GUESSING AND COGNITIVE DIAGNOSTICS: A GENERAL MULTICOMPONENT LATENT TRAIT MODEL FOR DIAGNOSIS

A Dissertation
Presented to
The Academic Faculty

By

Megan E. Lutz

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Psychology

Georgia Institute of Technology

May 2014

Guessing and Cognitive Diagnostics: A General Multicomponent Latent Trait Model for Diagnosis

Approved by:

Susan E. Embretson
School of Psychology
Georgia Institute of Technology

Thomas Morley
School of Mathematics
Georgia Institute of Technology

Christopher Hertzog
School of Psychology
Georgia Institute of Technology

Jonathan Templin
Department of Psychology and Research in Education
The University of Kansas

Daniel Spieler
School of Psychology
Georgia Institute of Technology

Date Approved: 28 March 2014

ACKNOWLEDGEMENTS

My first thanks go to Dr. Embretson for her guidance, knowledge, and suggestions as I navigated my way through this maze of a process. I am grateful to my friends William Horowitz and Jonathan Duggins, and their incredible facility with calculus. Our late-night sessions over the phone and online, drawing up complicated derivatives are legend and, though I hope never to repeat them, are something I look back on fondly. They are truly great friends to go the distance in checking and rechecking, and even triple-checking my math with me. I am thankful for Kristin Morrison's camaraderie and friendship, her helpfulness when I was at my wit's end, and her willingness to serve as a sounding board when I invariably hit a roadblock in my work. Many thanks to the Ivan Allen College and our talented IT support team, led by Erik Brown, for going above and beyond in getting me access to the computing resources I needed when they were not otherwise available; I quite literally could not have completed the required analyses without them.

I am deeply indebted to my parents, without whose love and support I could not have made it this far. Their guidance and encouragement—at all hours!—have been unconditional and played no small part in the size of the goals I set for myself. I hope to continue to make them proud on the next step of my journey. Finally, I am grateful for the quiet patience and support that Graham has given me for the last three years; this has been a much longer process than I originally envisioned. He has ridden out many peaks and troughs as I struggled through this project and school, and I am glad that we finish this together on a high. Thanks for sticking with me, things can only get better from here.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

SUMMARY

A common issue noted by detractors of the traditional scoring of Multiple Choice (MC) tests is the confounding of guessing or other false positives with partial knowledge and full knowledge. The current study provides a review of classical test theory (CTT) approaches to handling guessing and partial knowledge. When those methods are rejected, the item response theory (IRT) and cognitive diagnostic modeling (CDM) approaches, and their relative strengths and weaknesses, are considered. Finally, a generalization of the Multicomponent Latent Trait Model for Diagnosis (MLTM-D; Embretson & Yang, 2013) is proposed. The results of a simulation study are presented, which indicate that, in the presence of guessing, the proposed model has more reliable and accurate item parameter estimates than the MLTM-D, generally yielding better recovery of person parameters. Discussion of the methods and findings, as well as some suggested directions for further study, is included.

CHAPTER 1

INTRODUCTION

Multiple choice (MC) testing in educational settings has been around since the early 20th century, and has been controversial for nearly as long. Concerns about the inclusion of guessing and exclusion of partial knowledge in raw scores (e.g., Chernoff, 1962; Potthoff & Barnett, 1932; Ramsay, 1968) and issues related to item quality (e.g., de Finetti, 1965; Jersild, 1929; Jones, 1928) sprang up shortly after the use of MC and other such "objective tests" came in vogue. Other potential drawbacks have been identified: as early as the 1970s, concerns about the broadening use of MC scores had entered the literature (e.g., Bligh, 1979; Hall, Carroll, & Comer, 1988; Rust, 2002). There is a lack of agreement on the optimal number of alternatives to provide for an MC item (e.g., Haladyna & Downing, 1993; Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2005). There is further disagreement as to how such an item should be scored or formulated: alternatives to the conventional, number-right (NR) MC scoring method as well as to the standard objective item format have been proposed and studied (e.g., Bickel, 2010; Boldt, 1971; Gibbons, Olkin, & Sobel, 1979; Haladyna, 1992; Ramsay, 1968; Searle, 1942; Wilcox & Wilcox, 1988; Wisner & Wisner, 1997; Yunker, 1999), each with their own advantages and disadvantages. This paper will explore several of these alternatives in turn and discuss their relative merits, with special consideration given to primary works that address guessing and partial knowledge in MC test taking.

Armed with all of the different arguments for and against MC testing, the purpose of this paper is to investigate various proposed methods of formulating and scoring MC

tests, leading up to a proposal of a generalized cognitive diagnostic model that accounts for guessing on MC items. Chapter 1 is both an introduction to the response methods underlying item responses to MC items, presenting the purpose and general overview of the following chapters, and an introduction to MC testing. In Chapter 2, various means of addressing the previously noted issues, item formulation and design, as well as other scoring strategies are covered, along with some empirical results for those schemes. Test-taking and test-writing strategies are investigated from a psychometric viewpoint, as are strategies that lend themselves more or less to guessing. Approaches from classical test theory and item response theory are discussed and several models from the cognitive diagnostic modeling (CDM) framework are introduced, examining various existing latent trait and latent class models. Chapter 3 outlines the proposed study, drawing on the information presented in the preceding chapters for justification and groundwork, defining the scope of the proposed simulation and real data analysis, as well as the theoretical underpinnings of the estimation methods for the items and persons. The results of two pilot studies are also included as an indicator of feasibility. Chapter 4 contains the results of the described study, beginning with a comparison of the original and new models and then delving into the results in more detail. Finally, the paper concludes with a discussion of the results and their implications in Chapter 5.

Motivation for the Proposal

This introduction reviews a variety of classical test theory- (CTT) and latent-trait theory-based approaches to modeling guessing, partial knowledge, and misinformation, as well as determining whether and how guessing manifests itself in test-taking. It then reveals that the popular interpretation of "guessing", while not entirely inaccurate, is

incomplete. IRT models have been in use for several decades, but it was with an improved estimation method in the early 1980s that the most general 3-parameter logistic model (3PL) has seen wider application. The 3PL contains a lower asymptote, which is meant to account for the probability of correctly guessing on item (Birnbaum, 1968), and is now commonly referred to as the guessing parameter. Chapter 2 will cover the relevant IRT models and CDMs more thoroughly, the questions to consider here are: though Birnbaum included it to represent the probability of correctly guessing on an item, is the lower asymptote accurately defined and interpreted as a guessing parameter? How may the lower asymptote better be interpreted: is it an artifact of test-taking or scoring strategies?

Birnbaum's (1968) theoretical value for the lower bound, based on basic probability theory, is often not obtained when freely estimating the lower asymptote from real data. At times, the estimated lower asymptote is quite a bit higher than theory would suggest, which may simply be indicative of fewer functional distractors for a given item. On the other hand, the estimated lower asymptote may be lower than theory predicts, and it is this scenario that suggests that something is occurring with the item or the examinees to make it less likely to get a correct answer than just chance would allow. It is important to more fully understand the "guessing parameter" as it appears in various models, so that its inclusion in the proposed model in Chapter 2 can be justified.

To study the issue thus addressed, one must look at what is meant by "guessing" and how it has been handled by other researchers. Some of the models discussed distinguish "random guessing" to be a truly random selection from among the alternatives present, where an alternative is selected based on no knowledge of the content of the item

in question; this definition is most similar to what is commonly referred to as guessing in the context of the lower asymptote of the 3PL model, as that is the theoretical probability of correctly responding to an item with a complete absence of knowledge. However, as previously mentioned, the interpretation of the lower asymptote as random guessing is not always supported by the actual estimate arrived at in the 3PL framework. Other definitions of guessing include the notion of partial knowledge, or "constrained guessing", which is often treated as a separate type of guessing. In the case of partial knowledge, it is assumed that the examinee has some working knowledge in the domain of the item, but is unable to fully identify any one alternative as the single best answer. In these cases, an examinee may be able to eliminate one or more alternatives as incorrect, leaving several alternatives remaining from which to choose: it is at this point that random guessing from the remaining alternatives may occur.

Achievement Testing in Education

Bligh (1979) focused on the criticism that achievement testing was being used for purposes other than those originally intended, even while people were calling for more standardized testing. By broadly defining achievement tests as those used for any evaluation and both norm- and criterion-referenced information, Bligh unifies the type of instrument one may be working with. Within the framework so defined, Bligh then notes that the "primary purpose [of achievement tests] is to provide relevant information to be used with other sources in decision making"(p. 2). The importance of this caveat cannot be understated when dealing with past and present criticism of the standardized test: such tests must be interpreted within the context of other student achievements and assessments, a major component of test validity. As the popularity of standardized

achievement testing continues to grow, and as the demand for such tests increases due to government incentives, one must also be cognizant of how the tests themselves are designed and scored, and what information is being gleaned from them.

Hall, Carroll, and Comer (1988) noted the issue that Bligh (1979) framed so well: standardized achievement tests should not be the only measure considered for decision making. Classroom teachers from different grade levels rated their use of three different levels of assessment: their own tests, national exams, and state competency tests. The primary interest was in how the teachers used the three different sources in making their own decisions for their classroom, as well as student learning and as a reflection on how they, themselves, were performing as teachers. The results of the survey revealed that all three sources contributed to teachers' decisions about academic progress, adequacy of instruction materials, diagnosis of student weaknesses, and other indicators. None of the three sources was weighted much more than any other, each coming in at roughly equal levels, with more consideration generally being given to the teacher-prepared assessments. The external sources are taken into account, but other factors are also considered, and are considered more important. Teachers of different levels (e.g., elementary, middle and high school) used the tests for different purposes, specifically the elementary level teachers used the tests less for student promotion and retention and for motivating student learning.

With the passage of the No Child Left Behind (NCLB) Act of 2002, the federal government attempted to improve nationwide literacy and scholarship of its primary and secondary education students via measures of accountability, specifically performance of K-12 students on standardized achievement examinations. The NCLB, then, made major

use of the national and state achievement tests to grade schools, teachers, and student promotion and retention, contradicting the teachers' own weighting preference (Hall, Carroll, & Comer, 1988). However, Duckworth, Quinn, and Tsukayama (2011) urge caution in interpreting achievement test scores as the ultimate predictor of success in and out of the classroom; report card grades serve a separate and distinct function in such decision making.

The difference in objectives and student outcomes between standardized tests and report card grades can likely be attributed to in-class curriculum. Less emphasis on some topics means a student will pick up the knowledge outside of class for standardized tests, while diligence on material emphasized by the teacher would naturally correspond with higher report card grades. In one study, self-control and intelligence were measured as two distinct constructs that contribute differentially to academic performance: intelligence was found to contribute significantly to standardized achievement test scores, and less so or not at all to a student's grade point average (GPA; Duckworth, Quinn, & Tsukayama, 2011). Conversely, self-control contributed significantly to GPA, but not at all to standardized achievement test scores. If standardized test scores are determined by intelligence and less so by self-control, which is linked to classroom learning, are standardized test scores the best measure of a classroom or school? Indeed, the teachers involved in Study 3 (Duckworth, Quinn, & Tsukayama, 2011), recognized the distinction between the two types of assessment and used them in appropriate, complementary ways, neither giving more credence to one nor the other.

CHAPTER 2

LITERATURE REVIEW

In this chapter, several scoring strategies from each of the classical test theory (CTT), item response theory (IRT), and cognitive diagnostic modeling (CDM) frameworks will be discussed. Each model has advantages and disadvantages, and some measure guessing better than others, if they do so at all. Alternatives to the traditional MC item design are also presented and considered. Ultimately, the traditional test administration and design are settled upon for moving forward with the proposed model, but some options are presented and weighed here.

Classical Test Theory Scoring Strategies

In this section, the various methods of scoring to accommodate guessing and partial knowledge will be reviewed. The following methods, except where noted, are all scored within the CTT paradigm, where the test score is an estimate of an examinee's true score, or ability, in a given domain. It is only in how the item scores are calculated that these methods differ. Table 1 outlines the major alternative strategies discussed here, identifying the examinee and administrative tasks that differ relative to NR scoring. Also included in Table 1 are some of the more salient disadvantages and recommendations against the alternative scoring methods when compared to traditional scoring.

Table 1

Summary of alternative scoring methods

Method	Additional Tasks	Disadvantages
Option weighting	• Administrative: *a priori* option weighting by experts • Administrative: *post hoc* option weighting as function of responses	• Additional test time required • Negligible gains in reliability/validity
Formula scoring	• Examinee: understanding of new test instructions • Examinee: assessment of true knowledge state • Examinee: learning of temporary test-wise behavior • Administrative: explanation of scoring rules	• Additional test time required for instructions • Benefits test-wise students • Dependent on risk-averse/seeking behavior • Instructions penalize low ability students • Decreased reliability
Confidence testing		
Probability testing	• Examinee: weighing of all alternatives for correctness • Examinee: understanding of complex scoring rules • Examinee: assessment of true knowledge state • Administrative: formulation and understanding of scoring rule • Administrative: explanation of scoring rule	• Possible score of -∞ • Dependent on psychological variables • Additional test time required • Inappropriate for mathematically unsophisticated • Negligible gains in reliability/validity examinees
Pick-one testing	• Examinee: assessment of confidence in correctness of single • Examinee: understanding of complex scoring rules alternative • Administrative: formulation of scoring table for given scale • Administrative: explanation of scoring rule	• Additional time spent scoring • Additional test time required
Elimination/subset testing	• Examinee: understanding of complex scoring rules • Examinee: assessment of true knowledge state • Administrative: formulation of scoring table for given scale • Administrative: explanation of scoring rule	• Negligible gains in reliability/validity • Additional test time required • Additional scoring time required • Benefits random guessing • Penalties inconsistent for confidence level
Immediate feedback	• Examinee: answer item until correct alternative selected • Examinee: understand scoring on sliding scale for varying number • Administrative: formulation of scoring rule • Administrative: item task analysis • Administrative: *a priori* item calibration of attempts	• Interacts with test anxiety • Penalizes low-ability students • Additional test time required • Task-analysis for some items • Item calibration required for option pair probabilities • Negligible gains in reliability

Number right scoring

The conventional MC scoring method proceeds in the following manner: each examinee indicates a single alternative of those available for an item that he feels is the most correct, yielding a binary pattern over the length of the test; if an item is correctly answered, the examinee gets one point or full credit, else no credit is awarded. Over the course of an entire examination, a pattern of ones and zeroes for each student is obtained—the item response pattern—which is used in both CTT and IRT applications for scoring exams. This conventional scoring method is often referred to as "number right" (NR) scoring, where the final score on the test is simply the number of items an examinee answered correctly. Two major arguments have been made against this allocation system, that the scores include additional credit for items on which the examinee guessed (e.g., Chernoff, 1962), and that the scores are not reflective of partial knowledge on items on which the examinee might have answered incorrectly (e.g., de Finetti, 1965). In the case of CTT, one's true score is estimated by his observed score on a given instrument, typically the NR score; as Chernoff (1962) argues, that true score estimate and measurement error are artificially inflated by those items on which the examinee successfully guessed. It can be argued that some alternatives may be "more right" than others (e.g., Wilcox & Wilcox, 1988, Yunker, 1999), or that an incorrect alternative was endorsed simply because there was roughly equal certainty between it and the correct one (e.g., Bickel, 2010), with the remaining alternatives ruled out as possibilities.

Option weighting

Chernoff (1962) identified the likelihood of guessing or uncertainty as a product of both an individual examinee's selection of the correct alternative and the relative proportion of the overall population selecting each alternative presented. In his example, a correct alternative selected at a rate approximately equal to that of the other alternatives present indicated that only a few examinees actually know the correct response; in contrast, a correct alternative selected at a much higher rate than the remaining alternatives indicates that more students know the answer and are not randomly guessing among all options. The proposed approach to handling guessing was more direct: it explicitly identified items on which guessing by an examinee was likely and differentially weighted.

Ramsay (1968) noted that the expected average score due to guessing could be comparatively large and addressed that issue and the missed measurement of partial knowledge with a statistically-based method. By assigning *post hoc* weights to the different alternatives for an item, one could separate groups of respondents based on their resulting group mean scores. The relative weights for the alternatives for a given item can be chosen to maximize the separation between groups of respondents of different ability classifications, which informs criterion scores for said separation. As the weights can be determined for the alternatives based on the sample proportions of the item alternatives for students of different criterion groups, students can be awarded partial credit from those weights based on their partial knowledge. That is, incorrect alternatives selected by students from higher-performing groups would receive higher weights because of their attraction to the more able students. As weights can also be allowed to be negative, random guessing is penalized by costing the examinee points on an item for a random

selection of an unpopular alternative. Like Chernoff (1962), Ramsay's method depends on the group's distribution among answer options for each item to help determine an individual's score. Other methods for weighting alternatives, by experts or consensus *a priori* have been proposed (e.g., Pascale, 1971).

It is the nature of weighting the alternatives (Ramsay, 1968) that allows for potential misclassifications, which may have severe ramifications for the examinee. Criterion scores for classifying test-takers must be continually updated as the class and material evolve, but even updated scores may yield misclassification. While MC tests are primarily considered to be objective measures of one's knowledge state, subjectivity can be introduced by non-uniform item weights; Potthoff and Barnett (1932) noted that teachers often disagreed with the marks given by an un-weighted, standard scoring system, and that such discordance is not predictable. One suggestion for improving the quality of option-weighted items is for the different distractors to aligned along the construct of interest, enabling diagnosis and a more valid assessment and utilization of the resulting weights: the distractors could then be weighted by the criterion scores of those examinees selecting the respective alternatives (Echternacht, 1973). An IRT variation of this method, the nominal response model (Thissen & Steinberg, 1984), will be discussed later in this chapter.

Several studies have investigated the psychometric properties of tests scored under option weighting. Chevalier (1998) conducted an extensive review of different partial-credit and correction-for-guessing scoring systems and found inconsistent effects on reliability and validity of those methods. In another study, comparing the validity of option-weighted tests and NR tests for making pass/fail decisions, such as in end-of-year

examinations for promotion to the next grade level, Haladyna (1984) noted that previous research noted a tradeoff in reliability and validity: increases in reliability for option-weighted tests generally led to no gains in validity. Haladyna's study partially confirmed these results, finding that option weighting effectively increased reliability and also improved pass/fail decisions with regard to misclassificationHaladyna suggested that, as option weighting must be regulated and requires well-designed items, it should only be utilized for large, well-controlled testing programs and not for teacher-developed or other in-house classroom tests.

Haladyna's results are refuted by a study conducted by Kansup and Hakstian (1975), in which the option weights were determined empirically from examinees' subjective rankings of the alternatives. The option-weighted scores for both verbal and mathematics items were used and no practical increase in internal consistency was identified: in fact, a decrease in said reliability was found for one of the testing conditions. Kansup and Hakstian did not find significant changes in validity for the scoring methods over traditional NR methods, though a significant decrease in validity was observed for one of the administered tests. Due to the inconsistent and generally insignificant changes in reliability and possible decreases in validity, the research findings do not support option weighting as improving psychometric properties.

In a review of a number of option weighting studies, Frary (1989) likewise concluded that validity of option-weighted tests is suspect and had been poorly measured in the past, though a consistent increase in reliability was found. Haladyna's admonition against option weighting in smaller examinations reduces the exposure of students to such items and may confound exam performance with anxiety over a different

administration, which must be considered in the context of Pascale's (1971) recommendation in general against non-conventional test administration methods for younger children. However, larger-scale tests involving more centralized administration may benefit from option weighting via the nominal response model (Thissen & Steinberg, 1984).

Formula scoring

Formula scoring is one alternative to option weighting methods and is often also referred to as "correction for guessing" or "correction for chance" (e.g., Chevalier, 1998; Cross, 1975; Foster & Ruch, 1927; Horst, 1932; Little & Creaser, 1966; Ruch & DeGraff, 1927). Instead of identifying items based on the overall population's performance, formula scoring looks at individual item responses and applies one of several formulae to account for guessing or partial knowledge. Kurz (1999) and Chevalier (1998) provide two reviews of several such methods, from both CTT and IRT perspectives. The impact of risk-aversion and non-compliance with instructions, however carefully given, and the unequal penalization of examinees across the ability continuum raise concerns for its implementation. The next two sections outline some of the more common formula scoring methods and the drawbacks associated with post-hoc score corrections, respectively.

Variants of formula scoring

Two CTT formula scoring models, the random-guessing model and the rights minus wrongs (RW) correction model, respectively award partial credit for omitted items and penalize examinees for incorrectly answered items, where the reward and penalty are each weighted by the number of alternatives provided for each item. In the case of the

random-guessing model, omitted items are awarded a fraction of full item credit, under the assumption that an answer would otherwise have been based on a random guess. The random-guessing model is considered to be a positive model: by omitting items, examinees are assumed to be aware of their own knowledge state and are rewarded for this awareness for each omitted item. The RW model assumes all incorrectly answered items are due to random guessing, so the examinee is penalized for attempting an item he did not know. Any correction for guessing based on a penalization for incorrect responses, such as RW, depends on the equal difficulty of the distractors for the weighting of the penalty to be valid (Horst, 1932).

Both of the random-guessing and RW models require additional instructions to the examinee, outlining the scoring method and how omits in the former are rewarded and guessing in the latter are penalized (e.g., Lord, 1975; Ruch & DeGraff, 1927). The mechanics of responding to an item are unchanged between these formula scoring systems and a standard MC exam. However, more understanding of the instructions is required for the formula scoring models, which may penalize lower ability examinees before they begin the exam (e.g., Kurz, 1999).

Formula scoring and students' cognitive processes

Lord (1975) premised the success of formula scoring on explicit instructions to the examinees, where it is explained how one may maximize his performance by guessing only on those items for which he is able to eliminate at least one alternative and otherwise omitting items on which he can do no better than chance, which is instruction in test-wiseness. Even when providing explicit instructions to students to relate test-taking strategies with different scoring outcomes (Lord, 1975), non-compliance and other

issues arise when tests are administered in the formula scoring paradigm. Cross and Frary (1977) and others (e.g., Bliss, 1980; Plake, Wise, & Harvey, 1988) tested Lord's suggested, explicit instructions (see Lord, 1975). Those studies found instead that formula scoring unduly penalizes able students, who perhaps better understand the instructions. Cross and Frary (1977) also found individual differences in interpreting the test instructions as well as in examinees' ability to assess their own partial knowledge and guessing behavior, supporting earlier findings by Granich (1931).

Formula scoring and psychological variables

Cross and Frary (1977) also identified the potential of personality factors, such as risk aversion, to influence formula scoring results. Frary (1989) argued that formula scoring belongs in the classification of confidence testing because of the need for examinees to recognize their own partial knowledge and relative likelihood of item alternatives to judge whether they have a better-than-chance probability of getting an item correct. Foster and Ruch (1927) found that, though formula scoring supplies more information on examinee abilities than NR scoring, RW scoring tends to over-penalize examinees due to excessive omissions or risk-seeking in guessing when one ought not. Burton (2004) showed no consistent increase in RW scores, though that finding is impacted by low-ability examinees. In another study, risk-seeking behavior was assessed and compared with scores from NR and RW scoring methods (Bliss 1980); in that study, RW scoring yielded a higher internal consistency, but more risk-averse students omitted items that they had a better-than-chance probability of getting correct, yielding a higher penalty in terms of true score estimate than those less risk-averse.

Little and Creaser (1966) found that examinees are be penalized unduly under RW scoring , as a student may identify the correct response to an item, but with low confidence, which would lead to that item's being omitted under formula scoring. Such a scenario would lead to a lower true score estimate for that examinee as a result. From an administrative standpoint, one has to consider that RW scoring has the possibility of a negative true score estimate, while positive corrections for omits may yield a non-zero score for a blank test (e.g., Chevalier, 1998). It would be up to the test administrator to determine how to interpret and report such scores, as the political ramifications of a negative score can be tremendous, while receiving a negative score may have demoralizing effect on an examinee.

Summary of formula scoring findings

Glass and Wiley (1964) showed mathematically how RW formula scores are generally less reliable than NR scores, while at the same time RW scores increase the validity of the scores and their interpretations. Due to many problems with formula scoring, including the reliance on examinees of all ability levels to fully understand the instructions and to recognize their own partial knowledge, and the small-to-negligible changes in test validity and reliability, formula scoring is not recommended for general use.

Confidence testing

Given the insensitivity to confidence in an alternative of conventionally scored MC tests, a number of confidence testing methods have been developed. In all such methods, additional work is required of both test-takers and administrators, and to varying degrees. There are a variety of confidence testing models that have been

proposed, requiring varying amounts of additional input from both test-takers and administrators. Whether explicitly assigning a confidence level to all alternatives or just to those selected, or rating the correctness of the various alternatives or the one selected on a Likert-type scale to indicate confidence, these test-administration schemes are all forms of "confidence testing" (Echternacht, 1971). There is a large number of different confidence-testing formats available for selection, all of which attempt to measure either one or both of guessing and partial knowledge. The current review will cover the more general cases of these models. Some aspects of option weighting, such as when students assign their own weights, tie in with the notion of confidence testing. However, confidence testing as a type of scoring strategy is conceptually different from option weighting, as the latter typically has weights assigned by the test administrator during the scoring process or during test development.

Advantages of confidence testing

Wisner and Wisner (1997) identified the advantages of confidence testing to include rewarding genuine knowledge, reflected by correct answers confidently given; penalizing guessing or attempts to game the system, reflected by incorrect answers confidently given; and through the first two, providing additional motivation for more thorough studying and understanding of the content in question. More coverage in this review will be given to Bickel (2010), as it describes the most general form of confidence testing. As other studies and methods are included in the review, it will become apparent that, though Bickel's work is more recent, the other studies describe more specific ways of addressing the problem. All possess the advantages above to some extent, as well as similar disadvantages discussed later.

In testing the situation may arise where an examinee is able to eliminate most, but not all, of the alternatives for a given MC item, which is the same assumption underpinning correct RW scoring instructions. Consider the example posed by Bickel (2010), where two alternatives remain: a student may have a confidence vector for those two remaining alternatives of $\mathbf{p}_1 = (0.85, 0.15)$, while a second student may have a vector of $\mathbf{p}_2 = (0.51, 0.49)$. In both cases, the standard MC scoring would prompt both to select the first alternative. A third student may have a confidence vector of $\mathbf{p}_3 = (0.49, 0.51)$ for the same two alternatives, thereby selecting the second alternative because of her marginally higher confidence. In the standard MC scoring, the first two students would receive full credit on the item, while the third student would receive no credit. Bickel (2010) argues that there is a dual insensitivity of the scoring of these three students: the student with the most confidence in the correct answer receives the same credit as the student who all but randomly chose that correct answer from the two that remained. The third student, who basically has the same knowledge state as the second student, gets no credit for the item. Thus, students who are aware of their own knowledge and ability with high confidence are not separated from those who are less confident in their knowledge, and students of similar low confidence are separated in scores; this is the major argument for confidence testing and other methods that handle partial knowledge.

Probability testing

The most complex method of confidence testing was described and tested by Bickel (2010), in which examinees assign to each alternative their confidence of that alternative's being correct for a given item. A student can therefore maximize his score by assigning his personal probability of correctness for each alternative when the

administrator utilizes a "strictly proper scoring rule" (p. 347). Bickel recommends the logarithmic scoring rule for the reason that it possesses the desirable properties of locality and an association between score and increased knowledge of the content. For a scoring rule to have the local property, the score must only depend on the confidence assigned to the correct alternative when multiple alternatives are possible for a given item (Bickel, 2010). Local scoring rules will always give a higher score to correct answers given greater confidence.

Pick-One testing

Boldt (1971) proposed a "Pick-One" scoring system, in which the examinee selects one alternative he believes to be correct and assigns it a value from a 4- or 5- point scale to indicate his certainty in that alternative's correctness. In this way, there is no concern for scores of negative infinity as with probability testing (Bickel, 2010), nor of the difficulty addressed by de Finetti (1965) of specifying specific personal probabilities for any alternative. By only having to rate one's confidence on a pre-determined scale for a single alternative, the examinee has more time to complete more items on an examination. Tables for scoring using the Pick-One system can be provided *a priori* so students can understand how confidence ratings on a correct alternative correspond with the score on the exam.

Other confidence indicators for single alternatives

Wisner & Wisner (1997) developed and tested two systems similar to Pick-One. In both cases, examinees indicated the alternative they believed to be correct and then noted their confidence in that selection. In the first system, a 3-point Likert scale, representing high, moderate, and low confidence was used. In the second system, an

examinee indicated high confidence by circling the item number on his answer sheet: un-circled items were deemed to be of moderate confidence. Correct, high-confidence answers received extra credit, whereas incorrect answers with high confidence were penalized. Correct answers given with moderate confidence were neither penalized nor rewarded and were used to determine the base score available for an item. Honesty in admitting low confidence was encouraged by awarding partial credit for correct responses and lower partial credit for incorrect responses. In the experimental stage of these systems, examinees who opted for the confidence-weighted tests also received a report of their conventional score and their overall confidence level, helping diagnose overconfidence, which in this case was interpreted as misinformation.

With a wider range of possible scores, both scoring systems had a higher variance of scores than the standard method, and confidence-weighted scores were higher than standard scores (Wisner & Wisner, 1997). A lengthy scoring time was reported for the 3-point confidence scale system, due to the six possible point values available for each item; the two-level confidence system was relatively easier to grade, but still more time-consuming than traditional electronic NR scoring. The students in the study who opted in generally found the confidence testing format to be more fair in awarding points and that it encouraged additional time spent with the material. The instructions and scoring for both scoring systems are fairly straightforward to explain and understand, but require additional work on the part of the student during the test administration.

Disadvantages of confidence testing

Bickel (2010) conducted his study on college-age students in a decision analysis program. As the other studies included in this review were also used college-aged participants, there may be an issue with broadly replacing MC exams at all education levels with confidence weighting. Bickel (2010) discussed the use confidence testing throughout the first week of the class, and included assignments to further illustrate the outcomes associated with different confidence allocations for the scoring system, as that was the standard used to determine final grades in the course. It was also emphasized that an infinitely negative score was possible for a given exam because of the nature of the logarithmic scoring rule, and that a withdrawal or a grade of an 'F' were the two outcomes possible should that situation arise. College-level students without strong mathematics backgrounds or ability would have a difficult time understanding the scoring system and its impact on their grade.

Although Bickel (2010) strongly recommends the logarithmic scoring rule, Hakstian and Kansup (1975) found that confidence tests in general, and the logarithmic scoring rule in particular, provide no major gains in test reliability and some losses in validity. The authors argued that adopting a more complex scoring system has no benefit, besides the approval of students noted by Bickel (2010). Hakstian and Kansup did, however, find some gains in internal consistency and stability of confidence tests over NR scoring.

Echternacht (1971) conducted a review of several confidence testing techniques available at the time, and drew similar conclusions to those identified here. In general, confidence testing requires more of both the examinee and test administrator, in terms of time spent on their respective tasks of taking and scoring the exam. Implementation of a

confidence testing protocol requires thorough explanation of the scoring rules and system in place, which may put lower-ability or younger test-takers at a disadvantage (Kurz, 1999). Some scoring systems are too complex for use in primary school grades (e.g., Bickel, 2010; Wilcox & Wilcox, 1988), as it is doubtful school-age children would be able to fully grasp the mathematical intricacy involved, or to fully understand the ramifications of assigning different confidence levels to the alternatives.

Elimination and subset testing

One alternative to probability testing is the subset selection technique (Gibbons, Olkin, & Sobel, 1979), or elimination test (ET; Coombs, Milholland, & Womer, 1956). Under these test systems, confidence is demonstrated by selecting a subset of alternatives from those provided, as either probable correct alternatives (subset selection) or probable incorrect alternatives (ET). By allowing the selection of multiple alternatives, the two techniques allow for the measurement of partial knowledge and discourage guessing (Chang, Lin, & Lin, 2007; Coombs, Milholland, & Womer, 1956; Cross, Thayer, & Frary, 1980; Gibbons, Olkin, & Sobel, 1979; Tollefson & Chung, 1986); any subset from none to all alternatives is allowable, and if the subset contains the correct response to the item, the whole subset is deemed correct or incorrect, depending on inclusion or elimination testing. Under subset selection, the maximum score for an item is obtained for correct subsets of size one, much like with confidence testing, indicating complete confidence in the correct response, with scores diminishing for correct subsets of larger sizes. No score is earned if the subset contains all possible alternatives, indicating complete lack of confidence in any subset or alternative. Similarly to the logarithmic scoring method endorsed by Bickel (2010), incorrect subsets receive increasingly

negative scores (Gibbons, Olkin, & Sobel, 1979), corresponding to subsets of larger sizes, thereby penalizing students for wild guessing.

Hakstian and Kansup (1975) found no consistent increase in validity or reliability of ET over NR testing and, due to the increase in testing and scoring time of ET, recommended against its adoption, which was corroborated by the findings of Cross (1975). However, Chang, Lin, and Lin (2007) found that ET does measure partial knowledge better than NR scoring, which was consistent with the findings of a previous study that ET controls guessing better than other methods (Cross, Thayer, & Frary, 1977). The first study, however, found that ET unduly advantages random guessers (Cross, Thayer, & Frary, 1977).

There is a potential penalization of examinees with low confidence and incorrect partial knowledge; the lowest score possible on an item occurs when an incorrect subset of size $k - 1$ is indicated (Gibbons, Olkin, & Sobel, 1979), revealing low confidence on the part of the examinee. A misapplication or misunderstanding of a rule could occur, but this practice of giving smaller penalization to high-confidence incorrect selections than to low-confidence incorrect selections is contrary to other confidence testing protocols.

Immediate feedback

Wilcox and Wilcox (1988) developed a scoring formula for the answer-until-correct (AUC) method of testing, which is facilitated by computer-based testing systems. In AUC situations, a student indicates an alternative for an item and is given immediate feedback via presentation of a new item if the student answered correctly or re-presentation of the current item—with the previously selected alternative removed—if the student answered incorrectly. Feedback of this simple nature can prove instructive to

the test-taker and improve ability and learning, but also help measure the extent to which that test-taker may be guessing by the number of chances he needs to answer the item correctly. In this case, two models are of interest: one where examinees are assumed to guess perfectly randomly, and the other where the probability of a second alternative's being selected is conditioned upon the first alternative. In the second model, the "second response conditional probability" model, partial knowledge may come into play when the conditional probabilities for given alternative pairs vary, determined by the proportion of ordered alternative pairs observed in a calibration study, common error is the difference between the two alternatives. Thus, there is a non-random pattern of second choice alternatives for incorrectly answered items: some second choices are more popular given the initial incorrect selection.

While the AUC paradigm is well-measured by the conditional probability model, it requires the task analysis of each item for appropriate modeling of the probabilities (Wilcox & Wilcox, 1988). This is a large burden to place on a test administrator, especially for long tests with broad content. Additionally, as the original study consisted of similar spatial reasoning items involving apparent rotations of a point-of-view, the appropriateness of such a scoring format for disparate constructs or items without observable tasks may be questionable. More research into AUC models and scoring functions must be done before widespread implementation. Early work discovered that examination on otherwise unknown material was an aid to learning in an academic environment (e.g., Jersild, 1929), so the very mechanics of eliminating alternatives with minimal feedback may further familiarize the student with the material.

In a review of different confidence testing methods, Frary (1989) identified AUC as one way of potentially measuring partial knowledge under the same assumptions of Wilcox and Wilcox (1988). However, that review found AUC scores to be generally lower than the corresponding NR scores, which may be due to the self-fulfilling nature of immediate feedback for lower-ability examinees measured by Arkin and Walts (1983), among others. That is, lower-performing students who receive immediate feedback regarding their poor performance continue to do worse than if no feedback had been received. The items for AUC must be well-constructed such that the alternatives fall along the continuum of the construct so that the order in which alternatives are selected can also provide diagnostic information as to the examinees' abilities (Frary, 1989). In addition to impacting examinees' performance, AUC tests further polarize the naturally occurring difference in scores for lower and higher-ability students. Arkin and Walts (1983) found a significant interaction between test anxiety and feedback. Specifically, examinees with low test-anxiety were more impacted by immediate feedback than high test-anxious students.

Concerns raised by Cross (1975) regarding ET scoring in the previous section led to the development of a modified AUC/ET method. In the scoring paradigm of Cross, Thayer, and Frary (1980), a higher penalty is imposed on misplaced confidence, as occurs in the case of misinformation. In the study, elimination testing was used, but immediate feedback was provided such that no more alternatives could be eliminated once the correct answer was chosen. The study found higher reliability coefficients than strict ET, but not to conclusively recommend the new method over ET.

Psychological Variables and Alternative Scoring Systems

All of the alternative scoring methods previously discussed may have implications for psychological, construct-irrelevant variables. A major potential drawback to confidence testing is that confidence is a personality variable, which is often not the construct of interest in classroom or achievement testing; some studies show that such a variable may be a factor in determining exam scores. Echternacht, Boldt, and Sellman (1972) found at least a tentative correlation between confidence levels and test scores; personality traits were assessed prior to the commencement of training in a technical course. Partial correlations between confidence level from the Pick-One (Boldt, 1971) and Distribute 100—an alternative rating system similar to that of Bickel (2010)—and personality indicators, including dogmatism, anxiety, rigidity, impulsiveness, and self-sufficiency (Echternacht, Boldt, & Sellman, 1972) were calculated. The study found that some partial correlations were significant, but none consistently across both testing formats. The authors they asserted that confidence level is something inherent to each person, and so they tentatively concluded that there is no impact of a person's confidence in general on performance on a confidence test.

Koehler (1974) found an association between overconfidence and risk-taking propensity on confidence tests, by inserting nonsense items into a standard test. In that study, it was found that over-confidence was associated with increased variability in confidence test scores, beyond variability related to knowledge. However, overconfidence was not equivalent to risk-taking behavior, as determined by the number of attempted nonsense items when strictly instructed not to guess. Thus, over-confidence

is a personality trait that contributes to increased variability in confidence test scores, which is a recommendation against confidence testing.

Zakay and Glicksohn (1992) found a link between overconfidence and test scores: specifically that overconfident students received lower grades. In their study, students assigned a confidence level from 0 to 100 for the alternative they selected for each item on an exam. The tests were NR scored and students with high grades were well-calibrated and reported high confidence on items they answered correctly. On the contrary, students who were overconfident were, by definition, reporting high confidence on items they had incorrectly answered: these findings contradict those of Walker and Thompson (2001), who determined that students are risk-neutral on MC exams, and that risk-seeking and risk-averse behaviors do not factor into test scores. Zakay and Glicksohn (1992) concluded that personality influences on overconfidence and MC test-taking should be explored further.

Arkin and Walts (1983) found that test-anxiety interacted with immediate feedback when feedback is given early on, indicating that scores can be impacted by extraneous factors and differences in test scores can be polarized beyond what otherwise would have been expected. Hansen (1971) found that certainty in an answer was significantly correlated with F-scale measures of an authoritarian personality as well as, in some cases, risk-seeking activity. Further, Tollefson and Chung (1986) found that examinees had difficulty adjusting to alternative testing systems, as the examinees reported that the new instructions were perceived to be more difficult than conventional testing. Plake, Wise, and Harvey (1988) warned against non-conventional scoring or test-taking situations, noting that examinees do not always behave according to the rules in

those situations, even when they express understanding of the rules in place and how to optimize their test score. Edgerton and Stoloff (1967) identified test-wiseness, or facility with MC tests, to be a factor leading to variability in test scores. And, to reiterate Cross and Frary's (1977) findings, there are clear individual differences in how students interpret scores and identify their own partial knowledge, so the assumptions of the alternative scoring methods are not necessarily upheld. DeMars (2009) showed that motivation wanes over the course of multiple assessments and even within an assessment, and she provides a model to account for the decrease in effort exerted for later items, as there was a clear correspondence in effort, test scores, and guessing.

Design Considerations in Item Format

As teachers note the importance of both classroom assessments and standardized achievement tests for big decisions about their classroom and students (e.g., Duckworth, Quinn, & Tsukayama, 2011; Hall, Carroll, & Comer, 1988), having well-constructed assessments at all levels is vital to the success of educational programs. Even before automated scoring of MC items and fill-in-the-bubble forms and electronic form readers, there was an advantage in scoring accuracy gained by the use of MC items, as well as in scoring speededness (Cuff, 1931). Chang, Lin, and Lin (2007) found corroborative evidence indicating that the cost of administering a test can be reduced further by implementing computer based testing (CBT) systems, eliminating the need for pencil-and-paper tests, as there is no difference in performance on the two test formats. Now, the debate is less over the accuracy of scores and more over how the tests can be constructed so that the scores are meaningful in terms of examinee knowledge and skills.

Item types

Students in the United States are familiar with the standard MC test format and scoring rules, but there are other ways to administer an objective test, some of those with desirable properties are reviewed in this section. Searle (1942) described one such test, in which each item would have a variable number of correct alternatives, from none of the options to all of them. In this way, the amount of information directly tested could increase with little extra cost in time to the item writer: "each single alternative can be made a differentiating unit in the scoring of the test" (p. 703). Searle (1942) recommended that the number of correct alternatives should be approximately equal to the number of incorrect alternatives over the course of the entire test; machine-assisted scoring was can quickly score multiple-answer MC questions. Edgerton and Stoloff (1967) and Scheideman (1931) were other early proponents of drafting MC items with a varying number of alternatives.

Another alternative that simultaneously minimizes guessing and the chance of correctly answering an item solely due to guessing was proposed by Kubinger et. al (2010), in which it was shown that the 2-of-5 item designs were superior to those of 1-of-6 testing. In 2-of-5, five alternatives are presented for each item, exactly two of which are correct. The *a priori* probability of correctly guessing on the item is 0.10; in 1-of-6 the traditional MC presentation of one correct response out of six possible alternatives is scored, with an *a priori* guessing probability of 0.17. Kubinger et al. found that this small change in item design resulted in large changes in item difficulty for otherwise identical items, where 2-of-5 was found to be more difficult than 1-of-6, and as difficult as free-response. It can be extrapolated to assume that 2-of-5 would perform even better when compared to a more traditional, 4-alternative MC item. However, constructing a test—say

a mathematics test—where each item has two correct responses may be burdensome or impossible for the item writer.

Of the different forms of MC item construction, a few stand out as better than the others. True-False (TF) items have the advantage of being short, enabling the inclusion of more items that can be tested in a given length of time, but they have lower reliability than other formats and their quality depends on the ability of the item writer (Haladyna, 1992). MC items may contain as few as two alternatives and still yield higher reliability than TF (Haladyna, 1992), and MC items are only as good as the number of functional distractors available (e.g., Haladyna, 1992; Haladyna & Downing, 1993). In Haladyna's (2004) estimation, progress has been made in the development of alternative objective item types, such as multiple true-false and alternate choice, but more research into their relative advantages and disadvantages is warranted.

Distractor properties

Consideration must be given to guidelines for item writing and formulation for those specific item types. If one is interested in the phenomenon of guessing, and if the theoretical probability of correctly answering an item is a function of the number of alternatives on said item, one must consider the appropriate number of distractors. There is some disagreement in the literature as to how many distractors should be used for an MC item (e.g., Haladyna, 2004; Haladyna & Downing, 1993; Rodriguez, 2005). Haladyna, Downing, and Rodriguez (2002) conducted a review of dozens of item-writing textbooks, and ultimately recommended that four alternatives be used. Haladyna and Downing (1993) found that most often there was only one functioning distractor in an MC test, which is an argument for fewer, rather than more, alternatives.

Rodriguez (2005) conducted a meta-analysis spanning eight decades of research, and found three alternatives to be optimal. Rodriguez suggested that three alternatives was in line with the suggestion of Haladyna, Downing, and Rodriguez (2002) that one use only as many distractors as feasible, insofar as there are usually only three plausible alternatives. Further, the meta-analysis found that distractors are not the sole contributor to item difficulty and discrimination, though Haladyna and Downing (1993) did find a relationship between the number of distractors and an item's discrimination. As the number of alternatives is also used in generating a start value for the lower asymptote for some IRT estimation software, using items with unnecessary and infeasible distractors will hinder that estimation.

Distractor functioning

Like with item design, the development of the alternatives must be a thoughtful process. Horst (1932b) discussed the use of well-crafted alternatives as contributing to item difficulty: if the alternatives are ordered along the construct, selection of each alternative can provide information about an examinee's ability beyond just the NR score. Horst's recommendation was a prelude to IRT item-person comparisons.

Thissen, Steinberg, and Fitzpatrick (1989) used trace lines of the distractors to identify how different alternatives function for different ability levels, regardless of location on the construct. The distractor trace plots can indicate good distractors: such as those with monotonically decreasing functions over increasing ability; non-functioning distractors, such as those with constant functions over ability; and non-monotonic functions, which may help discriminate between moderate and high ability levels as well as indicate to whom that particular distractor is attractive. The authors argue against such

trace analyses for person ability estimation, however, indicating that the traces work best in large, well-controlled assessment programs with thousands of examinees, and are meant only for item analysis and improvement.

Test length

It has long been known that longer tests are more reliable, per the Spearman-Brown prophecy formula, but Glass and Wiley (1964) found an added benefit to lengthier tests: the reduction in differences in guessing behavior between examinees of different ability levels, which was upheld in Wang and Calhoun's (1997) construction of test-score critical values, recommending corrections for guessing for shorter assessments. However, longer tests—especially those perceived as low-stakes by the examinees—are prone to poorer estimates of examinee abilities due to lower effort or fatigue exerted on later items (DeMars, 2007). If a longer test is also timed, poorer person item parameter estimates under the IRT paradigm will be obtained due to the impact of test speededness on guessing and time spent on later items (e.g., DeMars, 2007; Goegebeur, DeBoeck, Wollack, & Cohen, 2008), and alternative item- and person-analyses must be implemented to account for those changes in test-taking behavior.

Test-taking strategies

The current review covers what studies of different testing methods have exposed about examinee behavior into guessing. Examinees are not very consistent in recognizing their own guessing behavior or partial knowledge (Cross & Frary, 1977), even when in full understanding of the testing scheme in use. Early in the MC literature, guessing—operationalized as willingness to attempt new and unfamiliar material—was shown to be independent from ability level (Granich, 1931). Thus, guessing itself arises due to

unfamiliarity with material, perhaps that which an examinee neglected to review before an examination; misapplication of a principle as in a mathematics or physics test resulting in an answer not present; an inability to identify the correct answer out of some distractors, rather than proceeding to eliminate known incorrect alternatives (e.g., Plake, Wise, & Harvey, 1988); or it may be due to the instruction and general rule that it is better to guess than leave an item blank, resulting in random guesses on speeded tests for slower students (e.g., Goegebeur, DeBoeck, Wollack, & Cohen, 2008); or a lack of motivation (e.g., DeMars, 2009). That is, construct-relevant and –irrelevant factors come into play when measuring guessing and test-taking strategies so the phenomenon of guessing is itself ill-defined.

Item Response Theory Models

Given the general recommendations against the alternative CTT scoring methods, a different approach to measuring guessing and partial knowledge must be considered. In this section, several dichotomous and polytomous IRT models are reviewed. Two IRT models for use with polytomous responses are explored for their utility in measuring partial knowledge and elimination of the impact of guessing, when used with items designed for those purposes. The IRT models are contrasted against CTT and some of the methods described earlier in this chapter.

The 3PL and the lower asymptote

As the 3PL model is the primary dichotomous IRT model of interest, the following sections describe the model, its derivation, and uses more thoroughly. As the 3PL is the only dichotomous IRT model that includes guessing as an item property,

empirical studies are also reviewed that describe issues with both the estimation of the overall model as well as the guessing parameter.

History of the 3PL.

Birnbaum (1968) furthered his logistic test model (LTM) theory and introduced guessing into the latent-trait models of item response theory.

> Even subjects of very low ability will sometimes give correct responses to multiple-choice items, just by chance. One model… assumes that if an examinee has ability θ, then the probability that he will *know* the correct answer is given by a normal ogive function… of exactly the kind considered [in the previous section]; it further assumes that if he does not know it he will guess, and, with probability [c_i], guess correctly (p. 404).

The 3PL models this probability by including c_i, a lower asymptote that accounts for the chance that an examinee of sufficiently low ability will still correctly answer an item. The 3PL model for the probability of correctly answering an item is provided in (2.1).

$$\Pr[X_i = 1 \mid \theta] = c_i + (1 - c_i)\frac{1}{1 + \exp\{-1.7a_i(\theta - b_i)\}} \tag{2.1}$$

In (2.1), a_i is descriptive of the information the item provides about person ability θ, or the discriminatory power of item i; b_i is the value of θ at which the point of inflection occurs, and is representative of the item difficulty, or the location of the item; and c_i defines the minimum probability of successfully answering the item, or the probability that a person completely lacking in ability ($\theta = -\infty$) gets the item correct. Although c_i can be justified as the psychological parameter for guessing, it need not mathematically or realistically be the case that guessing has occurred, or that it has occurred at random as in the CTT methods (Birnbaum, 1968).

For MC items where alternatives are laid out and among which exactly one is explicitly stated to be correct, the only way to respond incorrectly with any certainty is to omit it altogether. In the 3PL the psychological and statistical probability of a correct item response is tied in with the logistic model. As one may expect, the value of c_i would be a function of k_i, the number of alternatives for an item (Birnbaum, 1968), that is $c_i = \frac{1}{k_i}$. However, as will be discussed in the next section, c_i can also be freely estimated from the data during the process of fitting the overall logistic test model.

Changing the lower asymptote of a logistic model has the drawback of changing the meaning of the item difficulty. If one has a better than zero chance of correctly getting an item correct by merely choosing an answer, then that will naturally increase the probability he correctly answers an item at his ability level, or where $\theta = \beta_i$. However, α_i and β_i maintain their interpretations as item discrimination and difficulty parameters, respectively.

Empirical findings

Although Birnbaum (1968) introduced the 3PL model to address the reality of guessing on MC items, the lower asymptote does not always hold up to that interpretation under scrutiny. In some instances, the disconnect between the theoretical lower asymptote of the 3PL and the empirically-derived value will be highlighted.

Rasch vs. 3PL.

Some studies have been conducted to investigate the utility of the 3PL over the Rasch model (e.g., Glas, 2009; Maris & Bechger, 2009; Parchev, 2009). Specifically, one's personal perspective into IRT and measurement impacts the model selected and the determination as to whether guessing has occurred. Two people with different frames of

reference can arrive at different conclusions given the same data, where expected scores did not matched the observed score distribution for low abilities (Maris & Bechger, 2009). In that case, someone with Rasch-leanings may interpret the problem as one of poor person selection, where enough low-ability examinees did not sit the test; so that no guessing has occurred when the 3PL is fit. The Rasch model could still be found to perfectly fit the data, and so truncating the distribution of abilities resolves the issue of poor person representation. However, another may be more inclined to believe that guessing has taken place, and rather than look to the sample of people to fix the problem, the model itself is changed; the expected and observed scores might not have matched at the low ability levels because people with low ability may still chance upon the correct response. The two divergent perspectives achieve perfect fit, but with two very different models based on two different sets of assumptions: a sampling problem where students do not guess, or a well-sampled population where guessing has occurred.

The 3PL versus the 1PL

Partchev (2009) raised an issue similar to that of Maris and Bechger (2009): by virtue of the freely estimated "guessing" parameter, and non-zero priors for that parameter, the 3PL will find guessing where it may not actually exist. In his simulations, Partchev found that when guessing didn't exist the 1PL nearly perfectly recaptured the true item difficulty, but the 3PL over-estimated it. When guessing does occur, however, the 1PL tended to shrink harder items' difficulty estimates in response to the increased number of correct responses, and the 3PL again overestimated item difficulty. Partchev's (2009) simulations helped illustrate the situation that Maris and Bechger (2009) discussed: assumptions about strategies or examinees that influence the choice of model

will then influence the model estimates. Again, the intention of an analysis and the examination itself are major factors in choice of model to use, even in the instance of significant improvement in model fit (e.g., Embretson & Reise, 2000).

The 3PL and local optima.

Samejima (1973) identified a peculiarity of the 3PL parameter estimation: the possibility of non-global maxima. While Bock and Aitkin (1981) were instrumental in bringing MML and EM estimation to the forefront for IRT model estimation, the problem addressed by Samejima (1973) can still arise with poorly selected priors. In the case of the 3PL, for example, the Bock-Aitkin algorithm as implemented by Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) may arrive at a local optimum if one uses incorrect priors for the lower asymptote, yielding very different and sometimes inaccurate item parameter estimates.

The lower asymptote and theoretical chance

Another issue that may arise with the 3PL is the apparent inconsistency between the theoretical value of the lower asymptote, based on the number of alternatives available in an item (Birnbaum, 1968) and actual estimates obtained using various IRT estimation programs. For example, one study of seventh-grade mathematics achievement items, which were drafted within the 3PL framework, revealed a wide range of estimates for the lower asymptote (Lutz & Embretson, 2012). All items had four alternatives, so the theoretical lowest probability of person correctly answering the item would be 0.25, using Birnbaum's (1968) logic and the rules of probability. While the average lower asymptote across all items on that test was 0.23, the minimum and maximum values were 0.092 and 0.50, respectively. It should be noted that those parameter estimates were found using

Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) with starting points for the lower asymptote equal to 1/4 = 0.25; the data led the estimation away from the theoretical value, sometimes drastically. In fact, 17 of the 84 items on that test had a lower asymptote below 0.15 and eight of the items had a lower asymptote estimated to be greater than 0.35; nearly 30% of the items had a guessing parameter estimated to be at least 0.10 away from the theoretical value of 0.25 (Lutz & Embretson, 2012).

In another study, an abstract reasoning test (ART; Embretson, 1998, as presented in Embretson & Reise, 2000) also had variation in its lower asymptote estimates, though to a lesser degree. The ART consisted of 30 items with eight alternatives each (Embretson, 1998), so the theoretical probability of correctly obtaining the right answer by guessing is relatively low (1/8 = 0.125). In that study, however, there was still some substantial variability among the lower asymptotes, with a minimum of 0.095 and a maximum of 0.226 (Embretson & Reise, 2000). The findings of the ART are surprising, as the 30 items were generated to fit different previously identified item structures and were all testing the same construct with the same basic item type (Embretson, 1998).

The wide variety of item types within the math achievement tests discussed, which was designed to be a comprehensive examination of a year's worth of math gains, leads one to be less surprised at the high variation in the guessing parameters' estimates because of the different nuances of the math achievement construct (e.g., number sense, algebra, geometry, probability) involved (Lutz & Embretson, 2012), as opposed to the narrowly defined construct of abstract reasoning and the tightly controlled item formats of the ART (Embretson, 1998). It should be noted here that the sample size for both of

the math achievement tests and for the ART were, respectively, 4,000 and 787, both of which are sufficiently large to yield reliable estimates.

For estimated asymptotes much larger than theory would suggest, one could argue that even low ability students are still able to eliminate one or several alternatives before then guessing from those remaining, which is to say that some distractors are more functional than others. Lower-than-expected asymptotes are less easy to explain. An asymptote nearly equal to zero could mean that there is no guessing occurring in the sample, as Maris and Bechger (2009) and Partchev (2009) suggested in their examples and simulations. Another possibility is that there may be too few people at sufficiently low an ability level that are guessing to accurately estimate the asymptote, so there may be a problem of sample sufficiency. A third possibility is that the distractors are functioning too well, and are more attractive to students who know the material and would otherwise correctly answer the item; in this case the difference between the correct and incorrect alternatives is nuanced such that only higher ability students are attracted the incorrect alternative.

To investigate the third scenario of functional distractors, one would need to perform a distractor analysis, including inspection of the biserial correlations for the distractors or examination of the distractor trace plots (Thissen, Steinberg, & Fitzpatrick, 1989). One thing is clear: there is something about those items or their alternatives that make it less likely to answer them correctly by chance. Andrich & Styles (2011) performed one such analysis with a partial credit Rasch model, based upon the hypothesis that not all distractors are equally incorrect when scoring an MC test. Distractors with information, or functional distractors, were identified using the NRM, and then those

alternatives selected at a rate greater than chance were counted as partially correct. The additional information gained by selection of partially correct, as opposed to the wholly correct, alternative can be modeled using the PCM, where the categories are incorrect, partially correct, and correct, with no limit on the number of alternatives included in either of the first two categories (Andrich & Styles, 2011).

In order to successfully handle the people who are of insufficient ability to arrive at either the partially or wholly correct, a minimum probability for successful item completion can be implemented and examinees who fall below that probability on a given item will have their response treated as omitted, rather than have spurious guessing data included in the analysis: in the original paper, a probability cutoff of 0.2, or the reciprocal of the number of alternatives (Andrich & Styles, 2011) was used. The authors concluded that having an item with a functional distractor identified in this manner and scored using a three-point PCM was the same as having two independent, dichotomously scored items where the most correct answer in each was either the wholly or partially correct alternative.

Polytomous models

While a number of polytomous IRT models have been developed over the years, two are appropriate in a discussion of measuring guessing and partial knowledge on standard MC tests. The confidence testing procedures address partial knowledge by having students directly report their confidence in one or all of the alternatives, depending on the scoring method used. Confidence testing requires a change in testing strategy on the part of the examinee, increasing time spent per item due to the additional task and introspective, and may introduce the influence of the psychological variable of

confidence. The partial credit model (Master, 1982) and the nominal response model (Thissen & Steinberg, 1984) are the only two polytomous IRT models included in the current study because they are best suited for addressing guessing and partial knowledge on an otherwise standard MC instrument.

The partial credit model (PCM).

The PCM (Masters, 1982; 1988) is a polytomous IRT model that requires nothing extra on the part of the examinee and can be easily estimated using IRT software like Parscale (Muraki & Bock, 1997). The PCM is a Rasch-family model, satisfying the requirements that person and all item parameters be separable, so sufficient statistics exist in the data for each parameter to be estimated (Masters, 1982). While the PCM can be used for rating-scale type surveys (Masters, 1982), it was more generally developed for items where there are inherent thresholds, or steps, one must successively achieve to maximize points on a given item. A common example is a math item that requires the appropriate application of order of operations, e.g. $2*(4-5)^2 = x$. In this example, the steps one must go through are: Parentheses, 4-5 = -1; Exponents, $-1^2 = 1$; and Multiplication, 2*1 = 2. Thus, three steps are involved in the example item and correctly solving the item depends on both proceeding through the steps in the right order (PEM) and applying the required operation in each step appropriately (addition, exponentiation, multiplication). An item with three steps has a total of three possible points that could be awarded in the following manner, seen in Table 2.

Table 2

Sample scoring for PCM categories

Response	Score
Failed	0
4-5 = -1	1
$-1^2 = 1$	2
2*1 = 2	3

The ability to complete the successive step, having achieved the current step, is rewarded increasing partial credit; the same can be done for MC mathematics items with well-crafted distractors that result from common mistakes at each step. The PCM is appropriate for non-math or step-based solutions, as in Masters' (1982) geography example (p. 151), reproduced in Table 3.

Table 3

Sample item for scoring in PCM

The capital of Australia is		Score
a.	Wellington	1
b.	Canberra	3
c.	Montreal	0
d.	Sydney	2

While there are no steps *per se* involved in the recall that Canberra is the capital of Australia, there is an increasing correctness of the alternatives: Montreal is the capital of Quebec, Canada and is not in or near Australia and is the least correct alternative; Wellington is the capital of New Zealand and so earns one point; Sydney is a major, recognizable city and is actually in Australia, and so earns two points. While some may

argue Table 3 outlines an all-or-nothing question, one can also argue that the recall of relevant facts about Sydney and Wellington in particular, and familiarity with Oceania in general, indicates partial knowledge about Australian geography and should be credited as such.

The nominal response model (NRM)

Like the PCM, the NRM (Thissen & Steinberg, 1984) allows for the measurement of information present in various distractors. However, in the case of the NRM, there is no requirement of ordinal responses. Indeed, the name nominal response model indicates that the responses to the different items may in fact be nominal: in the case of an MC item, this means that the distractors may not be steps toward solving a problem or be subjectively more or less correct than others. Thissen and Steinberg (1984) addressed the issue by noting that, while standard scoring gives points only for the correct alternative when chosen, there is a great information loss when lumping the remaining alternatives together simply as "wrong". The NRM models the information from the distractors without any assumption of order among them. The NRM allows all alternatives of an MC test to be modeled directly as functions of the latent ability of interest (Thissen & Steinberg, 1984). Thus, the NRM is a response to the standard all-or-nothing scoring method of MC tests because it models the probability of nominally scaled alternatives.

The interpretation of the different parameters of the NRM can be a challenge (Thissen & Steinberg, 1984): it is instead the item response curves, and not the parameters, that reveal about the functioning of and information in the distractors over a given trait level, rather than the parameters themselves. Finally, there is a cost to the inclusion of information gained from this initial model, arising from inconsistent ability

estimates given a response near a narrow ability range, where selecting the correct response may actually penalize one's estimate. The use of graphical techniques with the NRM is imperative as it can illustrate such potential pitfalls and inform item design. Baker (1993) demonstrated that vertical and horizontal equating are both possible under the NRM. Further tests can be scored for partial credit and NRM enables identification of informative parameters (e.g., Andrich & Styles, 2011; Penfield, 2008).

Comparison to CTT

The fundamental difference between CTT and IRT lies within the assumptions made for the two paradigms' models' validity. The CTT model of an examinee's performance is a function of the observed raw score for that examinee; indeed the true score Tj for examinee j is the expected value of his raw score, E[Xj]. As the raw score is then a point estimate for the true score, the CTT true score formula is a basic means model, shown in (2.2) (e.g., Crocker & Algina, 2008).

$$T_j = X_j + \varepsilon_j \tag{2.2}$$

In (2.2), ε_j is a random variable representing the error in estimating one's true score with his observed score, which is effectively the bias of the test, normally assumed in CTT to be zero.

If T_j is considered to be analogous to ability, then one's ability estimate, X_j is the total number of items correctly answered on a test. As has been shown with the Rasch family of models (e.g., Andrich, 1988; Masters, 1982; Rasch, 1960), the raw score in IRT contains a lot of information about a person's ability—it is a sufficient statistic for the ability estimate—but CTT does not account for the nature of the items themselves. CTT does estimate item difficulty, which is the sufficient statistic for item difficulty in the

Rasch family of models: the proportion of students correctly answering an item. The CTT item difficulty is often referred to as the item's *p* value. The CTT estimate of item difficulty—or item facility, as higher *p* values indicate easier items—is calculated without consideration for the abilities of the people in the sample.

As Embretson and Reise (2000) illustrate when outlining their "'Rules of Measurement", under CTT, "unbiased estimates of item properties depend on having representative samples" (p. 15), so the consideration for population abilities is included in the assumption of the model, and not in the mathematics of the model itself. However, as IRT models all include simultaneous estimation of person and item characteristics, unbiased estimation of item properties—like difficulty and discrimination—can be obtained with even unrepresentative samples. As unbiased estimates for IRT model-based item parameters can be obtained with variable samples, those items are said to be calibrated and can be used to assess student ability and obtain highly reliable scores for examinees from other populations and samples, through computerized adaptive testing (e.g., Embretson & Reise, 2000). The CTT requirement for test equating is parallel forms and an equal number of items: each examinee has the same *T* for each form and the error variances for each form are equal (Crocker & Algina, 2008). The assumptions of IRT are easier for the practitioner to meet, facilitating test equating even when using heterogeneous sample of examinees. It is for this reason that IRT and other latent trait models are recommended.

Cognitive Diagnostic Models

Cognitive diagnostic models (CDMs) are an alternative to the IRT approach to latent trait modeling. The aim of IRT models is to simultaneously locate persons and

items on the same interval scale for assessment of ability. CDMs also allow for the comparison of persons to items, but on a different basis. Unlike the binary IRT models discussed, which all assume the unidimensionality of the measure, CDMs require an assessment of the different dimensions of an item, be they the result of differing strategies, skills, or steps required to solve the item, as with the PCM. Other dimensions within an item may arise from different cognitive requirements, which may or may not be relevant to the construct. In the following sections, a brief summary of CDMs is described, followed by an introduction of three basic CDMs. General CDMs, which encompass a wide variety of models, are also discussed. An alternate CDM, where the person parameters are latent traits on multiple dimensions instead of latent classes of mastery or non-mastery of skills, is described. The chapter concludes with a comparison between CDMs and IRT models.

Background of CDMs

In unidimensional IRT each person is given a scale score, which is compared against other persons for criterion-referenced testing, and against items for item selection in CAT (e.g., Embretson & Reise, 2000). In cognitive diagnostic modeling, persons are often assessed as masters or non-masters of the item dimensions; the latent scale of IRT is dichotomized on each item dimension, though this is not the case for latent trait models, some of which will also be covered in this chapter. CDMs do not necessarily assess item difficulty directly. Instead, the dimensions are represented in the model via a Q-matrix, which, when properly specified, indicates the pattern of dimension representation on each of the items, and comparisons can be made between dimension mastery and item requirements.

The **Q**-matrix identifies constraints on the model, and its misspecification can have serious repercussions for estimation of both person and item parameters (Rupp & Templin, 2007). For the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model a misspecified **Q-**matrix may make itself evident in poor model fit or in extreme values for the slip and guess parameters (e.g., Embretson & Yang, 2013; Rupp & Templin, 2007). As now the sets of both person parameters and item parameters are dichotomized for the DINA and other classification models, the dimensions of interest in the items will be referred to as components or attributes, the respective presence or mastery of which can be indicated using the binary scoring system for items and persons.

A number of CDMs exist with approaches ranging from the form of logistic item response models (e.g., Embretson & Yang, 2013; Henson, Templin, & Willse, 2009; von Davier, 2005) to cluster analysis (e.g., Chiu, Douglas, & Li, 2009; Nugent, Dean, & Ayers, 2010). As in the coverage of the IRT models, discussion will be limited to only a few CDMs, with focus on those that model guessing either explicitly or implicitly. The core CDMs selected are those that can be parameterized in the framework of the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009). Although some discussion is included in that paper, the current review will include an introduction to those models from the original perspectives as well. The models here discussed fall into one of two categories, compensatory and non-compensatory. In the case of compensatory models, which are represented here by the compensatory reparameterized unified model (C-RUM; Hartz, 2002, as cited in Rupp, Templin, & Henson, 2010), an abundance of one attribute is said to make up for, or compensate, a lack in another for a

given examinee. That is, an item requiring two attributes can still be successfully answered by an examinee who is a master of only one of those attributes.

Non-compensatory models generally assume that all attributes, or specific subsets of attributes, must be mastered for an examinee to successfully answer an item; having one attribute does not compensate for the lack of the others in this case. Mathematics achievement items are often thought to be non-compensatory. Consider, for example, an item that asks "How many ways can a committee of three people be chosen from a group of 5?" The answer, ${}_5C_3 = 10$ ways, is arrived at by correctly setting up and applying the ratio for combinations. This is a non-compensatory item because a student may be perfectly able to perform the required arithmetic, but if he does not recognize that the situation calls for a combination he will not successfully answer the problem: arithmetic does not compensate for a lack of mastery in basic combinatorics.

Core models

The core CDMs were chosen for inclusion in the current paper because they are in the LCDM family of models. Each of the models selected represents one of the noncompensatory or compensatory model classification. The reason for their associated classification will be discussed, along with parameter interpretation and the LCDM equivalent. All of the core CDMs in the current paper and the LCDM are binary skills models (Haertel, 1989), as the attributes are themselves dichotomously scored skills or abilities.

The DINA.

The DINA (Haertel, 1989; Junker & Sijtsma, 2001) is a latent classification model, and a fairly simple one (e.g., Rupp, Templin, & Henson, 2010). The DINA is a

conjunctive, non-compensatory model, meaning that an examinee must possess all skills required of an item to correctly answer an item. Although Haertel (1989) introduced the DINA model, Junker and Sijtsma (2001) provide cleaner notation and parameterization, and so it is their form that is utilized here.

The DINA models both latent and manifest response patterns of an examinee on an item. The latent responses are deterministic: 1 if the examinee has mastered all required attributes and 0 otherwise. Furthermore, all examinees with the same attribute mastery pattern have the same latent response pattern (Haertel, 1989; Junker & Sijtsma, 2001). The "noisy" portion of the model occurs because latent responses are not necessarily reproduced by a manifest response: just because a student should be able to answer an item correctly does not mean he will (Haertel, 1989).

As there are two possible responses to a dichotomous item, so there are two possible mismatches between a manifest response and a latent response, introducing noise into the system because the mismatches are probabilistic. In Haertel's (1989) terms, a false negative occurs when a student has mastered the required attributes but has an incorrect manifest response, which is often mnemonically referred to as a "slip", with probability s_i (Junker & Sijtsma, 2001). There is also the possibility of a false positive, which occurs when a student has not mastered the required attributes but still correctly answers the item. Junker and Sijtsma (2001) refer to the mnemonic "guessing" for false positive, but warn that both a true slip and a true guess are not necessarily represented by the DINA model. Thus, as with the 3PL, the DINA can be said to model something similar to guessing, though whether that is the true strategy involved in such false positives is unknown. Given a calibrated item, the intent of the DINA is to assess the true

latent state of the examinee, accounting for the noisy "and" gates that lead to false positives and false negatives (Haertel, 1989).

The C-RUM

The compensatory reparameterized unified model (C-RUM; Hartz, 2002, as cited in Rupp, Templin, & Henson, 2010) is a CDM that explicitly outlines the additional gain in the probability of successfully answering an item due to mastery of additional attributes required of that item. In the definition of the C-RUM, the mastery of an attribute contributes uniquely to the probability of endorsing an item, independently of other mastered or non-mastered attributes. The C-RUM treats the item-required attributes as mutually exclusive of one another, thus eliminating any need to consider their interactions or intersections. Both CDMs discussed so far involve some sort of lower bound for probability of correct item endorsement, which may or may not reflect the strategy of guessing.

As cognitive diagnostic modeling becomes more popular in testing, a number of general models have been derived in an effort to unify the models that currently exist and to provide a basis for flexible new models to be determined at the item level. For example, von Davier's (2005) general diagnostic models (GDMs) are flexible enough to accommodate both compensatory and non-compensatory models, as well as some IRT models. The GDM encompasses polytomous IRT models, which lifts the restriction of other CDMs that responses and classification be binary assignment. Dimitrov and Atanasov (2012) extended the conjunctive least squares distance model (LSDM; Dimitrov, 2007) into two models with looser restrictions on the relationship between the attributes and items: the LSDM-C, the conjunctive model that looks at patterns or

minimum subsets of item attributes required, and the LSDM-D, a disjunctive version (Dimitrov & Atanasov, 2012) of the original LSDM. Finally, de la Torre (2011) developed the generalized DINA (G-DINA) to address the relationship between the attributes and the items using three different link functions: the identity, the logit, and the log. The G-DINA is a general model that has been shown to include the DINA, DINO, and A-CDM under its umbrella of previously defined models. One notable advantage of such general models is that, once implemented, they allow for easy model comparisons among candidate models on an item-by-item level (e.g., Henson, Templin, & Willse, 2009).

The LCDM

In the interest of containing the scope of the current study, the remainder of the discussion of general CDMs will be limited to the LCDM (Henson, Templin, & Willse, 2009). The LCDM considers the relationship between attributes and the item directly, in the framework of log-linear models with a latent variable, α. The defining feature of such models is that the discrete observations are related to one another only through the latent variable, they are otherwise independent of one another. This is similar to the assumption of conditional independence of items in the IRT framework, in which the relationship between one item and another (or between the selection of one response option over another) is defined entirely by the person parameter. In IRT, it is the items that are conditionally independent; in LCDM, it is the item-required attributes that are conditionally independent (Henson, Templin, & Willse, 2009).

The LCDM is a general CDM that, via reparameterization, can represent the three core models described in the previous sections, as well as a variety of other models, both

defined in the literature and otherwise unspecified. The LCDM can be used for both exploratory and confirmatory purposes, depending on the requirements of the instrument and the theory underlying its use (e.g., Rupp, Templin, & Henson, 2010).

By incorporating the spectrum of CDMs, the LCDM defines a family of diagnostic models and is flexible enough to allow for different item types to be estimated even within a given test. That is, a test may consist either or both of non-compensatory or compensatory items, and the general nature of the LCDM can handle that (Henson, Templin, & Willse, 2009). Furthermore, the saturated LCDM, which contains all main effects and all possible interactions, can be used in an exploratory and theory-driven manner to investigate the behavior of items and their components.

The LCDM can be estimated using an EM algorithm, albeit with some constraints on the number of latent classes (e.g., Rupp, Templin, & Henson, 2010), as well as via Markov chain Monte Carlo (MCMC) methods, with uniform priors on the item parameters and a dichotomized multivariate normal prior on the latent variable side (Henson, Templin, & Willse, 2009). The saturated model is a useful diagnostic in and of itself, but limitations in the current state of the art, as well as computational time, means that for some assessments the full benefits of the model cannot be realized (e.g., Lutz, 2012). As algorithms and processing speeds improve, however, the LCDM will likely prove to be a more useful model across a wider range of applications.

The MLTM-D

The MLTM-D (Embretson & Yang, 2013) is a non-compensatory, hierarchical model. When item attributes can be considered to be finer measures of a larger construct, or component, the MLTM-D is an appropriate diagnostic model to use. Situations where

attributes may be nested within components may arise on broad scale tests of competency or achievement (e.g., Embretson & Yang, 2013; Lutz, 2012). In those cases, due to perhaps a limited test length, each attribute will have only a few items devoted to it, whereas in a more tailored classroom assessment of one construct, more items of a given skill can be included. The MLTM-D, then, identifies the probability of successful item completion to be a function of both the attributes and the higher-level components present in the item. The two levels of item-feature relationships are described by two different scoring matrices, **Q** and **C**.

As in the previous CDMs, the attribute-item relationships, or constraints, are identified by a **Q**-matrix, but in the case of MLTM-D those entries are not restricted to binary scores. For a test measuring M components consisting of K_m components each, if each item is assumed to measure at least one component, there are $B = 2^M\text{-}1$ possible component combinations, or blocks, that the items may be categorized into (Embretson & Yang, 2013). It follows that each block of items must have its own **Q**-matrix to represent the relationship of the K_m attributes for the components defining the items within that block. The item-component relationships are represented in an B x M **C**-matrix, which contains binary indicators of involvement of the mth component on the bth item block.

While the previously discussed CDMs were all restricted latent class models in which person parameters were a probability of attribute mastery or non-mastery, the MLTM-D reduces the parameter estimation load by instead locating the person on each of the higher level components. Thus, the MLTM-D is not a latent class model but is instead a latent trait model, more like the IRT models previously discussed. The two

levels of the MLTM-D are illustrated (2.3) and (2.4), emphasizing the hierarchical nature of the model in defining the probability that examinee *j* correctly answers item *i*.

$$\Pr[X_{ij} = 1] = \prod_{m=1}^{M} P_{ijm}^{c_{im}} \tag{2.3}$$

where

$$P_{ijm} = \Pr[X_{ijm} = 1 \mid \theta_{jm}, \mathbf{q}_{im}, \boldsymbol{\eta}_m] = \frac{1}{1 + \exp\left\{-1.7\left(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}\right)\right\}} \tag{2.4}$$

In (2.4), the coefficient of 1.7 in the exponent is used to scale the normal ogive model to the LTM, η_{mk} is the weight of the *k*th item feature on component *m*, and η_{m0} is the intercept for component *m*. The c_{im} in (2.3) are the elements of the **C** matrix, indicating the involvement of component *m* on item *i*. Equation (2.4) can be interpreted as the probability that examinee *j* correctly responds to the portions of item *i* relating to component *m*, or that examinee *j* has sufficient ability on component *m*.

One can see that the MLTM-D enables the common scaling of items and persons within a component, allowing for the item-person comparison possible in IRT. In (2.4), the attributes comprising the components involved on the item contribute differentially to an item's difficulty, so their location on the component scale can also be compared to a person's latent trait, allowing for diagnosis of what an examinee can and cannot do. Modeling item difficulty is not new to the MLTM-D, as it is the defining feature of the linear logistic test model (LLTM; Fischer, 1973), an extension of the Rasch model (1960) that also used qualitative item features to model item difficulty.

As MLTM-D requires estimation of only *M* person parameters for each examinee, there is a big advantage in estimation over the other CDMs that estimate 2^k latent classes,

corresponding to the attribute mastery patterns. The MLTM-D is preferable for longer tests that are broad in scope and cover a large number of attributes, as only locating persons on the component scale means that there is no cost to adding narrower or more finely defined attributes within the subsuming components (Embretson & Yang, 2013). Additionally, if a **Q**-matrix is not specified for the items within each component, (2.4) simplifies to the Rasch model, where $\beta_{im} = \eta_{m0}$ is the ith item difficulty on component m.

Noncompensatory IRT models

While not often formally referred to as cognitive diagnostic models, several models under the IRT umbrella fit with the CDM paradigm and bear mentioning here. An unnamed early noncompensatory model for dichotomous, multidimensional item, is similar in form to the MLTM-D and the model proposed later in this paper (Sympson, 1977), and is presented in (2.5).

$$\Pr[X_{ij} = 1] = \gamma_i + (1 - \gamma_i) \prod_{m=1}^{M} \frac{1}{1 + \exp\left\{-1.7\alpha_{im}\left(\theta_{jm} - \beta_{im}\right)\right\}} \tag{2.5}$$

One can see the major difference between (2.5) and the MLTM-D is the inclusion of a lower asymptote and the exclusion of the component indicator. In this way, the MLTM-D and the subsequent proposed model are more general, as they allow the dimensionality of the items to vary throughout the test. Other latent trait models can also be construed as CDMs, such as the linear logistic test model (LLTM; Fischer, 1973), the multidimensionsional IRT models (e.g., Hattie, 1981; Reckase, 1997a), the MLTM (Whitely, 1980) and the GLTM (Embretson, 1984). The LLTM, like the MLTM-D, models the item difficulty as a function as item attributes; unlike the MLTM-D, however, it is a unidimensional item and persons and items are still aligned along a single component. The multidimensional IRT models are similar to the MLTM-D in that they

treat person abilities and items as arrayed along multiple dimensions; however, the multidimensional IRT models typically do not model item difficulties as a function of item features. The MLTM and GLTM are the two direct precursors to the MLTM-D.

Comparison to IRT

Cognitive diagnosis models and item response theory models are both classes of latent trait models. All of the models discussed here are full-information models, meaning that the scored item responses are used directly in the estimation of both person and item parameters. A major goal of IRT is to provide a basis for equating across heterogeneous forms and populations, allowing for the reporting of a proficiency or interpretable ability score for the persons. Calibrated IRT items can be "banked" for use in computerized adaptive testing, wherein items selected for presentation to a given examinee are based on a rough estimate of the examinee's ability; further item exposure is based on an examinee's item responses to fine-tune the final estimation of the examinee's ability. IRT items can also be used to equated test scores and person abilities across different forms, when the psychometric properties of those forms are known. However an IRT-based test is delivered the end goal is to understand the behavior of items in the population so that a single proficiency score can be estimated for the test-takers.

Cognitive diagnostic models may also be used to bank items and to compare students' performance across forms. The key difference lies in the level at which said comparisons can be made: for IRT it is at the level of ability; for CDMs it is at both the ability level and the attribute level. That is, CDMs enable comparisons to be made between two respondents with the same overall proficiency level (e.g., Embretson & Yang, 2013). By identifying and measuring the features of items that contribute to item

difficulty (e.g., Embretson & Yang, 2013; Fischer, 1973; Henson, Templin, & Willse, 2009), or by classifying students according to what skills they have or have not mastered (e.g., Haertel, 1989; Junker & Sijtsma, 2001; Templin & Henson, 2006), one can draw distinctions between examinees that would otherwise be treated the same, as either proficient or not-proficient on the construct as a whole. The major benefit of CDMs is the estimation of attribute profiles, so teachers, administrators, parents, and students can see where remediation might best be focused for each individual. Such efforts are already underway at the component level (Embretson & Yang, 2013), and are made easier with the increased access to and speed of computers used in educational testing.

As with IRT, for a test to yield the best diagnostic information about the examinees, the items must be designed with diagnosis in mind. While CDMs have been applied to currently existing assessments (e.g., Embretson & Yang, 2013; Haertel, 1989; Templin & Henson, 2006), like any valid assessment and measurement, the diagnostic test must be grounded in theory (e.g., Gierl & Cui, 2008; Rupp & Templin, 2008) as CDMs are inherently confirmatory in nature. In the sense that CDMs are confirmatory, based on the model constraints outlined in the Q-matrix, poor model fit or suspect parameter estimates can be treated as evidence against the current model specifications (Rupp & Templin, 2008), which goes back to the theory underpinning the design of the items. With that in mind, the Q-matrix should be constructed to reflect only those item-attribute relations theorized to strongly influence item difficulty (e.g., Embretson & Yang, 2013; Henson, Templin, & Willse, 2009). IRT items, however, can be applied to current test forms and the model can be determined based on fit and substantive theory post-hoc. Tests for proficiency must cover a wide range of possible abilities, with items

clustered around areas on the latent scale of interest, such as cutoff points for minimum proficiency. Tests for diagnosis must cover a range of tasks or attributes of interest, and depend on the type of model believed to apply to the item-attribute relationships (Gierl & Cui, 2008).

Henson and Douglas (2005) extensively considered the problem of test construction for diagnosis—specifically reliability and information, two key concepts for developing an IRT-based assessment—within the context of the DINA and several other conjunctive models not discussed in the current review. In that paper, a cognitive diagnostic information index (CDI) was developed to help discriminate between attribute mastery patterns for examinee classification. For IRT, the most information occurs at the item's difficulty, as that is where the item response function in (2.1) has the steepest rate of change. Items with higher discrimination have correspondingly higher information, in terms of the Fisher information, at their location for this reason: the most information about differences between examinees with similar trait levels occurs where the discrimination parameter is fully realized (e.g., Embreston & Reise, 2000), thus the reference to the "discrimination" parameter. The CDI is a measure similar to the Fisher information, but for discrete classes and not a continuous trait. More information is desired and needed to discriminate classifications between people with similar ability patterns, so in terms of Euclidean distance those patterns that are "near" each other will be weighted more heavily in the CDI function than those patterns that are already disparate. Then the principle of selecting items to populate a test based on desired information (e.g., Eignor & Douglass, 1982) can be utilized for diagnostic assessments. using traditional likelihood methods.

Research Proposal

If one is trying to answer the question "is guessing occurring?" neither this paper nor the models discussed can answer with a definitive yes or no. In the context of probabilistic testing, guessing is a risk-seeking behavior so the estimated true scores are a confounding of actual ability and the risk-seeking/risk-averse psychological variable. In formula scoring, where the instructions explicitly state that guessing is not associated with a positive gain in the expected score, one still sees incorrect responses, indicating that the penalty scores for each student are a confounding of guessing in the absence of knowledge and or misinformation. In CTT, guessing is not assessed directly, though it impacts the standard error of measurement, or the reliability, of the instrument by artificially inflating one's true score estimate.

The IRT models and the CDMs discussed do have some means of handling guessing: by including either a guessing parameter, as in the 3PL and DINA, or a reference group probability, as in the LCDM and C-RUM. The polytomous IRT models allow for the possibility of guessing by allowing for alternative strategies toward reaching the correct answer. In the context of the CDMs and IRT, guessing also includes the notion of a false positive (e.g., Haertel, 1989), which may be achieved by other means, such as highly able students selecting an attractive, though incorrect, distractor that would not appeal to students with lower ability. Especially in the case of the 3PL, estimation of the guessing parameter does not always line up with theory, making the claim of purely random guessing on those items less reliable.

Random guessing may not be a constraining definition, it is only relevant for those who are interested in the phenomenon of truly random guessing by low-ability

examinees, or those interested in describing the strategy as an alternative to other approaches to test-taking. Recognizing that the guessing parameter is actually a confounded measurement of guessing and other person- and item-relevant aspects that may generate a false positive on an item may be enough to complete the analysis of interest, either continuing with the models as defined, or accounting for more item variability by the inclusion of additional, valid, and theoretically-derived variables, such that guessing is arguably the only thing remaining in the error.

Proposed model

If one considers the MLTM-D to be sufficiently generalizable to include a so-called guessing parameter, one must also consider where that parameter belongs and what it may look like. At the component level, the MLTM-D resolves to the Rasch model. As the 3PL is a generalization of the Rasch model with unique discrimination and lower asymptote for each item, one conceptualization of a generalized MLTM-D has a lower asymptote at this lower, component-model level. The inclusion of a lower asymptote at the component level (2.6) would mean that an examinee has some non-zero probability of γ_{im} for a positive latent response to the mth component involved in item i. The situation that would therefore arise is not entirely unlike that of the compensatory multidimensional IRT model (Reckase, 1997b; Sijtsma & Junker, 2006). As discussed in Chapter 2, examinees who are lacking in one dimension but are high in another can still perform well overall on an item under a compensatory model. Similarly, an item with higher probabilities of "guessing" at the component level, as in (2.6), would compensate for lower ability on that component, thereby increasing that component level's

probability and, multiplicatively, increasing the overall probability of successful item completion.

$$P_{ijm} = \Pr[X_{ijm}=1 \mid \theta_{jm}, \mathbf{q}_{im}, \boldsymbol{\eta}_m] = \gamma_{im} + (1-\gamma_{im}) \frac{1}{1+\exp\left\{-1.7\alpha_m\left(\theta_{jm} - \sum_{k=1}^{K}\eta_{mk}q_{imk} + \eta_{m0}\right)\right\}} \tag{2.6}$$

The drawback to (2.6) is that the lower asymptote then contributes multiplicatively to the overall probability for successful item completion, resulting in an interpretation that is inconsistent with that of Birnbaum (1968). For an item with c alternatives, Birnbaum claimed that the theoretical minimum probability of a correct response would be $1/c$. If a component-level lower asymptote were based on the number of alternatives presented in the item, the item-level lower minimum probability would, through multiplication, be on the order of $1/c^M$. Unlike with the 3PL, it is hard to conceptualize what value γ_{im} might theoretically take on, again due to the latent nature of the component-level model. Even fixing a common value for γ within or across components does not resolve the interpretability issue, and Birnbaum's (1968) justification for such a lower bound is lost.

The lower asymptote, if truly it were to represent a theoretical minimum probability for a correct MC item response, must be interpretable in that regard. For this reason a more appropriate parameterization of a generalized MLTM-D would take the form of (2.7)

$$\Pr[X_{ij}=1] = \gamma_i + (1-\gamma_i)\prod_{m=1}^{M} P_{ijm}^{c_{im}} \tag{2.7}$$

where P_{ijm} is as originally formulated in (2.4), and repeated in here for ease of reference:

$$P_{ijm} = \Pr[X_{ijm} = 1 \mid \theta_{jm}, \mathbf{q}_{im}, \boldsymbol{\eta}_m] = \frac{1}{1 + \exp\left\{-1.7\alpha_m \left(\theta_{jm} - \sum_{k=1}^{K_m} \eta_{mk} q_{imk} - \eta_{m0}\right)\right\}}$$

A more restricted version of (2.7) may be obtained by constraining the lower asymptote to be equal for all items, such that $\gamma_i = \gamma$; $i = 1, 2, \ldots, I$. For $\gamma_i = 0$, the MLTM-D as described by Embretson and Yang (2013) is obtained. Therefore, the proposed model is the MLTM-D, generalized for guessing, and so will be hereby referred to as the gMLTM-D.

Whether γ_i is unique or common across all items, the lower asymptote in (2.7) is applied to the probability of a positive manifest item response, given the components involved in the item and the examinees' ability relative to those components and the attributes comprising those components. The current study proposes further investigation into gMLTM-D and the implications for estimation of such lower asymptotes. The study will identify testing conditions in which the general model is and is not appropriate. The gMLTM-D should be estimable in a manner similar to the 3PL or 3PL multidimensional IRT model, in which the number of alternatives for each item is used to determine a starting point for the item lower asymptotes.

Hypothesis

It is hypothesized that in cases where guessing does actually occur, the gMLTM-D will result in better item and person ability estimates, both in terms of precision and accuracy, than the MLTM-D. In the absence of guessing, the gMLTM-D resolves to the MLTM-D, so estimates from both models should be fairly similar, and the MLTM-D would likely be recommended for its parsimony.

CHAPTER 3

RESEARCH METHODS

The model proposed in Chapter 2 will be estimated in two Feasibility Studies, outlined in the current chapter, to determine the extent to which further investigation can be undertaken for the resulting proposed study. The current chapter first outlines the parameter estimation methods for items and persons. A description of the constraints on the two feasibility studies already conducted for this proposal is next provided. The chapter concludes with a description of the proposed real data analysis and conditions for the proposed simulation study.

Parameter Estimation

Estimation of both the MLTM-D and gMLTM-D will be conducted via a two-step process: the item parameters γ_i, α_m, and η_{mk}, will be estimated first. The item parameter estimates will then be used to estimate each person's ability vector, θ_j.

Item parameter estimation

The 3PL item parameters can be estimated via the Bock-Aitkin MML-EM procedure (Bock & Aitkin, 1981). MML-EM treats each person as an observational unit, and all examinees with the same item response pattern $\vec{X}$ are grouped together, reducing the effective number of observations involved (Johnson, 2007). That is, MML-EM estimates items based on unique response patterns, $\vec{X}_p$, which are assumed to be a random sample from the population, and the number of examinees with that response pattern, n_p (Embretson & Reise, 2000). The probability of obtaining a response pattern is dependent on the ability levels, θ_q, present in the population, the relative frequency of

those Q ability levels in the populations, and the probability of observing $\vec{X}_p$ on items with parameters $\vec{\beta}$, given those abilities and their frequencies (Embretson & Reise, 2000).

Although abilities are unknown at the outset, in order to hone in on a unique set of parameter estimates, the abilities are assumed to follow a known distribution, typically the standard normal; ability parameters are estimated once the item estimation has stabilized (Birnbaum, 1968; Bock & Aitkin, 1981), discussed in the following section. Rather than assume the latent ability θ_q to be a discrete random variable, as is a requirement of the LCDM-family of models, one may prefer to think of ability as located along a continuum and better represented by a continuous probability distribution. By assuming a known distribution, such as the standard normal, one can choose latent trait levels over which to integrate the response pattern likelihood as a representation of what is actually present in the population (Bock & Aitkin, 1980).

For the gMLTM-D, the probability of obtaining the jth item response pattern, $\mathbf{x}_j$, given ability vector $\boldsymbol{\theta}_j$ is given in (3.1), and based on the assumption of local item independence.

$$\begin{aligned} \Pr[\mathbf{X} = \mathbf{x}_j] &= \int_{\boldsymbol{\theta}} \prod_{i=1}^{I} P_{ij}^{x_{ij}} (1 - P_{ij})^{1-x_{ij}} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned} \tag{3.1}$$

Here, $g(\boldsymbol{\theta})$ is the underlying distribution of person abilities, and is typically assumed to be the multivariate standard normal distribution such that $\boldsymbol{\theta}_j \sim N(\mathbf{0}, \mathbf{I}_M)$. After one has obtained a set of item response patterns for J examinees, the likelihood equation for the component-level attribute weights, η_{mk}, is

$$\frac{\partial L}{\partial \eta_{mk}} = \sum_h \sum_{i=1}^{I} \frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} \frac{\partial P_{ij}}{\partial \eta_{mk}} A\left(\tilde{\boldsymbol{\theta}}_h\right)$$
$$= -1.7\alpha_m \sum_h \sum_{i=1}^{I} \frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} \left(c_{im} q_{imk} \left(P_{ij} - \gamma_i\right)\left(1 - P_{ijm}\right)\right) A\left(\tilde{\boldsymbol{\theta}}_h\right) \tag{3.2}$$

where $\tilde{\boldsymbol{\theta}}_h$ is the *h*th *M*-dimensional quadrature point with weight $A\left(\tilde{\boldsymbol{\theta}}_h\right)$ derived from the standard multivariate normal distribution, and where

$$\tilde{N} = \frac{\sum_{j=1}^{J} L(\mathbf{x}_j \mid \boldsymbol{\theta})}{P(\mathbf{x}_j)} \quad \text{and} \quad \tilde{r}_i = \frac{\sum_{j=1}^{J} x_{ij} L\left(\mathbf{x}_j \mid \boldsymbol{\theta}\right)}{P(\mathbf{x}_j)}$$

represent the expected number of examinees at a given ability level and the expected number of correct responses to the *i*th item for students at that ability level, respectively. The likelihood equations for the remaining item parameters are provided in (3.3) through (3.5).

$$\frac{\partial L}{\partial \eta_{m0}} = 1.7\alpha_m \sum_h \sum_{i=1}^{I} \frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} \left(c_{im} \left(P_{ij} - \gamma_i\right)\left(1 - P_{ijm}\right)\right) A\left(\tilde{\boldsymbol{\theta}}_h\right) \tag{3.3}$$

$$\frac{\partial L}{\partial \gamma_i} = \sum_h \sum_{i=1}^{I} \frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} \left(1 - \prod_{m=1}^{M} P_{ijm}^{c_{im}}\right) A\left(\tilde{\boldsymbol{\theta}}_h\right) \tag{3.4}$$

$$\frac{\partial L}{\partial \alpha_m} = 1.7 \sum_h \sum_{i=1}^{I} \frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} \left(\left(P_{ij} - \gamma_i\right)\left(1 - P_{ijm}\right)\left(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} - \eta_{m0}\right)\right) A\left(\tilde{\boldsymbol{\theta}}_h\right) \tag{3.5}$$

For this paper, item parameters were estimated using the SAS software's NLMIXED procedure, copyright SAS Institute Inc. The estimation method used in the NLMIXED procedure is maximum likelihood with non-adaptive Gaussian quadrature for integral approximations; it differs from the MML-EM algorithm by using the full information matrix instead of estimated response frequencies. Sample source code and model definition is provided in Appendix A and Appendix B.

Person parameter estimation

The M-dimensional person ability vector will be estimated using EAP estimation, under the assumption of a multivariate standard normal distribution, such that $\boldsymbol{\theta}_j \sim N(\mathbf{0}, \mathbf{I}_M)$, utilizing a simple extension of existing SPSS code for estimating MLTM-D abilities (Embretson & Yang, 2013) to account for the inclusion of item asymptotes and component discriminations. SPSS version 21 (SPSS IBM, New York, U.S.A) was used to perform the analysis. In this case, 11 quadrature points from each dimension for the stated distribution for $\boldsymbol{\theta}_j$ were used.

Simulation Design

The simulation was designed to meet several ends: to demonstrate the appropriateness of the gMLTM-D under different test constructions and the accuracy of the parameter estimates from the gMLTM-D relative to the MLTM-D in those situations.

For each test condition, a total number of 40 replications was performed. All tests consisted of $I = 75$ items and $M = 3$ components, and all components were assumed to be independent (i.e., $\boldsymbol{\Sigma} = \mathbf{I}_3$). The same $\mathbf{C}$-matrix was used for all test conditions (Table 4), which was based on having unidimensional items make up 80% of the test, and the multi-component involvement was balanced among the remaining 20% of the items. The proportions of single- to multi-component items was based on the ratio found in the real-world test. The three manipulated conditions are the $\mathbf{Q}$-matrix, the mean lower asymptote, and the sample size. The experimental conditions are outlined in Table 5. Using a full factorial design for the variables at the levels defined in Table 5, there are a

Table 4

***C**-matrix for all simulated tests*

Item	c.1	c.2	c.3
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	1	0	0
11	1	0	0
12	1	0	0
13	1	0	0
14	1	0	0
15	1	0	0
16	1	0	0
17	1	0	0
18	1	0	0
19	1	0	0
20	1	0	0
21	0	1	0
22	0	1	0
23	0	1	0
24	0	1	0
25	0	1	0
26	0	1	0
27	0	1	0
28	0	1	0
29	0	1	0
30	0	1	0
31	0	1	0
32	0	1	0
33	0	1	0
34	0	1	0
35	0	1	0
36	0	1	0
37	0	1	0
38	0	1	0
39	0	1	0

Table 4 Continued

40	0	1	0
41	0	0	1
42	0	0	1
43	0	0	1
44	0	0	1
45	0	0	1
46	0	0	1
47	0	0	1
48	0	0	1
49	0	0	1
50	0	0	1
51	0	0	1
52	0	0	1
53	0	0	1
54	0	0	1
55	0	0	1
56	0	0	1
57	0	0	1
58	0	0	1
59	0	0	1
60	0	0	1
61	1	1	0
62	1	1	0
63	1	1	0
64	1	1	0
65	1	0	1
66	1	0	1
67	1	0	1
68	1	0	1
69	0	1	1
70	0	1	1
71	0	1	1
72	0	1	1
73	1	1	1
74	1	1	1
75	1	1	1

total of 18 different experimental combinations. Throughout the analysis a test condition is referred to as the combination of the **Q**-matrix and mean lower asymptote, for a total of 6 test conditions, regardless of sample size.

Table 5

Simulation study conditions

Variable	Description	Value/Distribution
I	Number of items	75
J	Number of examinees	1,200 3,000 4,800
K	Maximum number of attributes per component	10
M	Maximum number of components per test	3
$\mathbf{C}$	Item-component involvement	see Table 4
$\boldsymbol{\eta}_m$	Attribute weights and intercept for component m	Saturated model: U(-1.8,0) $\forall$ k Attribute model: • U(-1.8,0) for k = 1, 2, …, 6 • U(-1.8, -0.25) for k = 7, 8, 9, 10
$\boldsymbol{\theta}_j$	Ability parameters for person j	MVN($\mathbf{0}$, $\mathbf{I}_M$)
γ_i	Lower asymptote: "guessing" parameter	0 Beta(13.55, 94.83); $\mu_\gamma = 0.125$ Beta(46.6, 139.8); $\mu_\gamma = 0.25$
$\mathbf{Q}_m$*	Item-attribute involvement for component m	$\mathbf{I}_I$ $\mathbf{Q} = [\mathbf{Q}_1 \quad \mathbf{Q}_2 \quad \mathbf{Q}_3]$

* The non-identity $\mathbf{Q}_m$ matrices were randomly generated once and then held constant across other variable test conditions.

Q-matrix

Two different **Q**-matrices were used for the simulations, representing a saturated model and an attribute model. The saturated model corresponds to a **Q**-matrix that results in a Rasch model for each P_{ijm}. That is, an item has a uniquely estimated η_{mk} on every component with which it is involved. Thus, $K_m = I_m$, the number of items involved on

component m. As the same **C**-matrix was used for all tests, by necessity all tests run under the saturated model utilized the same **Q**-matrix.

In this case of the attribute model, $K_m < I_m$, and the η_{mk} are estimated based on the attributes that comprise the components, so that an item's difficulty is dependent not on the item but on the attributes of which it consists. In this case, the **Q**-matrix is necessarily smaller than that of the saturated model. For the purpose of the simulation study, all K_m = 10, and the same $\mathbf{Q} = \left[\mathbf{Q_1} \quad \mathbf{Q_2} \quad \mathbf{Q_3}\right]$ was used for tests run under the attribute model (Table 6 through Table 8). For the attribute model, **Q** was designed to guarantee as equal a representation among the attributes and attribute pairs across the items as possible to ensure good parameter estimation.

Table 6

***Q**-matrix for first component attributes*

Item	q1.1	q1.2	q1.3	q1.4	q1.5	q1.6	q1.7	q1.8	q1.9	q1.10
1	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0
6	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0
11	0	0	0	1	0	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0	0
13	0	0	0	1	0	0	0	0	0	0
14	0	0	0	0	0	1	0	0	0	0
15	1	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	1	0	0	0	0
17	0	0	1	0	0	0	0	0	0	0
18	0	1	0	0	0	0	0	0	0	0
19	0	1	0	0	0	0	0	0	0	0
20	0	0	1	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0

Table 6 Continued

Item	q1.1	q1.2	q1.3	q1.4	q1.5	q1.6	q1.7	q1.8	q1.9	q1.10
40	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0
53	0	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	1	0
62	0	0	0	0	0	0	0	0	0	1
63	0	0	0	0	0	0	0	1	0	0
64	0	0	0	0	0	0	1	0	0	0
65	0	0	0	0	0	0	0	0	1	0
66	0	0	0	0	0	0	0	0	0	1
67	0	0	0	0	0	0	0	1	0	0
68	0	0	0	0	0	0	1	0	0	0
69	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0
72	0	0	0	0	0	0	0	0	0	0
73	0	0	0	0	0	0	0	0	1	0
74	0	0	0	0	0	0	1	0	0	0
75	0	0	0	0	0	0	0	0	0	1

Table 7

***Q**-matrix for second component attributes*

Item	q2.1	q2.2	q2.3	q2.4	q2.5	q2.6	q2.7	q2.8	q2.9	q2.10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0
21	0	0	1	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0
23	0	1	0	0	0	0	0	0	0	0
24	0	0	0	1	0	0	0	0	0	0
25	0	0	1	0	0	0	0	0	0	0
26	0	0	0	0	1	0	0	0	0	0
27	0	0	0	0	1	0	0	0	0	0
28	0	0	1	0	0	0	0	0	0	0
29	0	0	0	0	0	1	0	0	0	0
30	0	0	0	0	0	1	0	0	0	0
31	0	0	0	1	0	0	0	0	0	0
32	0	0	0	1	0	0	0	0	0	0
33	0	0	0	0	1	0	0	0	0	0
34	0	0	0	0	1	0	0	0	0	0
35	0	1	0	0	0	0	0	0	0	0
36	1	0	0	0	0	0	0	0	0	0
37	0	1	0	0	0	0	0	0	0	0
38	0	0	0	0	0	1	0	0	0	0
39	1	0	0	0	0	0	0	0	0	0
40	0	0	1	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0

Table 7 Continued

Item	q2.1	q2.2	q2.3	q2.4	q2.5	q2.6	q2.7	q2.8	q2.9	q2.10
42	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	0
53	0	0	0	0	0	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0	0	0
60	0	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	1	0	0
62	0	0	0	0	0	0	0	0	0	1
63	0	0	0	0	0	0	1	0	0	0
64	0	0	0	0	0	0	0	1	0	0
65	0	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0	0
67	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	1	0	0
70	0	0	0	0	0	0	0	0	1	0
71	0	0	0	0	0	0	0	0	0	1
72	0	0	0	0	0	0	0	0	1	0
73	0	0	0	0	0	0	1	0	0	0
74	0	0	0	0	0	0	0	0	0	1
75	0	0	0	0	0	0	0	0	1	0

Table 8

***Q**-matrix for third component attributes*

Item	q3.1	q3.2	q3.3	q3.4	q3.5	q3.6	q3.7	q3.8	q3.9	q3.10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	1	0	0	0	0	0

Table 8 Continued

Item	q3.1	q3.2	q3.3	q3.4	q3.5	q3.6	q3.7	q3.8	q3.9	q3.10
42	1	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	1	0	0	0	0
44	0	0	1	0	0	0	0	0	0	0
45	0	1	0	0	0	0	0	0	0	0
46	0	0	0	1	0	0	0	0	0	0
47	1	0	0	0	0	0	0	0	0	0
48	0	0	1	0	0	0	0	0	0	0
49	0	0	0	0	1	0	0	0	0	0
50	0	1	0	0	0	0	0	0	0	0
51	1	0	0	0	0	0	0	0	0	0
52	0	0	1	0	0	0	0	0	0	0
53	0	0	0	1	0	0	0	0	0	0
54	0	0	0	0	0	1	0	0	0	0
55	0	0	0	0	1	0	0	0	0	0
56	0	0	0	0	0	1	0	0	0	0
57	0	1	0	0	0	0	0	0	0	0
58	0	0	0	0	1	0	0	0	0	0
59	1	0	0	0	0	0	0	0	0	0
60	0	0	0	1	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0	0
63	0	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	1	0	0
66	0	0	0	0	0	0	0	0	1	0
67	0	0	0	0	0	0	0	1	0	0
68	0	0	0	0	0	0	1	0	0	0
69	0	0	0	0	0	0	1	0	0	0
70	0	0	0	0	0	0	0	0	0	1
71	0	0	0	0	0	0	0	1	0	0
72	0	0	0	0	0	0	0	0	1	0
73	0	0	0	0	0	0	0	0	0	1
74	0	0	0	0	0	0	0	0	1	0
75	0	0	0	0	0	0	1	0	0	0

Lower asymptote

For all test conditions, the most general version of the gMLTM-D was implemented, so the lower asymptote was simulated and then estimated at the item level. The different levels of γ_i were chosen to reflect the absence of guessing (i.e., $\gamma_i = 0$, $i = 1, 2, \ldots, 75$) as well as the theoretical lower bounds for two common forms of MC items: 4- and 8- alternatives. The Beta distributions used to generate those lower asymptotes have corresponding means of 0.125, and 0.25, with variances of 0.001 for each distribution, or B(13.55, 94.83) and B(46.6, 139.8), respectively.

Sample

The sample sizes were based on the rule of thumb that 1,200 examinees are required for the 3PL unidimensional IRT model and the subsequent choices of 3,000 and 4,800 were equally spaced so that polynomial contrasts could be estimated across sample sizes if desired. The sample size of 4,800 was chosen as the minimum size required to reliably obtain estimates for all non-zero lower asymptotes, as determined by pilot studies.

Data generation

The simulation was conducted entirely in R versions 2.14.0 (R Development Core Team, 2011). The code for simulating item responses was adapted from that found in the MAT package (Choi, 2011), and all R code can be found in Appendix C. All η_{mk} for the saturated models were randomly generated from a uniform distribution, U(-1.8, 0), yielding relatively easy items (average $P_{i|j=0} = 0.822$ for single-component items), with new parameters generated for each replication and each test condition. For the three attribute model test conditions, the single-component η_{mk} were randomly generated from

the U(-1.8, 0) distribution. Empirical results have demonstrated that multi-component items are easier than single-component items; to mimic that, the η_{mk} for the multi-component attributes were drawn from the U(-1.8, -0.25) distribution. New parameters were generated for each replication and each test condition.

All person abilities were assumed to be independent, mirroring the independence of item components. Person abilities were therefore drawn from the MVN(**0**, **I**) distribution, with new abilities sampled for each replication and each test condition. In the end, a total of 720 tests were simulated.

Analysis

All tests were simulated under the gMLTM-D, but item parameters were estimated using both gMLTM-D and MLTM-D models, so each simulated test had two sets of item parameter estimates, enabling comparison and facilitating conclusions as to when the gMLTM-D is appropriate. This comparison is be similar to that of the 3PL and Rasch model of Maris and Bechger (2009) discussed in Chapter 2. The two models will be evaluated according to several criteria, outlined in Table 9.

Table 9

Analysis criteria

Criterion	Description
1	Root mean squared error (RMSE) of item parameter estimates
2	Bias of item parameter estimates
3	Bias-adjusted RMSE of item parameter estimates
4	Correlation of item parameter estimates with true values
5	RMSE of ability parameter estimates for a small sample of tests
6	Bias of ability parameter estimates for a small sample of tests
7	Correlation of ability parameter estimates with their true values

The models were also evaluated by comparing the root mean squared errors (RMSE) against the root mean squared standard errors (RSSE) of the estimates, and is useful in determining the precision of an estimate.

RMSE is a function of both variance and bias, and is calculated as the root mean squared deviation between parameters and their estimators for each replication, *r*, where *r* = 1, 2, …, 40 in the simulation study. Using the attribute weights as an example, the RMSE for replication *r* is calculated using (3.6), where *m* is the component, *k* is the attribute within the component, and K_m is the maximum number of attributes on component *m*..

$$RMSE_r = \sqrt{\frac{\sum_{k=1}^{K_m} \left(\hat{\eta}_{rmk} - \eta_{rmk}\right)^2}{K_m}} \tag{3.6}$$

The empirical bias was also calculated to complement the RMSE; while RMSE is a measure of an estimator's precision, bias indicates the presence and direction of inaccuracy in an estimator when it exists, as illustrated in (3.7) and again using attribute weights as an example.

$$\hat{\eta}_{rmk} = \eta_{rmk} + BIAS_r(\hat{\eta}_{mk}) + e_{rmk} \tag{3.7}$$

The empirical bias was also calculated to complement the RMSE; while RMSE points toward precision, bias indicates the presence and direction of any inaccuracy in an estimate, when any inaccuracy exists. The empirical bias for an estimator was calculated using equation (3.8), in which the mean of the simulated "true" values is subtracted from the mean of the estimated values.

$$Bias_r(\hat{\eta}_{mk}) = \frac{\sum_{k=1}^{K_m} \hat{\eta}_{rmk} - \eta_{rmk}}{K_m} \tag{3.8}$$

If an estimator is biased, per (3.7), one can adjust it by re-centering the estimator around the true value by subtracting off the bias. This is done in (3.9): for a better estimate of precision, the bias was parceled out from the estimate when calculating the RMSE, resulting in a bias-adjusted RMSE (RMSEadj), which was also compared against the RSSE.

$$RMSEadj_r = \sqrt{\frac{\sum_{k=1}^{K_m} \left[\left(\hat{\eta}_{rmk} - BIAS_r(\hat{\eta}_{mk}) \right) - \eta_{rmk} \right]^2}{K_m}} \tag{3.9}$$

Numerical and graphical summaries of the criteria were examined, followed by statistical tests to determine whether and where significant differences occurred. With only 40 replications at each design point, the hypothesis tests are particularly important for drawing statistical conclusions about any difference between the models, sample sizes, asymptote levels, or attribute types.

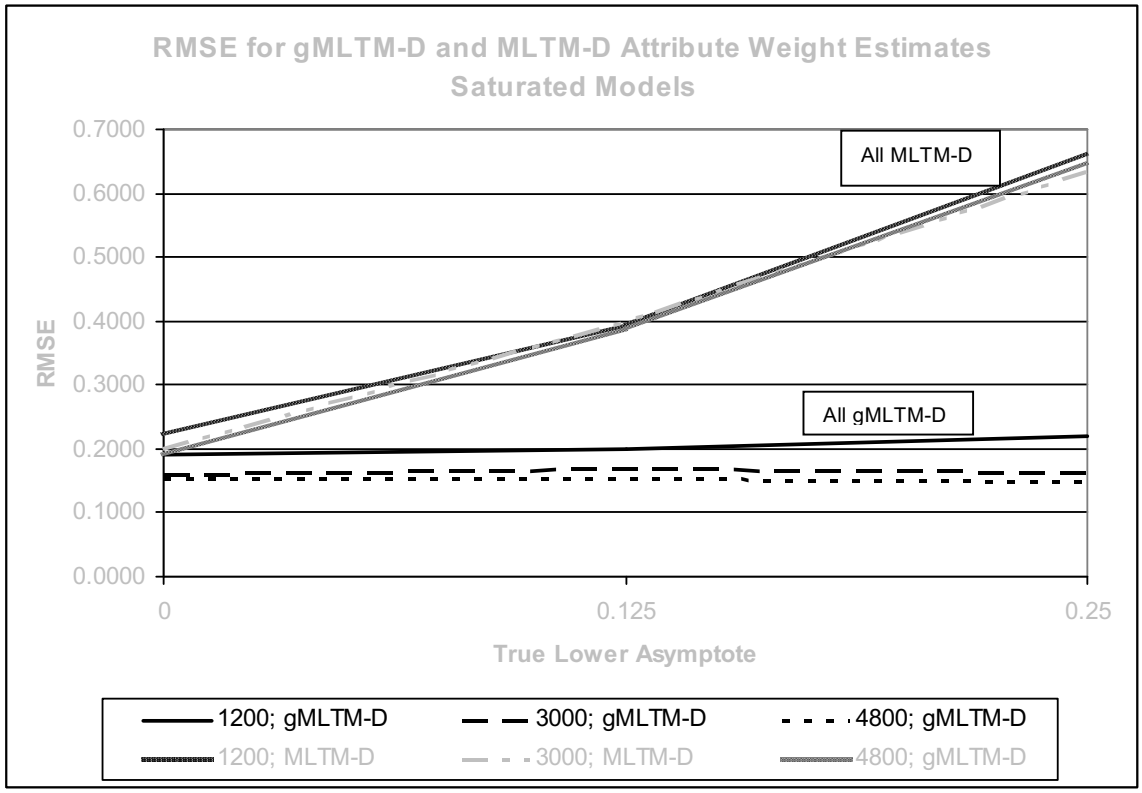
CHAPTER 4

RESULTS

This chapter begins with a presentation of some global results to provide perspective for subsequent results. The major findings for comparisons between the gMLTM-D and MLTM-D are presented, then the seven analytical criteria are covered in turn.

Global Results

To establish a basis for comparing the relative utility of the gMLTM-D and the MLTM-D, the RMSE of the estimated attribute weights, η_{mk}, are illustrated in Figure 1 and Figure 2; one can see the RMSE values increase as the lower asymptote increases for the MLTM-D estimates, while the RMSE values stay relatively constant for the gMLTM-D estimates; the RSSEs are smaller than both model RMSEs. The RMSE and RSSE are are higher for both models for the smaller sample sizes, which is consistent with statistical theory. There is clear visual evidence for a difference in the precision of the attribute weight estimates from the two model specifications, particularly for different levels of guessing.



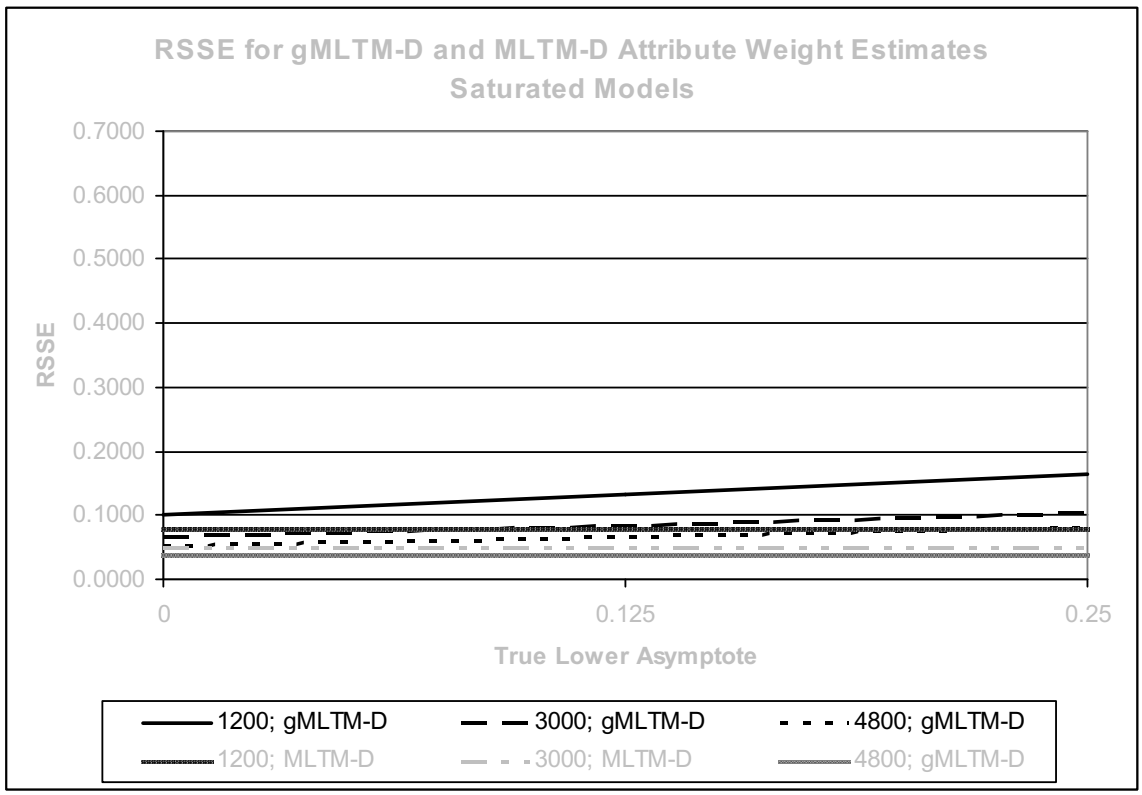
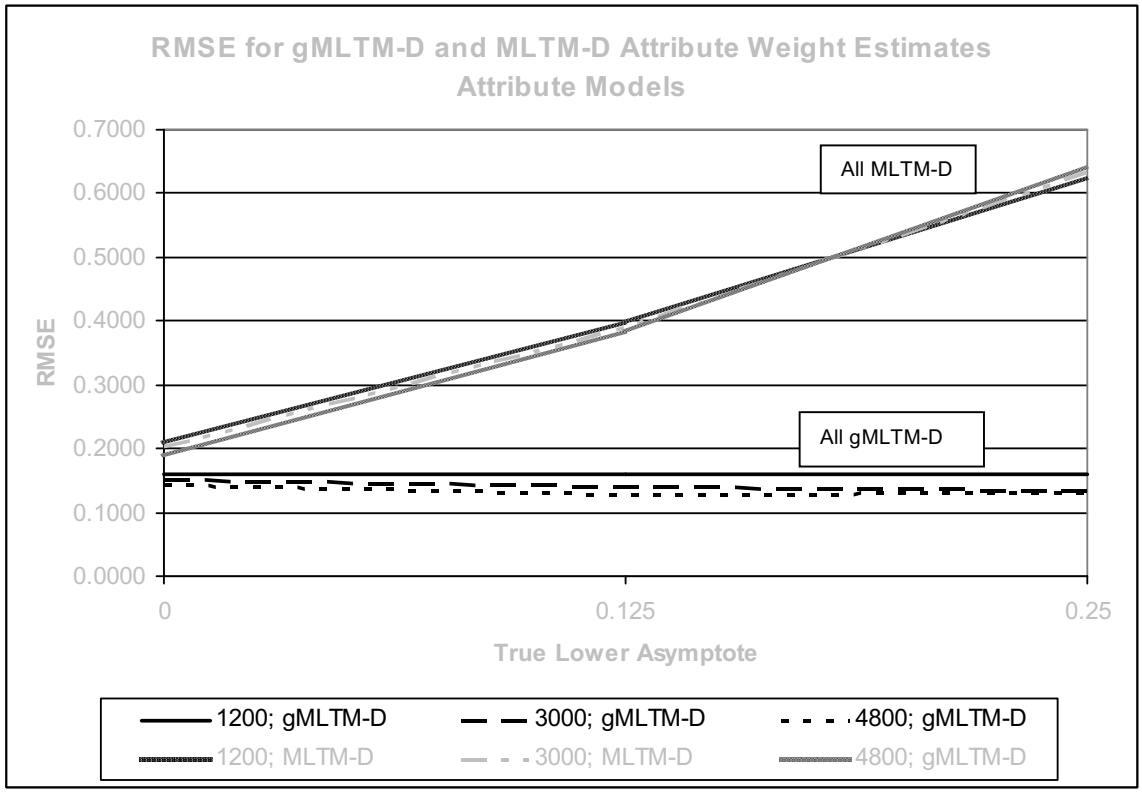


Figure 1. RMSE and RSSE for attribute weights of saturated models for gMLTM-D and MLTM-D estimates.



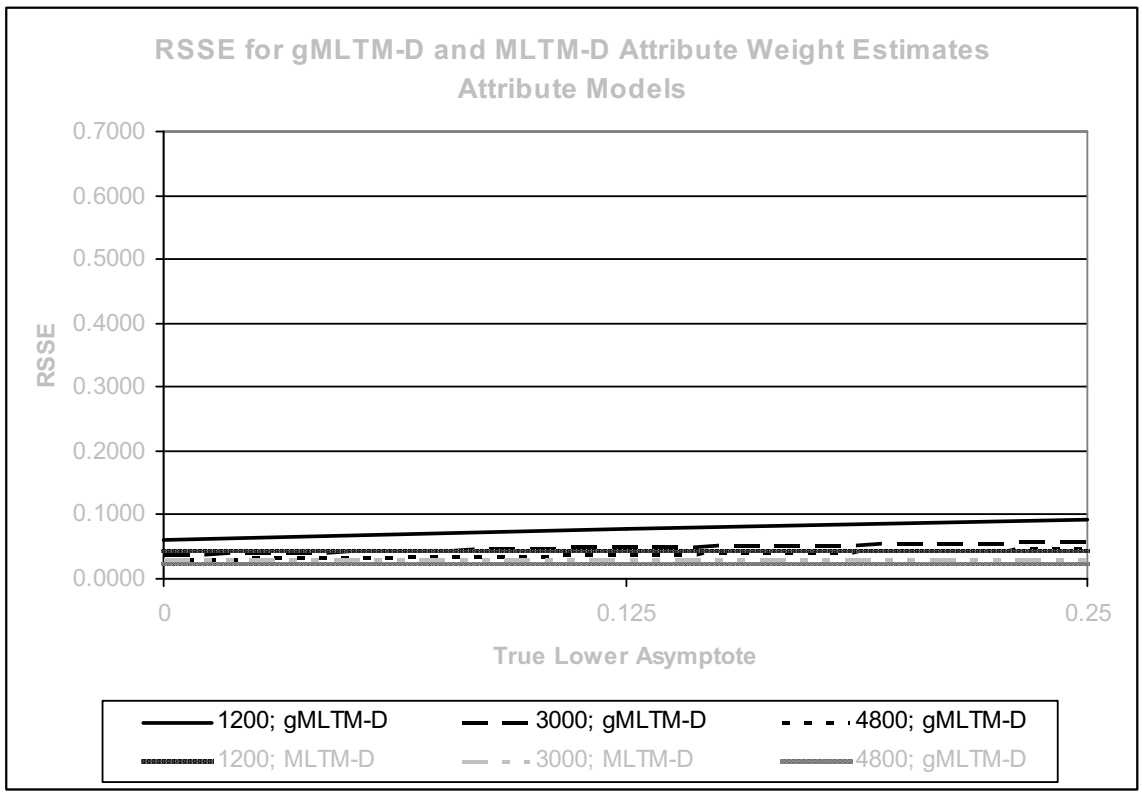


Figure 2. RMSE and RSSE for attribute weights of attribute models for gMLTM-D and MLTM-D estimates.

As RMSE is influenced in part by the bias of the parameter estimates, and the MLTM-D estimates were expected to be biased for increasing levels of guessing, the same plots are provided for the bias of the parameter estimates in Figure 3. One can see from inspection of the plots that the MLTM-D parameter estimates are, and increasingly so for higher levels of the mean lower asymptote. The two models perform approximately equally, in terms of precision and bias, when there is no guessing involved. This is to be expected, as the gMLTM-D resolves to the MLTM-D.

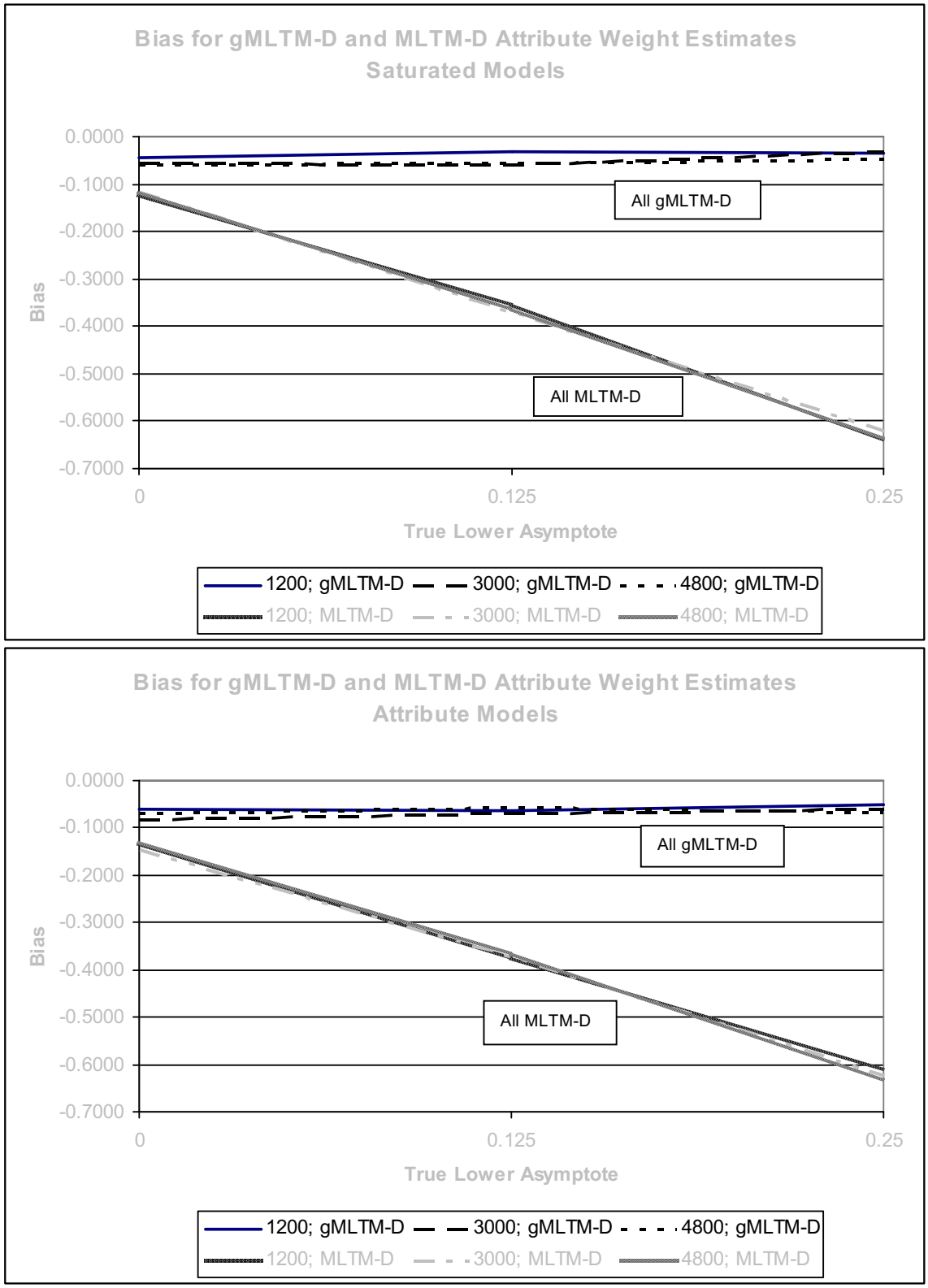


Figure 3. Bias for attribute weights of saturated and attribute models for gMLTM-D and MLTM-D estimates.

Item Parameter Results

The results for the item and component parameter estimates will be discussed in this section. The criteria for evaluating the three levels of item parameter estimates are RMSE, bias, bias-adjusted RMSE, and the correlation between the parameter estimates and simulated values. For each criterion, an ANOVA of the results is first presented, indicating whether there are any significant relationships to investigate. As the gMLTM-D and MLTM-D were estimated from the same simulated test results, that model comparison is a repeated measure, and sample size, lower asymptote, and attribute type are between effects, yielding a mixed-effect design. The repeated-measures tests are reported first, as it is the comparison between the gMLTM-D and MLTM-D that are of primary interest, and the tests include all interactions with sample size, lower asymptote, and attribute type. The remaining comparisons are between-subjects, where the factors are the three levels of sample size (random effect), the three levels of the true lower asymptote (random effect), and the two attribute types (fixed effect): the study looked at main effects and all two-way interactions due to these factors, and the between-subjects effects are reported second. A summary table of the criterion means are then presented and discussed. Each of the item parameters is considered in turn.

Component discriminations

The simulated component discriminations were all set to unity: the tests at the component level were generated from the Rasch model for all saturated model runs. The results for the four evaluation criteria will be discussed in turn in the following sections.

<u>RMSE</u>

A summary of the significant findings is provided in Table 10 and Table 11, which indicates that there are significant main effects due to model (i.e., gMLTM-D vs. MLTM-D), sample size, and attribute type (i.e., attribute or saturated model) on RMSE for component discrimination. Additionally, the effects of sample size, lower asymptote, and attribute type are all impacted by the model used to estimate the item parameters, as they all interact significantly with model. There are significant main effects for each of sample size, attribute type, and lower asymptote.

Table 10

RMSE: Tests for repeated measures contrasts for component discrimination estimates

Source	df	SS	MS	F_{obs}	*p*-value	η^2_p
Model	1	7.927	7.927	90289.7	<0.001	0.992
Model * Sample Size	2	0.004	0.002	24.489	<0.001	0.065
Model * Asymptote	2	2.874	1.437	16370.27	<0.001	0.979
Model * Attribute Type	1	0.009	0.009	101.238	<0.001	0.125
Model * Sample Size * Asymptote	4	0.001	<0.001	2.155	0.072	0.012
Model * Sample Size * Attribute Type	2	0.001	0.001	6.073	0.002	0.017
Model * Asymptote * Attribute Type	2	0.005	0.003	30.293	0.002	0.079
Error(Model)	706	0.062	8.78E-05	--	--	--

Table 11
RMSE: Tests for between-subjects effects for component discrimination estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	1	35.344	35.344	305233	<0.001	0.998
Sample Size	2	0.004	0.002	17.875	<0.001	0.048
Asymptote	2	0.065	0.032	279.163	<0.001	0.442
Attribute Type	1	0.001	0.001	12.924	<0.001	0.018
Sample Size * Asymptote	4	<0.001	4.17E-05	0.36	0.837	0.002
Sample Size * Attribute Type	2	2.24E-05	1.12E-05	0.097	0.908	0
Asymptote * Attribute Type	2	<0.001	9.91E-05	0.856	0.425	0.002
Error	706	0.082	0	--	--	--

Table 12 contains the average RMSE and RSSE of the component discriminations recovered from the gMLTM-D and MLTM-D model estimates. The mean RMSE values tend to decrease for the gMLTM-D as the lower asymptote increases, and the RMSEs are approximately equal for the saturated and attribute models simulated under the same guessing condition for the gMLTM-D. The same decreasing pattern is observed for each of the three sample sizes. The mean RMSE increases slightly as the sample size increases, but it is relatively stable within a test condition for the gMLTM-D.

All RMSEs for the gMLTM-D discriminations are lower than those for the corresponding MLTM-D estimates. In general, the RMSEs for the MTLM-D component discriminations display a different pattern, and increase as the lower asymptote decreases. Like those for the gMLTM-D, the mean RMSE with test condition increase slightly as sample size increases, but are relatively stable: the major change occurs within sample size as the level of guessing increases for the saturated- or attribute-type model. The RMSE for both the gMLTM-D and MLTM-D discrimination estimates are inflated relative to the RSSE, or the root mean squared standard errors of the parameter estimates,

indicating that the recovery of the true parameter values are less precise than the parameter estimates on average.

Table 12

Average RMSE and RSSE of component discrimination estimates

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	Mean RMSE	1,200	0.1879	0.1595	0.1328	0.1601	0.1919	0.1727	0.1402	0.1683
		3,000	0.1966	0.1707	0.1357	0.1677	0.1979	0.1754	0.1462	0.1731
		4,800	0.1978	0.1708	0.1412	0.1699	0.2033	0.1747	0.1484	0.1755
	Mean RSSE	1,200	0.0180	0.0246	0.0312	0.0246	0.0173	0.0232	0.0291	0.0232
		3,000	0.0114	0.0153	0.0195	0.0154	0.0111	0.0146	0.0182	0.0146
		4,800	0.0090	0.0122	0.0153	0.0122	0.0087	0.0116	0.0143	0.0116
gMLTM-D Mean RMSE			**0.1941**	**0.1670**	**0.1366**	**0.1659**	**0.1977**	**0.1743**	**0.1449**	**0.1723**
gMLTM-D Mean RSSE			**0.0128**	**0.0173**	**0.0220**	**0.0174**	**0.0124**	**0.0165**	**0.0205**	**0.0165**
MLTM-D	Mean RMSE	1,200	0.2193	0.2782	0.3234	0.2736	0.2207	0.2799	0.3138	0.2715
		3,000	0.2202	0.2811	0.3215	0.2743	0.2221	0.2801	0.3188	0.2736
		4,800	0.2190	0.2816	0.3247	0.2751	0.2251	0.2823	0.3207	0.2760
	Mean RSSE	1,200	0.0147	0.0152	0.0155	0.0152	0.0146	0.0150	0.0156	0.0151
		3,000	0.0093	0.0095	0.0099	0.0096	0.0092	0.0096	0.0098	0.0095
		4,800	0.0073	0.0075	0.0077	0.0075	0.0073	0.0075	0.0077	0.0075
MLTM-D Mean RMSE			**0.2195**	**0.2803**	**0.3232**	**0.2743**	**0.2226**	**0.2808**	**0.3177**	**0.2737**
MLTM-D Mean RSSE			**0.0105**	**0.0108**	**0.0111**	**0.0108**	**0.0104**	**0.0107**	**0.0110**	**0.0107**

Bias

As with the RMSEs, there are significant differences in the mean bias, as illustrated in Table 13. Most notably, there is again a significant effect due to model, and the same two-way interactions are also significant, indicating that the model specification influences the bias of the component discrimination estimates beyond the simple effects of each of the lower asymptote, attribute type, and sample size. Table 14 shows that, as with RMSE, the sample size, lower asymptote, and attribute type all significantly impact the bias of the component discrimination estimates.

The average empirical bias of the component discriminations is summarized in Table 15. As the lower asymptote increases, the bias for the gMLTM-D discrimination estimates decreases, regardless of sample size. The same pattern holds for both the saturated model and the attribute models. All gMLTM-D estimates are, on average, less biased than the corresponding estimates from the MLTM-D. A similar pattern to the RMSE emerges, where bias sizably increases for the MLTM-D estimates as the lower asymptotes increase. For both saturated and attribute models there is evidence of some increase in bias for gMLTM-D and MLTM-D estimates as sample size increases. These findings are consistent with what was predicted; the MLTM-D would yield biased item parameter estimates in the presence of guessing. The gMLTM-D, because it models the possibility of guessing, would be less biased in other item parameter estimates.

Table 13

Bias: Tests for repeated measures contrasts for component discrimination estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Model	1	8.04	8.035	88633.73	<0.001	0.99
Model * Sample Size	2	0.01	0.003	32.52	<0.001	0.08
Model * Asymptote	2	2.91	1.457	16070.08	<0.001	0.98
Model * Attribute Type	1	0.01	0.009	104.596	<0.001	0.13
Model * Sample Size * Asymptote	4	<0.001	<0.001	1.448	0.217	0.01
Model * Sample Size * Attribute Type	2	<0.001	0.001	7.036	0.001	0.02
Model * Asymptote * Attribute Type	2	0.01	0.003	31.836	<0.001	0.08
Error(Model)	706	0.06	9.07E-05	--	--	--

Table 14

Bias: Tests for between-subjects effects for component discrimination estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	1	35.067	35.067	295423.39	<0.001	0.998
Sample Size	2	0.006	0.003	23.486	<0.001	0.062
Asymptote	2	0.061	0.03	255.489	<0.001	0.420
Attribute Type	1	0.001	0.001	11.901	0.001	0.017
Sample Size * Asymptote	4	<0.001	4.21E-05	0.354	0.841	0.002
Sample Size * Attribute Type	2	1.41E-05	7.03E-06	0.059	0.943	0.000
Asymptote * Attribute Type	2	<0.001	8.98E-05	0.757	0.479	0.002
Error	706	0.084	<0.001	--	--	--

Table 15

Mean bias of component discrimination estimates

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1,200	-0.1871	-0.1577	-0.1290	-0.1579	-0.1909	-0.1709	-0.1372	-0.1663
	3,000	-0.1962	-0.1697	-0.1342	-0.1667	-0.1974	-0.1745	-0.1449	-0.1722
	4,800	-0.1974	-0.1701	-0.1404	-0.1693	-0.2028	-0.1739	-0.1473	-0.1747
gMLTM-D Mean		**-0.1936**	**-0.1659**	**-0.1345**	**-0.1647**	**-0.1970**	**-0.1731**	**-0.1431**	**-0.1711**
MLTM-D	1,200	-0.2187	-0.2777	-0.3226	-0.2730	-0.2199	-0.2791	-0.3126	-0.2705
	3,000	-0.2200	-0.2808	-0.3210	-0.2739	-0.2217	-0.2797	-0.3183	-0.2732
	4,800	-0.2188	-0.2814	-0.3243	-0.2748	-0.2249	-0.2819	-0.3199	-0.2755
MLTM-D Mean		**-0.2192**	**-0.2800**	**-0.3226**	**-0.2739**	**-0.2221**	**-0.2802**	**-0.3169**	**-0.2731**

Bias-adjusted RMSE

The change in the bias noted for both the MLTM-D and the gMLTM-D model estimates impacts the chosen estimate of precision. As seen in Table 16, there is still a significant main effect due to model on the precision of the estimates once bias has been accounted for; model also still interacts significantly with each of sample size, asymptote, and attribute type. Sample size, lower asymptote, and attribute type all still have significant main effects, averaged over the source of the estimates, as seen in Table 17. While adjusting the RMSE for the component discriminations for the bias makes them more precise, it does not eliminate the effects due to the different aspects of the simulation study.

The average bias-adjusted RMSE values are provided in Table 18. The removal of bias from the RMSEs yields values much closer to the RSSEs calculated from the standard errors of the model estimates for both models. The mean bias-adjusted RMSE increases as the mean lower asymptote increases for both the gMLTM-D and MLTM-D discrimination estimates. Additionally, as the sample size increases, the mean bias-adjusted RMSE tend to decrease within a given test condition for both models, which is consistent with statistical theory.

Table 16

Bias-adjusted RMSE: Tests for repeated measures contrasts for component discrimination

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Model	1	0.003	0.003	46.808	<0.001	0.062
Model * Sample Size	2	0.001	0.001	8.849	<0.001	0.024
Model * Asymptote	2	0.001	<0.001	6.727	0.001	0.019
Model * Attribute Type	1	<0.001	<0.001	6.323	0.012	0.009
Model * Sample Size * Asymptote	4	0.001	<0.001	3.258	0.012	0.018
Model * Sample Size * Attribute Type	2	<0.001	<0.001	1.553	0.212	0.004
Model * Asymptote * Attribute Type	2	<0.001	<0.001	1.655	0.192	0.005
Error(Model)	706	0.052	7.39E-05	--	--	--

Table 17

Bias-adjusted RMSE: Tests for between-subjects effects for component discrimination estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	1	0.173	0.173	3138.951	<0.001	0.816*
Sample Size	2	0.007	0.003	59.46	<0.001	0.144*
Asymptote	2	0.005	0.003	49.675	<0.001	0.123*
Attribute Type	1	0.001	0.001	10.247	0.001	0.014*
Sample Size * Asymptote	4	<0.001	5.71E-05	1.038	0.386	0.006
Sample Size * Attribute Type	2	<0.001	7.48E-05	1.36	0.257	0.004
Asymptote * Attribute Type	2	9.97E-06	4.99E-06	0.091	0.913	<0.001
Error	706	0.039	5.50E-05	--	--	--

Table 18

Average bias-adjusted RMSE and RSSE of component discrimination estimates

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	Mean Adjusted RMSE	1,200	0.0157	0.0214	0.0266	0.0213	0.0173	0.0227	0.0257	0.0219
		3,000	0.0101	0.0167	0.0180	0.0149	0.0129	0.0151	0.0172	0.0151
		4,800	0.0100	0.0130	0.0134	0.0121	0.0118	0.0146	0.0162	0.0142
	Mean RSSE	1,200	0.0180	0.0246	0.0312	0.0246	0.0173	0.0232	0.0291	0.0232
		3,000	0.0114	0.0153	0.0195	0.0154	0.0111	0.0146	0.0182	0.0146
		4,800	0.0090	0.0122	0.0153	0.0122	0.0087	0.0116	0.0143	0.0116
gMLTM-D Adjusted RMSE Mean			**0.0119**	**0.0170**	**0.0193**	**0.0161**	**0.0140**	**0.0175**	**0.0197**	**0.0171**
gMLTM-D RSSE Mean			**0.0128**	**0.0173**	**0.0220**	**0.0174**	**0.0124**	**0.0165**	**0.0205**	**0.0165**
MLTM-D	Mean Adjusted RMSE	1,200	0.0132	0.0147	0.0199	0.0159	0.0162	0.0190	0.0241	0.0198
		3,000	0.0086	0.0124	0.0155	0.0122	0.0113	0.0123	0.0154	0.0130
		4,800	0.0093	0.0094	0.0149	0.0112	0.0098	0.0128	0.0202	0.0143
	Mean RSSE	1,200	0.0147	0.0152	0.0155	0.0152	0.0146	0.0150	0.0156	0.0151
		3,000	0.0093	0.0095	0.0099	0.0096	0.0092	0.0096	0.0098	0.0095
		4,800	0.0073	0.0075	0.0077	0.0075	0.0073	0.0075	0.0077	0.0075
MLTM-D Adjusted RMSE Mean			**0.0104**	**0.0122**	**0.0168**	**0.0131**	**0.0124**	**0.0147**	**0.0199**	**0.0157**
MLTM-D RSSE Mean			**0.0105**	**0.0108**	**0.0111**	**0.0108**	**0.0104**	**0.0107**	**0.0110**	**0.0107**

Correlation with true values

The simulated component discriminations were all set to unity, so they did not have any variance. Therefore, a correlation between the simulated and estimated component discriminations cannot be calculated, nor would such a statistic be valuable in this instance.

Discrimination summary

The MLTM-D estimates of discrimination are consistently more biased than those of the gMLTM-D. Furthermore, the accuracy of the MLTM-D estimates grows worse as the lower asymptote increases, for both the saturated and attribute models. The gMLTM-D estimates were fairly stable, regardless of sample size. The accuracy of the discrimination estimates improved on average for the gMLTM-D improved as the mean lower-asymptote increased, indicating that the gMLTM-D has better discrimination parameter recovery than the MLTM-D, particularly when there is an increasing chance of false positives on an item.

Attribute weights

The attribute weights for both the saturated and attribute models were generated from a Uniform distribution with a mean less than zero to simulate an achievement test with relatively easy items. It was anticipated that increased guessing would increase the bias in MLTM-D attribute weight estimates, which are directly linked to an item's component difficulty.

RMSE

All effects involving the difference between the gMLTM-D and MLTM-D estimates are significant, including two- and three-way interactions with sample size,

attribute type, and lower asymptote level, indicating that the precision of the attribute weight estimates are influenced by the model used for estimating the item parameters (Table 19). There are also significant main effects due to sample size, asymptote, and attribute type, and the two way interactions involving attribute type, as seen in Table 20. The RMSE of the attribute weight estimates, therefore, is influenced by a variety of design factors.

The RMSE and RSSE values for the attribute weights are provided in Table 21. A similar pattern emerges for the attribute weight estimates as for the discrimination estimates in the previous section. For the attribute model estimates, the RMSEs for the gMLTM-D decreases as the mean lower asymptote increases. Under the saturated model, on average the RMSEs increase with the lower asymptote, but that pattern does not hold true at each sample size. This is likely due to the fact that the lower asymptotes were often poorly estimated for the gMLTM-D in the 1,200 simulee cases—and always poorly estimated in the absence of guessing (i.e., when $\mu_\gamma = 0$)—impacting the remaining item parameter estimates. Unlike with the component discrimination RMSEs, the precision increased as the sample size increased within each test condition, which is expected under statistical theory.

The MLTM-D estimates fared worse, with consistently higher RMSE values. As with the discrimination estimates, some of this was due to the impact of the bias on the estimates, discussed in the next section, increasing the measure of RMSE. The precision within each test condition stayed relatively stable, regardless of sample size for the MLTM-D estimates.

Table 19

RMSE: Tests for repeated measures contrasts for attribute weight estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Model	1	46.112	46.112	74070.21	<0.001	0.991
Model * Sample Size	2	0.103	0.051	82.479	<0.001	0.189
Model * Asymptote	2	23.082	11.541	18538.86	<0.001	0.981
Model * Attribute Type	1	0.074	0.074	119.609	<0.001	0.145
Model * Sample Size * Asymptote	4	0.051	0.013	20.526	<0.001	0.104
Model * Sample Size * Attribute Type	2	0.006	0.003	5.003	0.007	0.014
Model * Asymptote * Attribute Type	2	0.007	0.004	5.905	0.003	0.016
Error(Model)	706	0.44	0.001	--	--	--

Table 20

RMSE: Tests for between-subjects effects for attribute weight estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	1	58.54	58.54	94137.66	<0.001	0.993
Sample Size	2	0.082	0.041	65.893	<0.001	0.157
Asymptote	2	5.769	2.884	4638.485	<0.001	0.929
Attribute Type	1	0.053	0.053	84.894	<0.001	0.107
Sample Size * Asymptote	4	0.005	0.001	1.867	0.114	0.01
Sample Size * Attribute Type	2	0.013	0.007	10.454	<0.001	0.029
Asymptote * Attribute Type	2	0.006	0.003	4.623	0.010	0.013
Error	706	0.439	0.001	--		

Table 21

Average RMSE and RSSE of estimated attribute weights

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	Mean RMSE	1,200	0.1895	0.1992	0.2195	0.2027	0.1597	0.1592	0.1607	0.1599
		3,000	0.1603	0.1690	0.1615	0.1636	0.1503	0.1402	0.1346	0.1417
		4,800	0.1520	0.1525	0.1469	0.1504	0.1423	0.1274	0.1302	0.1333
	Mean RSSE	1,200	0.1008	0.1323	0.1634	0.1322	0.0595	0.0767	0.0935	0.0766
		3,000	0.0650	0.0827	0.1036	0.0837	0.0384	0.0485	0.0588	0.0486
		4,800	0.0511	0.0659	0.0814	0.0661	0.0302	0.0386	0.0465	0.0384
gMLTM-D RMSE Mean			**0.1672**	**0.1736**	**0.1760**	**0.1723**	**0.1508**	**0.1423**	**0.1418**	**0.1450**
gMLTM-D RSSE Mean			**0.0723**	**0.0936**	**0.1161**	**0.0940**	**0.0427**	**0.0546**	**0.0663**	**0.0545**
MLTM-D	Mean RMSE	1,200	0.2227	0.3938	0.6611	0.4259	0.2089	0.3978	0.6248	0.4105
		3,000	0.2002	0.3978	0.6331	0.4104	0.2041	0.3895	0.6332	0.4089
		4,800	0.1920	0.3885	0.6473	0.4092	0.1902	0.3836	0.6417	0.4052
	Mean RSSE	1,200	0.0768	0.0766	0.0785	0.0773	0.0444	0.0436	0.0439	0.0439
		3,000	0.0489	0.0480	0.0492	0.0487	0.0283	0.0278	0.0278	0.0280
		4,800	0.0381	0.0381	0.0386	0.0383	0.0223	0.0218	0.0220	0.0221
MLTM-D RMSE Mean			**0.2049**	**0.3934**	**0.6472**	**0.4152**	**0.2011**	**0.3903**	**0.6333**	**0.4082**
MLTM-D RSSE Mean			**0.0546**	**0.0542**	**0.0555**	**0.0548**	**0.0316**	**0.0311**	**0.0312**	**0.0313**

Bias

Contrast tests for differences in mean bias due to the estimating model reveal a significant main effect and multiple two- and three-way interactions, which may contribute to the relationships found in the preceding section (Table 22). Bias in the attribute estimates is also significantly contributed to by lower asymptote and attribute type, though not by sample size, indicating that, averaged across the two models, even smaller samples can yield equally accurate attribute weight estimates (Table 23).

Inspection of Table 24 shows the gMLTM-D attribute weight estimates are uniformly less biased than those of the MLTM-D, regardless of sample size and test condition, as expected,. That both the MLTM-D and gMLTM-D estimates have non-zero bias for the attribute and saturated models where $\mu_\gamma = 0$ can be partially explained by the bias in the discrimination parameters for those test conditions and the model specification in the SAS program. More discussion into these causes is included in Chapter 5. The bias and model specification may also explain why the mean empirical biases for tests in the absence of guessing are not equal for the MLTM-D and gMLTM-D, indicating very different parameter estimates from what should be two identical models under that condition.

The MLTM-D attribute weight estimates become dramatically more biased as the lower asymptote increases, though the estimates are fairly stable within a test condition regardless of sample size. The small bias present in the gMLTM-D estimates is relatively constant across sample sizes and levels of guessing.

Table 22

Bias: Tests for repeated measures contrasts for attribute weight estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Model	1	72.974	72.974	110820.12	<0.001	0.994
Model * Sample Size	2	0.021	0.01	15.785	<0.001	0.043
Model * Asymptote	2	31.601	15.801	23995.328	<0.001	0.986
Model * Attribute Type	1	0.023	0.023	34.96	<0.001	0.047
Model * Sample Size * Asymptote	4	0.005	0.001	1.781	0.131	0.01
Model * Sample Size * Attribute Type	2	0.007	0.003	5.188	0.006	0.014
Model * Asymptote * Attribute Type	2	0.03	0.015	22.542	0.000	0.06
Error(Model)	706	0.465	0.001	--	--	--

Table 23

Bias: Tests for between-subjects effects for attribute weight estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	1	33.301	33.301	27149.95	<0.001	0.975
Sample Size	2	0.007	0.004	2.918	0.055	0.008
Asymptote	2	7.077	3.538	2884.779	<0.001	0.891
Attribute Type	1	0.023	0.023	18.853	<0.001	0.026
Sample Size * Asymptote	4	0.01	0.002	1.95	0.100	0.011
Sample Size * Attribute Type	2	0.002	0.001	1.015	0.363	0.003
Asymptote * Attribute Type	2	0.004	0.002	1.441	0.237	0.004
Error	706	0.866	0.001	--	--	--

Table 24

Average bias of attribute weight estimates

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1,200	-0.0451	-0.0325	-0.0346	-0.0374	-0.0594	-0.0637	-0.0506	-0.0579
	3,000	-0.0558	-0.0614	-0.0317	-0.0496	-0.0812	-0.0709	-0.0589	-0.0703
	4,800	-0.0604	-0.0576	-0.0473	-0.0551	-0.0702	-0.0577	-0.0668	-0.0649
gMLTM-D Mean		**-0.0538**	**-0.0505**	**-0.0379**	**-0.0474**	**-0.0703**	**-0.0641**	**-0.0588**	**-0.0644**
MLTM-D	1,200	-0.1241	-0.3562	-0.6393	-0.3732	-0.1346	-0.3753	-0.6120	-0.3740
	3,000	-0.1176	-0.3708	-0.6196	-0.3693	-0.1469	-0.3715	-0.6247	-0.3810
	4,800	-0.1158	-0.3641	-0.6352	-0.3717	-0.1292	-0.3664	-0.6332	-0.3763
MLTM-D Mean		**-0.1191**	**-0.3637**	**-0.6314**	**-0.3714**	**-0.1369**	**-0.3710**	**-0.6233**	**-0.3771**

Bias-adjusted RMSE

Due to the large, non-constant bias of the MLTM-D estimates, the bias-adjusted RMSE is a useful measure for ascertaining the precision of the estimates themselves. Analysis of the repeated-measures contrasts due to model on the adjusted RMSEs reveals a significant difference, as well as all significant two-and three-way interactions due to change in estimating model (Table 25). The signifant difference due to model was hypothesize and is expected, particularly based on the results for the RMSE and bias from the previous sections. It is interesting to note that, after removing bias from the RMSE, all main effects and interactions due to asymptote, sample size, and attribute type significantly impact the precision of the attribute weight estimates as well, as seen in Table 26.

When the bias of the MLTM-D estimates is removed, one can see that the precision is on-par with that of the gMLTM-D estimates, as shown in Table 27. The bias-adjusted gMLTM-D RMSEs are mostly unchanged from the original RMSEs, as the estimates themselves were fairly accurate. However, neither the gMLTM-D nor the MLTM-D estimates are very precise relative to the true values, as measured by RMSE, when compared to the RSSEs from the estimates; both models yield high mean RMSEs on average, even after adjusting for bias. The pattern observed for the RMSEs in the previous section still exist for the bias-adjusted RMSEs of the attribute weight estimates.

Table 25

Bias-adjusted RMSE: Tests for repeated measures contrasts for attribute weight estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Model	1	0.02	0.02	106.59	<0.001	0.131
Model * Sample Size	2	0.029	0.015	77.649	<0.001	0.18
Model * Asymptote	2	0.135	0.068	357.95	<0.001	0.503
Model * Attribute Type	1	0.018	0.018	94.331	<0.001	0.118
Model * Sample Size * Asymptote	4	0.006	0.002	8.509	<0.001	0.046
Model * Sample Size * Attribute Type	2	0.003	0.002	9.011	<0.001	0.025
Model * Asymptote * Attribute Type	2	0.003	0.001	7.301	0.001	0.02
Error(Model)	706	0.133	<0.001	--	--	--

Table 26

Bias-adjusted RMSE: Tests for between-subjects effects for attribute weight estimates

Source	df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	1	14.189	14.189	67740.25	<0.001	0.99
Sample Size	2	0.151	0.075	359.822	<0.001	0.505
Asymptote	2	0.021	0.011	50.456	<0.001	0.125
Attribute Type	1	0.177	0.177	845.779	<0.001	0.545
Sample Size * Asymptote	4	0.009	0.002	10.214	<0.001	0.055
Sample Size * Attribute Type	2	0.015	0.007	34.884	<0.001	0.09
Asymptote * Attribute Type	2	0.008	0.004	18.633	<0.001	0.05
Error	706	0.148	<0.001	--	--	--

Table 27

Average bias-adjusted RMSE of attribute weight estimates

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	Mean Adjusted RMSE	1,200	0.1760	0.1903	0.2126	0.1930	0.1416	0.1381	0.1453	0.1417
		3,000	0.1483	0.1540	0.1549	0.1524	0.1239	0.1173	0.1162	0.1191
		4,800	0.1379	0.1391	0.1377	0.1382	0.1223	0.1103	0.1087	0.1138
	Mean RSSE	1,200	0.1008	0.1323	0.1634	0.1322	0.0595	0.0767	0.0935	0.0766
		3,000	0.0650	0.0827	0.1036	0.0837	0.0384	0.0485	0.0588	0.0486
		4,800	0.0511	0.0659	0.0814	0.0661	0.0302	0.0386	0.0465	0.0384
gMLTM-D Adjusted RMSE Mean			**0.1541**	**0.1612**	**0.1684**	**0.1612**	**0.1293**	**0.1219**	**0.1234**	**0.1249**
gMLTM-D RSSE Mean			**0.0723**	**0.0936**	**0.1161**	**0.0940**	**0.0427**	**0.0546**	**0.0663**	**0.0545**
MLTM-D	Mean Adjusted RMSE	1,200	0.1796	0.1669	0.1672	0.1712	0.1549	0.1296	0.1232	0.1359
		3,000	0.1608	0.1433	0.1296	0.1446	0.1395	0.1162	0.1023	0.1193
		4,800	0.1520	0.1348	0.1242	0.1370	0.1384	0.1130	0.1038	0.1184
	Mean RSSE	1,200	0.0768	0.0766	0.0785	0.0773	0.0444	0.0436	0.0439	0.0439
		3,000	0.0489	0.0480	0.0492	0.0487	0.0283	0.0278	0.0278	0.0280
		4,800	0.0381	0.0381	0.0386	0.0383	0.0223	0.0218	0.0220	0.0221
MLTM-D Adjusted RMSE Mean			**0.1641**	**0.1483**	**0.1403**	**0.1509**	**0.1443**	**0.1196**	**0.1098**	**0.1245**
MLTM-D RSSE Mean			**0.0546**	**0.0542**	**0.0555**	**0.0548**	**0.0316**	**0.0311**	**0.0312**	**0.0313**

Correlation with true values

The correlation between the estimated attribute weights and the known, simulated values were uniformly high for both the gMLTM-D and MLTM-D estimates, as shown in Table 28. As the sample size increase, the mean correlation within a test condition consistently increases, indicating a more reliable ordering of the attributes along the ability scale with a large number of simulees. The minimum correlation observed across all replications was 0.9256, which was obtained for a replication of the saturated model with a mean lower asymptote of 0.25 with 1,200 simulees estimated under the gMLTM-D. This corresponds with the results in Table 28, which indicate that for such a model, involving many parameters, adequate recovery along the continuum would be difficult with relatively few people.

In terms of mean correlation, the MLTM-D estimates of attribute weights perform better than those of the gMLTM-D for all sample sizes under the saturated model in the absence of guessing, and across all test conditions for 1,200 simulees. The gMLTM-D orders the attribute weight estimates better, however, as the sample size increases, particularly for the attribute models. This is further evidence that the relative parsimony of the MLTM-D makes it more efficient for smaller samples, while better recovery is possible as both sample size and the probability of false positives increase under the gMLTM-D.

Table 28

Average correlations between known and estimated attribute weights

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1,200	0.9736	0.9642	0.9495	0.9625	0.9868	0.9812	0.9755	0.9812
	3,000	0.9881	0.9812	0.9769	0.9821	0.9937	0.9918	0.9887	0.9914
	4,800	0.9916	0.9884	0.9850	0.9883	0.9957	0.9938	0.9925	0.9940
gMLTM-D Mean		**0.9845**	**0.9779**	**0.9705**	**0.9776**	**0.9921**	**0.9889**	**0.9855**	**0.9889**
MLTM-D	1,200	0.9832	0.9740	0.9630	0.9734	0.9905	0.9848	0.9768	0.9840
	3,000	0.9924	0.9843	0.9776	0.9848	0.9961	0.9918	0.9876	0.9919
	4,800	0.9944	0.9874	0.9805	0.9875	0.9972	0.9930	0.9877	0.9926
MLTM-D Mean		**0.9900**	**0.9819**	**0.9737**	**0.9819**	**0.9946**	**0.9899**	**0.9840**	**0.9895**

Attribute weight summary

In terms of RMSE and bias, the gMLTM-D estimates outperform the MLTM-D estimates, particularly as the lower asymptote increases. However, once the RMSE is adjusted for bias, the MLTM-D attribute weight estimates generally are more precise on average, with the exception of the saturated model in the absence of guessing. On the final criterion, the correlation between the estimates and true values, the MLTM-D estimates also marginally outperformed the gMLTM-D estimates across all test conditions and sample sizes. The model used to estimate the item parameters significantly impacts both the accuracy and the precision of the attribute weight estimates, and interacts significantly with random test features such as guessing, sample size, and attribute type. Test and item design features, such as the **Q**-matrix and amount of guessing, significantly contribute to the accuracy and precision of the attribute weights, when averaged across the model used to estimate the item parameters.

Lower asymptotes

Unlike the other item parameters, the mean of the lower asymptotes was manipulated in the design of the experiment across test conditions. For the saturated models, estimation of the lower asymptotes with 1,200 simulees was often unreliable. All replications of test conditions where $\mu_\gamma = 0$ could never estimate all lower asymptotes: many estimates would get stuck at one of the estimation constraints, regardless of sample size. The more parsimonious attribute models were better at estimating the lower asymptotes at all sample sizes. In the case where a lower asymptote estimate reached a constraint, no actual estimate or standard error was obtained. The following sections cover the analysis criteria for the estimated lower asymptotes in turn, excluding the items

whose lower asymptotes could not be estimated as well as the reference item on each test. As only the gMLTM-D estimates a lower asymptote, that is the pertinent model for the ensuing discussion.

RMSE

Tests for the mean RMSE obtained for the lower asymptote estimates shows significant main effects due to both sample size and attribute type, as well as their interaction, shown in Table 29. The mean RMSE and RSSE for the estimated lower asymptotes for the different test conditions and sample sizes are provided in Table 30. Relative to the attribute weight estimates, the lower asymptote estimates are fairly precise, when one compares the mean RMSE and RSSE. As expected, the mean RMSE decreases as the sample size increases within a given test condition, and the best RMSEs within a sample size are observed for the three attribute models. An interesting relationship among test conditions for both the saturated and attribute models can be seen, where the tests with $\mu_\gamma = 0.125$ have lower mean RMSEs than the other two tests at each sample size. As the distribution for the attribute weights was unchanged across the test conditions, the increased precision for the lower asymptotes at the middle level of guessing may reflect an optimal matching of the items' difficulty with the probability of guessing.

Table 29

RMSE: Tests of between-subjects effects for lower asymptote estimates

Source		df	SS	MS	F_{obs}	*p*-value	η^2_p
Intercept	Hypothesis	1	2.19	2.192	40.97	0.007	0.929
	Error	3.15	0.17	0.053[a]			
Attribute Type	Hypothesis	1	0.03	0.026	12.01	0.032	0.777
	Error	3.45	0.01	0.002[b]			
Sample Size	Hypothesis	2	0.08	0.042	29.96	0.035	0.969
	Error	1.94	0	0.001[c]			
Asymptote	Hypothesis	2	0.02	0.01	13.07	0.079	0.933
	Error	1.88	0	0.001[d]			
Attribute Type * Asymptote	Hypothesis	2	0	0.001	10.98	<0.001	0.023
	Error	945	0.07	7.083E-005[e]			
Sample Size * Asymptote	Hypothesis	4	0	4.79E-05	0.676	0.608	0.003
	Error	945	0.07	7.083E-005[e]			
Attribute Type * Sample Size	Hypothesis	2	0	0.001	20.08	<0.001	0.041
	Error	945	0.07	7.083E-005[e]			

a. 0.971 MS(Sample Size) + 1.294 MS(Asymptote) - 0.150 MS(Attribute Type * Asymptote) - 0.977MS(Sample Size * Asymptote) + 0.144 MS(Attribute Type * Sample Size) - 0.283 MS(Error)
b. MS(Attribute Type * Asymptote) + MS(Attribute Type * Sample Size) - MS(Error)
c. MS(Sample Size * Asymptote) + MS(Attribute Type * Sample Size) - MS(Error)
d. MS(Attribute Type * Asymptote) + MS(Sample Size * Asymptote) - MS(Error)
e. MS(Error)

Table 30

RMSE and RSSE of item lower asymptote estimates

		Saturated Models				Attribute Models			
Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
Mean RMSE	1,200	0.0873	0.0768	0.0890	0.0843	0.0715	0.0604	0.0696	0.0672
	3,000	0.0640	0.0561	0.0687	0.0629	0.0573	0.0447	0.0537	0.0519
	4,800	0.0545	0.0455	0.0616	0.0539	0.0513	0.0387	0.0489	0.0463
Mean RSSE	1,200	0.0678	0.0836	0.1002	0.0839	0.0465	0.0569	0.0661	0.0565
	3,000	0.0452	0.0544	0.0659	0.0552	0.0306	0.0357	0.0411	0.0358
	4,800	0.0358	0.0439	0.0525	0.0441	0.0238	0.0284	0.0324	0.0282
RMSE Grand Mean		**0.0686**	**0.0595**	**0.0731**	**0.0670**	**0.0600**	**0.0479**	**0.0574**	**0.0551**
RSSE Grand Mean		**0.0496**	**0.0607**	**0.0729**	**0.0610**	**0.0336**	**0.0403**	**0.0465**	**0.0402**

Bias

The mean RMSEs do not indicate that the lower asymptotes are very inaccurate, but investigation into possible bias is still worthwhile. Unlike with the RMSEs, the bias of the lower asymptote is only significantly influenced by the true lower asymptote of the items, as well as the interactions of the sample size and lower asymptote and sample size and attribute type, as indicated in Table 31. As the lower asymptote was, anecdotally, difficult to estimate for the smaller sample sizes, particularly for the saturated model, the findings in Table 31 are not surprising.

The mean bias of the lower asymptote estimates is detailed in Table 32, and one can see that some bias does exists for some test conditions and sample sizes. The bias in the lower asymptote estimates is the smallest for all sample sizes under testing conditions with the highest rate of guessing (i.e., $\mu_\gamma = 0.25$), and bias is relatively stable as sample size increase. The bias stabilizes at the middle level of guessing for both the saturated and attribute models after 3,000 simulees. There is a decline in the bias for tests in the absence of guessing as the sample size increases, as well, though the bias in the estimates was large enough to begin with that the same absolute gain yielded some bias in the end.

Table 31

Bias: Tests of between-subjects effects for lower asymptote estimates

Source		df	SS	MS	F_{obs}	*p*-value	η^2_p
Intercept	Hypothesis	1	0.148	0.148	0.686	0.490	0.244
	Error	3.15	0.46	.216[a]			
Attribute Type	Hypothesis	1	0.002	0.002	1.197	0.382	0.36
	Error	3.45	0.003	.001[b]			
Sample Size	Hypothesis	2	0.016	0.008	3.098	0.139	0.569
	Error	1.94	0.012	.003[c]			
Asymptote	Hypothesis	2	0.324	0.162	129.07	<0.001	0.984
	Error	1.88	0.005	.001[d]			
Attribute Type * Asymptote	Hypothesis	2	0	0	1.623	0.198	0.003
	Error	945	0.079	8.405E-005[e]			
Sample Size * Asymptote	Hypothesis	4	0.005	0.001	14.307	<0.001	0.057
	Error	945	0.079	8.405E-005[e]			
Attribute Type * Sample Size	Hypothesis	2	0.003	0.001	16.972	<0.001	0.035
	Error	945	0.079	8.405E-005[e]			

a. 0.971 MS(Sample Size) + 1.294 MS(Asymptote) - 0.150 MS(Attribute Type * Asymptote) - 0.977MS(Sample Size * Asymptote) + 0.144 MS(Attribute Type * Sample Size) - 0.283 MS(Error)

b. MS(Attribute Type * Asymptote) + MS(Attribute Type * Sample Size) - MS(Error)

c. MS(Sample Size * Asymptote) + MS(Attribute Type * Sample Size) - MS(Error)

d. MS(Attribute Type * Asymptote) + MS(Sample Size * Asymptote) - MS(Error)

e. MS(Error)

Table 32

Mean bias of lower asymptote estimates

	Saturated Models				Attribute Models			
Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
1,200	0.0628	0.0251	0.0030	0.0303	0.0540	0.0164	-0.0054	0.0217
3,000	0.0463	0.0113	-0.0036	0.0180	0.0431	0.0122	-0.0060	0.0164
4,800	0.0400	0.0099	-0.0054	0.0148	0.0388	0.0131	-0.0052	0.0156
Grand Mean	**0.0497**	**0.0154**	**-0.0020**	**0.0210**	**0.0453**	**0.0139**	**-0.0055**	**0.0179**

Bias-adjusted RMSE

Removal of the empirical bias observed in some of the lower asymptote estimates yields a better representation of the precision with which the estimates match the simulated values. Consistent with the results from the original RMSEs, analysis of the bias-adjusted RMSEs reveals significant effects due to sample size, attribute type, lower asymptote, and all two- and three-way interactions (Table 33). However, inspection of Table 34 reveals a different pattern present in the bias-adjusted RMSEs than that of the RMSEs in the previous section, which increased as the true lower asymptote increases.

One can see that there is marked improvement in the RMSE values relative to RSSE, particularly for the attribute and saturated models simulated in the absence of guessing, which exhibited the most bias in the lower asymptote estimates. It is the attribute and saturated models simulated in the absence of guessing that were the most troublesome in terms of lower asymptote estimation, as even at the largest sample size at least one estimate got caught at a constraint and could not be estimated in each simulated test.

Table 33

Bias-adjusted RMSE: Tests of between-subjects effects for lower asymptote estimates

Source		df	SS	MS	F_{obs}	p-value	η^2_p
Intercept	Hypothesis	1	1.729	1.729	26.5	0.007	0.871
	Error	3.15	0.257	.065[a]			
Attribute Type	Hypothesis	1	0.027	0.027	12.81	0.003	0.77
	Error	3.45	0.008	.002[b]			
Sample Size	Hypothesis	2	0.063	0.031	25.05	0.002	0.943
	Error	1.94	0.004	.001[c]			
Asymptote	Hypothesis	2	0.054	0.027	19.67	0.003	0.931
	Error	1.88	0.004	.001[d]			
Attribute Type * Asymptote	Hypothesis	2	0.002	0.001	24.81	<0.001	0.05
	Error	945	0.043	4.515E-005[e]			
Sample Size * Asymptote	Hypothesis	4	0.001	<0.001	6.726	<0.001	0.028
	Error	945	0.043	4.515E-005[e]			
Attribute Type * Sample Size	Hypothesis	2	0.002	0.001	22.03	<0.001	0.045
	Error	945	0.043	4.515E-005[e]			

a. 0.971 MS(Sample Size) + 1.294 MS(Asymptote) - 0.150 MS(Attribute Type * Asymptote) - 0.977MS(Sample Size * Asymptote) + 0.144 MS(Attribute Type * Sample Size) - 0.283 MS(Error)

b. MS(Attribute Type * Asymptote) + MS(Attribute Type * Sample Size) - MS(Error)

c. MS(Sample Size * Asymptote) + MS(Attribute Type * Sample Size) - MS(Error)

d. MS(Attribute Type * Asymptote) + MS(Sample Size * Asymptote) - MS(Error)

e. MS(Error)

Table 34

Mean bias-adjusted RMSE and RSSE for lower asymptote estimates

		Saturated Models				Attribute Models			
Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
Mean Adjusted RMSE	1200	0.0603	0.0719	0.0884	0.0735	0.0466	0.0566	0.0679	0.0570
	3000	0.0440	0.0541	0.0677	0.0553	0.0375	0.0420	0.0519	0.0438
	4800	0.0368	0.0440	0.0610	0.0473	0.0335	0.0352	0.0478	0.0388
Mean RSSE	1200	0.0678	0.0836	0.1002	0.0839	0.0465	0.0569	0.0661	0.0565
	3000	0.0452	0.0544	0.0659	0.0552	0.0306	0.0357	0.0411	0.0358
	4800	0.0358	0.0439	0.0525	0.0441	0.0238	0.0284	0.0324	0.0282
Adjusted RMSE Grand Mean		**0.0471**	**0.0567**	**0.0724**	**0.0587**	**0.0392**	**0.0446**	**0.0559**	**0.0466**
RSSE Grand Mean		**0.0496**	**0.0607**	**0.0729**	**0.0610**	**0.0336**	**0.0403**	**0.0465**	**0.0402**

Correlation with true values

Two test conditions were simulated such that all lower asymptotes were set equal to zero, to represent the situation where no guessing occurs; in those cases, the correlation between the estimated values and the true values is necessarily zero, as the true values do not vary. The other two test conditions were simulated such that the variance of the true lower asymptote was 0.001, meaning that correlation analysis would not be meaningful for the lower asymptote estimates. Discussion of parameter recovery for the lower asymptotes therefore, is limited to bias and RMSE, or the accuracy and precision, of the estimates, which has been covered in the previous sections.

Lower asymptote summary

Per the correlation and both RMSE summaries of the lower asymptote estimates, the best recovery of the true lower asymptotes occurred for test conditions at the middle level of guessing ($\mu_\gamma = 0.125$). The non-monotonic association between correlation and RMSEs as the true lower asymptote increases is different from that observed for both the discrimination and attribute weights, discussed in the previous sections. As the MLTM-D does not estimate a lower asymptote, a model comparison can not be conducted for lower asymptote estimates.

Person Parameter Results

Person parameters were estimated using the MLTM-D and gMLTM-D item parameter estimates obtained for a selection of 18 tests, based on several criteria for appropriateness: representation across the design and plausibility of the item estimates. The tests were chosen to represent all design points from the simulation, while the specific replications of the tests were selected based on the gMLTM-D item parameter

estimation. Whenever possible, a replication where all lower asymptotes were estimated was selected; when such a case did not exist, a replication was chosen where the fewest lower asymptotes were constrained at the lower boundary during the estimation process. The rationale behind the item parameter criterion was the more item parameters were successfully recovered, the better they would be, and the better the resulting person estimates.

The mean and standard deviation for the true and estimated component abilities for each sample size and model for the saturated model are provided in Table 35. As the true lower asymptote increases, the mean of the gMLTM-D estimates tends to be closer to the true mean of 0 than MLTM-D means, and means of both estimates decrease as the sample size increases. The person estimates obtained for the attribute model tests follow a similar pattern, where the larger samples generally yield means closer to zero.

Table 35

Saturated model descriptive statistics for person parameters and estimates

Test	True θ_1	True θ_2	True θ_3	gMLTM-D θ_1	gMLTM-D θ_2	gMLTM-D θ_3	MLTM-D θ_1	MLTM-D θ_2	MLTM-D θ_3
Saturated Model $\gamma = 0$; 1,2000	0.0139	0.0095	-0.0001	0.0151	0.0869	0.1127	0.0202	0.0649	0.1129
	(0.9722)	(0.9880)	(1.0069)	(1.0213)	(1.043)	(1.024)	(1.0227)	(1.0499)	(1.0274)
Saturated Model $\gamma = 0$; 3,000	0.0100	-0.0132	-0.0209	0.011	-0.0239	0.0544	0.0111	-0.024	0.0598
	(0.9983)	(0.9901)	(1.0008)	(1.0367)	(1.053)	(1.0276)	(1.0395)	(1.0557)	(1.0295)
Saturated Model $\gamma = 0$; 4,800	-0.0268	-0.0157	-0.0121	0.023	0.0247	0.0331	0.0219	0.0283	0.0375
	(0.9641)	(0.9937)	(1.0162)	(1.0115)	(1.0355)	(1.0207)	(1.0164)	(1.0376)	(1.0231)
Saturated Model $\gamma = 0.125$; 1,200	-0.0287	-0.0268	0.0007	0.0265	-0.0054	0.1014	0.0407	0.0131	0.1237
	(0.9896)	(1.0145)	(0.9865)	(0.9738)	(1.0042)	(0.9358)	(0.9711)	(0.9973)	(0.9241)
Saturated Model $\gamma = 0.125$; 3,000	0.0046	-0.0182	-0.0179	-0.0164	-0.0021	0.0173	-0.0023	-0.0086	0.0093
	(1.0034)	(1.0186)	(0.9806)	(0.9901)	(1.0001)	(0.985)	(0.9936)	(1.0035)	(0.9888)
Saturated Model $\gamma = 0.125$; 4,800	-0.0090	-0.0234	-0.0081	0.0376	0.0202	-0.02	0.025	0.0313	-0.0158
	(1.0156)	(0.9979)	(0.9936)	(0.9998)	(0.9951)	(0.9601)	(1.0034)	(0.9946)	(0.9659)
Saturated Model $\gamma = 0.25$; 1,200	0.0301	0.0160	0.0671	0.0264	0.0057	0.0458	0.0254	-0.0053	0.0748
	(1.0087)	(1.0179)	(0.9792)	(0.9527)	(0.9663)	(0.9229)	(0.9418)	(0.9658)	(0.9046)
Saturated Model $\gamma = 0.25$; 3,000	-0.0126	-0.0043	-0.0204	0.0124	0.0208	0.0093	0.0044	0.0254	0.0089
	(1.0159)	(1.0108)	(0.9884)	(0.9667)	(0.9439)	(0.9549)	(0.9638)	(0.9365)	(0.9497)
Saturated Model $\gamma = 0.25$; 4,800	0.0318	0.0038	0.0153	0.0102	-0.001	0.007	0.0104	-0.0158	0.0166
	(0.9857)	(0.9884)	(1.0046)	(0.9466)	(0.9526)	(0.9453)	(0.939)	(0.9525)	(0.9377)

Table 36

Attribute model descriptive statistics for person parameters and estimates

Test	True θ_1	True θ_2	True θ_3	gMLTM-D θ_1	gMLTM-D θ_2	gMLTM-D θ_3	MLTM-D θ_1	MLTM-D θ_2	MLTM-D θ_3
Attribute Model $\gamma = 0$; 1,200	-0.0293	0.0151	0.0098	0.1152	0.0952	-0.024	0.1436	0.1121	-0.0224
	(1.0079)	(1.0071)	(1.0009)	(0.9927)	(1.0397)	(0.9812)	(0.9810)	(1.0374)	(0.9871)
Attribute Model $\gamma = 0$; 3,000	0.0182	0.0293	-0.0116	0.0089	-0.0239	-0.0187	0.0114	0.0230	-0.0096
	(0.9863)	(0.9858)	(0.9696)	(1.0673)	(1.0010)	(1.0024)	(1.0699)	(1.0091)	(1.0043)
Attribute Model $\gamma = 0$; 4,800	-0.0142	0.0077	0.0243	-0.0029	0.0048	-0.0251	-0.0036	0.0106	-0.0238
	(1.0034)	(0.9930)	(1.0214)	(1.0280)	(1.0091)	(1.0162)	(1.032)	(1.0107)	(1.0196)
Attribute Model $\gamma = 0.125$; 1,200	0.0493	-0.0455	0.0040	-0.0235	-0.0148	-0.0256	-0.0262	-0.0210	-0.0431
	(0.9707)	(0.9526)	(0.9770)	(0.9767)	(0.9523)	(0.9898)	(0.9837)	(0.9598)	(1.0009)
Attribute Model $\gamma = 0.125$; 3,000	0.0205	-0.0426	0.0541	-0.0040	0.0112	-0.0330	0.0086	0.0083	-0.0151
	(0.9772)	(0.9960)	(1.0060)	(0.9639)	(1.0235)	(0.9926)	(0.9639)	(1.0241)	(0.9916)
Attribute Model $\gamma = 0.125$; 4,800	-0.0107	0.0102	-0.0186	0.0399	0.0339	0.0496	0.0610	0.0225	0.0475
	(0.9959)	(1.0085)	(1.0214)	(0.9653)	(0.9964)	(0.9571)	(0.9597)	(1.0018)	(0.9567)
Attribute Model $\gamma = 0.25$; 1,200	0.0189	-0.0159	0.0216	-0.0381	-0.0523	0.0044	-0.0174	-0.0165	0.0021
	(1.0066)	(0.9825)	(0.9683)	(0.9520)	(0.9467)	(0.9295)	(0.9430)	(-0.0165)	(0.9290)
Attribute Model $\gamma = 0.25$; 3,000	0.0281	-0.0134	0.0222	0.0084	-0.0066	-0.0278	0.01720	-0.0052	-0.0216
	(0.9775)	(1.0209)	(1.0079)	(0.9233)	(0.9769)	(0.9621)	(0.9197)	(0.9746)	(0.9563)
Attribute Model $\gamma = 0.25$; 4,800	-0.0308	0.0017	0.0237	0.0224	0.0077	0.0203	0.0241	0.0138	0.0272
	(0.9900)	(0.9893)	(0.9984)	(0.9478)	(0.9475)	(0.9439)	(0.9467)	(0.9419)	(0.9400)

RMSE

The RMSE and RSSE of the person estimates for each component are provided in Table 37 through Table 39: one must bear in mind the table cells each represent a single replication of the simulation. Both RMSE and RSSE are fairly stable across sample size within a test condition, there is no clear relationship for the values for estimates obtained from either model. For a given test, however, the RMSE values are uniformly lower for the person estimates obtained from the gMLTM-D items than those obtained from the MLMT-D items, which is a relationship that generally holds for the RSSEs, as well. There is evidence that the more parsimonious attribute models yield less precise estimation of the person abilities, regardless of the source of the item parameter estimates: the RMSE and RSSEs for the attribute models than for tests under the saturated models with the same mean lower asymptote. Within tests of saturated models or attribute model types, as the lower asymptote increased, the RSSE of the person estimates tends to increase for both the gMLTM-D and MLTM-D; there is no consistent trend for the RMSE relative to increasing lower asymptote.

Table 37

RMSE and RSSE for first component person estimates

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	RMSE	1200	0.3977	0.4184	0.4264	0.4142	0.4381	0.4368	0.4343	0.4364
		3000	0.3924	0.4064	0.4187	0.4058	0.3802	0.4357	0.4366	0.4175
		4800	0.4017	0.3961	0.4216	0.4065	0.4381	0.4368	0.4343	0.4364
	RSSE	1200	0.4129	0.4283	0.4478	0.4297	0.4300	0.4556	0.4641	0.4499
		3000	0.4079	0.4381	0.4399	0.4286	0.3895	0.4633	0.4588	0.4372
		4800	0.4229	0.4122	0.4486	0.4279	0.4300	0.4556	0.4641	0.4499
gMLTM-D Mean RMSE			**0.3973**	**0.4070**	**0.4222**	**0.4088**	**0.4103**	**0.4346**	**0.4325**	**0.4258**
gMLTM-D Mean RSSE			**0.4146**	**0.4262**	**0.4454**	**0.4287**	**0.4160**	**0.4524**	**0.4545**	**0.4410**
MLTM-D	RMSE	1200	0.3976	0.4201	0.4313	0.4163	0.4455	0.4404	0.4353	0.4404
		3000	0.3930	0.4078	0.4238	0.4082	0.3810	0.4363	0.4395	0.4190
		4800	0.4030	0.3987	0.4259	0.4092	0.4140	0.4340	0.4318	0.4266
	RSSE	1200	0.4157	0.4412	0.4718	0.4429	0.4295	0.4700	0.4830	0.4608
		3000	0.4111	0.4521	0.4584	0.4405	0.3923	0.4746	0.4814	0.4494
		4800	0.4280	0.4248	0.4674	0.4400	0.4318	0.4492	0.4575	0.4462
MLTM-D Mean RMSE			**0.3978**	**0.4089**	**0.4270**	**0.4112**	**0.4135**	**0.4369**	**0.4355**	**0.4286**
MLTM-D Mean RSSE			**0.4183**	**0.4393**	**0.4659**	**0.4412**	**0.4179**	**0.4646**	**0.4740**	**0.4521**

Table 38

RMSE and RSSE for second component person estimates

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	RMSE	1200	0.3755	0.3902	0.4187	0.3948	0.3944	0.4407	0.4436	0.4262
		3000	0.3791	0.4097	0.4575	0.4154	0.4022	0.3860	0.4025	0.3969
		4800	0.4035	0.3987	0.4304	0.4109	0.4217	0.4032	0.4304	0.4185
	RSSE	1200	0.3891	0.4111	0.4353	0.4118	0.3898	0.4851	0.4621	0.4457
		3000	0.3912	0.4254	0.4650	0.4272	0.4246	0.4036	0.4234	0.4172
		4800	0.4130	0.4196	0.4520	0.4282	0.4338	0.4159	0.4560	0.4352
gMLTM-D Mean RMSE			**0.3860**	**0.3995**	**0.4355**	**0.4070**	**0.4061**	**0.4100**	**0.4255**	**0.4139**
gMLTM-D Mean RSSE			**0.4062**	**0.4496**	**0.4408**	**0.4322**	**0.4161**	**0.4349**	**0.4472**	**0.4327**
MLTM-D	RMSE	1200	0.3736	0.3928	0.4228	0.3964	0.3967	0.4441	0.4435	0.4281
		3000	0.3799	0.4119	0.4610	0.4176	0.4045	0.3868	0.4067	0.3994
		4800	0.4040	0.4016	0.4375	0.4144	0.4220	0.4062	0.4338	0.4207
	RSSE	1200	0.3951	0.4199	0.4540	0.4230	0.3911	0.5039	0.4745	0.4565
		3000	0.3938	0.4388	0.4797	0.4374	0.4322	0.4100	0.4352	0.4258
		4800	0.4164	0.4300	0.4744	0.4402	0.4366	0.4303	0.4741	0.4470
MLTM-D Mean RMSE			**0.3858**	**0.4021**	**0.4405**	**0.4095**	**0.4077**	**0.4124**	**0.4280**	**0.4160**
MLTM-D Mean RSSE			**0.4017**	**0.4296**	**0.4694**	**0.4336**	**0.4199**	**0.4481**	**0.4613**	**0.4431**

Table 39

RMSE and RSSE for third component person estimates

			Saturated Models				Attribute Models			
Model	Measure	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	RMSE	1200	0.4201	0.4660	0.4369	0.4410	0.4532	0.4238	0.4429	0.4400
		3000	0.3868	0.4062	0.4231	0.4054	0.4155	0.4296	0.4368	0.4273
		4800	0.4093	0.4560	0.4345	0.4333	1.7527	1.5828	1.3931	1.5762
	RSSE	1200	0.4041	0.4443	0.4469	0.4317	0.4213	0.4414	0.4346	0.4324
		3000	0.4003	0.4339	0.4365	0.4236	0.4335	0.4421	0.4585	0.4447
		4800	0.4144	0.4707	0.4389	0.4413	0.4382	0.4452	0.4540	0.4458
gMLTM-D Mean RMSE			**0.4054**	**0.4427**	**0.4315**	**0.4265**	**0.4300**	**0.4316**	**0.4381**	**0.4333**
gMLTM-D Mean RSSE			**0.4062**	**0.4496**	**0.4408**	**0.4322**	**0.4446**	**0.4412**	**0.4604**	**0.4487**
MLTM-D	RMSE	1200	0.4201	0.4721	0.4380	0.4434	0.4543	0.4310	0.4466	0.4440
		3000	0.3877	0.4087	0.4278	0.4080	0.4159	0.4275	0.4413	0.4283
		4800	0.4099	0.4585	0.4402	0.4362	0.4220	0.4414	0.4383	0.4339
	RSSE	1200	0.4103	0.4541	0.4626	0.4423	0.4680	0.4547	0.4930	0.4719
		3000	0.4029	0.4494	0.4524	0.4349	0.4369	0.4509	0.4767	0.4548
		4800	0.4177	0.4854	0.4595	0.4542	0.4420	0.4612	0.4709	0.4580
MLTM-D Mean RMSE			**0.4059**	**0.4464**	**0.4354**	**0.4292**	**0.4307**	**0.4333**	**0.4421**	**0.4354**
MLTM-D Mean RSSE			**0.4103**	**0.4630**	**0.4582**	**0.4438**	**0.4489**	**0.4556**	**0.4802**	**0.4616**

Bias

Although the RMSE values closely aligned with the RSSE calculated from the standard errors of the person estimates, the bias of the person estimates for each component are presented in Table 40 through Table 42. One can see that, altogether, both the gMLTM-D and MLTM-D were fairly accurate in the person ability estimation. The two model sources of item parameter estimates performed about equally in terms of bias, regardless of sample size and attribute type. No consistent pattern emerges within attribute type for either model as sample size or lower asymptote increases, though bias generally appears to decrease as the lower asymptote increases, and to increase as the sample size increases. On average, the gMLTM-D estimates were less biased than the MLTM-D estimates, but marginally so.

Table 40

Bias for first component person estimates

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1200	0.0011	0.0552	-0.0036	0.0176	0.1446	-0.0729	-0.0570	0.0049
	3000	0.0010	-0.0210	0.0250	0.0017	-0.0093	-0.0245	-0.0197	-0.0178
	4800	0.0464	0.0546	0.0021	0.0344	0.0049	0.0371	0.0407	0.0276
gMLTM-D Mean		**0.0162**	**0.0296**	**0.0078**	**0.0179**	**0.0467**	**-0.0201**	**-0.0120**	**0.0049**
MLTM-D	1200	0.0063	0.0694	-0.0046	0.0237	0.1729	-0.0756	-0.0363	0.0203
	3000	0.0011	-0.0069	0.0170	0.0037	-0.0068	-0.0119	-0.0109	-0.0099
	4800	0.0453	0.0419	0.0023	0.0298	0.0041	0.0582	0.0424	0.0349
MLTM-D Mean		**0.0176**	**0.0348**	**0.0049**	**0.0191**	**0.0567**	**-0.0098**	**-0.0016**	**0.0151**

Table 41

Bias for second component person estimates

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1200	0.0774	0.0214	-0.0103	0.0295	0.0801	0.0307	-0.0364	0.0248
	3000	-0.0107	0.0161	0.0251	0.0102	-0.0036	0.0538	0.0067	0.0190
	4800	0.0362	0.0238	-0.0118	0.0161	-0.0096	0.0335	0.0119	0.0119
gMLTM-D Mean		**0.0343**	**0.0204**	**0.0010**	**0.0186**	**0.0223**	**0.0393**	**-0.0059**	**0.0185**
MLTM-D	1200	0.0554	0.0399	-0.0212	0.0247	0.0969	0.0245	-0.0007	0.0402
	3000	-0.0107	0.0096	0.0297	0.0095	-0.0063	0.0508	0.0082	0.0176
	4800	0.0398	0.0349	-0.0266	0.0161	0.0289	0.0325	0.0085	0.0233
MLTM-D Mean		**0.0282**	**0.0281**	**-0.0060**	**0.0168**	**0.0256**	**0.0359**	**0.0013**	**0.0209**

Table 42

Bias for third component person estimates

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1200	0.1129	0.1007	-0.0213	0.0641	-0.0338	-0.0296	-0.0172	-0.0268
	3000	0.0753	0.0352	0.0297	0.0467	-0.0071	-0.0871	-0.0500	-0.0481
	4800	0.0471	-0.0140	-0.0170	0.0054	-0.0251	0.0527	0.0211	0.0162
gMLTM-D Mean		**0.0784**	**0.0407**	**-0.0029**	**0.0387**	**-0.0220**	**-0.0213**	**-0.0154**	**-0.0196**
MLTM-D	1200	0.1131	0.1230	0.0077	0.0812	-0.0321	-0.0470	-0.0195	-0.0329
	3000	0.0807	0.0272	0.0293	0.0457	0.0020	-0.0692	-0.0438	-0.0370
	4800	0.0515	-0.0097	-0.0073	0.0115	-0.0179	-0.0219	-0.0118	-0.0172
MLTM-D Mean		**0.0817**	**0.0468**	**0.0099**	**0.0462**	**-0.0200**	**-0.0216**	**-0.0136**	**-0.0184**

Correlation with true values

The correlations of the person ability estimates for each component with the simulated true values are summarized in Table 43 through Table 45. For both models, the correlation decreases as the true lower asymptote increases, regardless of attribute type. The gMLTM-D and MLTM-D items yield person estimates that correlate roughly equally with the true person abilities for each design point; the correlations are never consistently better from one model or the other, even for a single test. On average, on tests with non-zero guessing (i.e., $\mu_\gamma = 0.125$, μ_γ 0.25), the correlations for the gMLTM-D person estimates exceed those of the MLTM-D estimates.

Summary of person estimates

All three criteria tend to indicate better person parameter recovery for the gMLTM-D, though with the limited sample in each table cell no formal tests can be conducted. Generally, the better abilities estimates coincide with those test conditions and models where the item parameters are also better estimated; similar trends in RMSE, bias, and correlations for both person and item estimates are observed under gMLTM-D and MLTM-D. Although only one test for each test condition and sample size was used to demonstrate person parameter recovery, the findings at each level are of practical use for determining whether and when to use the gMLTM-D or MLTM-D.

Table 43

Correlations between first component person estimates and true values

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1200	0.9131	0.9215	0.9108	0.9070	0.9146	0.9021	0.9048	0.9071
	3000	0.9186	0.9263	0.9172	0.9123	0.9345	0.8996	0.8962	0.9101
	4800	0.9165	0.9197	0.9240	0.9056	0.9176	0.9050	0.9054	0.9093
gMLTM-D Mean		**0.9160**	**0.9225**	**0.9173**	**0.9083**	**0.9176**	**0.9050**	**0.9054**	**0.9093**
MLTM-D	1200	0.9123	0.9218	0.9108	0.9044	0.9151	0.9015	0.9029	0.9065
	3000	0.9176	0.9264	0.9166	0.9098	0.9345	0.8991	0.8945	0.9093
	4800	0.9152	0.9198	0.9227	0.9032	0.9175	0.9044	0.9031	0.9083
MLTM-D Mean		**0.9150**	**0.9226**	**0.9167**	**0.9058**	**0.9175**	**0.9044**	**0.9031**	**0.9083**

Table 44

Correlations between second component person estimates and true values

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1200	0.9359	0.9255	0.9122	0.9245	0.9292	0.8934	0.8955	0.9061
	3000	0.9330	0.9179	0.8930	0.9146	0.9181	0.9287	0.9198	0.9222
	4800	0.9225	0.9206	0.9040	0.9157	0.9118	0.9189	0.9032	0.9113
gMLTM-D Mean		**0.9305**	**0.9213**	**0.9031**	**0.9183**	**0.9118**	**0.9189**	**0.9032**	**0.9113**
MLTM-D	1200	0.9360	0.9246	0.9106	0.9238	0.9296	0.8924	0.8937	0.9052
	3000	0.9331	0.9171	0.8911	0.9137	0.9180	0.9283	0.9179	0.9214
	4800	0.9227	0.9197	0.9010	0.9145	0.9118	0.9178	0.9015	0.9104
MLTM-D Mean		**0.9306**	**0.9205**	**0.9009**	**0.9173**	**0.9118**	**0.9178**	**0.9015**	**0.9104**

Table 45

Correlations between third component person estimates and true values

		Saturated Models				Attribute Models			
Model	Sample Size	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean	$\mu_\gamma = 0$	$\mu_\gamma = 0.125$	$\mu_\gamma = 0.25$	Mean
gMLTM-D	1200	0.9207	0.8892	0.8963	0.9020	0.8961	0.9076	0.8919	0.8985
	3000	0.9303	0.9152	0.9062	0.9173	0.9117	0.9114	0.9039	0.9090
	4800	0.9199	0.8924	0.9024	0.9049	0.9136	0.9030	0.9027	0.9064
gMLTM-D Mean		**0.9236**	**0.8989**	**0.9016**	**0.9081**	**0.9136**	**0.9030**	**0.9027**	**0.9064**
MLTM-D	1200	0.9210	0.8881	0.8948	0.9013	0.8961	0.9063	0.8901	0.8975
	3000	0.9306	0.9143	0.9037	0.9162	0.9118	0.9109	0.9013	0.9080
	4800	0.9201	0.8916	0.8993	0.9037	0.9136	0.9028	0.9010	0.9058
MLTM-D Mean		**0.9239**	**0.8980**	**0.8993**	**0.9070**	**0.9136**	**0.9028**	**0.9010**	**0.9058**

CHAPTER 5

DISCUSSION

This final chapter includes discussion of the major findings of this study. The main concentration is on the relative merits of the MLTM-D and the gMLTM-D in different test and item design contexts. The implications of the findings is then discussed, followed by an outline of the limitations of the current study. The paper concludes with recommendations for areas of future study, as directed and identified by the current study and findings.

Discussion of Findings

It was hypothesized that the gMLTM-D would produce better item and person estimates than the MLTM-D, particularly for tests with a non-zero probability for obtaining a false positive, and that item parameter estimation would improve, regardless of model, as the sample size increased. It was further hypothesized that better person estimates would be obtained from gMLTM-D item parameter estimates under the same conditions. The results support these hypotheses to an extent. As a whole, the gMLTM-D outperformed the MLTM-D in terms of item parameter estimation, with better results on all decision criteria except for correlations between true and estimated attribute weights. The gMLTM-D-estimated component discriminations and attribute weights were both significantly less biased and more precise than the corresponding MLTM-D estimates, even under the two test conditions where the models are functionally equivalent.

Recovery of the lower asymptotes was less successful than that of the other item parameters, regardless of the metric used. Although the inclusion of the asymptote

estimate in the model specification improved the estimation of all gMLTM-D item parameter estimates, the lower asymptote estimates themselves were poorly estimated, when they were estimated at all. Particularly for the saturated models, there were replications for 4,800 simulees where all lower asymptotes were not successfully estimated.

In the absence of guessing the gMLTM-D and the MLTM-D are functionally equivalent, and one would expect the two models to yield equal item parameter estimates. Despite the noted difficulty estimating $\gamma = 0$, however, the gMLTM-D discrimination and attribute weight estimates were significantly less biased and significantly more precise than the corresponding estimates from the MLTM-D. The unexpected, extreme difference in parameter estimates from two equivalent models leads the author to believe that the estimation algorithm within the NLMIXED software is the cause, and not something inherent in the model. As the RSSEs values of the MLTM-D estimates are all smaller than those of the gMLTM-D, there must be some other cause for the discrepancy between the two model's estimates for the non-guessing conditions.

Although the gMLTM-D outperformed the MLTM-D on all metrics in item parameter estimation, there was little difference between the two models in recovery of person parameters on all criteria. The 16 tests chosen for person measurement were selected based on the gMLTM-D asymptote estimation, where the fewest asymptotes were held at a boundary condition during the estimation process. The individual tests selected were representative of all simulated tests in terms of bias and RMSE of item parameter estimates. Firm statistical conclusions about ability parameter recovery cannot be made, as only one test was selected from each design point.

Implication of Findings

Administrators who are interested in latent trait diagnostic information but are also concerned about the impact of guessing and partial knowledge on item calibration and their subsequent ability estimation should view these findings as a positive, promising first step. The results of the real data analysis indicate that the gMLTM-D is practical for use on currently existing tests that have items that can be scored with a sparse, hierarchical component and attribute structure. It should be noted that item calibration under the gMLTM-D, particularly for quality estimation of lower asymptotes, is only feasible with very large sample sizes under the current technology. Smaller samples can produce some lower asymptote estimates under the condition of the attribute model, but as the simulation study revealed, those estimates do not correlate highly with the true values.

The noted relationship between increasing sample size and increasing correlation for the lower asymptote estimates should be considered when implementing the gMLTM-D, and it should only be used for large-scale testing. If one uses the MLTM-D in the presence of guessing, one must be cognizant of the increasing impact on the person ability estimates as the guessing probability increases.

The simulation study was designed to reproduce the conditions of a test administered in an academic setting, specifically using the seventh-grade mathematics achievement test as a blueprint for the basic design. The items, with an average easiness of approximately 0.7, mimic tests of academic achievement, which are generally easier than tests of aptitude. One outcome of simulating relatively easy items is difficulty estimating a lower asymptote, particularly with a standard normal distribution. If there is

poor representation of the lower abilities then there will be fewer simulees with sufficiently low abilities who need to guess on any item. When one considers the matching of the simulated person distribution, with a latent trait mean greater than the single-component item means, one realizes that few students would guess on any item, even those with small lower asymptotes; it is little surprise that there was poor lower asymptote recovery for the smaller sample sizes, particularly for the saturated models. Aptitude tests are intended to gauge a person's intelligence or ability to learn, not what one has already learned in school, and so items tend to be more difficult; had the current study modeled more difficult items like would typically be found on an aptitude or intelligence test, more simulees from the sampled population would have had to guess, and the lower asymptotes would have been better recovered even at the smaller sample sizes.

Limitations

As with any simulation study, one must be careful about generalizing these results to other testing scenarios. The test designs were tightly controlled, and the **C** and **Q** matrices did not vary at all throughout the simulation. In normal testing administrations, forms seldom have identical structures so this is an unlikely scenario to encounter outside the simulation. Due to time constraints, the simulation only consider a standard multivariate normal distribution for the examinee abilities; it is certainly possible on achievement tests of a unified construct that abilities may be correlated and not independent. The current study did not investigate the possibility of correlated abilities, however; nor did it investigate the possibility of non-normal ability distributions, which

would further impact the estimation of the lower asymptote, depending on the skewness of the population.

A major limitation for this study was computation time and power. Estimation of a single replication of the MLTM-D item parameter estimates could take 1.5 to 30 hours, with smaller sample sizes and attribute models taking less time. Estimation of the gMLTM-D—particularly for the saturated models—took noticeably longer, because it is a less parsimonious model. The gMLTM-D estimation could take 12 to 300 hours, where the smaller sample sizes and attribute models took less time. The two test conditions that took the longest to run under the gMLTM-D were the attribute and saturated model with a true lower asymptote of zero. As mentioned in Chapter 4, many estimates for the lower asymptotes in those cases got held at a boundary constraint and were never estimated, regardless of sample size. Some investigation into the issue has led the author to believe this is an algorithmic problem with the software and is not specific to the model itself.

Recommendations for Future Study

The results and limitations of the current study point to some interesting areas for future research.

- Performance of the gMLTM-D on different test lengths. The current study was time-limited and could only investigate one test length. For completeness, shorter and longer tests should be investigated.
- As discussed in the literature review, the recommended number of alternatives for an MC item is three. The current study simulated items with eight and four alternatives; simulating a test under the recommended scenario would be warranted.

- Investigation into larger sample sizes. Recovery of the lower asymptote was better on tests administered to larger samples; for recommendation or rule-of-thumb to administrators, more study is needed to determine a minimum sample size for lower asymptote estimation for the gMLTM-D.
- Improved matching of lower asymptotes to item difficulties. Non-zero lower asymptotes would likely be better estimated on tests with harder items, where guessing is more likely to occur. The ability distribution should also be matched to the item distribution so that the persons taking the test are likely to guess when guessing is expected.
- More understanding is needed of the NLMIXED algorithm and start values. The gMLTM-D and MLTM-D are the same model with the lower asymptote is zero, yet different estimates were obtained under the model specification. Alternative specifications result in identical estimates, but no lower asymptote estimates for the gMLTM-D. Both resulted in biased estimates when there should be no bias.

The current results show that the gMLTM-D is promising and offers advantages over the MLTM-D, but that more study is warranted before it is implemented in a testing program. The gMLTM-D provides unbiased item and person estimates in the presence of guessing on comprehensive, multidimensional tests. However, the requirement of large sample sizes, particularly for estimating tests designed under saturated model, and longer estimation times are a drawback to the gMLTM-D. As guessing and partial knowledge are a concern and are only addressed by a handful of latent trait models, the generalization of the MLTM-D is an important step in measurement.

APPENDIX A

Sample MLTM-D Item Parameter Estimation

```
PROC NLMIXED DATA = mltmd.allTC6r27J1,200 NOAD QPOINTS = 5 MAXITER = 1000 TECH=quanew;
     PARMS       a1-a3 = 1 *start values for item parameters, based on simulated values;
                 b1_1-b1_10
                 b2_1-b2_10
                 b3_1-b3_10 = 0.9;
     BOUNDS  a1-a3 > 0;

     int1 = b1_1*q1_1 + b1_2*q1_2 + b1_3*q1_3 + b1_4*q1_4 + b1_5*q1_5 + b1_6*q1_6 + b1_7*q1_7 +
             b1_8*q1_8 + b1_9*q1_9 + b1_10*q1_10;

     int2 = b2_1*q2_1 + b2_2*q2_2 + b2_3*q2_3 + b2_4*q2_4 + b2_5*q2_5 + b2_6*q2_6 + b2_7*q2_7 +
             b2_8*q2_8 + b2_9*q2_9+b2_10*q2_10;;

     int3 = b3_1*q3_1 + b3_2*q3_2 + b3_3*q3_3 + b3_4*q3_4 + b3_5*q3_5 + b3_6*q3_6 + b3_7*q3_7 +
           b3_8*q3_8 + b3_9*q3_9 + b3_10*q3_10;

*model specification based on intercept model, not difference model;
     eta1 = 1.7*(a1*theta1 + int1);
     eta2 = 1.7*(a2*theta2 + int2);
     eta3 = 1.7*(a3*theta3 + int3);

     p1 = 1/(1+exp(-eta1));
     p2 = 1/(1+exp(-eta2));
     p3 = 1/(1+exp(-eta3));

      p = min(max(exp(c_1*log(p1)+c_2*log(p2)+c_3*log(p3)),0.00001),1-0.00001);

     MODEL score ~ BINARY(p);
     RANDOM theta1 theta2 theta3 ~ NORMAL([0,0,0],[1,0,1,0,0,1]) SUBJECT = ID;
RUN;
```

APPENDIX B

Sample gMLTM-D Item Parameter Estimation

```
PROC NLMIXED DATA = gMLTMD.allTC6r1J1,200 NOAD QPOINTS = 5 MAXITER = 1000 TECH=quanew;

     PARMS       a1-a3 = 1 *start values for item parameters, based on simulated values;
                 g1-g5 g7-g75 = 0.25
                 b1_1-b1_10
                 b2_1-b2_10
                 b3_1-b3_10 = 0.9;
     BOUNDS  a1-a3 > 0, 0 <= g1-g5 g7-g75 <=.5;

     g6=0.25; *reference item for lower asymptotes ;


     g = g1*item1 + g2*item2 + g3*item3 + g4*item4 + g5*item5 + g6*item6 + g7*item7 + g8*item8 + g9*item9
           + g10*item10 + g11*item11 + g12*item12 + g13*item13 + g14*item14 + g15*item15 + g16*item16
                 + g17*item17 + g18*item18 + g19*item19
           + g20*item20 + g21*item21 + g22*item22 + g23*item23 + g24*item24 + g25*item25 + g26*item26
                 + g27*item27 + g28*item28 + g29*item29
           + g30*item30 + g31*item31 + g32*item32 + g33*item33 + g34*item34 + g35*item35 + g36*item36
                 + g37*item37 + g38*item38 + g39*item39
           + g40*item40 + g41*item41 + g42*item42 + g43*item43 + g44*item44 + g45*item45 + g46*item46
                 + g47*item47 + g48*item48 + g49*item49
           + g50*item50 + g51*item51 + g52*item52 + g53*item53 + g54*item54 + g55*item55 + g56*item56
                 + g57*item57 + g58*item58 + g59*item59
           + g60*item60 + g61*item61 + g62*item62 + g63*item63 + g64*item64 + g65*item65 + g66*item66
                 + g67*item67 + g68*item68 + g69*item69
           + g70*item70 + g71*item71 + g72*item72 + g73*item73 + g74*item74 + g75*item75;
```

```
     int1 = b1_1*q1_1 + b1_2*q1_2 + b1_3*q1_3 + b1_4*q1_4 + b1_5*q1_5 + b1_6*q1_6 + b1_7*q1_7 +
            b1_8*q1_8 + b1_9*q1_9 + b1_10*q1_10;

     int2 = b2_1*q2_1 + b2_2*q2_2 + b2_3*q2_3 + b2_4*q2_4 + b2_5*q2_5 + b2_6*q2_6 + b2_7*q2_7 +
            b2_8*q2_8 + b2_9*q2_9 + b2_10*q2_10;

     int3 = b3_1*q3_1 + b3_2*q3_2 + b3_3*q3_3 + b3_4*q3_4 + b3_5*q3_5 + b3_6*q3_6 + b3_7*q3_7 +
            b3_8*q3_8 + b3_9*q3_9 + b3_10*q3_10;

*model specification based on intercept model, not difference model;

     eta1 = 1.7*(a1*theta1 + int1);
     eta2 = 1.7*(a2*theta2 + int2);
     eta3 = 1.7*(a3*theta3 + int3);

     p1 = 1/(1+exp(-eta1));
     p2 = 1/(1+exp(-eta2));
     p3 = 1/(1+exp(-eta3));

     p = min(max(g + (1 - g)*(exp(c_1*log(p1)+c_2*log(p2)+c_3*log(p3))),0.00001),1-0.00001);

     MODEL score ~ BINARY(p);
     RANDOM theta1 theta2 theta3 ~ NORMAL([0,0,0],[1,0,1,0,0,1]) SUBJECT = ID;

RUN;
```

APPENDIX C

Test Simulation

```
simGMod<-function(ipar,cors, M, J, D=1.7, easiness=T, seed=1){
    call<-match.call()
    I<-nrow(ipar)
    if(M == 1){
        sigma <-1
    } else if (M >1){
        sigma <- as.matrix(cors)
        sigma[upper.tri(sigma)]<-t(sigma)[upper.tri(sigma)]
        if(dim(sigma)[1]!=dim(sigma)[2]||M!=dim(sigma)[1])
            stop("ERROR: number of dimensions and covariance matrix
            non-conforming")
        } else stop("ERROR: number of dimensions and covariance matrix
            non-conforming")
    if(M == 1){
        sigma.inv=1
    } else sigma.inv<-solve(sigma)
    gg<-matrix(ipar[,1])
    aa<-matrix(ipar[,2])
    bb<-matrix(ipar[,M:(3+M-1)],ncol=M)
    cc<-matrix(ipar[,(3+M):(3+M+M-1)],ncol=M)
    if(!easiness)
        bb<--bb
    if(!is.null(seed))
        set.seed(seed)
    TH<-matrix(rnorm(J*M),J,M) #create an initial theta matrix
    random<-matrix(runif(J*I),J,I)
    if(all(sigma!=0)&&M>1)
        TH<-TH %*%chol(sigma)
    resp<-matrix(0,J,I)
    P<-matrix(NA,J,I)
        for (j in 1:J){ ##added
            for (i in 1:I){
                Pm.comp<-matrix(0,nrow=M)
                for(m in 1:M){
                    Pm.comp[m]<-Pm.comp[m]+(1+exp(-D*(TH[j,m]-
                        bb[i,m])))^(-cc[i,m])
                }
#calculate the probability of correctly solving item i for examinee j
                P[j,i]<-gg[i]+(1-gg[i])*
                  Reduce("*",Pm.comp,accumulate=FALSE)
               resp[j,i]<-resp[j,i]+ifelse(random[j,i]<P[j,i],1,0)
            }
      } ##end loop through J people
    resp<-as.data.frame(resp)
    names(resp)<-paste("S",1:I,sep="")
    out<-list(call=call,theta=TH,resp=resp)
    return(out)
}#end function simGMod
```

REFERENCES

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1*(4), 363-378.

Andrich, D., & Styles, I. (2011). Understanding Rasch measurement: Distractors with information in multiple choice items: A rationale based on the Rasch model. *Journal of Applied Measurement, 12*(1), 67-95.

Arkin, R.M., & Walts, E.A. (1983). Performance implications of corrective testing. *Journal of Educational Psychology, 75*(4), 561-571.

Baker, F.B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17*(3), 236-251.

Bickel, J.E. (2010). Scoring rules and decision analysis education. *Decision Analysis, 7*(4), 346-357.

Birnbaum, A. (1968). Some latent trait models. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison–Wesley. Chapters 17-20.

Bligh, H. (1979). Achievement testing--A look at trends. Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April).

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement, 17*(2), 147-153.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *47*(2), 443-459.

Boldt, R.T. (1971). A simple confidence testing format. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED056098

Burton, R.F. (2004). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment and Evaluation in Higher Education,* 30(1), 65-72.

Chang, S., Lin, P., & Lin, Z. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology & Society, 10*(4), 95-109.

Chernoff, H. (1962). The scoring of multiple choice questionnaires. *The Annals of Mathematical Statistics, 33*(2), 375-393.

Chevalier, S.A. (1998). A review of scoring algorithms for ability and aptitude tests. Paper presented at the Annual Meeting of the Southwestern Psychological Association (New Orleans, LA, April).

Chiu, C., Douglas, J.A., Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika,74*(4), 633-665.

Choi, S.W. (2011). MAT: Multidimensional adaptive testing (MAT). R package version 0.1-3. http://CRAN.R-project.org/package=MAT

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16*, 13–37.

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

Cross, L.H. (1975). An investigation of a scoring procedure designed to eliminate score variance due to guessing in multiple-choice tests. Paper presented at the Annual

Meeting of the National Council on Measurement in Education (Washington, D.C., April).

Cross, L.H., & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement, 14*(4), 313-321.

Cross, L.H., Thayer, N.J., & Frary, R.B. (1980). A new method for administering and scoring multiple-choice tests: theoretical considerations and empirical results. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 1980).

Cuff, N.B. (1932). Scoring objective tests. *Journal of Educational Psychology, 23*(9), 681-686.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *The British Journal of Mathematical and Statistical Psychology, 18*(1), 87-123.

de Gruijter, D.N. (1984). A comment on 'some standard errors in item response theory.' *Psychometrika, 49*(2), 269-272.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.

DeMars, C.E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23-45.

Dimitrov, D.M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement, 31*(5), 367-387.

Dimitrov, D.M., & Atanasov, D.V. (2012). Conjunctive and disjunctive extensions of the least squares distance model of cognitive diagnosis. *Educational and Psychological Measurement, 72*(1), 120:138.

Echternacht, G. (1971). The use of confidence testing in objective tests. Retrieved from http://www.eric.edu.gov/contentdelivery/servlet/ERICServlet?accno=ED058307

Echternacht, G. (1973). A note on the variances of empirically derived option scoring weights. Retrieved from http://www.eric.ed.gov/PDFS/ED073173.pdf

Echternacht, G.J., Boldt, R.F., & Sellman, W.S. (1972). Personality influences on confidence test scores. *Journal of Educational Measurement, 9*(3), 235-241.

Edgerton, H.A., & Stoloff, P.H. (1967). A note on test item difficulty. *Educational and Psychological Measurement, 27*(2), 261-265.

Eignor, D.R., & Douglass, J.B. (1982). A comparison of the one-, the modified three-, and the three-parameter item response theory models in the test development item selection process. Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, March).

Embretson, S.E. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175-186.

Embretson, S.E. (1998). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson. S. E., & Yang. X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*(1), 14-36.

Feasel, K., Henson, R., & Jones, L. (2004). *Analysis of the Gambling Research Instrument (GRI).* Unpublished manuscript.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Foster, R.R., & Ruch, G.M. (1927). On corrections for chance in multiple-response tests. *Journal of Educational Psychology, 18*(1), 48-51.

Frary, R.B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*(1), 79-96.

Gibbons, J.D., Olkin, I., & Sobel, M. (1979). A subset selection technique for scoring items on a multiple choice test. *Psychometrika, 44*(3), 259-270.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment.

Glass, G.V., & Wiley, D.E. (1964). Formula scoring and test reliability. *Journal of Educational Measurement, 1*(1), 43-49.

Goegebeur, Y., De Boeck, P., Wollack, J.A., & Cohen, A.S. (2008). A speeded item response model with gradual process change. *Psychometrika, 73*(1), 65-87.

Granich, L. (1931). A technique for experimentation on guessing in objective tests. *Journal of Educational Psychology, 22*(2), 145-156.

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

Hakstian, A.R., & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple choice tests: II. Testing procedures. *Journal of Educational Measurement, 12*(4), 231-239.

Haladyna, T.M. (1984). Increasing information from multiple-choice test items. Paper presented at the annual meeting of the American Educational Research Association (New Orleans, LA, April).

Haladyna, T.M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*(1), 73-88.

Haladyna, T.M., & Downing, S.M. (1993). How many options is enough for a multiple choice test item? *Educational & Psychological Measurement, 53*(4), 999-1010.

Hansen, R. (1971). The influence of variables other than knowledge on probabilistic tests. J*ournal of Educational Measurement, 8*(1), 9-14.

Hartz, S. (2002) *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation.

Hattie, J.A. (1981). Decision criteria for determining unidimensionality. *Open Access Dissertations and Theses.* Retrieved from http://digitalcommons.mcmaster.ca/opendissertations/2212/

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277.

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.

Horst, P. (1932). The chance element in the multiple choice test item. *Journal of General Psychology, 6*, 209-211.

Horst, P. (1932b). The difficulty of multiple choice test item alternatives. *Journal of Experimental Psychology, 15*(4), 469-472.

IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Jersild, A.T. (1929). Examination as an aid to learning. *Journal of Educational Psychology, 20*(8), 602-609.

Johnson, M.S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software, 20*(10), Retrieved from http://www.jstatsoft.org/v20/i10/paper.

Jones, H.E. (1929). A comparison of objective examination methods. *Journal of Educational Method,* 273-276.

Jun, H. (2013). Diagnostic measurement from a standardized math achievement test using multidimensional latent trait models. (Unpublished master's thesis). Georgia Institute of Technology: Atlanta, GA.

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.

Kansup, W., & Hakstian, A.R. (1975). A comparison of several methods of assessing partial knowledge in multiple choice tests: I. Scoring procedures. *Journal of Educational Measurement, 12*(4), 219-230.

Kubinger, K.D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection & Assessment, 18*(1), 111-115.

Kurz, T.B. (1999). A review of scoring algorithms for multiple-choice tests. Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January).

Little, E., & Creaser, J. (1966). Uncertain responses on multiple-choice examinations. *Psychological Reports, 18*(3), 801-802.

Lord, F.M. (1965). A note on the normal ogive or logistic curve in item analysis. *Psychometrika*, *30*(3), 371-372.

Lord, F.M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*(1), 7-11.

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores.* Oxford, England: Addison-Wesley.

Lutz, M. (2012). Identifying and measuring cognitive aspects of a mathematics achievement test. (Unpublished master's thesis). Georgia Institute of Technology: Atlanta, GA.

Lutz, M., & Embretson, S.E. (2012). Predicting the psychometric properties of seventh grade mathematics achievement items from standards-based and cognitive-based models of item complexity. Report IES1002A-2012 for Institute of Educational Sciences Grant R305A100234. *Cognitive Measurement Laboratory*, Georgia Institute of Technology: Atlanta, GA.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives, 7*(2), 75-88.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Masters, G.N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education, 1*(4), 279-297.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied* Psychological Measurement 16(2), 159-176.

Muraki, E., & Bock, R.D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data [computer software]. Chicago, IL: SSI.

Mokken, R.J, & Lewis, C. (1982). A nonparameteric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417-430.

Nugent, R., Dean, N., Ayers, E. (2010). Empty K-means: A flexible skill set profile clustering method. Paper presented at the 75th meeting of the International Meeting of the Psychometric Society (Athens, Georgia, United States, July 2010).

Partchev, I. (2009). 3PL: A useful model with a mild estimation problem. Measurement: Interdisciplinary Research and Perspectives, 7(2), 94-96.

Pascale, P.J. (1971). Innovations in item scoring procedures. Retrieved from http://www.eric.edu.gov/contentdelivery/servlet/ERICServlet?accno=ED056096

Penfield, R.D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement, 45*(3), 247-269.

Plake, B.S., Wise, S.T., & Harvey, A.L. (1988). Test-taking behavior under formula and number-right scoring conditions. *Bulletin of the Psychonomic Society, 26*(4), 316-318.

Potthoff, E.F., & Barnett, N.E. (1932). A comparison of marks based upon weighted and unweighted items in a new-type examination. *Journal of Educational Psychology, 23*(2), 92-98.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J.O. (1968). A scoring system for multiple choice tests. *British Journal of Mathematical and Statistical Psychology, 21*(2), 247-263.

Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambelton (Eds.), *Handbook of modern item response theory*. (pp. 271-286). New York: Springer-Verlag.

Reckase, M. D. (1997b). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-27.

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues And Practice*, *24*(2), 3-13.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal Of Educational Measurement*, *44*(4), 293-311.

Ruch, G.M., & Degraff, M.H. (1926). Corrections for chance and "guess" vs. "do not guess" instructions in multiple response tests. *Journal of Educational Psychology, 17*(6), 368-375.

Rupp, A. A., & Templin, J. (2007). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model.

Samejima, F. (1972). A paradox in the knowledge of guessing model for the multiple-choice item. *Proceedings of the Annual Convention of the American Psychological Association, 7*(1), 61-62.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38*(2), 221-233.

Scheidemann, N.V. (1933). Multiplying the possibilities of the multiple choice form of objective question. *Journal of Applied Psychology, 17*(3), 337-340.

Searle, L.V. (1942). Scoring formulae for a modified type of multiple-choice question. *Journal of Applied Psychology, 26*(5), 702-710.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33*(1), 75-102.

Sympson, J.B. (1977). A model for testing with multidimensional items. Paper presented at the 1977 Computerized Adaptive Testing Conference, Minneapolis, MN. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/

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

Thissen, D., Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*(2), 161-176.

Tollefson, N., & Chung, J. (1986). A comparison of two methods of assessing partial knowledge on multiple-choice tests. Technical Report for Kansas University – General Research Fund.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16).

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement*, *7*(1), 67-74.

Wang, J., & Calhoun, G. (1997). A useful function for assessing the effect of guessing on true–false and multiple-choice tests. *Educational and Psychological Measurement, 57*(1), 179-185.

Walker, D.M., & Thompson, J.S. (2001). A note on multiple choice exams with respect to students' risk preference and confidence. *Assessment & Evaluation in Higher Education*, *26*(3), 260-267.

White, P.O. (1976). A note on Keats' generalization of the Rasch model. *Psychometrika, 41*(3), 405-407.

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479-494.

Wilcox, R.R., & Wilcox, K. (1988). A note on decisionmaking processes for multiple-choice test items. *Journal of Educational Measurement, 25*(3), 247-250.

Wisner, J.D., & Wisner, R.J. (1997). A confidence-building multiple-choice testing procedure. *Business Education Forum, 51*(4), 28-31.

Yunker, B.D. (1999). Adding authenticity to traditional multiple choice test formats. *Education, 120*,(1), 82-87.

Zimmerman, D.W. (1969). A simplified probability model of error of measurement. *Psychological Reports, 25*(1), 175-186..

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). Bilog-MG (Version 3.0) [computer software]. Lincolnwood, IL: SSI.