributions for each class are plotted in Figure 4.3. Although there was an overlap between latent classes, those in Class 1 were higher in reading literacy ability than those in Class 2. The difference in mean ability between classes was 1.35. Approximately 75 % of the examinees belonged to Class 1, and about 25 % in Class 2 (see Table 4.6). Particularly, as comparing with raw score distribution

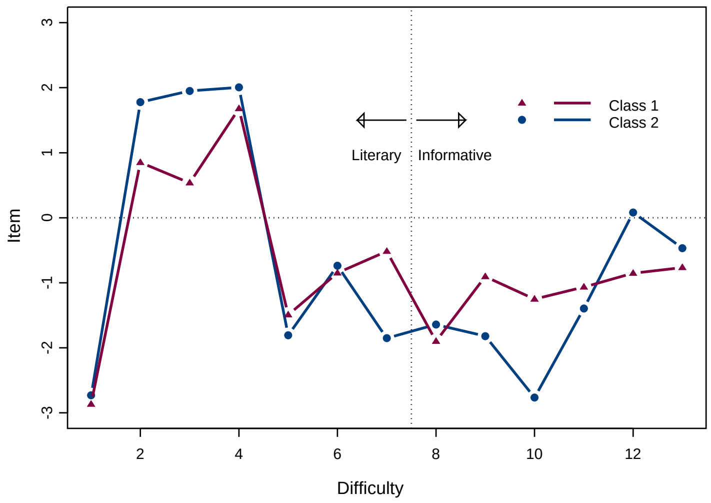


Figure 4.1: Item Difficulty Patterns of the 2-Class DCMixRM for PIRLS 2006

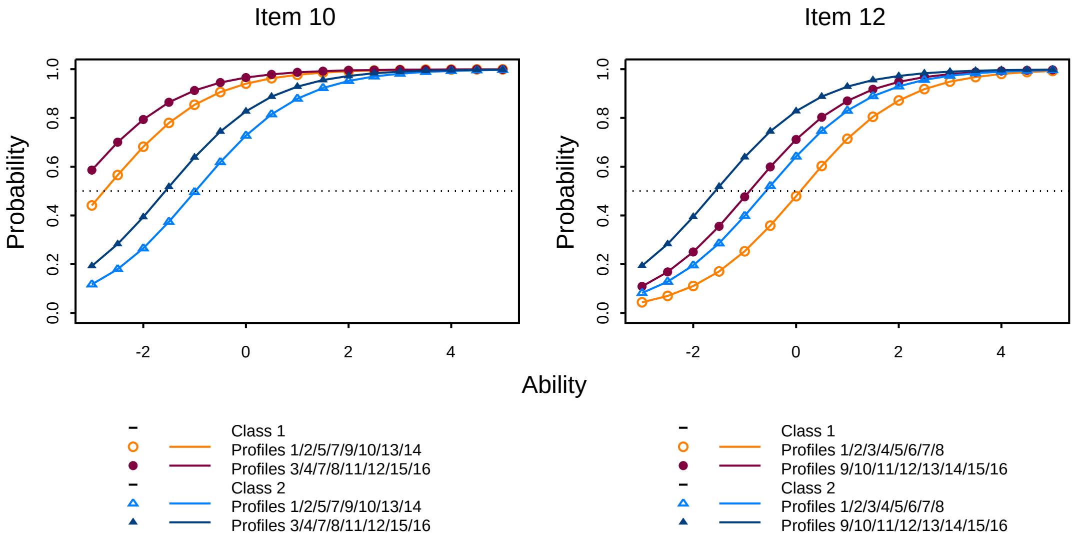


Figure 4.2: Item Characteristic Curves of PIRLS 2006

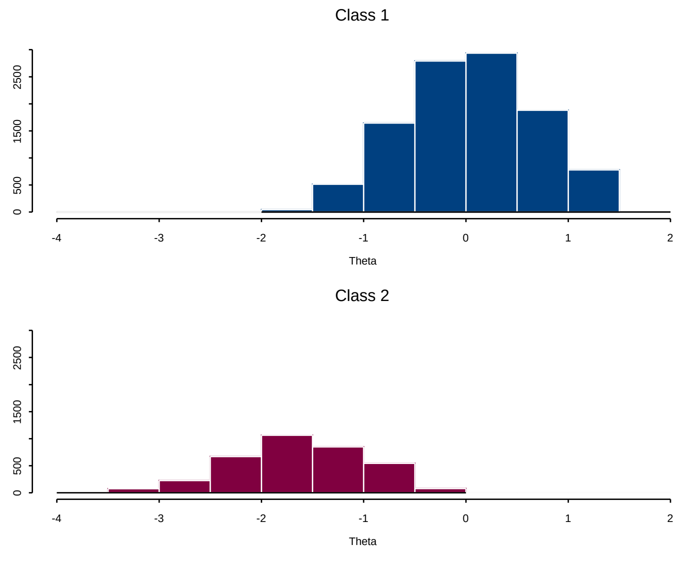


Figure 4.3: Ability Distributions of Two Classes for PIRLS 2006

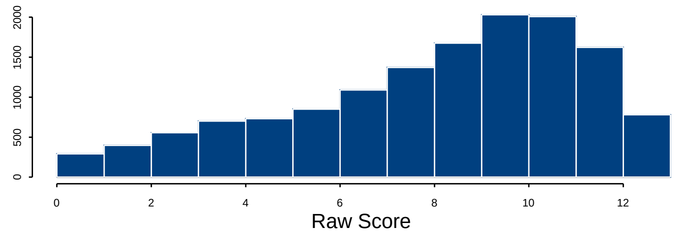


Figure 4.4: Distribution of Raw Scores of PIRLS 2006

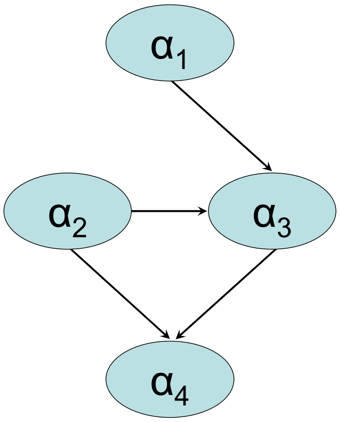


Figure 4.5: Hypothetical Relationship among Attributes in PIRLS 2006

as shown in Figure 4.4, it appears that the 2-class DCMixRM solution well described ability distribution.

Classification of Mastery Profiles. The mastery profiles detected in these data lead to some possibly interesting observations. As can be seen in the Total column in Table 4.6, Profiles 3, 4, 7 and 13 were completely empty and Profiles 8 and 14 had only a few cases. These results indicate three patterns in mastery profiles. First, if an examinee had mastered α_3 (integration), then he/she also must have mastered α_1 (retrieval), because the mastery profiles of ($\underline{0}0\underline{1}0$), ($\underline{0}0\underline{1}1$),($\underline{0}1\underline{1}0$) and ($\underline{0}1\underline{1}1$) were empty. This may also help explain why the correlation between these two attributes was so high (see Table 4.7).

Second, if an examinee had mastered both α_1 and α_2, he/she also must have mastered α_3. Therefore, if an examinee answers correctly to Items 5, 6, 7, 12, or 13, it is likely that Items 3, 4, 9, 11 or 10 were also answered correctly.

Table 4.6: Examinee Classification Results with PIRLS 2006

	Frequency			Percent		
Profile	Class 1	Class 2	Total	Class 1	Class 2	Marginal
1 (0000)	1380	1694	3074	44.9	55.1	21.8
2 (0001)	1211	685	1896	63.9	36.1	13.4
3 (0010)	0	0	0	0.0	0.0	0.0
4 (0011)	0	0	0	0.0	0.0	0.0
5 (0100)	1047	332	1379	75.9	24.1	9.8
6 (0101)	933	111	1044	89.4	10.6	7.4
7 (0110)	0	0	0	0.0	0.0	0.0
8 (0111)	32	0	32	100.0	0.0	0.2
9 (1000)	0	411	411	0.0	100.0	2.9
10 (1001)	4	80	84	4.8	95.2	0.6
11 (1010)	881	20	901	97.8	2.2	6.4
12 (1011)	1412	0	1412	100.0	0.0	10.0
13 (1100)	0	0	0	0.0	0.0	0.0
14 (1101)	0	23	23	0.0	100.0	0.2
15 (1110)	1261	69	1330	94.8	5.2	9.4
16 (1111)	2439	84	2523	96.7	3.3	17.9
Total	10600	3509	14109	75.1	24.9	100.0
Mean[a]	0	-1.35				

[a] Mean ability of each class.

Table 4.7: Tetrachoric Correlation between Attributes of PIRLS 2006

	α_2 Inference	α_3 Integration	α_4 Evaluation
α_1 Retrieval	.384	.997	.316
α_2 Inference		.479	.212
α_3 Integration			.389

Third, the mastery profile results indicate there are two different groups of struggling readers. In one group, there is a lack of mastery of α_1 (retrieval) and in the other group, there is a lack of mastery of α_2 (inference). These attributes appear to be required in order to acquire α_3 (integration). Based on this information, those examinees who were classified as having Profiles 5 or 6 might benefit from an instructional intervention focusing on acquisition of α_1 (retrieval) because this might help them master α_3 (integration). Of those who were classified as having Profiles 9 or 10, an intervention related to α_2 (inference) might have the same benefit. The other implication of the results is that, in an adaptive testing situation, if an examinee misses an item related to α_3 (integration), it might be a good idea to provide sets of items which relate to α_1 and α_2 sequentially to diagnose which of these attributes the examinee has not yet mastered.

These patterns observed from results here might be an indication of a hierarchical relationship among attributes (see Figure 4.5). That is, there is not a hierarchical relationship between Attributes 1 and 2, but both attributes are prerequisite to master Attribute 3.

Relationship of Latent Classes and Mastery Profiles. Results presented in Table 4.6 show that those examinees who mastered more than three attributes tended to belong to Class 1. This class was also higher in ability than Class 2. What this suggests is that the more attributes examinees have mastered, the higher their ability is likely to be. Someone who was able to master only α_1 was likely to be classed into Class 2. Mastery of a single attribute did not always result in someone being classified into Class 2, however, as someone who mastered either α_2 or α_4 tended to belong to Class 1.

The association between attributes and latent class membership was modeled in this study using a multinomial logit model. As discussed in Chapter 3, to construct this model, we needed reference groups which in this case were Class 2 and Profile 16. Recall that Profiles 3, 4, 7 and 14 were discarded because they contained no observations. The results can be found in Table 4.8.

The coefficients for Profiles 8, 9, 11, 12 and 14 were not significant at the .05 level. This was because most of the examinees belonging to Class 1 had Profiles 8, 11 or 12, and most of the examinees belonging to Class 2 had Profiles 9 or 14. α_3 (integration) was a critical attribute which appeared to operate to separate Class 1 and Class 2. Not only did Class 1 have higher ability than Class 2, but the members of Class 1 also mastered α_3 (integration). This can be seen in the mastery profiles that were more likely for Class 1 than for Class 2. That is, examinees with Profiles 8, 11, and 12 mastered α_3, and examinees with Profile 14 did not, even though examinees with Profile 14 did master the other attributes. Further, if an examinee mastered all of the attributes, he/she was much more likely (i.e., $e^{3.369}$=29 times odds ratio) to belong to Class 1 versus Class 2. If an examinee mastered only α_1 and α_4, he/she had only $e^{-6.364}$=.0017 times odds of belonging to Class 1 versus Class 2.

Table 4.9 summarizes class membership and mastery profile for each country. This is a unique feature of DCMs and can help to identify country-level strengths and weaknesses. Results presented in this table show that country may not be the best proxy for latent class membership. For instance, students in the United States and Austria presented the same reading literacy ability on average, and yet the components of mastery profiles in two countries were very different from each other (see Figure 4.6). In the United States, Profile 11 may be considered as a target group for providing a supplementary instructional program. Since the target group has not mastered α_2, the program may be designed to focus on acquiring α_2 (integration). On the other hand, in Austria a group of students classified as Profile 2 appear to be struggling readers who have not mastered α_1, and educators may consider providing a supplementary instructional program focusing on this attribute (i.e., on retrieval).

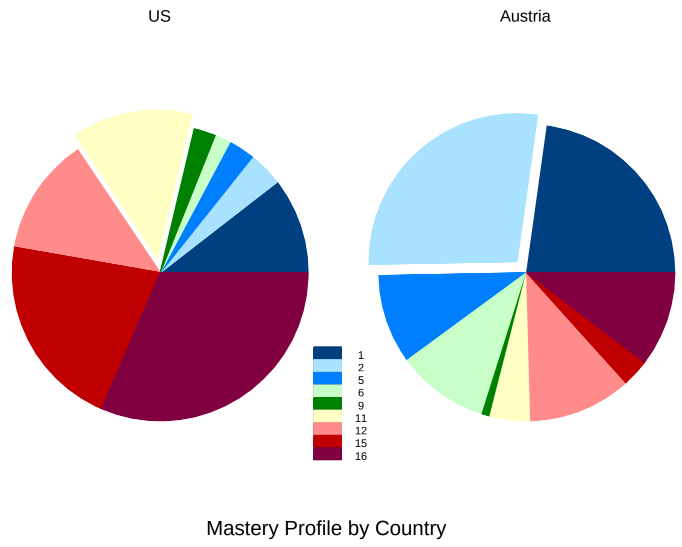


Figure 4.6: Mastery Profiles of United States and Austria of PIRLS 2006

Table 4.8: Effects of Mastery Profiles on Latent Class Membership of PIRLS 2006

Profile	Estimate	Std.Dev.	χ^2	Pr> χ^2
1 (0000)	-3.573	.117	937.0	<.000
2 (0001)	-2.799	.121	536.5	<.000
5 (0100)	-2.220	.128	302.7	<.000
6 (0101)	-1.240	.150	68.6	<.000
8 (0111)	14.202	1155.400	.0	.990
9 (1000)	-21.368	399.500	.0	.957
10 (1001)	-6.364	.524	147.4	<.000
11 (1010)	.417	.252	2.8	.098
12 (1011)	14.202	173.900	.0	.935
14 (1101)	-21.368	1688.900	.0	.999
15 (1110)	-.463	.166	7.8	.005
16 (1111)	3.369	.111	921.4	<.000

Note. Class 2 was the reference group.

Table 4.9: Distribution of Latent Classes and Mastery Profiles of Each Country

	Class		Profile												
Country	1	2	1	2	5	6	8	9	10	11	12	14	15	16	Total
Austria	305	34	77	93	33	34	1	3	0	15	38	0	10	35	339
Bulgaria	225	30	30	16	29	16	1	6	0	15	26	1	40	75	255
Taipei	248	49	57	52	46	31	0	5	2	10	23	2	22	47	297
Denmark	236	30	40	28	16	20	1	5	0	24	55	0	27	50	266
France	258	27	47	36	42	24	0	2	0	9	25	0	29	71	285
Georgia	237	70	71	34	40	32	0	10	1	19	26	2	22	50	307
Germany	471	54	109	105	38	31	1	2	0	40	97	0	34	68	525
Hong Kong	275	34	37	35	37	34	0	5	2	28	29	0	30	72	309
Hungary	222	21	43	31	31	30	1	5	0	15	17	0	21	49	243
Iceland	204	33	61	46	33	44	1	3	0	4	9	0	8	28	237
Indonesia	85	217	117	55	44	12	0	30	4	2	12	1	9	16	302
Iran	222	144	153	40	25	3	0	21	3	41	24	0	25	31	366
Israel	201	61	31	30	21	18	3	4	3	14	19	3	31	85	262
Italy	221	20	33	40	33	44	2	0	0	6	24	0	15	44	241
Kuwait	41	213	104	70	20	19	0	18	5	1	3	0	2	12	254
Latvia	258	26	48	38	24	32	4	4	1	25	32	0	27	49	284
Lithuania	277	36	54	20	62	39	1	1	0	17	14	0	52	53	313
Luxembourg	305	27	62	81	20	29	0	2	0	34	54	0	13	37	332
Moldova	180	86	79	50	40	44	1	3	0	5	14	1	12	17	266
Morocco	55	158	100	47	16	10	0	11	6	2	3	1	6	11	213
Netherlands	270	15	31	20	16	16	0	4	0	34	63	0	38	63	285
New Zealand	333	87	59	31	24	14	0	20	3	50	66	0	55	98	420
Norway	202	45	69	30	48	33	0	9	0	8	12	0	16	22	247
Poland	269	49	86	42	31	27	1	8	0	24	40	1	23	35	318
Qatar	119	321	176	121	37	23	0	22	10	8	15	1	10	17	440
Romania	221	54	79	41	22	19	2	12	0	21	27	1	16	35	275
Russia	292	28	28	26	45	47	3	4	2	10	10	0	50	95	320
Singapore	370	54	32	17	10	7	0	8	2	46	92	0	74	136	424
Slovakia	294	63	90	47	50	37	2	6	1	13	15	1	31	64	357
Slovenia	308	44	67	58	31	16	1	7	2	39	45	0	27	59	352
South Africa	173	702	424	142	71	38	0	75	21	14	35	4	22	29	875
Spain	225	34	68	33	36	20	1	3	0	18	25	0	12	43	259
Sweden	255	28	36	21	22	31	0	3	1	23	28	0	39	79	283
Trinidad & Tobago	164	103	62	35	17	9	0	21	6	19	41	1	15	41	267
Macedonia	126	136	68	35	30	15	0	15	4	7	6	1	32	49	262
United States	296	48	36	13	10	6	1	8	0	45	44	0	73	108	344
England	218	49	29	24	14	4	0	6	1	24	41	0	29	95	267
Scotland	211	37	30	23	8	6	0	5	1	24	38	1	36	76	248
Belgium Flemish	284	25	28	36	34	32	2	4	1	25	41	0	33	73	309
Belgium French	250	49	56	58	65	45	0	5	0	3	13	0	13	41	299
Canada Ontario	219	46	32	17	26	12	1	6	0	29	32	0	48	62	265
∼ Quebec	222	30	48	32	27	15	1	1	1	13	32	0	28	54	252
∼ Alberta	248	23	22	13	9	9	0	6	0	27	35	0	56	94	271
∼ British Columbia	252	28	25	14	19	6	0	4	1	28	40	1	64	78	280
∼ Nova Scotia	253	41	40	20	27	11	0	9	0	23	32	0	55	77	294

4.2 DCMixRM with Mathematics: State's Accountability Test

Description of the Data. A sample of 4,000 examinees was randomly drawn for purposes of this example from a tenth-grade mathematics test administered in a midwestern state. The test consisted of 84 dichotomously scored items, and responses from 4,000 examinees were randomly sampled. The test was used as part of a state assessment program and was designed to provide information about whether or not state standards had been met in mathematics. Results were reported at the state level, the district level, and the school level, and were intended to be used to support both school improvement and accountability.

As was the case for the PIRLS 2006 test, this mathematics test was designed to provide a single overall measure of proficiency in mathematics. The test was not designed to classify examinees into mastery profiles and the test was not designed to provide diagnostic information. The test was specifically designed to measure four standards. For purposes of this example, therefore, we used these four standards to construct a Q-matrix for the test. The standards included Number and Computation (α_1), Algebra (α_2), Geometry (α_3), and Data Analysis (α_4). Number and Computation consists of number sense, number systems and their properties, estimation, and computation; Algebra includes patterns, variables, equations, and inequalities, functions, and models; Geometry consists of geometric figures, measurement and estimation, transformational geometry, and geometry from an algebraic perspective; and Data analysis includes statistics and probability.

The entries of the Q-matrix of items are presented in Table 4.10 along with descriptive statistics. As indicated in this table, 40 of the 84 items were selected. The resulting Q-matrix had a simple structure as each item of the 40 items measured a single attribute. These 40 items resulted in a test that consisted of 10 items measuring each of the attributes.

Model Selection. The same model fitting process that was used with the PIRLS 2006 test was used with the mathematics test. Three DCMixRMs, consisting of models with one- to three-classes, were fit to the data. AIC, BIC, and entropy values were calculated for each model. The results for these indices are presented in Table 4.11 for each solution. Among the

Table 4.10: Q-matrix with Mean and SD of Items in the Mathematics Test

Item	α_1	α_2	α_3	α_4	Mean	SD	Item	α_1	α_2	α_3	α_4	Mean	SD
1	1	0	0	0	.63	.483	21	0	0	1	0	.62	.486
2	1	0	0	0	.43	.495	22	0	0	1	0	.66	.475
3	1	0	0	0	.60	.491	23	0	0	1	0	.72	.451
4	1	0	0	0	.53	.499	24	0	0	0	1	.53	.499
5	1	0	0	0	.81	.396	25	0	0	0	1	.56	.496
6	1	0	0	0	.58	.493	26	0	0	0	1	.74	.441
7	1	0	0	0	.59	.492	27	0	0	0	1	.65	.476
8	1	0	0	0	.44	.497	28	0	0	0	1	.75	.435
9	1	0	0	0	.72	.451	29	0	0	0	1	.57	.495
10	0	1	0	0	.66	.475	30	1	0	0	0	.60	.489
11	0	1	0	0	.49	.500	31	0	0	1	0	.53	.499
12	0	0	1	0	.58	.494	32	0	1	0	0	.57	.496
13	0	0	1	0	.63	.483	33	0	1	0	0	.57	.495
14	0	0	1	0	.53	.499	34	0	1	0	0	.53	.499
15	0	0	1	0	.62	.486	35	0	1	0	0	.59	.491
16	0	0	0	1	.59	.492	36	0	1	0	0	.49	.500
17	0	0	0	1	.64	.480	37	0	1	0	0	.64	.479
18	0	1	0	0	.68	.467	38	0	0	1	0	.59	.492
19	0	1	0	0	.53	.499	39	0	0	0	1	.60	.489
20	0	0	1	0	.68	.468	40	0	0	0	1	.48	.500

Note. α_1 Number & Computation; α_2 Algebra; α_3 Geometry; and α_4 Data Analysis. Means are proportions of examinees who answered correctly to the item.

Table 4.11: Goodness-of-Fit Statistics of DCMixRMs for the Mathematics Test

	Class 1	Class 2	Class 3[a]
Number[b]	95	152	209
AIC	186303	**184692**	
BIC	186901	**185649**	
Entropy	.58	**.67**	

[a] No convergence.

[b] Number of parameters that were estimated in the model.

three models, the three-class model was not converged and AIC, BIC, and entropy suggested that the two-class DCMixRM fit the best for this data set. There was no conflict among fit indices, and as a result, we decided to fit the two-class model to these test data. Recall that in this model, 16 mastery profiles were estimated for each latent class.

Item Parameter Estimates. As with the PIRLS data set, in order to compare the class-specific item difficulties, item difficulties were placed onto the scale of Class 1 through the mean and sigma method as described in Equation 3.5. The equated class-specific item difficulties and attribute effects are presented in Table 4.12. Also, item difficulty patterns of each class can be found in Figure 4.7. First of all, Items 12, 14 and 15 were easier for Class 1, whereas Items 20, 21, 22, and 23 were easier for Class 2. Interestingly, all these items required α_3 (geometry). Two items that had high difficulties for both classes were Items 24 and 25. These items were related to α_4 (data analysis).

Two points are noteworthy regarding attribute effects. First, there were 10 items that involved α_1 (number and computation), and yet only for 2 of these 10 items, the attribute main effects for α_1 (number and computation) were significant at the .05 level. The α_3 (geometry) effects of 5 items were not significant either. This suggests that either this test may not be adequate as a measure of α_1 and α_3 or that the Q-matrix may be misspecified

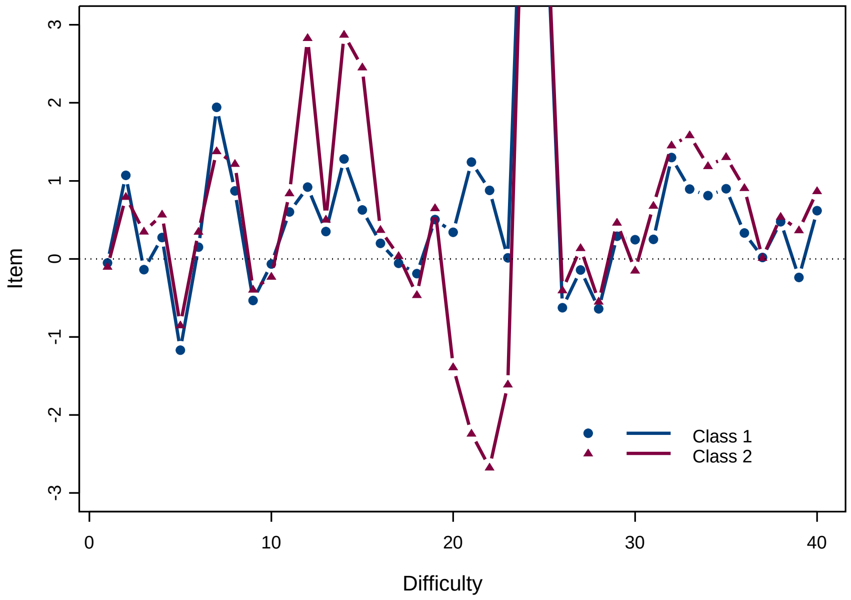


Figure 4.7: Item Difficulty Patterns of the 2-Class DCMixRM for the Mathematics Test

for those items. Also, Items 32, 33 and 35 had main effects for α_2 (algebra) that were very high (i.e., greater than 2).

Ability Distribution and Latent Classes. The ability distribution for each class is shown in Figure 4.8. The difference in means of ability of classes was 0.152 indicating that although examinees in Class 2 had a slightly higher level of proficiency in mathematics, this difference is not meaningful. In other words, it is possible to conclude that the two classes had essentially the same level of mathematics ability. Approximately 46 % of examinees belonged to Class 1 and 54 % to Class 2, which also indicates examinees were evenly distributed in terms of class membership (see Table 4.13).

Also, as shown in Figure 4.9, the raw score distribution appeared square-shaped rather than bell-curved. The 2-class DCMixRM with similar means seems to detect this shape

Table 4.12: Item Difficulty and Attribute Effect of the Mathematics Test

	Item Difficulty		Attribute Effect			
Item	Class 1	Class 2	λ_1	λ_2	λ_3	λ_4
1 (1000)	-.051	-.101	.000			
2 (1000)	1.072	.795	.000			
3 (1000)	-.137	.349	.000			
4 (1000)	.273	.569	.000			
5 (1000)	-1.170	-.852	.192			
6 (1000)	.151	.347	.184			
7 (1000)	1.944	1.381	5.192*			
8 (1000)	.870	1.218	.487*			
9 (1000)	-.532	-.394	.161			
10 (0100)	-.066	-.228		.169		
11 (0100)	.601	.841		.230*		
12 (0010)	.921	2.831			4.221*	
13 (0010)	.350	.504			1.256*	
14 (0010)	1.281	2.873			3.777*	
15 (0010)	.628	2.453			4.305*	
16 (0001)	.200	.370				.272*
17 (0001)	-.056	.037				.204*
18 (0100)	-.188	-.463		.069		
19 (0100)	.504	.650		.348*		
20 (0010)	.341	-1.389			.084	
21 (0010)	1.241	-2.239			.000	
22 (0010)	.879	-2.674			.000	
23 (0010)	.013	-1.609			.000	
24 (0001)	6.757	6.451				8.238*
25 (0001)	5.084	5.259				7.480*
26 (0001)	-.625	-.404				.242*
27 (0001)	-.143	.139				.332*
28 (0001)	-.639	-.547				.235*
29 (0001)	.289	.464				.261*
30 (1000)	.246	-.149	.000			
31 (0010)	.250	.681			.101	
32 (0100)	1.299	1.454		2.503*		
33 (0100)	.895	1.584		2.238*		
34 (0100)	.810	1.190		1.169*		
35 (0100)	.900	1.305		2.257*		
36 (0100)	.331	.908		.000		
37 (0100)	.018	.012		.372*		
38 (0010)	.477	.538			.903*	
39 (0001)	-.238	.367				.000
40 (0001)	.617	.868				.143

Note. λ_1 Number & Computation; λ_2 Algebra; λ_3 Geometry; λ_4 Data Analysis.
* p <.05 for attribute effects.

Table 4.13: Examinee Classification Results with the Mathematics Test

	Frequency			Percent		
Profile	Class1	Class2	Total	Class 1	Class 2	Marginal
1 (0000)	368	48	416	88.5	11.5	10.4
2 (0001)	276	33	309	89.3	10.7	7.7
3 (0010)	25	86	111	22.5	77.5	2.8
4 (0011)	26	70	96	27.1	72.9	2.4
5 (0100)	92	25	117	78.6	21.4	2.9
6 (0101)	52	74	126	41.3	58.7	3.2
7 (0110)	16	246	262	6.1	93.9	6.6
8 (0111)	46	524	570	8.1	91.9	14.3
9 (1000)	332	35	367	90.5	9.5	9.2
10 (1001)	215	55	270	79.6	20.4	6.8
11 (1010)	41	61	102	40.2	59.8	2.6
12 (1011)	39	228	267	14.6	85.4	6.7
13 (1100)	133	66	199	66.8	33.2	5.0
14 (1101)	105	106	211	49.8	50.2	5.3
15 (1110)	25	43	68	36.8	63.2	1.7
16 (1111)	39	470	509	7.7	92.3	12.7
Total	1830	2170	4000	45.8	54.3	100.0
Mean[a]	.00	.15				

[a] Mean ability of each class.

properly. Although the 2-class DCMixRM fit the best to both PIRLS 2006 and this data set, in PIRLS 2006 two classes presented different ability means, but here the mean abilities in mathematics of two classes were very similar.

Classification of Mastery Profiles. Table 4.13 also summarizes a cross-classification of latent class and mastery profile of examinees. The most noticeable observation is that with respect to the mastery profiles, Profile 8 was the profile with the most cases (N = 570)

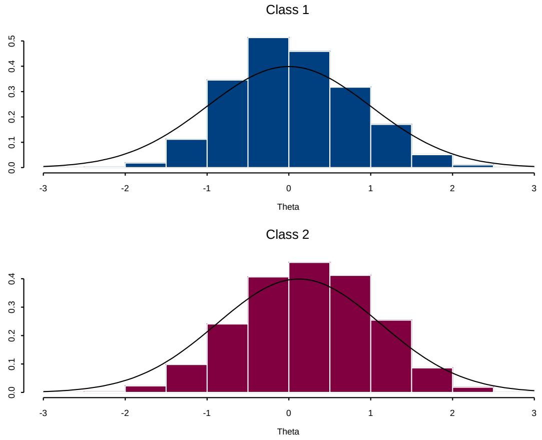


Figure 4.8: Ability Distributions of Two Classes for the Mathematics Test

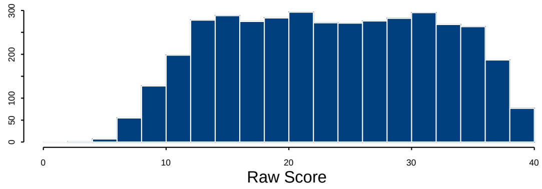


Figure 4.9: Distribution of Raw Scores of the Mathematics Test

Table 4.14: Tetrachoric Correlations between Attributes of the Mathematics Test

	α_2 Algebra	α_3 Geometry	α_4 Data[b]
α_1 Number [a]	-.064	-.068	.132
α_2 Algebra		.570	.318
α_3 Geometry			.425

[a] Number and Computation; [b] Data Analysis.

followed by Profiles 16 (N=509) and 1 (N = 416). Profiles 4 (N=96), 11 (N = 102) and 15 (N = 68) had the fewest cases. This suggests there may be three patterns of mastery profiles and some possible interpretations: mastery state of α_1 (number and computation) was independent from mastery states of other attributes; if an examinee had mastered α_2 (algebra), it was more likely that that person would also have mastered α_3 (geometry); if an examinee had mastered α_2 (algebra) and α_3 (geometry), that person would also have mastered α_4 (data analysis). This was also supported by tetrachoric correlation patterns in Table 4.14 in which correlations between α_2 and α_3 and α_3 and α_4 were substantially positive.

Second, since Class 1 was slightly lower in ability than Class 2, it is reasonable to find that more examinees in the non-mastery group (i.e., Profile 1) appeared in that class. Further, more examinees in the all mastery group (i.e., Profile 16) were expectedly found in Class 2. However, if an examinee had mastered only α_3 (geometry), that individual had higher odds of belonging to Class 2 (see Profile 3 in Table 4.13). In fact, every profile which included α_3 (geometry) had more examinees belonging to Class 2. This includes Profiles 3, 7, 8, 12, and 16. That is, α_3 (geometry) appears to be a critical attribute in determining membership in Class 1 or Class 2 for these data.

Relationship between Latent Class and Mastery Profile. The association between latent class and mastery profile was rigorously modeled with a multinomial logit model.

Table 4.15: Effects of Mastery Profiles on Latent Class Membership of Mathematics

Profile	Estimate	SD	χ^2	Pr> χ^2
1 (0000)	2.235	.151	217.896	<.0001
2 (0001)	2.322	.179	168.475	<.0001
3 (0010)	-1.037	.218	22.657	<.0001
4 (0011)	-.792	.220	12.949	.0003
5 (0100)	1.501	.216	48.115	<.0001
6 (0101)	-.155	.176	.773	.3794
7 (0110)	-2.534	.246	106.077	<.0001
8 (0111)	-2.235	.152	217.030	<.0001
9 (1000)	2.448	.173	200.052	<.0001
10 (1001)	1.561	.149	109.346	<.0001
11 (1010)	-.199	.195	1.044	.3070
12 (1011)	-1.568	.169	85.950	<.0001
13 (1100)	.899	.149	36.466	<.0001
14 (1101)	.189	.138	1.883	.1700
15 (1110)	-.344	.240	2.054	.1518
16 (1111)	-.198	.048	16.934	<.0001

Note. Class 2 was reference.

Consistent with other analyses in this study, Class 2 and Profile 16 were used as baselines for the multinomial regression. The results are given in Table 4.15. Based on these results, for instance, examinees with Profile 9 had approximately $e^2.45 \doteq 11.59$ times higher odds of belonging to Class 1 than Class 2. In contrast, those who mastered attributes 2 and 3, or attributes 2, 3 and 4 tended to have higher odds of being in Class 2 (see Profiles 7 and 8 in Table 4.15). These associations did not appear to be related to the number of attributes examinees had mastered. For illustration, even when an examinee possessed computation, algebra, and data analysis, if the examinee did not also master geometry, (i.e., as required by Profile 14), it was not likely that the individual would belong to Class 2. If, however, an examinee mastered the other three attributes of algebra, geometry, and data analysis (i.e., had Profile 8), then that individual would tend to be classified into Class 2.

An important objective of this analysis was to explain what caused the qualitative differences in examinees response behaviors on the mathematics test. In this analysis, members of Class 1 tended to lack knowledge related to geometry. Tetrachoric correlations also support the unique contribution of geometry as can be seen in the relationship between algebra and geometry and between data analysis and geometry (see Table 4.14).

4.3 Summary and Conclusions

In this chapter, we presented empirical examples of how to fit the DCMixRM model to real test data. To do this, we used data from a reading test and from a mathematics test. The reading test was a part of the PIRLS 2006 international assessment program and the mathematics test was a part of a state assessment program.

These analyses demonstrated the kinds of information that the DCMixRM is capable of providing about examinees, items and attributes based only on item responses. The diagnostic level information about the attributes appears to provide useful information about some of the factors that may be causing the latent classes to form.

In order to determine the best fit model for each data set, AIC, BIC, and entropy indices were compared among candidate models. Although there is no research reported specifically focusing on the use of any of these indices with DCMs, previous research with LCM and MixIRT models suggested that BIC was useful in identifying the most accurate model given the data. As a result, in this study BIC was used to inform model selection. The principle of parsimony was also considered. Based on these considerations, a 2-class DCMixRM was taken as the solution for both tests.

The two-class DCMixRM solution appeared to be the best fit for both data sets. The ability means for the two latent classes were different in one of the two data sets. For reading, the two classes showed a large difference whereas for mathematics two classes exhibited similar mean ability. The characteristics of two latent classes were also discussed in terms of item difficulty patterns for both tests.

For each test, the distribution of mastery profiles was also examined. In theory 16 mastery profiles were possible because 4 attributes were specified in both tests. Frequencies of individuals with some profiles were large while other profiles were seldom observed. In particular, in reading literacy no observations were found for a profile that included mastery of α_1 but nonmastery of α_3. The resulting pattern of profiles suggested one possible inference about the relationship among the attributes in reading literacy: α_1 (retrieval) might be prerequisite for α_3 (integration). On the mathematics test, the relationship among α_2 (algebra), α_3 (geometry), and α_4 (data analysis) was discussed.

The mastery profile patterns in reading suggest two distinct profiles of struggling readers: one profile indicated the students had not mastered *inference* and the other profile indicated the students had not mastered *retrieval*. This information can be used to design future remedial intervention.

However, the most unique contribution of the DCMixRM was that it related mastery profile to latent class. This was also translated into an attribute level. In reading literacy,

α_3 (integration) was the critical attribute that distinguished Class 1 from Class 2 and in mathematics, it was α_3 (geometry).

In conclusion, the DCMixRM can yield rich and various perspectives for persons and items. First, taking advantage of item information, we can verify whether or not items or a test measured correctly the construct of interest as it should. By investigating the significance of the attribute effects for a given Q-matrix, we can ascertain whether or not the test included enough items to measure all relevant attributes. If the test cannot measure some critical attributes, then we may want to revise the Q-matrix or add relevant items in the test. For instance, the 13 reading items may not have adequately measured the mastery state for attribute α_2 (inference). Likewise, the 40 mathematics items may not have adequately measured mastery status on α_1 (number and computation). The DCMixRM allows for making inferences about why some examinees' response patterns differed from others based on mastery profiles.

Chapter 5

CONCLUSIONS AND DISCUSSION

IRT models have been widely used to make inferences regrading item difficulty and examinee ability in psychometrics. These models depend heavily on the local independence such that the observed items are independent of each other when conditioning on ability. As with other statistical models, inferences drawn from IRT models are valid if and only if the assumption holds.

However, it is often the case that even after conditioning on ability, association between item responses still exists. This is an indicator that local dependence is violated that threatens the validity of inferences. The local dependence can be caused from difference among examinees' cognitive patterns (i.e., heterogeneity) or notable closeness of several items within a test (i.e., multidimensionality).

MixIRT models have been proposed to handle examinees' heterogeneity by combining an IRT model with an LCM. These models have been useful for detecting latent classes while still providing a model-based scale for items and examinees.

In spite of the usefulness of these models, they do not lead researchers to explanations as to what causes the latent classes to form and how they can describe the characteristics of latent classes identified in their study.

Some research has been conducted using a covariate(s) such as age, gender, and ethnicity to predict latent class membership hoping that it can provide a clue about latent classes. Although this is somewhat informative, it may not be as informative as would be needed for describing the reasons behind the differences in response patterns between the latent classes.

Previous research suggests that heterogeneity in item response patterns may be related to individuals' differences in knowledge states. However, without a rigorous theory explaining why these patterns may be expected to occur, it is not easy to confirm what causes latent classes to form.

Alternatively, this dissertation has suggested that mastery profiles can be used for describing latent classes in MixIRT models. To do so, the focus of this dissertation was, in part, on the use of mastery profiles to explain latent classes, and we proposed the DCMixRM in which a Rasch model, an LCM, and an LCDM were combined. The Rasch model component captures a general latent ability, or a continuous variable; the LCM is used in an exploratory way to detect latent classes; and the LCDM classifies mastery profiles in a confirmatory way by the use of prescribed relationships between items and attributes explicitly expressed in the Q-matrix. It was assumed that the inclusion of mastery profiles as covariates might provide useful information to help explain what may be causing the latent classes to form. In other words, this model takes into account not only heterogeneity in a population but also multidimensionality of a test resulting from multiple attributes required for correctly answering test items; and the model specifies association between mastery states of attributes and heterogeneity of a population in order to provide information that could be useful in understanding what may have caused the latent classes to form.

The DCMixRM has several advantages: this model provides information about both a global ability and mastery profiles formed over the set of attributes; it provides a way to detect heterogeneity in the population; it may yield more accurate classification of latent classes by the use of covariates; more importantly, it is possible to provide a rigorous explanation about features of latent classes. In addition, the model helps overcome drawbacks of each model. The major drawback of LCDM is that the misspecification of the Q-matrix can have a serious impact on parameter estimation. The inclusion of a continuous latent variable (ability) can handle this possible incompleteness of the Q-matrix.

To evaluate whether or not to realize these advantages, the model and estimation methods were investigated with a simulation study in terms of convergence rates and parameter recovery under varying conditions. Also, empirical data analyses were conducted with two sets of large scale data to demonstrate how to use the model in practice. The data sets were drawn from an international reading test and from a statewide mathematics test. For both the simulation and the empirical studies, a standard latent variable modeling software package, M*plus* Version 5.21, was used to estimate the model parameters.

Based on the results from the simulation study, the following conclusions may be drawn. First, the estimation process converged well in estimating parameters when latent classes have the same levels of mean ability.

Second, the item difficulty and the ability parameters in the DCMixRM were well recovered. This result was consistent with other IRT studies in which item difficulties and abilities were relatively well recovered.

Third, the attribute effect parameters appeared to be overestimated under some conditions, but as the sample size increased, the bias in attribute effect parameter estimates decreased. In this study, the magnitudes of attribute effects were set relatively smaller than those of item difficulties. In general, when coefficients are small, they can be difficult to accurately be estimated. This might be the case with this result.

Fourth, the class membership was reasonably accurately detected. The inclusion of mastery profiles as covariates seemed to prevent label switching. From these observations, the inclusion of the mastery profiles as covariates did appear to yield more accurate class membership and enabled a clearer description of latent classes.

Identifying the correct mastery profile was, however, difficult with this estimation algorithm. This may have been because there are too many mastery profiles (i.e., 16 mastery profiles), one of which each examinee has to be classified into. Clearly, it would be important to study this issue further before making inferences about mastery profiles. Also, inferences of attribute relationships through the use of tetrachoric correlations may not be completely accurate.

However, the estimation algorithm appeared to be capable of providing a general picture of the associations between latent classes and mastery profiles even though results of the simulation study suggest it may tend to underestimate the strength of that relationship. In particular, when the number of items per attribute was small, it appeared to have greater difficulty in estimating this relationship. Further research on this issue is needed.

We also illustrated how to apply the DCMixRM to the empirical data set using two large-scale testing programs, such as an international reading literacy testing program, the PIRLS 2006, and a statewide mathematics testing program. Even though neither test was developed using either a DCM or a MixRM, it was shown that the model could be used to evaluate whether the test was appropriate for measuring a general ability as well as mastery states for a set of attributes. The model does this by estimating both item difficulties and attribute effects through their significance levels. In particular, coefficients of attribute effects can be used to test the impact of the attributes on item responses. Furthermore, the significance levels of the interaction effects provides information about compensatory relationships among attributes for the individual items.

Second, the model appeared to be capable of simultaneously yielding information about examinees' strengths and weaknesses with respect to both a general ability and mastery profile. As with other DCMs, this kind of information about the mastery profiles may have the potential to be useful for designing individualized interventions focusing on strengths and weaknesses in mastery states of attributes.

The model also provided information about the composition of the examinee population in terms of mastery profiles. This information may be useful in identifying strengths and weaknesses of the population identified by a particular latent class, potentially leading to a means of specifying at-risk groups in the examinee population and possibly informing decision making about supplementary instructional programs.

Third, one could examine whether expert knowledge, in fact, matches examinees' learning strategies by comparing the Q-matrix with resulting mastery profiles. This kind of information can be used in revising the Q-matrix or possibly for better understanding the discrepancy between learning theory and individual learning strategies. For instance, in the PIRLS 2006 analysis, several mastery profile patterns were missing. This can be an indicator of a plausible association between attributes. This information may be used for investigating whether there exists hierarchical attribute structures or not.

Finally, the relationship between latent classes and mastery profiles provided a cognitive explanation as to what may have caused latent classes to form. This information is contained in the mastery profiles and potentially can be used to indicate which examinees responded differentially to test items and why they may have done so. Mastery states on *integration* and *geometry*, for example, seemed to cause latent classes to form in the reading and the mathematics test data, respectively.

It is believed that the DCMixRM developed in this dissertation has the potential to be useful for many practical testing applications. The proposed DCMixRM is a complex model measuring high dimensional latent space. It was only partly explored in this study, and future research is needed to examine applications of this type of model. First, there was no clear explanation as to the cause(s) of the large biases in attribute effect estimates. This may have been because the magnitudes of attribute effects in the simulation study were set too small relative to the item difficulties. Different sets of values may improve recovery in attribute effect parameters.

Second, selecting the best fitting model is an important concern for latent class models, DCMs, MixIRT models, as well as for this particular model. Misspecification of a Q-matrix and determining the correct number of latent classes are very important issues and factors with DCMs and finite mixture latent models. Research on these issues is warranted to make valid applications of the models.

Third, estimation of the DCMixRM was only studied using a maximum likelihood estimation algorithm. Bayesian estimation algorithms offer an alternative for estimating parameters of such complicated models. Bayesian with Markov chain Monte Carlo (MCMC) methods, for example, offer some important benefits, particularly for estimating high dimensional models like the DCMixRM. This is due in part to the capability of Bayesian for taking advantage of prior knowledge and to the efficiency of MCMC for estimating a large number of dimensions. It might be informative, for example, to see whether Bayesian estimators yield more efficient estimates for this model as has sometimes been shown with other models.

Fourth, although the study did not include longitudinal data, the model can also be used for longitudinal data sets. To do so, vertical equating for item ability and attribute effect should be investigated. Also, since the model includes two discrete latent variables and one continuous variable, the way to combine latent growth and latent transition modeling can be another research area for this model.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Aitkin, M., Bennett, N., & Hesketh, J. (1981). Teaching styles and pupil progress: A re-analysis. *British Journal of Educational Psychology*, *51*(2), 170-186.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bandalos, D. (2006). The use of Monte Carlo studies in structural equation modeling research. , 385–426.

Bolt, D. M. (1999, April). Applications of an IRT mixture model for cognitive diagnosis. Paper presented at the AERA Annual Meeting, Montreal, Quebec, Canada.

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

Bonnefon, J. F., Eid, M., Vautier, S., & Jmel, S. (2008). A mixed Rasch model of dual-process conditional reasoning. *The Quarterly Journal of Experimental Psychology*, *61*(5), 809-824.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Brown, M., Askew, M., Baker, D., Denvir, H., & Millett, A. (1998). Is the national numeracy strategy research-based? *British Journal of Educational Studies*, *46*(4), 362-385.

Bucholz, K. K., Heath, A. C., Reich, T., Hesselbrock, V. M., Krarner, J. R., Nurnberger, J. I., et al. (1996). Can we subtype alcoholism? a latent class analysis of data from relatives of alcoholics in a multicenter family study of alcoholism. *Alcoholism: Clinical and Experimental Research*, *20*(8), 1462-1471.

Buck, G., & Tatsuoka, K. K. (1998). Application of the Rule-Space procedure to language testing: examining attributes of a free response listening test. *Language Testing*, *15*(2), 119-157.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2006, April). An investigation of priors on the probabilities of mixtures in the mixture rasch model. Paper presented at the annual meeting of the Psychometric Society, Montreal, CN.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. A. (2007, April). Latent transition analysis with a mixture item response theory measurement model. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Chung, H. (2003). *Latent-class modeling with covariates.* Unpublished doctoral dissertation, Pennsylvania State University.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133-148.

Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disablilities Research & Practice*, *20*(4), 225-233.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum, Lawrence Associates.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton University Press.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173-178.

Dayton, C. M., & Macready, G. B. (2007). Latent class analysis in psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (p. 421-446). Amsterdam, Boston: Elsevier North-Holland.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115-130.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

Deb, P., & Trivedi, P. K. (2002). The structure of demand for health care: Latent class versus two-part models. *Journal of Health Economics*, *21*(4), 601-625.

DiBello, L. V., & Crone, C. (2001, April). Technical methods underlying the PSAT/NMSQT enhanced score report. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

DiBello, L. V., & Crone, C. (2002, April). Skill-based scoring models for the PSAT/NMSQT. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (p. 361-389). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*(4), 249-262.

Eckstein, Z., & Wolpin, K. I. (1999). Why youths drop out of high school: The impact of preferences, opportunities, and abilities. *Econometrica*, *67*(6), 1295-1339.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*, 20-30.

Feddag, M. (2008). Statistical inference for the multidimensional mixed Rasch model. *Communications in Statistics: Simulation and Computation*, *37*(9), 1732-1749.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A Latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading*, *10*(3), 301-322.

Fu, J. (2005). *A polytomous extension of the Fusion model and its Bayesian parameter estimation.* Unpublished doctoral dissertation, University of Wisconsin-Madison.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT (Research Report No. 2005-11).* New York: College Examination Board.

Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research*, *14*(5), 475-483.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301-321.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis.* Cambridge, UK: Cambridge University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale: Incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement*, *67*, 280-299.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*, 39-59.

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Commissioned paper prepared for the Committee on the Foundations of Assessment, National Research Council*.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Kaiser, F. G., & Keller, C. (2001). Disclosing situational constraints to ecological behavior: A confirmatory application of the mixed Rasch model. *European Journal of Psychological Assessment*, *17*, 212-221.

Karelitz, T. M. (2004). *Ordered category attribute coding framework for cognitive assessments.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* Springer.

Kruidenier, J. (2002). Literacy assessment in adult basic education. , *3*, 84-151.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*, *14*(4), 671–694.

Laumann, E. O., Paik, A., & Rosen, R. C. (1999). Sexual dysfunction in the united states prevalence and predictors. *JAMA*, *281*(6), 537-544.

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

Lee, O., Fradd, S. H., & Sutman, F. X. (1995). Science knowledge and cognitive strategy use among culturally and linguistically diverse students. *Journal of Research in Science Teaching*, *32*, 797-816.

Li, F., Cohen, A. S., & Bottge, B. A. (2007, April). A latent transition analysis model for assessing change in cognitive skills across repeated measures. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353-373.

Linn, R. L. (1986). Testing and assessment in education: Policy issues. *American Psychologist*, *41*, 1153-1160.

Lord, F. M. (1975). The ‘ability’ scale in item characteristic curve theory. *Psychometrika*, *40*(2), 205–217.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*(8), 611-631.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187-212.

McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park: Sage Publications.

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

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure: A mixture-distribution rasch analysis. *European Journal of Psychological Assessment*, *24*, 27-34.

Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations. *Methods of Psychological Research Online*, *3*, 75-93.

Meyer, J. P. (2008, March). A mixture Rasch model with item response time components. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, New York.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, *15*(4), 363-389.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

Mislevy, R. J., Wingersky, M. S., Irvine, S. H., & Dann, P. L. (1991). Resolving mixtures of strategies in spatial visualization tasks. *British Journal of Mathematical and Statistical Psychology*, *44*, 265-288.

Moseley, D. (2004). The diagnostic assessment of word recognition and phonic skills in five-year-olds. *Journal of Research in Reading*, *27*(2), 132-140.

Mroch, A. A., Bolt, D. M., & Wollack, J. A. (2005, April). A new multi-class mixture Rasch model for test speededness. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Mullis, I. V. S., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Amsterdam, the Netherlands: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary school in 40 countries.* Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Muthén, L. K., & Muthén, B. O. (1998-2004). *Mplus technical appendices.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Nichols, P., Chipman, S., & Brennan, R. (1995). *Cognitively diagnostic assessment.* Lawrence Erlbaum.

Nitko, A. J. (1995). Curriculum-based continuous assessment: A framework for concepts, procedures and policy. *Assessment in Education: Principles, Policy & Practice*, *2*(3), 321-337.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36.

Rijkes, C. P. M., & Kelderman, H. (2006). Latent-response rasch models for strategy shifts in problem-solving processes. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (p. 311-329). New York: Springer.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349–359.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (p. 324-332). Münster: Waxmann.

Rost, J., & von Davier, M. (1993). Measuring different traits in different populations with the same items. In J. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology: Proceedings of the 7th european meeting of the psychometric society in trier.* Stuttgart: Gustav Fischer Verlag.

Rupp, A. A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78-96.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment methods: Theory and application.* New York: The Guilford Press.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective.* Unpublished doctoral dissertation, University of Maryland, College Park.

Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007). Mini-mental status examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the mmse. *Journal of Clinical Epidemiology*, *60*(3), 268-279.

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

Smit, A., Kelderman, H., & van der Flier, H. (2000). The Mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research*, *5*(4), 31-43.

Smit, A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an eysenck personality questionnaire. *Methods of Psychological Research Online*, *8*(3), 23-50.

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer et al. (Eds.), *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, *44*(4), 313-324.

Tatsuoka, K. K. (1983). Rule Space: An Approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnoses. In R. L. Fredericksen, A. M. Glaser, Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Templin, J. L. (2004). *Generalized linear mixed proficiency models.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychology Methods*, *11*(3), 287-305.

Templin, J. L., Henson, R. A., & Douglas, J. (in press). General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates. *Manuscript under* *review*.

Thacher, J., & Morey, E. (2003). Using individual characteristics and attitudinal data to identify and characterize groups that vary significantly in their preferences for monument preservation: A Latent-Class Model.

Thomas, H., & Horton, J. J. (1997). Competency criteria and the class inclusion task: Modeling judgments and justifications. *Developmental Psychology*, *33*(6), 1060-73.

Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling*, *6*, 84–98.

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement*, *23*(4), 283-297.

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

von Davier, M. (2005). *A General Diagnostic Model applied to language testing data.* ETS Research Report No. RR-05-16. Princeton, NJ: Educational Testing Service.

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution rasch models: Extensions and applications.* New York: Springer Verlag.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics.* Amsterdam: Elsevier.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Wang, C., & Gierl, M. J. (2007, April). Investigating the cognitive attributes underlying student performance on the SAT critical reading subtest: An Application of the attribute hierarchy method. Paper presented at the annual meeting of the National Council on

Measurement in Education, Chicago, Illinois.

Wilson, M. (1989). Saltus: A Psychometric model for discontinuity in cognitive development. *Psychological Bulletin*, *105*, 276-289.

Yamamoto, K. (1987). *A Model that combines IRT and latent class models.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.