ABSTRACT

Title of Dissertation:	A MIXED-STRATEGIES RASCH TESTLET MODEL FOR LOW-STAKES TESTLET-BASED ASSESSMENTS
	Ying-Fang Chen, Doctor of Philosophy, 2013
Directed By:	Dr. Hong Jiao Department of Human Development and Quantitative Methodology

In low-stakes assessments, a lack of test-taking motivation inevitably occurs because test scores impose inconsequential effects on test takers' academic records. A common occurrence is that some test takers are unmotivated and simply apply random guessing strategy rather than solution strategy in taking a test. Testlet effects also arise because educational assessment items are frequently written in testlet units. A challenge to psychometric measurement is that conventional item response theory models do not sufficiently account for test-taking motivation heterogeneity and testlet effects. These construct-irrelevant variances affect test validity, accuracy of parameter estimates, and targeted inferences. This study proposes a low-stakes assessment measurement model that can simultaneously explain test-taking motivation heterogeneity and testlet effects. The performance and effectiveness of the proposed model are evaluated through a simulation study. Its utility is demonstrated through an application to a real standardized low-stakes assessment dataset. Simulation results show that overlooking test-taking motivation heterogeneity and testlet effects adversely affected model–data fit and model parameter estimates. The proposed model improved model–data fit and classification accuracy and well recovered model parameters under test-taking motivation heterogeneity and testlet effects. For the real data application, the item response dataset, which was originally calibrated with the Rasch model, was fitted better by the proposed model. Both test-taking motivation heterogeneity and testlet effects were identified in the real dataset. Finally, a set of variables selected from the real dataset is used to explore potential factors that characterize the latent classes of test-taking motivation. In the science assessment, science proficiency was associated with test-taking motivation heterogeneity.

A MIXED-STRATEGIES RASCH TESTLET MODEL FOR LOW-STAKES TESTLET-BASED ASSESSMENTS

By

Ying-Fang Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Hong Jiao, Chair
Professor Robert Croninger
Professor Jeffrey R. Harring
Professor Robert W. Lissitz
Professor George Macready

Dedication

To my dear parents and beloved husband,

for their endless love and support.

Acknowledgments

My deepest gratitude goes to Dr. Hong Jiao, whose sound guidance, expertise, and continual support and encouragement makes this dissertation possible. Doing research with her is among the most invaluable experiences of my life!

I thank all the dissertation committee members for their careful review and insightful feedback toward the improvement of the quality of this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Statement of the Problem

Low-stakes educational assessments, such as National Assessment of Educational Progress, Progress in International Reading Literacy Study, Programme for International Student Assessment (PISA), and Trends in International Mathematics and Science Study, have been increasingly directed toward recording and monitoring students' academic progress for the past several years. These assessment measures are intended to collect information about student achievement and performance in targeted domains. The outcomes of these assessments inform how well students are prepared for the future and determine the accountability of school systems, institutes, programs, and teacher instruction. Test results also serve as essential references for creating educational policies that intend to improve students' overall competence. Achieving these goals necessitates a valid measurement tool that is free from construct-irrelevant noise. Consequently, the adequacy of psychometric measurements has become crucial to formulating better measures of student skills, knowledge, and abilities.

The common current practice is to use item response theory (IRT; Lord, 1980) for developing standardized educational assessments. IRT models define the mathematical relation between observable item performance and an examinee’s unobservable ability. An examinee’s probability of providing a correct response is predicted by his/her ability and the characteristics of the item (Hambleton, Swaminathan, & Rogers, 1991). IRT models are grounded on two key assumptions: (1) unidimensionality and (2) local independence (de Ayala, 2009; Embretson & Reise, 2000; Hambleton et al., 1991; Hambleton & Swaminathan, 2010; Yen & Fitzpatrick, 2006). Unidimensionality pertains to the idea that only a single ability underlines the respondents’ performance on a set of items. Local independence holds “when the relationship among items (or persons) is fully characterized by the IRT model” (Embretson & Reise, 2000, p. 48). According to Reckase (2009), the assumption of local independence is that “the probability of a collection of responses (responses of one person to the items on a test, or the responses of many people to one test item) can be determined by multiplying the probabilities of each of the individual responses” (p. 13). This statement implicitly suggests that no clustering dependence among items and that no clustering effects among persons. Local item independence and local person independence can be mathematically represented by Equations (1) and (2), respectively, as follows (Jiao, Kamata, Wang, & Jin, 2012; Reckase, 2009).

$$P(U = u \mid \theta) = \prod_{i=1}^{I} P(u_i \mid \theta) = P(u_1 \mid \theta)P(u_2 \mid \theta)...P(u_I \mid \theta), \quad (1)$$

where $P(\mathrm{U} = u \mid \theta)$ is the probability of an item response vector u ($u = [u_1,.., u_\mathrm{I}]$) for a respondent with ability θ, and $P(u_i \mid \theta)$ denotes the probability of an individual item response u_i to item i for a respondent with ability θ. This expression shows that an examinee's joint probability of responses to a set of items is equal to the product of probabilities of individual items at a given ability. Equation (1) also indicates that a respondent's responses to a set of items are statistically independent.

$$P(U_i = u_i \mid \theta) = \prod_{j=1}^{n} P(u_{ij} \mid \theta_j) = P(u_{i1} \mid \theta_1) P(u_{i2} \mid \theta_2) \ldots P(u_{in} \mid \theta_n), \tag{2}$$

where U_i is the response vector to the item i by n respondents with abilities θ_j, and u_{ij} is the response of respondent j to the i^{th} item. This expression indicates that the probability of the responses to a single item i by n respondents is equal to the product of probabilities of individual respondents' responses (with abilities in the θ vector) to the item i.

In practice, however, the assumption of local independence usually cannot be stringently satisfied. The present study highlights two nuisance factors that can violate local independence but are often overlooked in the analysis of test results for standardized educational assessment data such as PISA. The first is person characterization—test-taking motivation—which makes test takers apply distinct test-taking strategies in taking a test and therefore introduces examinee heterogeneity into the data. The second factor is test item characterization—testlet effects—which refer to a group of homogeneous items that cluster dependently around a common stimulus (i.e., item clustering). When person heterogeneity and item dependence exist in the data, conventional IRT models are misspecified. If such misspecification is not taken into account, it results in inaccurate parameter estimates and targeted inferences (Jiao et al., 2012).

A basic requirement for obtaining accurate test results and obtaining valid inferences is that individual respondents should be motivated to take the test. Some standardized assessments, such as PISA, are low-stakes measures for test takers. That is, assessment results are used to draw inferences and have inconsequential effects on individual test takers' academic records. It is common that some test takers in low-stakes assessments are unmotivated and simply guess randomly during a test; this therefore makes their test scores an invalid reflection of the actual levels of knowledge, skills, and abilities. Although researchers have proposed different methods for enhancing test-taking motivation in low-stakes assessments (e.g., provision of incentives; see Cole, 2007; Wise & DeMars, 2005), a lack of test-taking motivation cannot be completely avoided. Besides, a number of studies have been devoted to direct and indirect measures of test-taking motivation, in which aberrant test takers are identified first and their item responses are excluded from subsequent data analyses. Examples of direct and indirect measures of test-taking motivation are motivation filtering (e.g., Sundre, & Wise, 2003; Swerdzewski, Harmes, & Finney, 2011) and person-fit indices (e.g., Armstrong & Shi; 2009; Cui, & Leighton, 2009; Glas & Dagohoy, 2007; Glas & Meijer, 2003; Karabatsos, 2003; Wise & Kong, 2005; see also Meijer &

Sijtsma, 2001). However, these approaches still have their own limitations which will be addressed in the next chapter.

Central to the issue of test-taking motivation heterogeneity in low-stakes assessments is that test takers are likely to behave differentially in taking a test; that is, adopt the solution strategy or random guessing strategy. A challenge to psychometric measurement is that conventional IRT measurement models including the Rasch model do not sufficiently account for such test-taking motivation heterogeneity. Given that test-taking motivation is unobservable, a latent class perspective permits one to treat test-taking motivation as a latent variable that distinguishes examinees into distinct latent examinee populations. Following this line of thinking, a different approach would be to incorporate a latent class model (Dayton, 1999; McCutcheon, 1987) for describing qualitative examinee heterogeneity in conventional IRT modeling; this approach is called IRT-based mixture modeling (Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990). A family of IRT-based mixture models has been applied by other researchers. These models effectively describe examinee attributes that point to qualitative heterogeneity indicators, such as latent differential item functioning (e.g., Cohen & Bolt, 2005; de Ayala, Kim, Stapleton, & Dayton, 2002; Kelderman & Macready, 1990; Maij-de Meij, Kelderman, & van der Flier, 2010; Samuelsen, 2005), heterogeneous test-taking strategies or motivation (e.g., Lau, 2009; Mislevy, & Verhelst 1990; Subedi, 2009), speededness (e.g., Bolt, Cohen, & Wollack, 2002; Meyer, 2010), faking or response style (e.g., Eid & Zickar, 2007; Zickar, Gibby, & Robie, 2004), and psychological attributes (e.g., Finch & Pierson, 2011; Smith, Ying, & Brown, 2012). Another appeal of IRT-based mixture modeling is that it offers the opportunity to predict latent class membership by using covariates (Samuselsen, 2005).

On these grounds, the present research (1) adopts the idea of IRT-based mixture modeling to capture test-taking motivation heterogeneity in low-stakes assessments; and (2) looks more closely into the potential factors that characterize latent class membership pointing to test-taking motivation heterogeneity in low-stakes assessment. The second objective is significant because there are few studies investigating the potential factors that are associated with unmotivated respondents after latent class members have been identified in previous studies.

In addition to test-taking motivation heterogeneity in low-stakes assessments, testlet effects may be another serious problem when applying the traditional IRT models. In most educational assessments, testlets that are deliberately constructed with a series of related items sharing a common stimulus are frequently used. The use of testlets (i.e., context-dependent items) introduces local item dependence in the estimation of IRT model parameters, which leads to an overestimation of test reliability and biased parameter estimates (Bradlow, Wainer, & Wang, 1999; Chen & Thissen, 1997; DeMars, 2006; Sireci, Thissen, &

Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993). To manage testlet effects, researchers have devoted considerable effort on expanded versions of IRT models, such as the polytomous IRT model (Cook, Dodd, & Fitzpatrick, 1999; Thissen et al., 1989) or testlet IRT model (Bradlow et al., 1999; Jiao, Wang, & Kamata, 2005; Wainer, Bradlow, & Du, 2000; Wainer & Wang, 2000; Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005). The above models that manage testlet effects are comprehensively reviewed in the next chapter. A deficiency in analyzing low-stakes testlet-based assessments, however, is that no adequate psychometric measurement model can simultaneously account for test-taking motivation heterogeneity (i.e., valid respondents and random guessers) and testlet effects.

1.2 The Purpose and Significance of the Study

This study aims to address the issues of test-taking motivation heterogeneity and testlet effects in low-stakes educational assessments; it also aims to propose a measurement model that simultaneously incorporates test-taking motivation heterogeneity and testlet effects of data. The development of the proposed measurement model focuses on resolving test-taking motivation heterogeneity and testlet effects to improve measurement accuracy. These efforts are worthwhile endeavors because the failure to incorporate examinee heterogeneity and item clustering (due to testlet effects) in traditional IRT models has become a critical concern that affects model parameter estimation accuracy and related inferences. Given that educational policies and related implications are typically data driven, to develop a measurement model for low-stakes assessments by this current research is important and necessary for improving parameter estimation intended for drawing inferences. Furthermore, this work will enable low-stakes assessment-based decisions and associated implications to be more fair and just. The proposed measurement model can also serve as a useful model-based filtering technique in low-stakes assessments because it is expected to identify unmotivated respondents who apply random guessing strategy based on item response data. In this research, the performance and effectiveness of the proposed model are evaluated through a simulation study and through an application to a real standardized low-stakes assessment dataset. Furthermore, this study has practical significance and implications for those who seek empirical evidence for the likelihood of factors associated with respondents who apply random guessing strategy in low-stakes assessments. As previously stated, very few researchers have investigated the empirical factors related to test-taking motivation heterogeneity in real low-stakes assessment data. In the current work, a follow-up exploratory investigation is conducted to empirically explore the potential factors that characterize heterogeneity, making the identification and interpretations of latent class membership more meaningful. Elucidating the qualitative interpretations of identified latent class membership is a worthwhile endeavor because understanding the attributes associated with unmotivated test takers who guess randomly can unravel the qualitative composition of latent classes and can help practitioners manage the issue of test-taking motivation in low-stake assessments. The findings of this research are expected to advance methodological knowledge on analyzing item response data from testlet-based low-stakes assessments, as well as to identify the attributes that are potentially associated with unmotivated test takers in low-stakes assessments. This research aims to answer the following questions:

1. What is the effect of overlooking heterogeneous test-taking motivation and testlet effects in low-stakes testlet-based assessments?
2. How well are model parameters recovered in the proposed model under the presence of heterogeneous test-taking motivation and testlet effects?
3. How does the proposed model perform in real low-stakes assessment data in terms of model–data fit? Are there unmotivated test takers and testlet effects identified empirically?
4. What are the potential factors that characterize test-taking motivation heterogeneity from real low-stakes assessment data?

1.3 Outline of Chapters

The remainder of the dissertation is divided into five chapters. Chapter Two reviews the literature that makes up the theoretical foundation of this dissertation. Chapter Three presents an overview of the proposed measurement model and comparison models, as well as the simulation study design, Bayesian estimation, data analyses, and empirical study. The results of the simulation are presented in Chapter Four, and the findings from the empirical study are provided in Chapter Five. Finally, Chapter Six summarizes and discusses the research results; this chapter also contains the limitations of the study and suggestions for future research.

Chapter 2: Literature Review

This research aims at developing a measurement model for low-stakes testlet-based assessments. This chapter reviews the issues that revolve around low-stakes assessments, such as (1) the problems related to testing-taking motivation and the practical approaches that are currently used to manage this psychological process; (2) different modeling approaches that can account for test-taking motivation heterogeneity; (3) testlet effects and the modeling approaches to accounting for such effects; and (4) the technique used to estimate parameters in this study: a Bayesian estimation with a Markov chain Monte Carlo algorithm.

2.1 Test-Taking Motivation in Low-Stakes Assessments

Test-taking motivation refers to "an examinee's drive to engage in and persist to the completion of a test" (Lau, 2009, p. 4). Such motivation can be conceptualized from the perspective of educational psychology—expectancy value theory (Pintrich & Schunk, 2002; Wigfield & Eccles, 2000; Wise & DeMars, 2005)—a test taker would be motivated if (s)he associates high expectancy or strong value beliefs with a particular assessment. *Expectancy* pertains to a test taker's evaluation of his/her ability to complete a test. *Value* encompasses attainment value (e.g., the importance of doing well on a test), intrinsic value (e.g., the enjoyment derived out of accomplishing a test), utility value (e.g., the benefits of doing well on a test), and perceived cost (e.g., the perceived cost of prioritizing a test over other more valued personal investments, such as time or energy).

In high-stakes assessments, test takers are universally motivated to devote committed efforts because results are linked to academic records which will be used for high-stakes decisions, such as admission or replacement in different instructional programs. In low-stakes assessments, however, test results carry no substantial consequence (e.g., low attainment and/or utility values) for individual examinees. Furthermore, test takers are compelled to sacrifice time or energy to take a test and in the process forgo other more valued activities (high perceived costs). Consequently, lack of test-taking motivation inevitably occurs in low-stakes assessments. For example, Mislevy and Verhelst (1990) indicated that in a college low-stakes reading assessment, proctors observed some examinees completing answer sheets without opening the test booklets.

In the literature, varied percentages of unmotivated test takers have been observed in low-stakes assessments. For example, in Brown and Gaxiola (2010) 4% of university students reported that they did not exert their best effort in a low-stakes information skills test. In Sundre and Wise (2003), 12.5% and 12.8% of university students were classified under the “very low” test-taking motivation category (scores below 20 out of 50), as measured by a self-report scale in a Nature World (NAW-5) test and in a quantitative reasoning quotient test, respectively. In Wise and DeMars (2005), 7.27% of test takers scored low (scores below 20 out of 50), as measured by a self-report motivation scale in a low-stakes US history and political science test. Wise and DeMars (2006) reported that in a university low-stakes Information literacy test, about 5% of the test takers were categorized as unmotivated on the basis of a post-test self-report motivation scale. Lau (2009) revealed that approximately 1.2% of university test takers were random responders, as identified by the mixed-strategies IRT modeling and a self-report test-taking motivation scale. Subedi (2009) found that approximately 5% of the test takers of a statewide mathematics assessment were unmotivated test takers, as classified via the mixed-strategies IRT modeling.

Unmotivated students who take low-stakes assessments deserve considerable attention and comprehensive investigation. The phenomenon of lack of motivation requires resolution because the test scores of unmotivated test takers provide inaccurate psychometric information for assessing test performance; they are also poor indicators of actual proficiency levels. In particular, test validity would be negatively influenced and average proficiency levels would be underestimated if respondents continue to be unmotivated. In an experimental study, for example, community college students performed significantly better in the graded exam than in the non-graded exam, with performance exhibiting a large effect size, Cohen's *d* = 1.27 (Napoli & Raymond, 2004). Wise and DeMars (2005) conducted a meta-analysis designed to compare 12 empirical and experimental studies. In their meta-analysis, test-taking motivation was measured by a self-report test-taking motivation scale or manipulated with external incentives during the experiment. Examples of external incentives include paying students to participate, awarding students with extra course points, or informing students about the importance of the test scores prior to test administration (see Table 1 in Wise & DeMars for more details). Their results indicate that the average test scores of unmotivated examinees were significantly lower than those of motivated examinees, with an average effect size of 0.59.

To mitigate the negative effects arising from the unmotivated completion of low-stakes assessments, researchers frequently use the following practical methods:

(1) Implementing treatments designed to increase test-taking motivation

Strategies for enhancing test-taking motivation are implemented prior to test administration. Examples of such treatments are offering incentives and explaining the importance/benefits of taking a test (e.g., Cole, 2007; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; Wise & DeMars, 2005). These treatments are practical and easily implementable, but nonetheless require monetary or time investments. Using treatments to increase test-taking motivation also prevents the accurate determination of test takers' preferences for incentives. Put it differently, although a selected incentive (e.g., money) effectively works for some test takers, such incentive may be ineffective or may even pose a negative effect on others (e.g., O'Neil et al., 2005).

(2) Direct and indirect measures of test-taking motivation

Motivation filtering and person-fit measures can be used to detect suspected unmotivated respondents. These methods are then succeeded by statistical adjustments, such as the exclusion of suspected unmotivated respondents from data analyses. A commonly used motivation filtering tool is the self-report test-taking motivation scale (e.g., Sundre, 2007; Sundre, & Wise, 2003; Swerdzewski et al., 2011; Wise & DeMars, 2005; Wise & Kong, 2005; Wise, Pastor, & Kong, 2009) which is implemented immediately after a given test. Test takers respond to questions or prompts (e.g., "Doing well on this test is important to me." or "I engaged in good effort throughout this test.") using a Likert scale with a score range of 1 (strongly disagree) to 5 (strongly agree). However, employing a self-report test-taking motivation scale casts doubt on the accuracy of the results because unmotivated test takers may also provide random or false responses on the test-taking motivation scale. Furthermore, the adequacy of the cutoff point for classifying unmotivated and motivated respondents requires more empirical evidence (Swerdzewski et al., 2011; Wise & DeMars, 2005; Wise & Kong, 2005).

An alternative to measuring test-taking motivation is determining the length of response time (e.g., Swerdzewski et al., 2011; Wise & Kong, 2005; Wise, Pastor, & Kong, 2009), a method that uses item response time as a proxy for test-taking motivation. Unfortunately, this particular feature is also the origin of this method's limitations. First, the supposed consistency (i.e., convergent validity) between the manifest response time and unobservable test-taking motivation requires strong empirical support. For example, in Wise and Kong (2005), the correlation between the self-report test-taking motivation scale and the total test time was very low ($r = 0.22$); moreover, the self-report test-taking motivation scale and the index developed by the authors (i.e., Response Time Effort) were weakly correlated ($r = 0.25$). In Wise et al. (2009), the correlation between self-report test-taking motivation scale and Response Time Effort was negligible ($r = 0.06$). Second, the effectiveness of a time threshold in distinguishing response strategies also requires more evidence and examination (Wise & Kong, 2005). Finally, the collection and accuracy of item-level response time heavily depend on the administration of computer-based testing, which is not implemented in most traditional testing scenarios.

An indirect approach to detecting unmotivated test takers is through person-fit statistics (e.g., Armstrong & Shi; 2009; Cui, & Leighton, 2009; Glas & Dagohoy, 2007; Glas & Meijer, 2003; Karabatsos, 2003; Wise & Kong, 2005; see also Meijer & Sijtsma, 2001). Aberrant item respondents are flagged because their observed item response patterns are inconsistent with expected item response patterns. However, the unanticipated item response patterns detected by person-fit statistics may be due to different factors such as cheating (or answer copying), item disclosure effects, guessing, inattention/carelessness, a lack of motivation, creative responses, test anxiety, tendency to choose extreme options, or ignoring reverse wording (Cui, & Leighton, 2009; de la Torre & Deng, 2008; Emons, 2008; Emons, Sijtsma, & Meijer, 2004; Glas & Meijer, 2003; Karabatsos, 2003). The person-fit indices are excessively sensitive to different types of aberrancy; thereby suspected misfit test takers may not necessarily be unmotivated test takers (Wise & DeMars, 2006; Wise & Kong, 2005; Wise et al., 2009). Wise and Kong (2005) found that person-fit statistics and the self-report test-taking motivation scale were weakly correlated ($r = –0.17$); the authors concluded that person-fit indices and the self-report test-taking motivation scale may measure different constructs.

Apparently, the use of direct and indirect measures of test-taking motivation presents numerous unresolved problems. Given that traditional IRT models have an implicit homogeneity assumption related to test-taking motivation (Hambleton et al., 1991), identifying a suitable psychometric model is important in managing test-taking motivation heterogeneity that is encountered in low-stakes assessments. This study considers a psychometric modeling approach to accounting for test-taking motivation heterogeneity. The succeeding section introduces and discusses a modeling approach that incorporates test-taking motivation heterogeneity in analysis.

2.2 Accounting for Heterogeneous Test-Taking Motivation with Mixture Modeling

Distinct test-taking motivation drives test takers to behave differently during a test. Motivated test takers respond to items in a way (i.e., solution strategy) that reflects actual knowledge, skills, and abilities, whereas unmotivated test takers are apt to respond in a random fashion (i.e., random guessing strategy) in low-stake assessments because of low-stakes test results. The heterogeneity of test-taking motivation in low-stakes assessments generates qualitatively heterogeneous item response patterns which may distinguish between motivated and unmotivated test takers.

Unfortunately, test-taking motivation heterogeneity is unobservable. Furthermore, the traditional IRT models commonly used to calibrate educational assessments cannot adequately account for such examinee heterogeneity. One solution to capturing unobservable test-taking motivation heterogeneity is to use mixed-strategies IRT models (Lau, 2009; Subedi, 2009; Mislevy & Verhelst, 1990), which incorporate IRT and random guessing strategy models in a model and allow for different item response functions for distinct latent classes at the examinee level. The mixed-strategies IRT models are special cases of the HYBRID model (Yamamoto, 1987, 1989; Yamamoto & Gitomer, 1993).

The original HYBRID model (Yamamoto, 1987, 1989; Yamamoto & Gitomer, 1993) combines an IRT model with an LCA model (LCA; Dayton, 1999; McCutcheon, 1987) into a single model. It incorporates multiple response strategies, i.e., the target strategy of solving an item and other strategies that test takers may employ. It is assumed that "correct solutions indicate that the student has acquired the cognitive skills necessary to solve a problem, and incorrect solutions indicate some deficit in that set of skills" (Yamamoto & Gitomer, 1993, p. 276). Test takers who employ the demanded response strategy of solving items are modeled by an IRT two-parameter logistic (2PL) model, whereas test takers who represent a unique understanding or misunderstanding of the material being measured are modeled by an LCA model. Each test taker belongs to either the IRT group or one of the LCA groups. Within the IRT group, local independence and unidimensionality are assumed. In the HYBRID model, ability parameter is only meaningful to test takers whose item responses are best fitted by the IRT model.

Mislevy and Verhelst (1990) further extended Yamamoto's work (1987, 1989) and proposed an IRT-based mixed-strategies measurement model. The authors illustrated how IRT mixture modeling can be used when test takers employ different strategies. An example presented in the study is a mixture IRT model that comprises two latent classes, enabling the simultaneous identification of valid respondents and random guessers in an examinee population. A test taker belongs to only one exhaustive and exclusive latent class. The item response patterns of unmotivated test takers who provide random responses are assumed distinctly and qualitatively different from those of valid respondents. The latent class of valid respondents corresponds to the Rasch measurement model, whereas that of random guessers corresponds to a function of the chance of success. Mislevy and Verhelst used marginal maximum likelihood (ML) estimates with an expectation–maximization (EM) algorithm for model estimation. The authors used item response probabilities as bases in estimating latent class membership. They did not conduct a simulation to evaluate model effectiveness under varied testing conditions, but their application to a real low-stakes dataset showed differences in item parameter estimates between the one-class Rasch model (i.e., all respondents' item responses were used for item parameter estimation) and mixed-strategies two-class model (i.e., random guessers corresponded to a chance model, and valid respondents' responses were used for item parameter estimation; see also Equations 3 and 4).

Lau (2009) has recently extended Mislevy and Verhelst's (1990) mixed-strategies Rasch model to mixture one-parameter logistic (1PL) and mixture 2PL IRT models, with estimation implemented in Mplus (Muthén & Muthén, 1998-2010). The results of her simulation study show that the mixed-strategies IRT models produced accurate model–data fit and model parameter estimates under the presence of random guessing respondents. The findings also indicate that the one-class model may still function when less than 1% of guessers were present.

Subedi (2009) also extended this modeling approach to a mixture 2PL IRT model (called the mixture IRT model with random guessing) to distinguish random guessers and valid respondents. Subedi implemented Bayesian estimation in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). The model effectively identified subsets of guessers and produced accurate model parameter estimates. In Subedi's sequential real data application, around 5% of the test takers were identified as random guessers; the final distribution of the proficiency classification decision (i.e., advanced, proficient, basic, and below basic) was only slightly influenced when a small proportion of test takers (e.g., less than 5%) engaged in random guessing.

Mathematically, the marginal probability of a correct answer in IRT-based mixed-strategies modeling and the probability of success by a test taker within a class are expressed in Equations (3) and (4), respectively (Lau, 2009; Mislevy & Verhelst, 1990; Subedi, 2009):

$$P(x=1) = \pi_1 \left[\frac{\exp[a_i(\theta_j - b_i)]}{1+\exp[a_i(\theta_j - b_i)]} \right] + \pi_2 \left[\frac{\exp(\tau_i)}{1+\exp(\tau_i)} \right], \tag{3}$$

$$P(x=1) = g_j \left[\frac{\exp[a_i(\theta_j - b_i)]}{1+\exp[a_i(\theta_j - b_i)]} \right] + (1-g_j) \times \left[\frac{\exp(\tau_i)}{1+\exp(\tau_i)} \right], \tag{4}$$

where a response vector is represented by $\boldsymbol{x} = (x_1,.., x_I)$. The latent class proportions for motivated and unmotivated classes are π_1 and π_2, respectively. The item discrimination and difficulty parameters for item i are denoted by a_i and b_i. The ability parameter for test taker j is represented by θ_j. The indicator of the latent class membership of an examinee j is g_j, in which $g = 1$ indicates the latent class of valid respondents and $g = 0$ refers to the latent class of random guessers. The item response patterns by motivated examinees that apply solution strategy correspond to an IRT model, whereas those by unmotivated examinees that apply random guessing strategy correspond to a random guessing strategy model. The item threshold for the random guessing strategy is a constant, τ_i. The probability of a correct response by guessing takes the value of the reciprocal of the number of options in a multiple-choice item. In a mixed-strategies 2PL model, a_i is allowed to vary across items; in a mixed-strategies 1PL model, a_i is constant across items; in a mixed-strategies Rasch model, a_i takes on a constant value of 1 across items.

An essential assumption underlying the mixed-strategies modeling is that categorical latent class membership can describe examinees' latent heterogeneous test-taking motivation. This modeling approach (Lau, 2009; Mislevy & Verhelst, 1990; Subedi, 2009) and aforementioned direct and indirect measures of test-taking motivation (i.e., motivation filtering, person-fit statistics) are governed by the same logic because their ultimate goals are to classify examinees into unmotivated and motivated groups and to facilitate follow-up statistical adjustment. Test-taking motivation is fundamentally regarded as a person characteristic that is kept constant in a test.

This study adopts the mixed-strategies modeling approach because it presents many desirable advantages. First, this modeling approach aids the identification of unmotivated respondents who guess randomly and therefore serves as a useful model-based motivation filtering technique for low-stakes assessments. Second, it does not require two-step modeling because both latent class membership and model parameters can be estimated on the basis of item response patterns rather than on external manifest variables (e.g., item response time). Third, this approach is suited for both pencil-and-paper and computer-based testing scenarios. Fourth, the mixed-strategies approach eliminates the need to consider the issues that arise from the use of motivation filtering (i.e., self-report test-taking motivation scale, item response time, or person-fit statistics). These issues are discussed earlier in the dissertation, such as validity concerns, appropriateness of a cutoff score or time threshold for identifying unmotivated test takers, or the consistency between test-taking motivation and item response time/person-fit statistics. More important, a latent class approach enables researchers to further use covariates to characterize latent class membership (Samuelsen, 2005). Such investigation considerably improves result interpretation and delineation of the implications of using mixed-strategies modeling.

A major limitation of the current applications of the mixed-strategies modeling is that those models disregard testlet effects that are often encountered in educational assessments. Failure to control for test-taking motivation heterogeneity and testlet effects can contribute to inaccurate estimation and invalid inferences. Another limitation of the current applications of the mixed-strategies modeling is that after the latent classes (i.e., motivated and unmotivated respondents) in the examinee population have been identified, previous studies did not further explore which indicators can potentially help the interpretation of the latent classes of test-taking motivation heterogeneity. The failure to interpret identified latent class membership makes an unobservable latent class variable much difficult to understand and diminishes the applicability of the mixed-strategies modeling to real-world testing scenarios. The current research is different from previous work in that it incorporates test-taking motivation heterogeneity and testlet effects into a single measurement model in testlet-based assessments and aims to empirically investigate the potential attributes that are associated with latent class membership from real data. The literature on testlet effects and the modeling approaches that explain such effects are reviewed in the next section.

2.3 Testlet Effects and Measurement Models that Manage Such Effects

A testlet is a commonly used item format in educational assessments. Frequently seen testlets comprise subsets of related items that correspond to a common stimulus, such as passages, graphic contexts, listening records, or laboratory tasks. Testlet-based items are desirable in educational assessments for several reasons. First, formulating testlet-based items is an economical and efficient option for test developers; the same holds true for test takers (i.e., examinees can answer several questions with one passage). Second, using testlet-based assessment conserves testing time, making it significantly more realistic and applicable to real-world testing scenarios. More important, testlet-based items can measure the higher level cognitive skills (e.g., *Evaluating* or *Creating* in Bloom's revised taxonomy; see Anderson & Krathwohl, 2001) that are often embedded in situational or authentic contexts.

The drawback to testlets is that the items are usually interdependent on one another, and the estimation and interpretation of an item are correlated to that of other items within the same testlet (i.e., testlet effects). Testlet effects can violate the assumption of local item independence in the IRT models. As stated earlier (see Chapter 1), local item independence is that an examinees' responses to different items are statistically independent after taking examinee ability into account (Hambleton et al., 1991); namely, a respondent's performance on one item is independent of his or her responses to any other items in the test. Previous studies have indicated that the presence of local item dependence caused an overestimation of test reliability and produced inaccurate parameter estimates (e.g., Bradlow et al., 1999; Chen & Thissen, 1997; DeMars, 2006; Sireci et al., 1991; Thissen et al., 1989; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993).

To manage testlet effects in the unidimensional IRT modeling framework, researchers have developed two major modeling approaches: the polytomous and testlet models. The polytomous modeling approach involves treating testlet-based items as polytomous items and then fitting data by using the polytomous model (Cook et al., 1999; Thissen et al., 1989). The rationale that underlies this approach is that "a summed score of items within a testlet" is regarded as "a super item with partial credits" because testlet-based items share a common stimulus. Polytomous items are treated as locally independent and thereby local item dependence caused by testlets can be absorbed (Yen & Fitzpatrick, 2006). This approach manages testlet effects (Cook et al., 1999; Sireci et al., 1991; Thissen et al., 1989; Yan, 1997) but still suffers from certain limitations. First, it requires coding of raw item response scores into testlet scores. The testlet effects are indirectly resolved and psychometric information at the individual item level is lost. For example, it is criticized that "a testlet score does not say anything about the response pattern that produced the score" (de Ayala, 2009, p. 132). Second, aggregating the item scores within a testlet to a testlet score dramatically shortens test length (e.g., if five items exist within each of four testlets, only four polytomous items remain after a re-coding procedure). Given that the testlet scores are used to estimate respondents' ability levels, a decrease in the total number of items may reduce the information for estimating person-related parameters, thereby potentially affecting the accuracy of person-related parameter estimates. Chen, Jiao, and von Davier (2013) compared the effectiveness of different approaches to dealing with testlet effects in the framework of mixture IRT modeling. The authors assessed model parameter recovery in the mixture Rasch model, mixture polytomous model, and mixture Rasch testlet model. The results indicate that the polytomous modeling approach produced poor classification accuracy for latent classes, biased estimates of mixing proportions, and low accuracy for ability parameter estimates. Their results also implicitly suggest that person-related parameters (i.e., latent class classification accuracy, mixing proportion estimates, and ability parameter estimates) were negatively affected by the polytomous scoring of a testlet as a single item.

The testlet modeling approach entails directly incorporating testlet parameters into IRT measurement models (Bradlow et al., 1999; Jiao et al., 2005; Wainer et al., 2000; Wainer & Wang, 2000; Wang et al., 2002; Wang & Wilson, 2005). This method explicitly accounts for testlet dependence and enables the evaluation of the magnitudes of testlet effects. The probability of correctly obtaining an item in a testlet model with three item parameters is mathematically expressed as (Wainer et al., 2000; Wainer & Wang, 2000)

$$P(x=1)=c_i+(1-c_i)\left[\frac{\exp[a_i(\theta_j-b_i+\gamma_{jd(i)})]}{1+\exp[a_i(\theta_j-b_i+\gamma_{jd(i)})]}\right], \qquad (5)$$

where c_i is the item guessing parameter, and $\gamma_{jd(i)}$ is the random-effects testlet parameter associated with item *i* within testlet *d* for examinee *j*. If c_i is 0 and testlet variances are constant across testlets, the model is reduced to the two-parameter testlet model developed by Bradlow et al. (1999). If c_i is 0 and a_i takes on a fixed value of 1, the model is reduced to the one-parameter Rasch testlet model proposed by Wang and Wilson (2005). Table 1 briefly summarizes the studies devoted to testlet models, outlining how previous researchers manipulated testlet effects, what kinds of item formats have been explored, and which testlet models have been investigated.

There are potential limitations of testlet modeling that may possibly restrict its applicability. First, in most cases, testlet units can be easily recognized on the basis of an observable common passage, table, graph, or diagram. However, identifying testlet units become difficult when the items formulated are based on unobservable stimuli. For example, items clusters are designed to measure different constructs, such as comprehension or critical thinking in a reading literacy assessment. In such a testing scenario, correctly identifying unobservable constructs that form testlet units necessitates expert content knowledge, making testlet modeling less applicable. Second, many educational assessments use conventional IRT models not only for test development and parameter estimation, but also for test scoring, equating, linking, and norming. Applying testlet modeling solely on test estimation can bring forth challenges to the aforementioned psychometric analyses including test equating, linking, and norming, thereby diminishing the applicability of testlet modeling.

The advantages of testlet modeling are that individual item-level psychometric information can be retained and test length does not decrease. These features are attractive because interpreting the results of dichotomously scored items is more straightforward and meaningful than elucidating those of polytomously scored items. Testlet models have shown to improve model fit and estimation accuracy (Bradlow et al., 1999; Wainer et al., 2000; Wainer & Wang, 2000; Wang et al., 2002; Wang & Wilson, 2005). For these reasons, this study adopts the testlet modeling to account for testlet effects in educational assessments. The proposed model that combines the mixed-strategies Rasch modeling with a testlet model will be comprehensively discussed in the next chapter.

Table 1

Studies of IRT-based Testlet Models

References	Testlet units	Testlet length	Testlet variance	Item format	Model	Study approach
Chen et al. (2013)	4	8	0.25, 1	Testlet-based	Mixture Rasch testlet model	Simulation
Bao (2007)	2	10	0.2, 1.5	Individual + testlet-based	2PL testlet model	Simulation
Bradlow et al. (1999)	3, 6	5, 10	0.5, 1, 2	Individual+ testlet-based	2PL testlet model	Simulation
Jiao et al. (2005)	6	5	0, 0.25, 1, 2.25	Testlet-based	Three-level 1PL testlet model	Simulation
Jiao et al. (2012)	4 4	9 7–9	0.25, 1 [0.1557, 0.2656]	Testlet-based	Four-level 1PL testlet model	Simulation Empirical
Jiao, Wang, & Lu (2009); Jiao, von Davier, & Wang (2010)	4	9	0.25, 1, 2.25	Testlet-based	Mixture Rasch testlet model	Simulation
Jiao, Wang, & He (2013)	6 6	9 9	0, 0.25, 0.5625, 1 [0.0510, 0.8630]	Testlet-based	Three-level 1PL testlet model and Rasch testlet model	Simulation Empirical

Table 1

Studies of IRT-based Testlet Models (continued)

References	Testlet units	Testlet length	Testlet variance	Item format	Model	Study approach
D. Li (2009)	6	5	0, 1, 2	Testlet-based	3PL testlet model	Simulation
	8	3–5	[0.245, 1.072]			Empirical
F. Li (2009)	3–12	2–16	0, 0.25, 1	Testlet-based	1PL testlet, 2PL testlet, and 3PL testlet models	Simulation
	8	3–6	[0.19, 0.71]			Empirical
Wainer & Wang (2000)	86	5, 10, 13	[0.025, 0.925]	Individual + testlet-based	3PL testlet model	Empirical
Wang et al. (2002)	2, 3, 6	3, 6, 9	0, 0.5, 1	Individual + testlet-based	3PL testlet model	Simulation
Wang & Wilson (2005)	2–8	3–10	0.25, 0.50, 0.75, 1	Testlet-based	Rasch testlet model	Simulation
	11	2–4	[0.07, 2.09]	Individual + testlet-based	Rasch testlet model	Empirical

2.4 Bayesian Estimation

A Bayesian estimation with Markov chain Monte Carlo (MCMC) algorithm and ML estimation are two major statistical techniques for model parameter estimation. An essential difference between MCMC and ML methods, as indicated by Kim and Bolt (2007), is that ML calculates parameters by finding the maximum likelihood of the observed data; MCMC, on the other hand, uses prior distributions to estimate model parameters, assuming that observations can be sampled from the parametric posterior distributions implied by the model. For the present research, the MCMC method is chosen over the ML algorithm for model estimation for a number of reasons. First, the MCMC method enables highly flexible implementation for complex models (Kim & Bolt, 2007). Yen and Fitzpatrick (2006) pointed out that MCMC methodology can easily accommodate complex data, such as item responses with complicated dependence. For these reasons, MCMC has been popular and useful in the estimation of IRT-based mixture models (e.g., Bolt et al., 2002; Cho, Cohen, & Kim, 2013; Dai, 2009; Cohen & Bolt, 2005; Jiao et al., 2012; Li, Cohen, Kim, & Cho, 2009; Meyer, 2010; Samuelsen, 2005; Subedi, 2009) and in the estimation of testlet models (e.g., Bradlow et al., 1999; Jiao et al., 2012, 2013; Wainer et al., 2000; Wang et al., 2002). By contrast, ML exhibits estimation efficiency (i.e., short estimation time) and has been widely used for mixture models (e.g., Alexeev, Templin, & Cohen, 2011; Cho, Jiao, & Macready, 2012; Finch & Pierson, 2012; Jiao et al., 2010; Mislevy & Verhelst, 1990; Rost, 1990) or testlet models (e.g., Glas, Wainer, & Bradlow, 2000; Jiao et al., 2010; Wainer & Wang, 2000). However, in complex models, ML method with the EM algorithm may suffer from unbounded likelihood functions or from the generation of multiple local maxima of likelihood (Kiefer & Wolfowitz, 1956). When respondents have perfect or all-zeros response patterns, ML estimation may pose critical challenges, such as a failure of convergence. Considering that this research focuses on developing a new measurement model that is also highly complex, MCMC method is chosen because it offers an opportunity for researchers to experiment with the proposed model. The second factor that drives the use of MCMC is its provision of more comprehensive information for describing model parameters. MCMC describes model parameters on the basis of their corresponding posterior distributions, whereas ML implements description in terms of point estimates. The third reason is that MCMC exhibits high estimation accuracy. Glas et al. (2000) compared the calibration performance of marginal ML and MCMC in testlet-based adaptive testing. The authors found that (1) the parameter estimates obtained from MCMC and ML were highly correlated, but ML tended to underestimate the width of the interval region; and (2) MCMC provided interval and point estimates that were closer to true values in the main. Jiao et al. (2013) compared parameter recovery by ML, MCMC, and six-order Laplace estimation in the one-parameter testlet model. Their results show that (1) MCMC generated the least amount of bias in item parameter estimates; (2) no discernible difference in terms of ability variance and ability parameter recovery was found between ML and MCMC; and (3) MCMC and Laplace estimation satisfactorily recovered testlet variance, whereas ML underestimated true testlet variance under large testlet effects. Another reason MCMC is chosen is that it offers both numerical and graphical tools that are useful for monitoring convergence. The accumulated evidence excellently facilitates the determination of appropriate cases. A major limitation of MCMC estimation, however, is that it demands a long iterative process, so that a substantial amount of time is needed for model parameter estimation.

Basically, MCMC computation involves obtaining posterior distributions on the basis of both prior distribution and the likelihood function. Bayes' theorem is expressed as follows:

$$f(\Omega \mid X) = \frac{f(X \mid \Omega) * f(\Omega)}{\int_{\Omega} f(X \mid \Omega) f(\Omega) d\Omega}, \qquad (6)$$

where X is a set of item response data, Ω is a set of model parameters, $f(\Omega)$ represents the prior of model parameters, $f(X \mid \Omega)$ denotes the likelihood of item response data given all the model parameters, and $f(\Omega \mid X)$ is the posterior density of model parameters given the data (Kim & Bolt, 2007). With an MCMC algorithm, the model is fitted to the item response dataset by simulating a random sample that can approximate the probability distribution of the model parameter (Sinharay, 2003, 2004).

The MCMC estimation requires the specification of priors. A suggested probability distribution is the *conjugate* prior, which causes posterior distribution to take on the same form as prior distribution (i.e., the posterior distribution has a known functional form), as well as facilitates more efficient sampling from the posterior. The strength of the prior can be reduced by specifying prior distribution (i.e., the mean and variance of a prior distribution) as noninformative; for example, a normal prior is assigned to the ability parameters, but the mean or variance of this prior can be specified with hyper priors (Kim & Bolt, 2007).

After a set of priors for model parameters are determined, a sampling mechanism is iteratively run. Gibbs sampling (German & German, 1984) and Metropolis Hastings algorithm (Hasting, 1970) are two popular samplers. The former is preferred when *conjugate* priors are used, so that samples can be directly simulated from a known form of posterior distribution; and the latter is useful when the distributional form of conditional distributions is unknown, so that samples are indirectly generated as candidate observations from proposal distributions (Kim & Bolt, 2007; Sinharay, 2003). MCMC estimation runs until Markov chains achieve convergence; then, inferences are drawn from the stationary posterior distribution of the targeted model parameters.

In this research, the WinBUGS software (Lunn et al., 2000) is used for MCMC estimation because the software is free and flexibly applied to complex models. WinBUGS is easy to implement because in the internal phase of the program, sampling algorithms are automatically selected and therefore do not require specification by users (Kim & Bolt, 2007). It also numerically and graphically provides multiple diagnostic tools that are useful for monitoring convergence. More details on WinBUGS can be found on its official website (http://www.mrc-bsu.cam.ac.uk/bugs/).

Chapter 3: Methodology

This study has multiple facets. First, the issues arising from overlooking test-taking motivation heterogeneity and testlet effects in low-stakes testlet-based assessments are addressed through a simulation study. Second, this study proposes a measurement model that incorporates test-taking motivation heterogeneity and testlet effects in analysis. The performance of the proposed model is evaluated with simulated data under varied testing conditions and is explored with an empirical dataset. Finally, this study empirically explores potential indicators for facilitating the explanation of heterogeneous test-taking motivation in real low-stakes testlet-based assessment data. The following section introduces the proposed and comparison models, simulation study design, estimation procedure, data analyses, and empirical study.

3.1 Models

As mentioned in Chapter 2, test-taking motivation can be conceptualized from the perspective of expectancy value theory. The existence of unmotivated test takers in low-stakes assessments stems from low-stakes test results and the time and energy costs incurred by test takers. In such a scenario, unmotivated test takers tend to provide random responses rather than respond in way that reflects their actual knowledge; motivated and unmotivated test takers therefore behave differently during a test. Test-taking motivation in low-stakes assessments is observed or measured through the manner by which test takers respond to items—via a solution strategy or random guessing strategy. In the proposed measurement model, test-taking motivation is operationalized and modeled through probability-based item response functions. A latent class variable categorizes test takers into motivated and unmotivated classes on the basis of item response patterns. Test takers whose item responses are best predicted by the IRT model belong to the motivated class that applies the solution strategy, whereas those whose item response patterns are best fitted by the random guessing function belong to the unmotivated class that applies the random guessing strategy. Test takers within the same latent class have qualitatively homogeneous item response patterns, whereas test takers between classes have qualitatively heterogeneous item response patterns.

The proposed measurement model incorporates both test-taking motivation heterogeneity and testlet effects in its analysis. The development of the proposed model (hereafter called the mixed-strategies Rasch testlet model) borrows the ideas from the HYBRID model (Yamamoto, 1987, 1989; Yamamoto & Gitomer, 1993). In essence, the proposed model is an extension of the following models: the mixture Rasch model with a combination of valid respondents and random guessers (Mislevy & Verhelst, 1990), the mixture two-parameter model with completely guessing behaviors (Subedi, 2009), and the

Rasch testlet model (Wang & Wilson, 2005). The marginal probability of a correct response in the proposal model is expressed as

$$P(x=1)=\pi_1\left[\frac{1}{1+\exp[-(\theta_j-b_i+\gamma_{jd(i)})]}\right]+\pi_2\left[\frac{1}{1+\exp(-\tau_i)}\right], \tag{7}$$

and the probability of getting an item correctly in the proposed model is expressed as

$$P(x=1)=(1-g_j)\times\left[\frac{1}{1+\exp(-\tau_i)}\right]+g_j\left[\frac{1}{1+\exp[-(\theta_j-b_i+\gamma_{jd(i)})]}\right], \tag{8}$$

where a response vector is represented by $\mathbf{x} = (x_1,.., x_I)$. The mixing proportions are π_1 and π_2, in which $\pi_1 + \pi_2 = 1$ and $0 < \pi_g < 1$. P(x = 1) refers to the probability of success for item i of examinee j in latent class g. The indicator of latent class membership for examinee j is g_j, which distinguishes latent classes in a population (i.e., the latent group membership of an examinee is a model parameter to be estimated). The categorical latent class variable has two categories: $g = 1$ for motivated item respondents (i.e., solution strategy) and $g = 0$ for unmotivated respondents (i.e., random guessing strategy). For unmotivated test takers across the entire proficiency continuum, their probabilities of obtaining correct responses can be expected by chance. Assuming that four options are available in a multiple-choice item, the probability of a correct response by guessing is 0.25—the reciprocal of the number of item options. Therefore, τ_i as a constant of item threshold for random guessing response is fixed as -1.0986, making 1/[1+exp(-τ)] to be equal to 0.25. For motivated test takers, their probabilities of success can be characterized by the Rasch testlet model, in which b_i is the difficulty for item *i,* θ_j is the ability parameter for examinee *j*, and $\gamma_{jd(i)}$ is the random-effects testlet parameter associated with item *i* within testlet *d* for examinee *j*. The testlet parameter describes the interaction between a test taker and an item nested within a testlet, and the strength of testlet effects is indicated by testlet variance $\sigma^2_{jd(i)}$. The integration of both item response functions enables the management of test-taking motivation heterogeneity and testlet effects in a single measurement model.

This study compares the results from the proposed model with the findings from the Rasch model and the mixed-strategies Rasch model. This analysis is to demonstrate the impact in disregarding test-taking motivation heterogeneity and testlet effects, as well as to assess the effectiveness of the mixed-strategies Rasch testlet model. The Rasch model, which has been widely used to analyze item response data in large-scale assessments, assumes zero testlet variance and conditionally independent items. It considers a one-class examinee population. The Rasch model disregards both test-taking motivation heterogeneity and testlet effects. The probability of a correct response in the Rasch model is expressed as

$$P(x=1)=\frac{1}{1+\exp[-(\theta_j-b_i)]}. \tag{9}$$

The mixed-strategies Rasch model incorporates test-taking motivation heterogeneity but disregards testlet effects in data. This model is included in the study because testlet-based items are often calibrated as though they were independent (Wainer et al., 2000). The probability of a correct response in this model is expressed by

$$P(x=1)=(1-g_j)\times\left[\frac{1}{1+\exp(-\tau_i)}\right]+g_j\left[\frac{1}{1+\exp[-(\theta_j-b_i)]}\right]. \tag{10}$$

3.2 Simulation Study Design

The simulation study mimics a real-world testing scenario that approximates a standardized educational assessment, the PISA assessment, which is constructed and calibrated under the Rasch measurement model (Rasch, 1960). In this simulation, data are generated under the Rasch model with testlet effects (see Equation 8, in which $g_j = 1$). Both item and ability parameters are simulated from standard normal distribution. In the generated data matrix, a proportion of item responses are replaced with unmotivated test takers' item response patterns. The item response probabilities of obtaining correct responses for unmotivated respondents are simulated under the random chance model (see Equation 8, in which $g_j = 0$).

To elicit test-taking motivation on the basis of expectancy value theory, low-stakes assessments should address at least one of the components of the expectancy value model (i.e., expectancy, attainment value, intrinsic value, utility value, perceived cost). In this simulation, six testlets are generated, each with six dichotomously scored multiple-choice items. In this design, test length enables test takers to complete all items within an appropriate duration. Thus, expectancy is high because a test taker believes (s)he can complete the test, and perceived cost is reasonable because a test taker exerts an acceptable level of energy. Items are also of appropriate difficulty (neither too easy nor too difficult), prompting test takers to deem the items intellectually challenging (thus, high intrinsic value).

Table 2

The Specification of the Simulation Design

Manipulated Factors	Levels
Sample size	1,000 3,000 5,000
Percentage of unmotivated respondents in the examinee population	1% 5% 15%
Magnitude of testlet effects	Testlet variance = 0.25 (small) Testlet variance = 1.00 (large)
Estimation model	The Rasch model The mixed-strategies Rasch model The mixed-strategies Rasch testlet model

Four factors are manipulated in the simulation: sample size, percentage of unmotivated respondents in the examinee population, testlet variance, and the estimation model. The specification of the simulation design is summarized in Table 2.

Sample size. Three sample sizes—1,000, 3,000 and 5,000—are considered because such variety facilitates observation in standardized educational assessments. Besides, the variety of sample sizes in the current study can provide evidence of how model parameters are recovered in terms of small-, moderate-, and large-sample conditions.

Percentage of unmotivated respondents. This study manipulates the percentage of random guessers in an examinee population at three levels: 1%, 5%, and 15%. This series of percentages are applied in accordance with the findings in previous empirical studies and with the levels used in earlier simulation studies. In real data analyses, the percentages of unmotivated test takers in low-stakes assessments may differ depending on testing conditions. As reviewed in Chapter 2, a range of 1.2% to 12.8% has been observed across empirical studies (Brown & Gaxiola, 2010; Lau, 2009; Subedi, 2009; Sundre & Wise, 2003; Wise & DeMars, 2005, 2006). In previous simulation studies, Subedi (2009) manipulated 0%, 5%, and 10% of random guessers in an examinee population, with 0% serving as the baseline and the other two levels functioning as benchmarks for assessing the effects of different guessing proportions on parameter estimation. Lau (2009) manipulated 0.9%, 9%, and 20% of an entire population and classified them as random guessers. Lau then added these groups to valid respondents (N = 5,000), thereby coming up with approximately 0.89% (i.e., 45/5045 × 100%), 8.26%, and 16.67% of random guessers in the examinee population, respectively. In Lau, 0.89% and 8.26%, respectively, represented the low- and high-end plausible estimates, respectively, of the random guessers observed in Wise and DeMars (2006); 16.67% represented the maximum allowable percentage of random guessers in an examinee population. In the current research, the choice of 5% was decided upon to reflect a middle frequency of random guessers in *real-world* low-stakes assessments—an approach consistent with the findings of Brown and Gaxiola (2010), Subedi (2009), and Wise and DeMars (2006). One percent corresponds to a minor effect of random guesses and 15% as equivalent to a considerable effect of random guesses on the accuracy of parameter estimates.

Testlet effects. Two magnitudes of testlet effects (i.e., testlet variance = 1.00 or 0.25) are included to represent large and small testlet effects; these levels are consistent with those applied in previous simulation studies (i.e., Jiao et al., 2012, 2013; D. Li, 2009; F. Li, 2009; Wang & Wilson, 2005). The manipulated testlet variances in the current research are reasonable and realistic when evaluated against the estimated testlet variances in previous empirical examples. For example, Wainer et al. (2000) reported that estimated testlet variances on SAT and GRE verbal tests ranged from 0.11 to 0.96; Jiao et al. (2013) revealed that six estimated testlet variances on a K-12 large-scale reading comprehension test ranged from 0.0510 to 0.8630; D. Li (2009) found that testlet variance estimates on Assessing Comprehension and Communication in English State-to-State for English Language Learners ranged from 0.245 to 1.072. In F. Li (2009), eight estimated testlet variances on a large-scale grade-three reading assessment ranged from 0.19 to 0.71.

Number of replications. Each testing condition is replicated 25 times. In selecting the number of replications, a primary consideration is the heavy computation required in MCMC estimation and the intensive time that a single estimation involves. For example, the proposed model spends 10 to 25 hours in implementing estimation for a single dataset with a sample of 5,000. The number of replications in the current research (i.e., 25) has been indicated as sufficient to generate good power with which to detect whether manipulated factors affect the precision of item difficulty parameters in a Monte Carlo IRT 2PL model study (Harwell, Stone, Hsu, & Kirisci, 1996). Following the post-hoc procedure introduced in Jiao et al. (2013), the current work shows that the magnitudes of estimation bias in item difficulty parameters (between –0.017 and 0.051) were only about 1.5% of the range of simulated values (between –2.140 and 2.307). In Wang and Wilson (2005, as cited in Jiao et al., 2013), the magnitudes of estimation bias in item difficulty parameters (between –0.063 and 0.050) over 100 replications were about 2.8% of the range of simulated values (between ±2.00). In the current study, another post-hoc analysis is conducted to assess the standard deviation of Monte Carlo errors (MC errors) of model parameter estimates across

25 replications under a given testing condition. Results indicate that the standard deviation values of the MC errors across replications were very small: at a range of 0.0015 to 0.0040 for the mixed-strategies Rasch testlet model; 0.0015 to 0.0037 for the mixed-strategies Rasch model; and 0.0004 to 0.0005 for the Rasch model across testing conditions. These results imply that the standard errors of model parameter estimates across 25 replications varied to a minimal extent; put it differently, such errors were highly stable and only slightly varied. The low estimation bias, as well as the small standard deviation of MC errors of model parameter estimates, across the replications in the current simulation study implicitly support the appropriateness of 25 replications. Furthermore, the number of replications in the present work is, in actuality, relatively larger than those used in previous Bayesian IRT-based simulation studies. For example, five replications (e.g., Cho & Cohen, 2010), 10 replications (e.g., Cho et al., 2013; Dai, 2009; Meyer, 2010), 15 replications (Subedi, 2009), and 20 replications (e.g., S. Frederickx, F. Frederickx, De Boeck, & Magis, 2010) are observed in IRT-based mixture models with MCMC estimation.

Estimation model. After datasets are generated, they are estimated by the Rasch model (Rasch, 1960), the mixed-strategies Rasch model (Mislevy & Verhelst, 1990), and the mixed-strategies Rasch testlet model. As previously stated, the mixed-strategies Rasch testlet model represents a combined measurement model that simultaneously manages test-taking motivation heterogeneity and testlet effects; the mixed-strategies Rasch model represents a measurement that accounts for a mixture of latent examinee populations, but disregards testlet effects; the Rasch model assumes a one-class population and zero testlet variance of data, which represents a commonly used approach for analyzing low-stakes testlet-based assessments (e.g., PISA assessment).

3.3 Estimation

The specification of priors. The Bayesian estimation with MCMC algorithm implemented in WinBUGS 1.4.3 (Lunn, et al., 2000) is used for estimation. To ensure convergence, the priors and hyper-priors are specified using the priors recommended by other researchers who applied comparable IRT-based mixture models (i.e., Cho et al., 2013; Dai, 2009; Jiao et al., 2009; Jiao, von Davier, Kamata, Chen, 2011; Subedi, 2009):

$$b_i \sim \textit{normal}\,(0, 1),\ i = 1,\ldots, \mathrm{I};$$

$$\theta_j \sim \textit{normal}\,(\mu_\theta,\ \sigma^2_\theta),\ j = 1,\ldots, \mathrm{J};$$

$$\mu_\theta \sim \textit{normal}\,(0, 1);$$

$$\sigma^2_\theta \sim \textit{inverse-gamma}\,(a_\theta\ , b_\theta);$$

$$\gamma_{jd(i)} \sim \textit{normal}\,(0,\ \sigma^2_{jd(i)});$$

$$\sigma^2_{jd(i)} \sim \textit{inverse-gamma}\,(a_\gamma, b_\gamma);$$

$$g_j \sim \textit{categorical}\ (\pi_g\ []),\ \ j = 1,\ldots,\ \mathrm{J};$$

$$(\pi_1, \pi_2) \sim \textit{dirichlet}\ (\mathrm{alpha}[]);$$

where item difficulty parameters are assumed from a standard normal distribution. The sum normalization of the item difficulty parameters is posited for scale identification. Two-stage normal priors are assigned to ability and testlet parameters. Ability parameters are assumed from a normal distribution, where the mean is from a standard normal distribution (*normal* [0, 1]) and the variance is from an inverse-gamma distribution. Testlet parameters are assumed from normal distributions (0, $\sigma^2_{jd(i)}$), in which testlet variances are assigned inverse-gamma distributions. According to Curtis (2010), "the inverse-gamma prior is the conjugate prior for a variance parameter from a normal likelihood, so the update in an MCMC algorithm is a simple random draw from an inverse-gamma distribution" (p. 12). Inverse-gamma priors are used for variance parameters primarily to achieve convergence (Curtis, 2010). On the basis of previous studies (i.e., Dai, 2009; Jiao et al., 2012, 2013; F. Li, 2009) and the preliminary analyses in the current work, the inverse-gamma distribution of ability and testlet variances is specified as *gamma* (1, 1) to ensure convergence. Class membership is estimated on the basis of the frequencies of an examinee being sampled into each class. As a conjugate prior for a categorical parameter *g*, the hyper-prior for latent class membership follows a Dirichlet distribution. Mathematically, Dirichlet distribution is expressed as follows (Spiegelhalter, Thomas, Best, & Lunn, 2003):

$$\frac{\Gamma(\sum_g \alpha_g)}{\Pi_g \Gamma(\alpha_g)} \prod_g^G \pi_g^{\alpha_g - 1}, \tag{11}$$

where G is the number of categories and $0 < \pi_g < 1$, $\sum_g \pi_g = 1$. The parameters are ($\alpha_1, \ldots, \alpha_g$), which take positive values. In Bayesian mixture models, Dirichlet distribution serves as a prior distribution and is the conjugate prior of the categorical distribution (i.e., a generalization of the Bernoulli distribution; the parameters are the probabilities for the categories given one trial) or the conjugate prior of the multinomial distribution (i.e., a generalization of the binomial distribution; the parameters are the probabilities for the categories given *n* trials). The alpha parameters for the mixing proportion distribution are (.5, .5) as starting values (Bolt et al., 2002; Cohen & Bolt, 2005; Li et al., 2009; Meyer, 2010). This study estimates latent class membership, item difficulty parameters, ability parameters, and variances of ability and testlet parameters.

Label switching. Label switching of latent classes is a potential problem in mixture models. Cho et al. (2013) comprehensively described two types of label switching in IRT-based mixture models. The first type arises across iterations within a single Markov chain, and the second occurs when labels switch over replications. Label switching problems frequently occur when different latent classes are respectively characterized by a common item response function (e.g., in the mixture Rasch model, a Rasch model is assumed for each class), in which no constraints (e.g., item difficulty) are posited on a certain class. In the current research, respondents from two classes de facto correspond to distinct item functions, and each latent class label is embedded in its item function (see Equation 8). A respondent's item response patterns are best fitted by either the chance model or the Rasch (testlet) model; thus, it is expected that no label switching occurs in this study. To verify this expectation, in the data analysis procedure, label switching is monitored; the latent class labels of the datasets in which labels switch are corrected.

Convergence assessment. In MCMC estimation, monitoring chain convergence is an important step that guarantees sampled observations from the algorithm can represent a sample from the posterior distribution of a model parameter (Kim & Bolt, 2007). In this study, two chains of iterations are run and chain convergence is diagnosed with multiple criteria. Diagnostic plots are used to examine whether Markov chains converge to a stationary distribution. The plots used include history plots (where convergence is achieved when Markov chains combine and become stationary after an initial burn-in), quantiles plots (where the mean and 95% confidence interval of a parameter should stabilize at the posterior mean), autocorrelation plots (in which autocorrelation dropping to zero as evidence of convergence refers to a lack of correlation among iterations in the chain), and density plots (in which a smooth density distribution shows satisfactory convergence). This study also monitors MC errors, which are estimates of the standard errors of the mean, to determine how many sampled states of the chain are needed. A suggested approach is to run simulations until the MC error of a parameter is less than 0.05 (Spiegelhalter et al., 2003). Once sufficient evidence of convergence is obtained, the burn-in iterations are disregarded and the remaining iterations are used as bases for drawing inferences on model parameters from the posterior distribution. If non-convergence is diagnosed in MCMC estimation, the solution used by this study is to remove non-converged datasets and replace them with new item response datasets from the same study condition. Non-converged cases are discarded because their estimates are merely random values which cannot represent a sample from the posterior distribution of a model parameter.

3.4 Data Analyses

The outcome statistics for evaluating the performance of the proposed model are model selection, latent class classification accuracy, and model parameter recovery. The outcome statistics are computed over replications.

Model selection. In this research, the Akaike Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), corrected Akaike Information Criterion (AICc; Burnham & Anderson, 2002), and sample size-adjusted Bayesian Information Criterion (SABIC; Sclove, 1987) are used for model selection. Multiple model-fit indices rather than a single index are used for latent class selection because the estimation of latent class membership can vary among model-fit indices under varied testing conditions.

This research uses the above-mentioned likelihood-based model-fit indices for three reasons. First, Bayesian-based model-fit indices, such as the deviance information coefficient (DIC; Spiegelhalter, Best, Carlin, & von der Linde, 2002) and posterior predictive model checks (PPMC; Gelman, Carlin, Stern, & Rubin, 1998), are favorable for Bayesian modeling with MCMC estimation; however, they have been demonstrated to perform disproportionally worse in the mixture one-parameter model, generating only about 0% to 50% accuracy on selecting the correct model across testing conditions (Li et al., 2009). Given the unreliability of the aforementioned Bayesian-based model-fit indices, they are disregarded in this research. Second, this research adopts AIC and BIC because they are widely used model-fit statistics and have been recommended for Bayesian modeling with MCMC estimation (Congdon, 2003). Li et al. (2009) investigated the efficacy of five model-fit indices (AIC, BIC, DIC, BF, & PPMC) in the mixture IRT models. The authors found that BIC exhibited the best performance, generating 100% accuracy in selecting correct models for all testing conditions; AIC came in the second, slightly tending to select a model with a high number of latent classes. Similar findings regarding the performance of AIC and BIC in IRT-based mixture models can also be found in Cho and Cohen (2010), Cho et al. (2012), and Preinerstorfer and Forman (2012). Given that the mixed-strategies Rasch testlet model is a newly constructed measurement model, and that model selection could vary depending on different testing conditions, it is important and necessary to evaluate how BIC and AIC function in the proposed model. Third, this study includes SABIC and AICc as well because they are suitable for numerous parameters or for small samples (Burnham & Anderson, 2002; Yang, 2006), issues that are often encountered in mixture models. In some related mixture models or latent class analysis models, SABIC has been indicated to accurately select the correct model (Nylund, Asparouhow, & Muthén, 2007; Tofighi & Enders, 2008; Yang, 2006).

The formulas for computing the model-fit statistics in this work are as follows

$$\mathrm{AIC} = -2\ln L + 2k, \tag{12}$$

$$\mathrm{BIC} = -2\ln L + k\ln(N), \tag{13}$$

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{N-k-1}, \tag{14}$$

$$\mathrm{SABIC} = -2\ln L + k\ln\left(\frac{N+2}{24}\right), \tag{15}$$

where lnL is the log-likelihood, k is the number of parameters, and N is the sample size. Lower values of the model-fit statistics indicate better fit; therefore, a model with the smallest value is selected as the best fitting model. For each of the model-fit indices, the percentages of replications in which a particular measurement model is chosen are summarized. Evaluating the performance of model-fit indices is valuable because the models selected could differ depending on varied testing conditions. The evaluation in this research can improve the understanding of how different model-fit indices behave when the proposed model is applied.

Latent class classification accuracy. Classification accuracy assesses how well a model assigns test takers to distinct latent classes. Put it differently, a test taker is classified into a particular latent class because his/her item response patterns are best fitted by a particular item response function embedded in a mixture model. A high classification accuracy indicates the capability of the mixed-strategies model to distinguish test takers in terms of test-taking motivation. In this research, latent class classification is evaluated in the mixed-strategies Rasch model and in the mixed-strategies Rasch testlet model because these two allow for heterogeneous examinee groups. This outcome statistic pertains to the percentage of examinees correctly classified as valid respondents and random guessers. It is expressed as follows:

Classification accuracy =

$$\frac{\text{Number of examinees correctly classified into the correct latent class}}{\text{Total number of examinees}}\times 100\%.$$

(16)

Model parameter recovery. After the estimated parameter distribution is adjusted to the equivalent scale as the true parameter distribution, bias and root mean square error (RMSE) are used to assess the recovery of item and ability parameters. These two statistics are used because they both can quantify the distance between estimated and simulated parameter values, and because they have been regarded as useful indices for evaluating parameter recovery in previous simulation studies. Bias refers to the difference between generated and estimated values across replications; it indicates an overestimation or underestimation of a model parameter estimate. RMSE is the square root of the average of the squared difference between generated and estimated values across replications; the squaring process makes it more sensitive to large biases. Small values of these statistics indicate good recovery of model parameters. They are expressed thus:

$$\text{Bias} = \frac{\sum_{r=1}^{R}(\eta_r - \eta)}{R}, \qquad (17)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R}(\eta_r - \eta)^2}{R}}, \qquad (18)$$

where η_r is the estimated model parameter for the r^{th} replication, η is the simulated model parameter for the r^{th} replication, and R is the number of replications. In terms of model parameter recovery, ANOVA analysis and effect sizes (a small effect size [$\eta^2 = 0.01$]; a medium effect size [$\eta^2 = 0.06$]; a large effect size [$\eta^2 = 0.14$]; see Cohen, 1988) are provided to determine which manipulated factors would significantly affect the precision of model parameter estimates. The current study also assesses how well the ability and testlet variances are recovered in the proposed measurement model under varied testing conditions.

3.5 Empirical Study

To answer the research question 3 (the performance of the proposed model in real data; see Chapter 1), a real item response dataset drawn from the PISA assessment is used. PISA is appropriate for this research for four reasons. First, PISA dichotomously scored items are constructed and calibrated under the Rasch measurement model. Second, the test results of the PISA assessment attach no consequence to examinees' academic records. Some unmotivated test takers therefore exist in the sample. Third, PISA cognitive assessment items are designed in testlet units. The PISA 2006 Technical Report (OECD, 2009) stated that “PISA items are arranged in units based around a common stimulus. Many different types of stimulus are used including passages of text, tables, graphs and diagrams, often in combination. Each unit contains up to four items assessing students’ scientific competencies and knowledge” (p. 28). Fourth, PISA data not only assess domain-specific knowledge and skills; but also provide rich information (from student surveys) on student background, learning strategies, traits, and attitudes. A panel of experts deliberately designed the assessment to include the aforementioned information, which is collected for use in analyzing PISA results.

The PISA assessment is held every three years with different targeted domains (i.e., reading, mathematics, and science) and is administered to 15-year-old students in 57 OECD countries (OECD, 2006). PISA 2006, a pencil-and-paper assessment with focus on science literacy, is selected for investigation in this research. OECD (2006) defines science literacy as “the ability to use scientific knowledge and processes not only to understand the natural world but to participate in decisions that affect it” (p. 12). A sample item response dataset is extracted from the 2006 PISA international science assessment data (OECD, 2007a), with 2,327 examinees (1,122 males, 48.2%; 1,205 females, 51.8%) corresponding to 21 dichotomously scored items. There are seven testlets, each with three multiple-choice items. Examinees with complete responses on the set of items are included. The extracted sample assessment dataset is estimated in

WinBUGS 1.4.3 (Lunn et al., 2000) with MCMC estimator. The priors and hyper-priors, as well as the convergence assessment in the real data analyses, are the same to those used in the simulation study. The model-fit indices used in the simulation study are used for model selection in the real data application. Model selection has been a critical issue in real data applications; the results from the simulation study provide useful information on selecting a model for the real data analyses. In addition to model-fit selection, the percentage of unmotivated test takers, estimates of item difficulty and ability parameters, and estimated testlet variances are summarized.

As stated earlier, previous studies that applied mixed-strategies IRT models have not further explored the factors that characterize test-taking motivation heterogeneity empirically. To answer the research question 4 (see Chapter 1), a follow-up exploratory investigation is conducted in the second stage of the current empirical study to explore the potential factors that characterize the heterogeneity of test-taking motivation. This investigation illustrates the way to empirically interpret latent class members (i.e., valid respondents and random guessers) after they are identified by the proposed model. The categorical latent class membership obtained from the first stage of the empirical study is then connected to a series of selected variables in the PISA 2006 student survey data (OECD, 2007a). In this case, the selected variables are gender, language, and science proficiency, as well as economic, social, and cultural status (ESCS), enjoyment of science, interest in science, self-efficacy in science tasks, self-concept of science, and motivation to learn science. These variables are possibly relevant to test-taking motivation in the domain of science achievement (Table 3). Gender and language are categorical variables, and the other variables are continuously scored (i.e., z score: positive scores indicate higher levels of the attribute). This study hypothesizes that unmotivated test takers are characterized by certain personal attributes related to the specific domain; i.e., science. For example, an examinee who minimally enjoys science may be more likely to exhibit no test-taking motivation in a low-stakes assessment.

Among these selected variables, gender and ability (i.e., mathematical ability was measured by SAT math and the Natural World Test) predicted test-taking motivation (i.e., test-taking motivation was measured by a self-report opinion scale) in a university-wide low-stakes quantitative test (Barry, Horst, Finney, Brown, & Kopp, 2010). Ability (i.e., ability was measured by math test scores) and language described aberrant item respondents (i.e., high-scoring students or second-language learners tended to provide aberrant item responses), as detected by person-fit statistics in a mathematical assessment (Petridou & Williams, 2007). Dodeen and Darabi (2009) investigated the correlations between a person-fit index and several variables in a mathematic achievement test, and results show that students' math attitudes and math motivation were negatively related to person-fit statistics; namely, students with low motivation to learn math or with low attributes toward math were more likely to give unusual response patterns.

This investigation is exploratory in nature. The statistical model of fitted logistical regression modeling can be expressed as

$$\ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = \beta_{0j} + \beta_{1j}Gender + \beta_{2j}Language + \beta_{3j}ScienceProficiency + \beta_{4j}ESCS$$
$$+\beta_{5j}Enjoyment + \beta_{6j}Interest + \beta_{7j}SelfEfficacy + \beta_{8j}SelfConcept + \beta_{9j}Motivation, \quad (19)$$

where the left-hand side of the equation represents the predicted log odds of success (i.e., $x = 1$, examinees that are random guessers) and $\hat{\pi}(x)$ shows the predicted probability of being an unmotivated respondent. The right-hand side of the expression lists intercepts (β_{0j}), as well as predictors and their corresponding regression coefficients (β_{1j}–β_{9j}). The set of variables include categorical and continuous covariates. Categorical variables are re-coded as dummy variables in the regression model. This investigation aims to facilitate the explanation of latent class membership that characterizes test-taking strategy heterogeneity in low-stakes assessments.

Table 3

Variables in the PISA 2006 Survey Used for Empirical Investigation

Variable	Definition and variable name in PISA student survey data	Coding scheme
Outcome variable		
Latent class membership		1: random guessers; 0: valid respondents
Predictors		
Gender	Gender “ST04Q01”	1: female; 0: male
Language	Language at home “ST12Q01”	1: language of the test; 0: others
Science proficiency	The first plausible value in science “PV1SCIE” - “the plausible values are random draws from the marginal posterior of the latent distribution” (OECD, 2009, p. 9).	Continuously scaled
ESCS	Index of economic, social and cultural status “ESCS” -a combined index of home possessions, the higher parental occupation, and the higher parental education expressed as years of schooling.	Continuously scaled
Enjoyment of science	Enjoyment of science “JOYSCIE” -students’ level of enjoy learning about science.	Continuously scaled
Interest of science	General interest in learning science “INTSCIE” - students’ level of interest in learning science and science-related topics	Continuously scaled
Self-efficacy of science	Science self-efficacy “SCIEEFF” - “how much students believe in their own ability to handle tasks effectively and overcome difficulties” (OECD, 2007b, p. 133).	Continuously scaled
Self-concept of science	Science self-concept “SCSCIE” - “students’ beliefs in their own academic abilities in science” (OECD, 2007b, p. 133).	Continuously scaled
Motivation to learn science	Instrumental motivation to learn science “INSTSCIE” -students’ perceptions of “the importance of learning science for either their future studies or job prospects” (OECD, 2007b, p. 146).	Continuously scaled

Note. More details can be found in OECD (2007b, 2009).

Chapter 4: Results of Simulation Study

This chapter presents the results of the simulation study. Section 4.1 provides the descriptive statistics of the simulated parameters, and Section 4.2 presents the convergence of parameter estimation. Sections 4.3, 4.4, and 4.5 present the findings on model selection, classification accuracy, and recovery of model parameter estimates, respectively.

4.1 Descriptive Statistics of the Simulated Parameters

The descriptive statistics of the simulated parameters are summarized in Table 4. The item difficulty parameters ranged from –2.140 to 2.307. The generated item parameters contained no items that are too easy or too difficult, as is observed in many practical low-stakes assessments. A complete list of the generated item difficulty parameters is provided in Appendix A. The generated ability parameters for three sample sizes exhibited mean values around 0 and standard deviation values around 1.

Table 4

Descriptive Statistics for the Simulated Item Parameters

Parameter	Test Condition	Minimum	Mean	Maximum	Standard Deviation
b	All conditions	-2.140	-0.150	2.307	0.991
θ	Sample = 1,000	-3.147	0.011	3.404	1.019
θ	Sample = 3,000	-3.090	-0.005	3.417	0.995
θ	Sample = 5,000	-3.587	-0.009	3.492	1.017

4.2 Evaluation of Parameter Convergence

Generally, the model parameters converged well, as indicated by the multiple criteria. Figures 1–4 show some examples of quantile, autocorrelation, density, and history plots, which indicate a good mixing of Markov chains and satisfactory convergence. Most model parameters reached convergence after 3,000 iterations, whereas testlet variances required more iterations for them to reach convergence. Numerically, all MC errors of the targeted parameters were less than 0.05 (Table 5), indicating convergence. For each Markov chain, a minimum of 10,000 iterations are generally required with a burn-in of 5,000 and a post-burn-in of 5,000 iterations for inferences (i.e., two Markov chains result in 10,000 iterations for inferences). For some datasets, 10,000 to 20,000 iterations are needed to guarantee convergence for all model parameters; for those datasets, additional 5,000 iterations are added to draw inferences when all model parameters converge. Non-converged datasets and label switching problems were not observed in the simulation study.

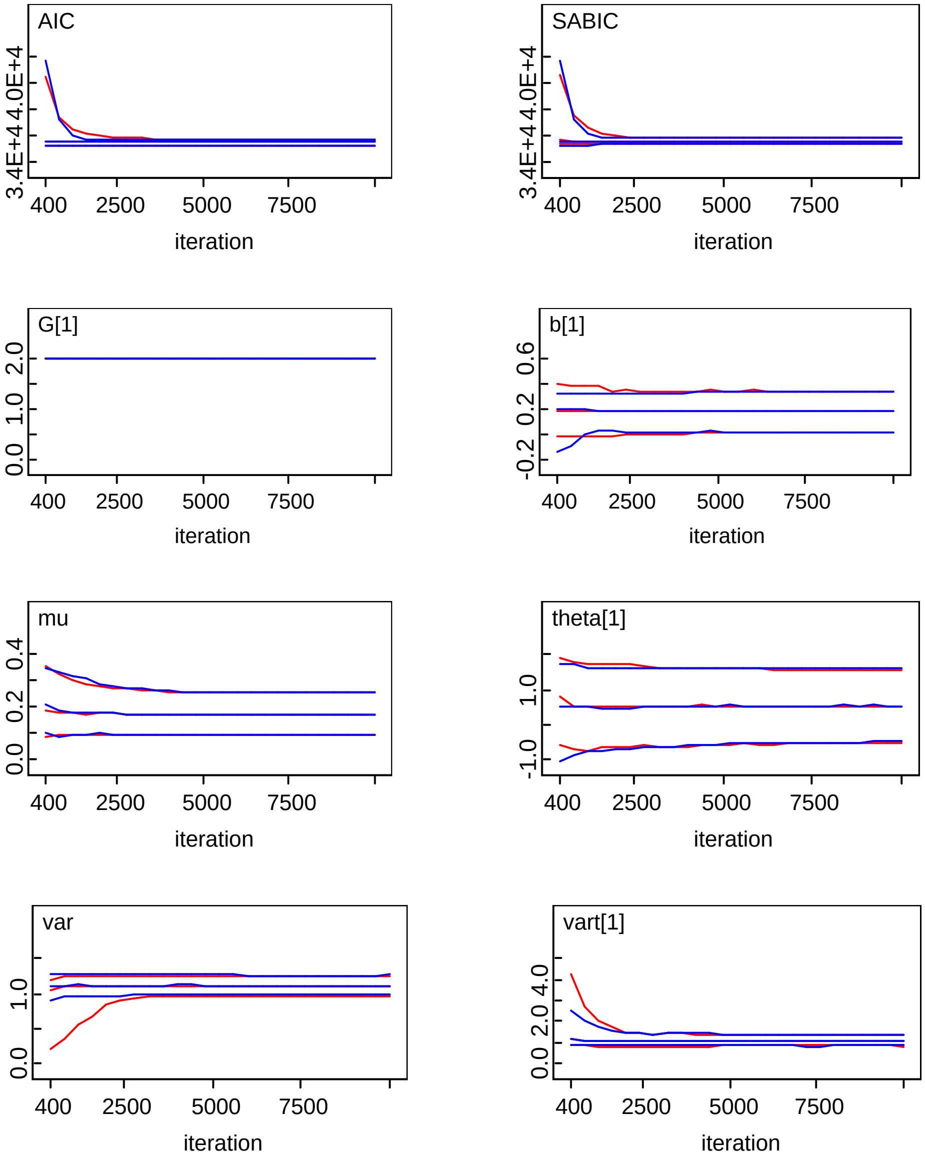

Figure 1. Sample Quantiles Plots.

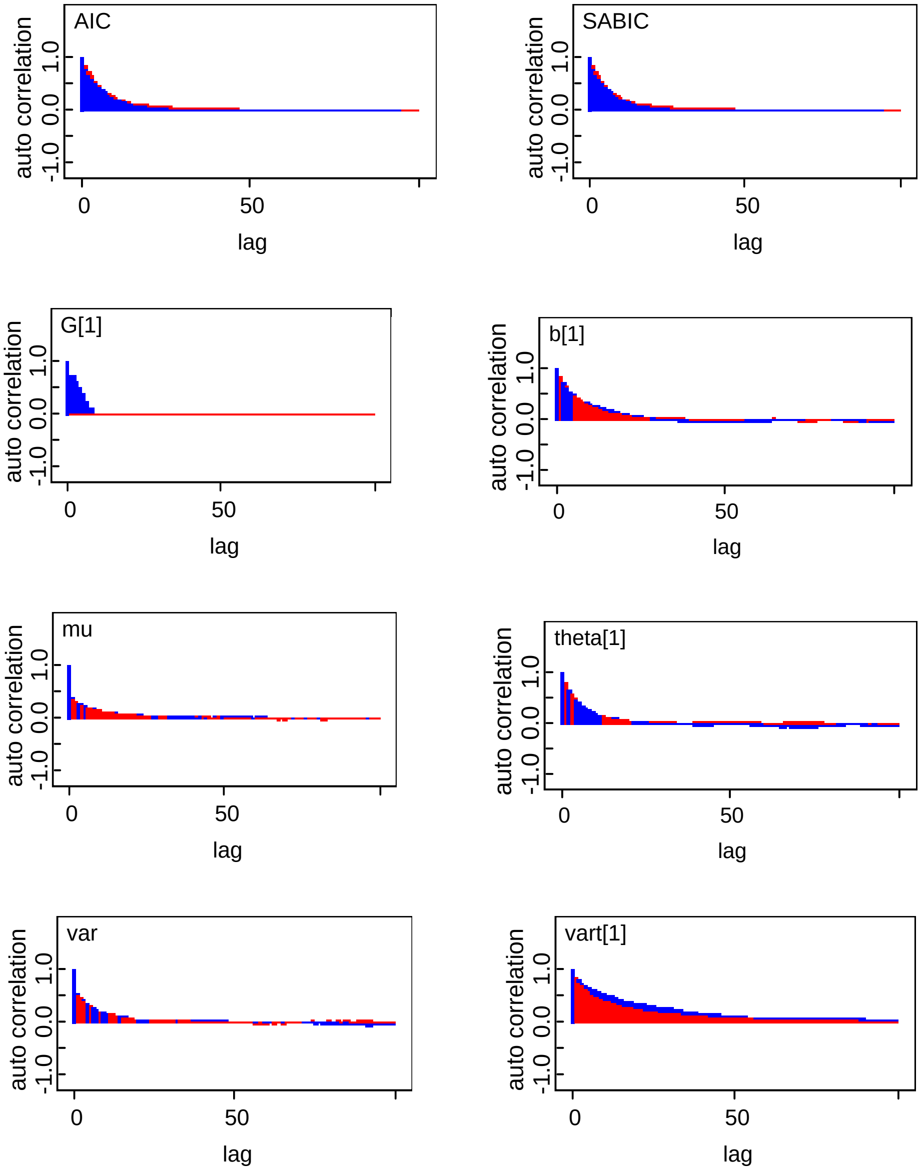

Figure 2. Sample Autocorrelation Plots.

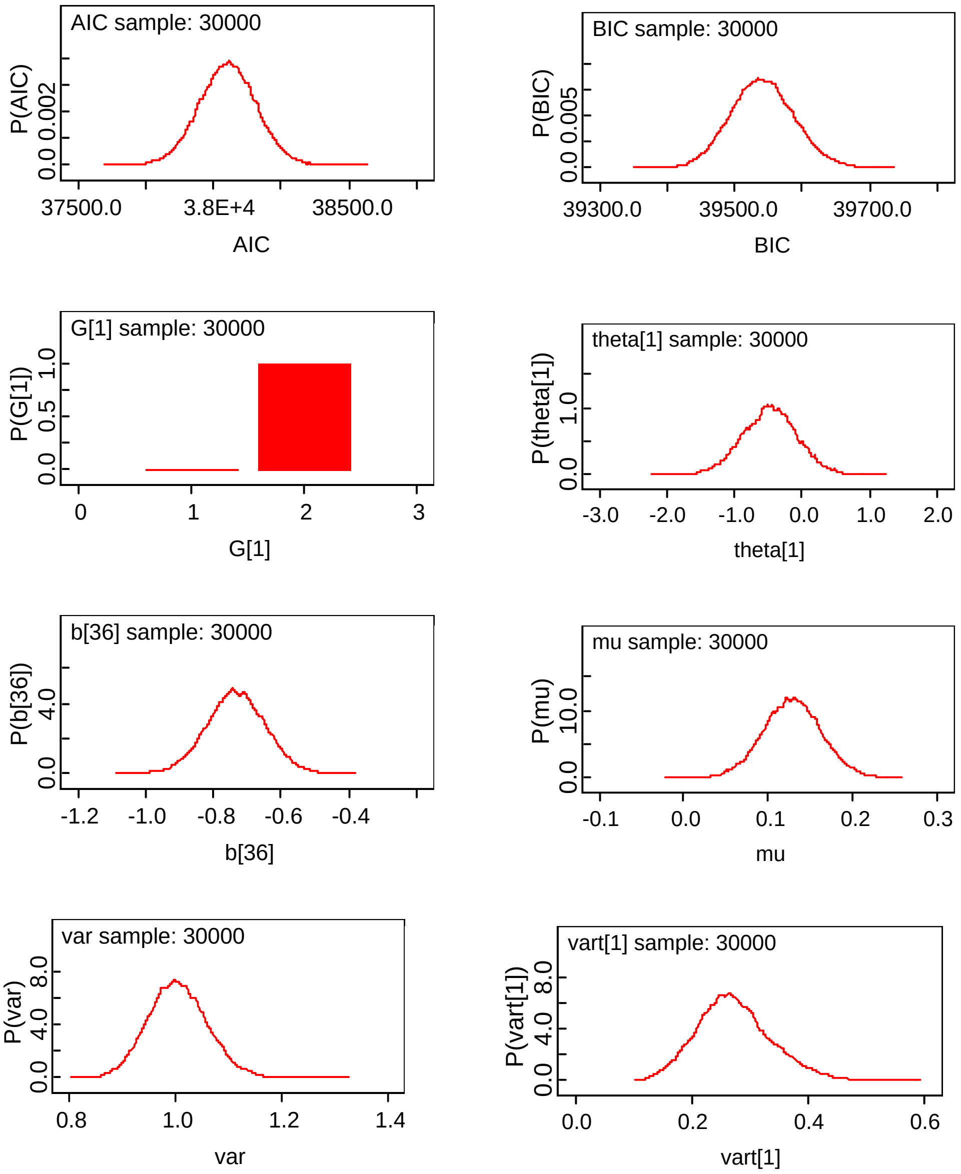

Figure 3. Sample Density Plots.

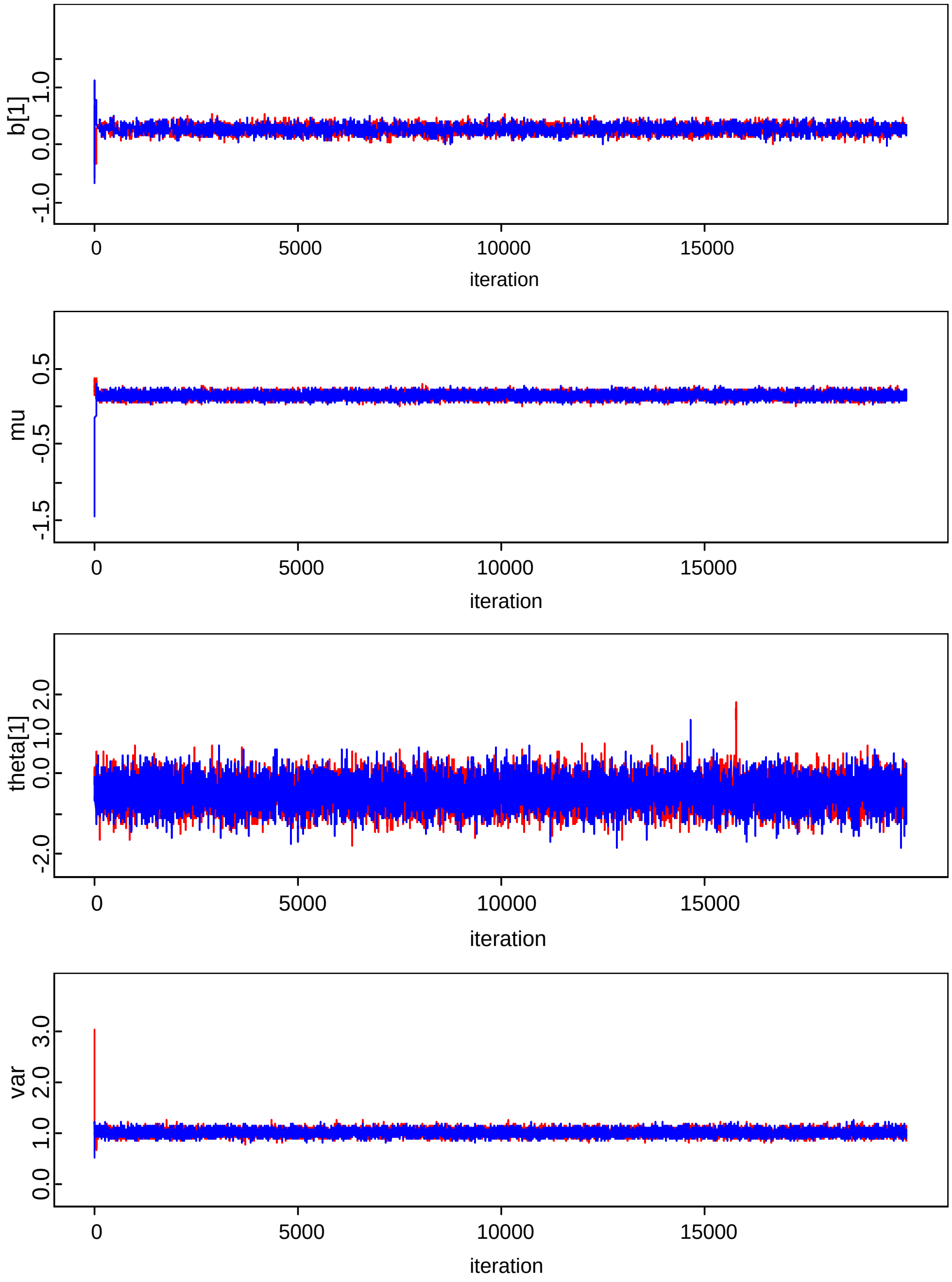

Figure 4. Sample History Plots.

Table 5

WinBUGS MC Errors for the Selected Parameters

Parameter	MC Error		
	Mixed-Strategies Rasch Testlet Model	Mixed-Strategies Rasch Model	Rasch Model
b[1]	0.0015	0.0009	0.0009
b[2]	0.0017	0.0012	0.0011
b[3]	0.0015	0.0011	0.0013
theta[1]	0.0174	0.0090	0.0082
theta[2]	0.0154	0.0069	0.0067
theta[3]	0.0152	0.0079	0.0089
mu	0.0004	0.0003	0.0002
var	0.0009	0.0005	0.0004
G[1]	0.0000	0.0000	−
G[2]	0.0000	0.0000	−
G[3]	0.0000	0.0000	−
vart[1]	0.0032	−	−
vart[2]	0.0033	−	−
vart[3]	0.0029	−	−
vart[4]	0.0031	−	−
vart[5]	0.0032	−	−
vart[6]	0.0035	−	−

Note. The parameters are selected from the first replicated dataset with sample = 5,000, testlet variance = 0.25, and guessers = 1 %.

4.3 Results for Model Selection

In this study, the model–data fit is compared in terms of AIC (Akaike, 1974), BIC (Schwarz, 1978), AICc (Burnham & Anderson, 2002), and SABIC (Sclove, 1987). The smallest value of the model-fit indices indicates the best fitting model. Results of model-fit statistics are shown in Table 6, which presents the number of replications favored by a particular model within a testing condition. In 25 replications, all the indices suggested the mixed-strategies Rasch testlet model as the best fitting model and the Rasch model as the worst. That is, under test-taking motivation heterogeneity and testlet effects, the mixed-strategies Rasch testlet model provided the best model–data fit. The model-fit statistics reflected consistent and equivalently effective performance across testing conditions, indicating that these model–data fit statistics are useful for the estimation models considered.

Table 6

Summary of Model Selection Frequency (Accuracy %)

			AIC			AIC$_C$		
Sample	Testlet Variance	Guessers %	Mixed-strategies Rasch testlet model	Mixed-strategies Rasch model	Rasch model	Mixed-strategies Rasch testlet model	Mixed-strategies Rasch model	Rasch model
1,000	0.25	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
	1	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
3,000	0.25	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
	1	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
5,000	0.25	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
	1	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)

Note. Frequency refers to the frequency at which a model is selected across replications.

Table 6

Summary of Model Selection Frequency (continued)

			BIC			SABIC		
Sample	Testlet Variance	Guessers %	Mixed-strategies Rasch testlet model	Mixed-strategies Rasch model	Rasch model	Mixed-strategies Rasch testlet model	Mixed-strategies Rasch model	Rasch model
1,000	0.25	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
	1	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
3,000	0.25	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
	1	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
5,000	0.25	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
	1	1%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		5%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)
		15%	25 (100 %)	0 (0 %)	0 (0 %)	25 (100 %)	0 (0 %)	0 (0 %)

Note. Frequency refers to the frequency at which a model is selected across replications.

4.4 Results for Classification Accuracy

The accuracy of latent class classification is reflected by the percentage of correctly classified test takers. Table 7 presents the results on classification accuracy for the mixed-strategies Rasch testlet model and for the mixed-strategies Rasch model. The Rasch model is excluded because it allows for only one latent class in an examinee population. The mixed-strategies Rasch testlet model and the mixed-strategies Rasch model exhibited satisfactory and comparably correct identification of latent class membership. The percentages of correct classification across varied testing conditions ranged from 95.48% to 99.33% for the mixed-strategies Rasch testlet model and from 94.61% to 99.31% for the mixed-strategies Rasch model. Subedi (2009) also reported high classification accuracy, with a range of 96.92% to 98.06% for the mixture IRT model with random guessing. The classification accuracy in the current work was uninfluenced by sample size or testlet effects, and accuracy was slightly higher as the percentage of unmotivated respondents increased. The latter result may be attributed to the fact that when the percentage of unmotivated respondents decreases (e.g., from 15% to 1%), the mixing proportion of latent classes in an examinee population becomes more extreme (e.g., from 0.85:0.15 to 0.99:0.01), making the partitioning of latent classes in the examinee population more easily achievable. Previous studies that investigated the varied mixing proportions in mixture IRT models also revealed similar findings, i.e., that more unbalanced composition of latent classes in the examinee population produced slightly better classification accuracy (e.g., Chen et al., 2013; Cho et al., 2012).

Table 7

Classification Accuracy

Sample	Testlet Variance	Guessers %	Mixed-strategies Rasch testlet model	Mixed-strategies Rasch model
1,000	0.25	1%	99.28	99.25
		5%	97.64	97.56
		15%	95.58	95.52
	1	1%	99.29	99.14
		5%	97.70	97.34
		15%	95.52	94.61
3,000	0.25	1%	99.29	99.29
		5%	97.65	97.62
		15%	95.62	95.52
	1	1%	99.32	99.22
		5%	97.73	97.32
		15%	95.78	94.85
5,000	0.25	1%	99.33	99.31
		5%	97.70	97.64
		15%	95.48	95.37
	1	1%	99.31	99.20
		5%	97.76	97.30
		15%	95.66	94.69

4.5 Results for Parameter Recovery

Item parameter recovery. To assess whether test-taking motivation heterogeneity and testlet effects influence the precision of item parameter estimation, this study evaluates the recovery of item difficulty parameters by comparing the estimated and simulated parameters in terms of bias (Table 8) and RMSE (Table 9). Across testing conditions, the variability of the bias in item difficulty parameters was lowest in the mixed-strategies Rasch testlet model and highest in the Rasch model. The Rasch model exhibited a noteworthy trend: an increase in testlet variance or an increase in the percentage of unmotivated test takers increased the variability of the bias in item difficulty parameters. Similarly, in the mixed-strategies Rasch model, the variability of the bias in item difficulty parameters was higher under large testlet variance conditions than it was under small testlet variance conditions. In the mixed-strategies Rasch testlet model, this variability decreased as sample size increased.

The RMSE values of the item difficulty parameters are numerically summarized in Table 9 and graphically depicted in Figures 5–12. The recovery of item parameters differed depending on estimation model, magnitude of testlet effects, percentage of unmotivated respondents, and sample size. The precision of item parameter estimation was most strongly affected by the estimation model fitting to data. For example, item parameters were recovered to the greatest extent by the mixed-strategies Rasch testlet model because the mean RMSE values approached zero across testing conditions (RMSE = 0.034 to 0.097). The

Rasch model exhibited the worst item parameter recovery (RMSE = 0.055 to 0.228).

Generally, an increase in testlet effects and in the percentage of unmotivated respondents, as well as a decrease in sample size, increased the RMSE values of the item difficulty parameters. RMSE visibly increased as testlet variance increased (Figures 5–10), particularly in the comparison models where testlet effects are not taken into account (i.e., the mixed-strategies Rasch model and Rasch model). Disregarding the heterogeneity of test-taking motivation and testlet effects in the Rasch model affected its precision in item parameter estimation; the RMSE of the item difficulty parameters increased as the magnitude of testlet effects and/or the percentage of unmotivated test takers in the examinee population increased (Figures 5–10). Under a large sample size, the RMSE values of the item difficulty parameters were typically lower, particularly those estimated by the mixed-strategies Rasch testlet model (Figures 11–12).

The ANOVA and effect sizes are analyzed for the RMSE in item difficulty ($\alpha = .05$), which evaluates how estimation model and testing conditions affected the recovery of item parameters. Significant main effects were found for estimation model ($F[2, 1890] = 221.76$, $p < .001$, a large effect size [$\eta^2 = 0.190$]), testlet effects ($F[1, 1890] = 162.12$, $p < .001$, a medium effect size [$\eta^2 = 0.079$]), percentage of guessers ($F[2, 1890] = 43.29$, $p < .001$, a small effect size [$\eta^2 = 0.044$]), and sample size ($F[2, 1890] = 40.96$, $p < .001$, a small effect size [$\eta^2 =$

0.042]). The post-hoc Turkey comparisons indicate that significant differences in terms of the RMSE in beta were found (1) between any pair of estimation models; (2) between small and the other two samples; (3) between any pair of percentages of guessers; and (4) between levels of testlet effects. In addition, two interaction effects on the RMSE in item difficulty were statistically significant: model*testlet effects ($F[2, 1890] = 29.49$, $p < .001$, a small effect size [$\eta^2 = 0.030$]) and model*guessers ($F[4, 1890] = 31.46$, $p < .001$, a medium effect size [$\eta^2 = 0.062$]). This finding indicates that one level of testlet effects had high RMSE in item difficulty within a certain estimation model, and that the other level of testlet effects showed high RMSE in item difficulty within other estimation model(s). Additionally, one level of guessers exhibited high RMSE in item difficulty within a certain estimation model, while the other level of guessers had high RMSE in item difficulty within other estimation models.

Table 8

Bias in Item Difficulty Parameter Estimates

Sample	1,000						3,000					
Testlet Variance	0.25			1			0.25			1		
Guessers %	1%	5%	15%	1%	5%	15%	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model												
Minimum	-0.044	-0.043	-0.025	-0.023	-0.049	-0.052	-0.018	-0.017	-0.019	-0.023	-0.021	-0.033
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.045	0.047	0.032	0.045	0.051	0.049	0.022	0.023	0.023	0.016	0.017	0.031
Standard Deviation	0.019	0.020	0.015	0.017	0.019	0.024	0.009	0.009	0.009	0.008	0.008	0.014
Mixed-strategies Rasch model												
Minimum	-0.136	-0.111	-0.109	-0.339	-0.353	-0.368	-0.094	-0.109	-0.091	-0.319	-0.335	-0.357
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.094	0.134	0.112	0.270	0.296	0.293	0.101	0.087	0.098	0.270	0.286	0.314
Standard Deviation	0.049	0.054	0.051	0.135	0.141	0.138	0.043	0.042	0.045	0.135	0.137	0.140
Rasch model												
Minimum	-0.150	-0.225	-0.478	-0.351	-0.440	-0.642	-0.114	-0.234	-0.456	-0.338	-0.431	-0.660
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.117	0.284	0.532	0.298	0.418	0.649	0.133	0.257	0.513	0.295	0.407	0.653
Standard Deviation	0.057	0.109	0.211	0.144	0.187	0.276	0.053	0.101	0.208	0.144	0.183	0.277

Table 8

Bias in Item Difficulty Parameter Estimates (continued)

Sample	5,000					
Testlet Variance	0.25			1		
Guessers %	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model						
Minimum	-0.021	-0.023	-0.012	-0.020	-0.021	-0.019
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.015	0.016	0.015	0.018	0.014	0.022
Standard Deviation	0.007	0.009	0.007	0.010	0.009	0.010
Mixed-strategies Rasch model						
Minimum	-0.096	-0.096	-0.095	-0.326	-0.346	-0.332
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.079	0.091	0.099	0.266	0.285	0.308
Standard Deviation	0.042	0.044	0.042	0.135	0.138	0.136
Rasch model						
Minimum	-0.117	-0.217	-0.460	-0.343	-0.443	-0.632
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.113	0.241	0.519	0.292	0.409	0.649
Standard Deviation	0.053	0.100	0.205	0.144	0.186	0.276

Table 9

RMSE of Item Difficulty Parameter Estimates

Sample	1,000						3,000					
Testlet Variance	0.25			1.00			0.25			1.00		
Guessers %	1%	5%	15%	1%	5%	15%	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model												
Minimum	0.048	0.054	0.062	0.061	0.067	0.067	0.032	0.034	0.037	0.036	0.035	0.037
Mean	0.076	0.079	0.083	0.084	0.088	0.097	0.044	0.046	0.048	0.047	0.049	0.055
Maximum	0.109	0.110	0.109	0.117	0.137	0.131	0.062	0.059	0.069	0.069	0.075	0.081
Standard Deviation	0.014	0.014	0.014	0.014	0.016	0.015	0.007	0.008	0.008	0.007	0.010	0.009
Mixed-strategies Rasch model												
Minimum	0.051	0.053	0.061	0.064	0.066	0.065	0.031	0.031	0.037	0.037	0.032	0.035
Mean	0.083	0.088	0.090	0.136	0.139	0.141	0.055	0.057	0.060	0.116	0.118	0.122
Maximum	0.161	0.151	0.155	0.345	0.371	0.375	0.111	0.118	0.116	0.324	0.339	0.361
Standard Deviation	0.023	0.026	0.024	0.070	0.076	0.075	0.020	0.020	0.022	0.077	0.078	0.081
Rasch model												
Minimum	0.053	0.055	0.055	0.062	0.064	0.054	0.030	0.031	0.032	0.037	0.035	0.034
Mean	0.086	0.114	0.179	0.141	0.168	0.228	0.061	0.091	0.163	0.123	0.150	0.217
Maximum	0.172	0.289	0.536	0.357	0.455	0.653	0.141	0.261	0.514	0.342	0.434	0.662
Standard Deviation	0.027	0.059	0.128	0.075	0.107	0.167	0.027	0.059	0.132	0.083	0.110	0.174

Table 9

RMSE of Item Difficulty Parameter Estimates (continued)

Sample	5,000					
Testlet Variance	0.25			1.00		
Guessers %	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model						
Minimum	0.025	0.025	0.028	0.025	0.026	0.023
Mean	0.034	0.035	0.037	0.037	0.038	0.040
Maximum	0.050	0.049	0.051	0.055	0.051	0.054
Standard Deviation	0.006	0.005	0.006	0.006	0.005	0.007
Mixed-strategies Rasch model						
Minimum	0.026	0.023	0.027	0.027	0.030	0.024
Mean	0.049	0.051	0.050	0.112	0.115	0.114
Maximum	0.107	0.104	0.106	0.328	0.348	0.335
Standard Deviation	0.020	0.020	0.020	0.079	0.082	0.081
Rasch model						
Minimum	0.028	0.023	0.025	0.026	0.029	0.032
Mean	0.055	0.086	0.158	0.119	0.149	0.214
Maximum	0.127	0.243	0.520	0.344	0.445	0.649
Standard Deviation	0.028	0.059	0.131	0.085	0.114	0.173

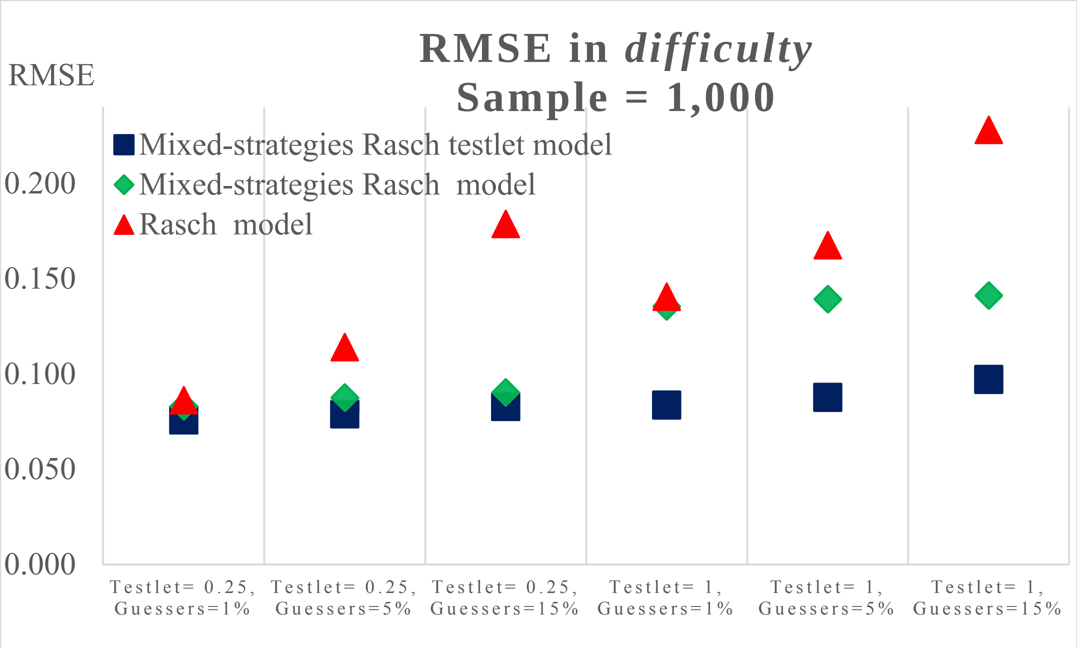

Figure 5. Plot of RMSE of Item Difficulty Estimates at a Sample = 1,000.

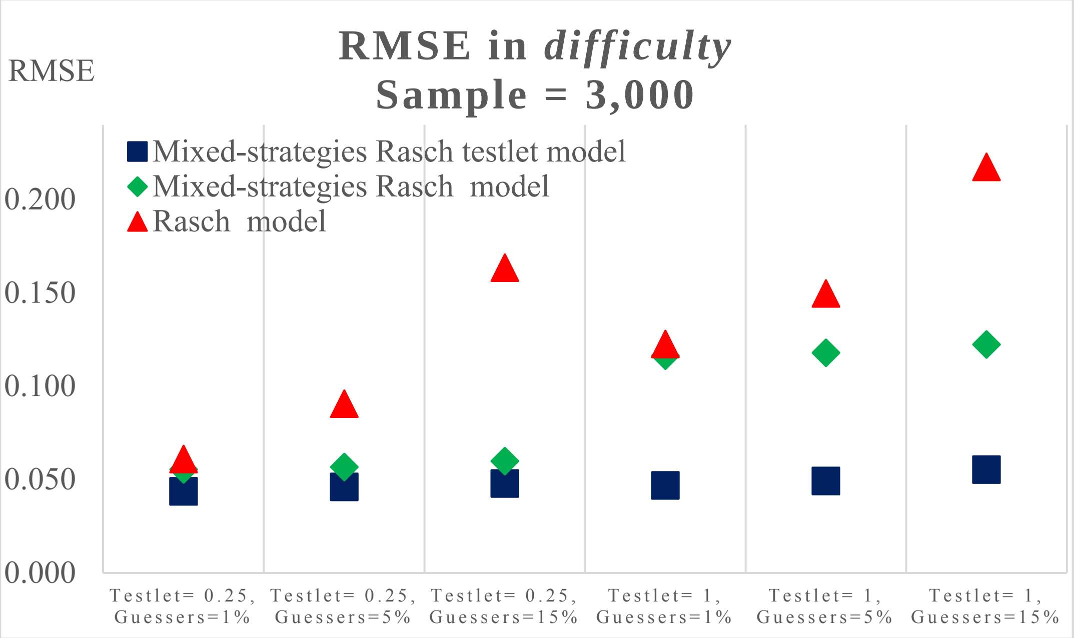

Figure 6. Plot of RMSE of Item Difficulty Estimates at a Sample = 3,000.

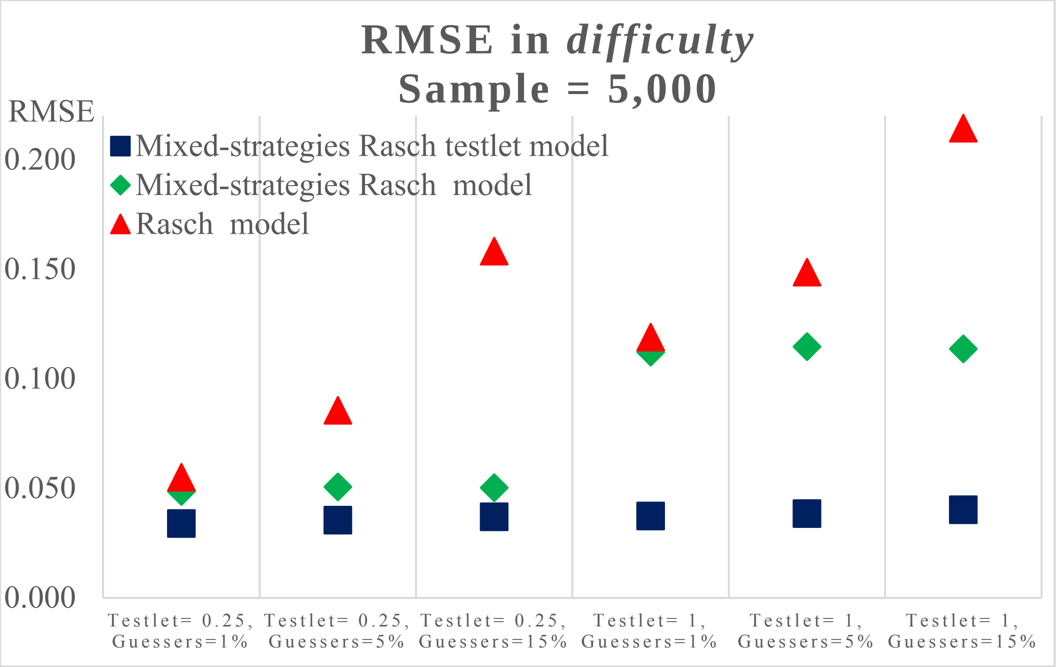

Figure 7. Plot of RMSE of Item Difficulty Estimates at a Sample = 5,000.

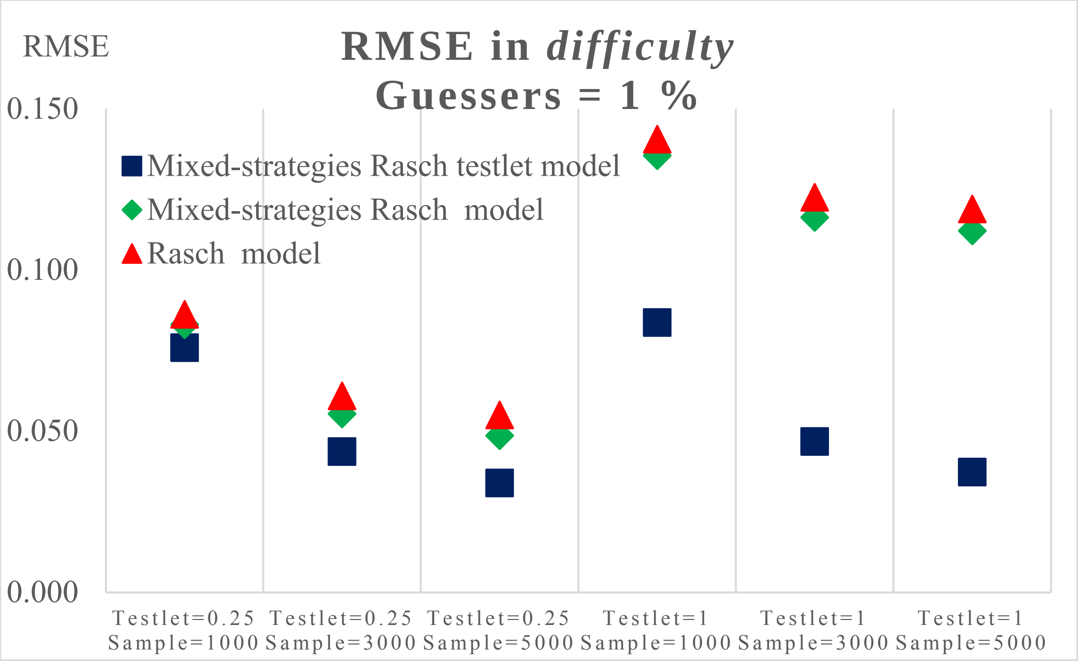

Figure 8. Plot of RMSE of Item Difficulty Estimates at Guessers = 1%.

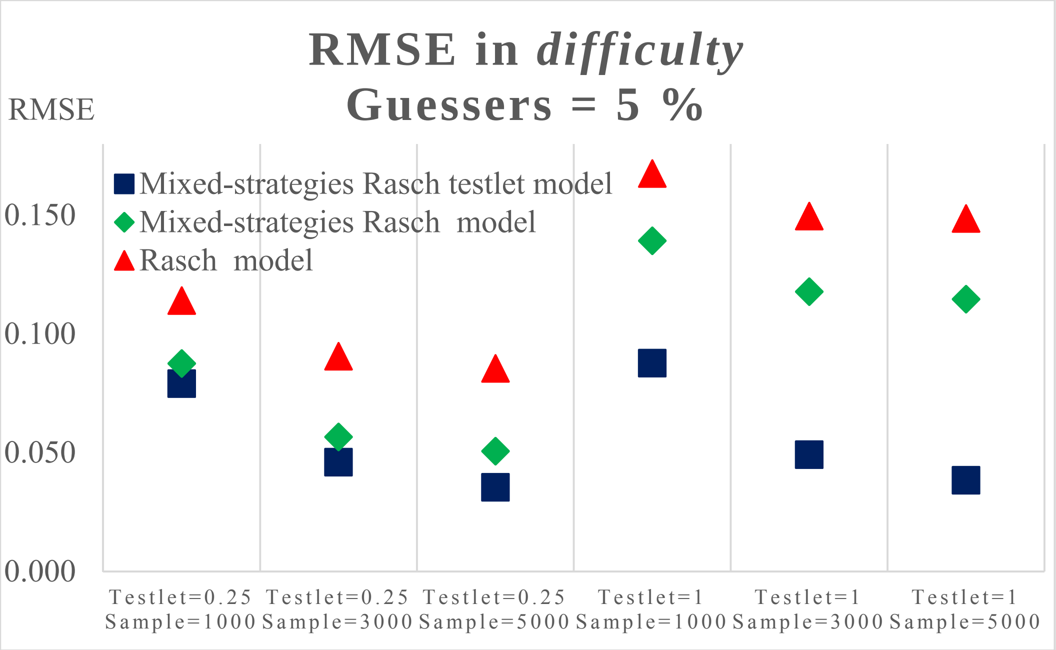

Figure 9. Plot of RMSE of Item Difficulty Estimates at Guessers = 5%.

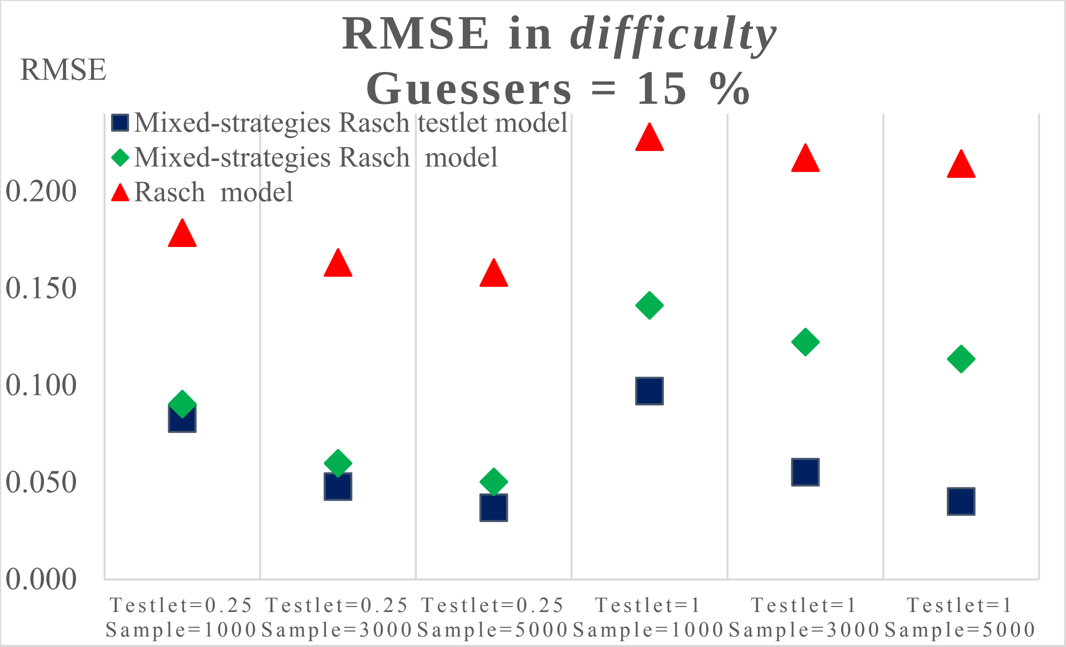

Figure 10. Plot of RMSE of Item Difficulty Estimates at Guessers = 15%.

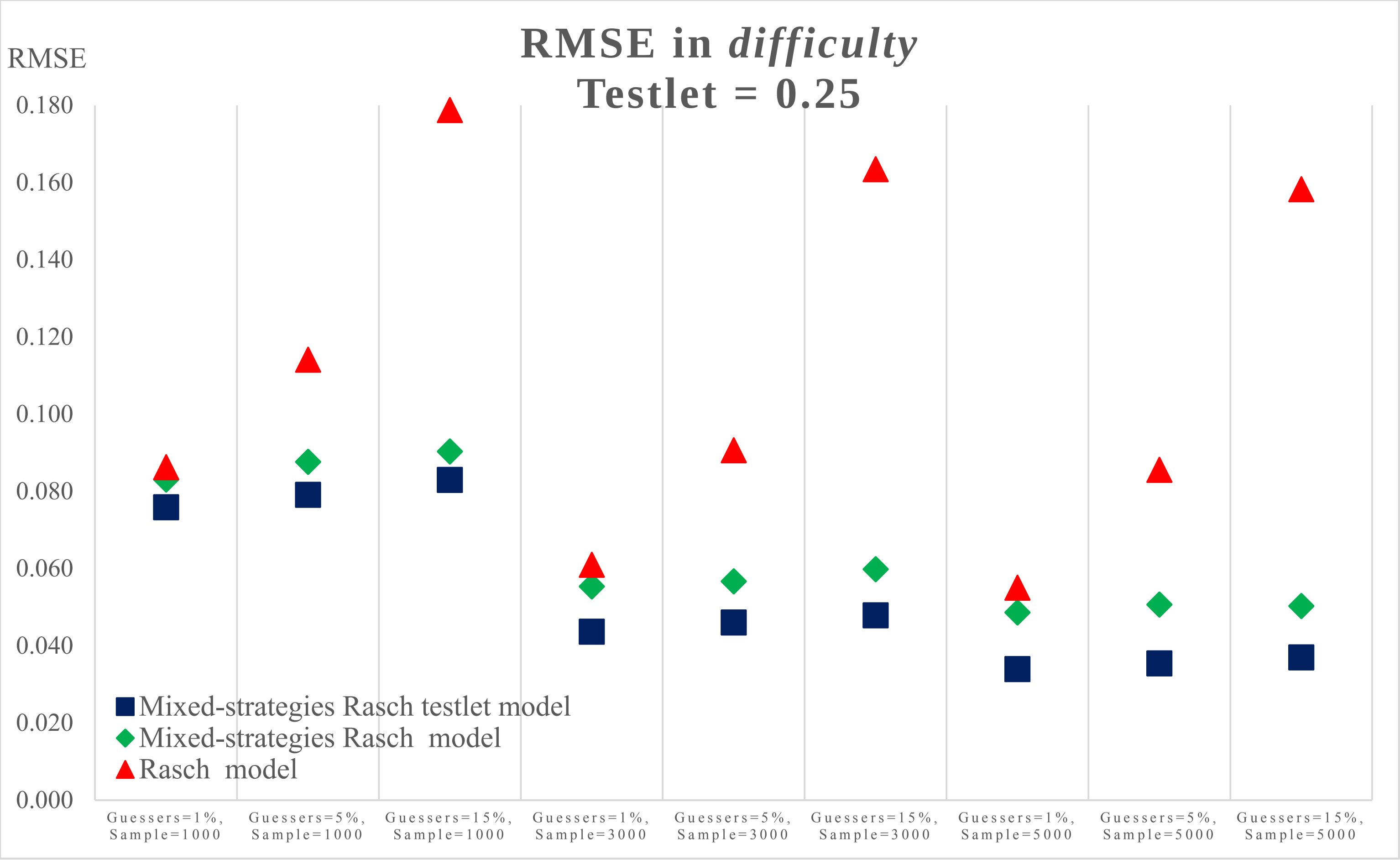

Figure 11. Plot of RMSE of Item Difficulty Estimates at a Testlet Variance = 0.25.

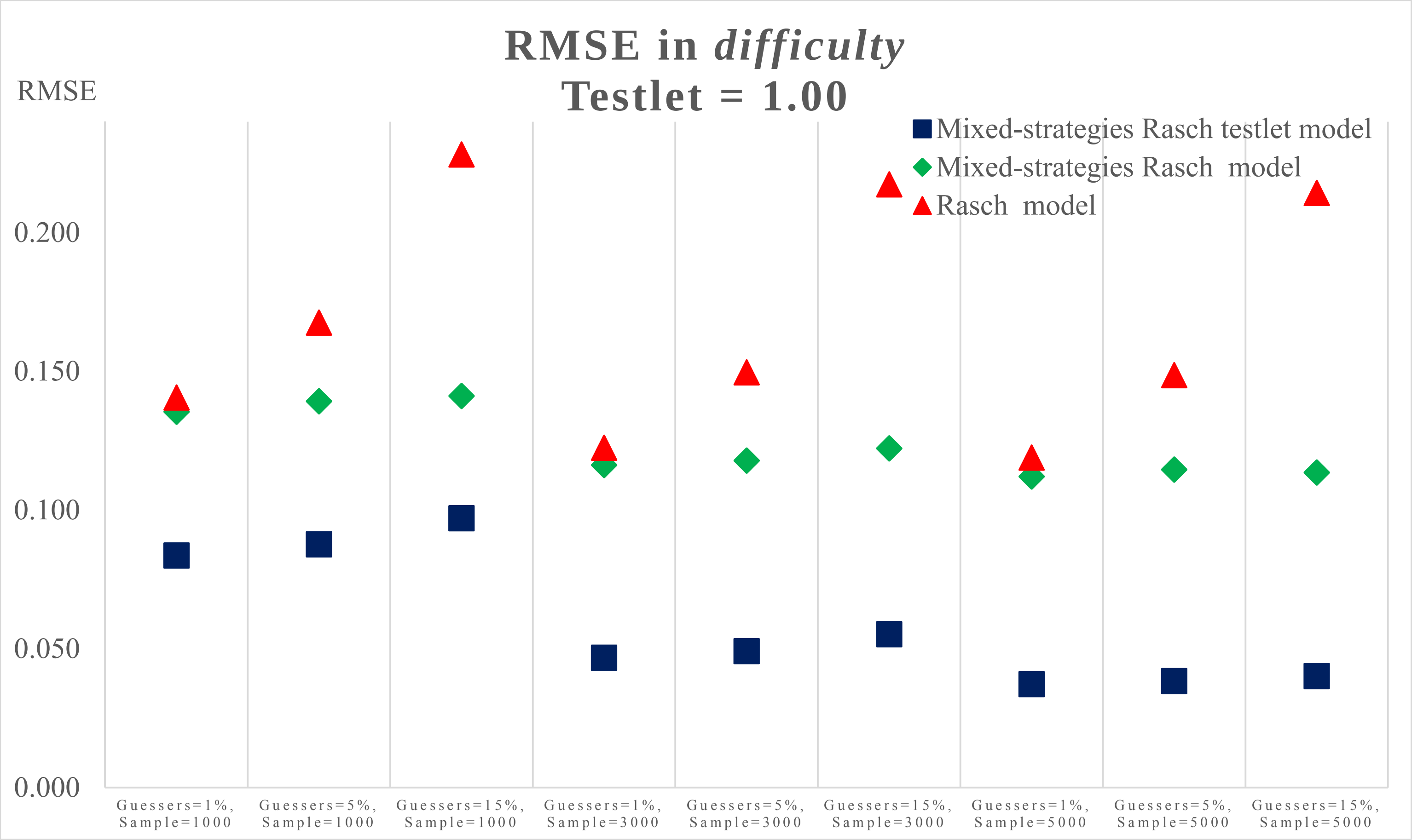

Figure 12. Plot of mean RMSE of Item Difficulty Estimates at a Testlet Variance = 1.00.

Ability parameter recovery. Table 10 presents the bias in ability parameter estimates. Three factors affected bias variability: percentage of unmotivated respondents, magnitude of testlet variance, and estimation model. An observed tendency is that the increasing magnitude of testlet effects and percentage of unmotivated respondents elevated bias variability. Across testing conditions, bias variability was generally smallest in the mixed-strategies Rasch testlet model and largest in the Rasch model.

The RMSE values of the ability parameter estimates are listed in Table 11 and graphically illustrated in Figures 13–20. The recovery of ability parameters differed in terms of estimation model, testlet effects, and percentage of unmotivated respondents. The precision of ability parameter estimation diminished as the magnitude of testlet effects and percentage of unmotivated respondents increased. The mixed-strategies Rasch testlet model and mixed-strategies Rasch model exhibited minimal difference in the RMSE values of the estimates. Chen et al. (2013) reported similar findings, revealing that the mixture Rasch model and mixture Rasch testlet model produced comparable ability parameter estimates when testlet effects were present in the data. In the current study, under numerous unmotivated respondents (i.e., 5% & 15%), the Rasch model provided the worst ability parameter recovery among the three estimation models.

The ANOVA is conducted for the RMSE in ability parameter estimates, which assesses which manipulated factors significantly contributed to the precision of ability parameter estimates. Three main effects exhibited statistical significance: (1) estimation model (*F*[2, 154386]= 944.62, $p < .001$, a small effect size [$\eta^2 = 0.012$]); (2) testlet effects (*F*[1, 154386]= 4519.04, $p < .001$, a small effect size [$\eta^2 = 0.028$]); and (3) percentage of unmotivated respondents (*F*[2, 154386]= 721.36, $p < .001$). The post-hoc Turkey comparisons indicate significant differences in the RMSE of ability parameters (1) between the Rasch model and the other estimation models; (2) between any pair of the percentages of unmotivated respondents; and (3) between levels of testlet effects. In addition, two significant interaction effects occurred: model*guessers (*F*[4, 154386]= 364.86, $p < .001$) and guessers*testlet effects (*F*[2, 154386]= 11.62, $p < .001$). This finding shows that estimation model and unmotivated respondents, as well as testlet effects and unmotivated respondents, exerted joint effects.

Table 10

Bias in Ability Parameter Estimates

Sample	1,000						3,000					
Testlet Variance	0.25			1.00			0.25			1.00		
Guessers %	1%	5%	15%	1%	5%	15%	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model												
Minimum	-1.020	-1.082	-1.008	-1.232	-1.257	-1.285	-1.035	-1.044	-1.116	-1.205	-1.214	-1.397
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.797	0.887	0.999	1.048	1.155	1.330	0.928	1.044	1.131	1.039	1.198	1.379
Standard Deviation	0.195	0.221	0.258	0.273	0.303	0.337	0.195	0.216	0.256	0.283	0.296	0.331
Mixed-strategies Rasch model												
Minimum	-1.024	-1.095	-1.026	-1.248	-1.275	-1.307	-1.044	-1.056	-1.128	-1.221	-1.225	-1.408
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.818	0.908	0.995	1.082	1.151	1.373	0.941	1.050	1.180	1.037	1.230	1.386
Standard Deviation	0.200	0.227	0.263	0.285	0.315	0.357	0.199	0.220	0.259	0.288	0.304	0.343
Rasch model												
Minimum	-2.193	-3.118	-3.429	-2.326	-2.951	-3.445	-3.452	-4.102	-3.938	-3.610	-4.021	-3.998
Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.797	1.492	1.843	1.038	1.738	1.849	1.087	1.444	1.833	1.198	1.387	1.846
Standard Deviation	0.224	0.372	0.626	0.296	0.419	0.649	0.264	0.405	0.637	0.331	0.441	0.653

Table 10

Bias in Ability Parameter Estimates (continued)

Sample	5,000					
Testlet Variance	0.25			1.00		
Guessers %	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model						
Minimum	-1.116	-1.104	-1.198	-1.331	-1.345	-1.349
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	0.991	1.193	1.210	1.384	1.282	1.624
Standard Deviation	0.192	0.217	0.255	0.280	0.297	0.330
Mixed-strategies Rasch model						
Minimum	-1.131	-1.120	-1.215	-1.354	-1.377	-1.384
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	1.010	1.230	1.243	1.395	1.292	1.638
Standard Deviation	0.198	0.223	0.262	0.289	0.310	0.350
Rasch model						
Minimum	-2.708	-3.082	-3.695	-2.694	-3.126	-3.701
Mean	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	1.782	1.721	1.819	1.765	1.765	1.913
Standard Deviation	0.240	0.395	0.598	0.314	0.433	0.616

Table 11

RMSE of Ability Parameter Estimates

Sample	1,000						3,000					
Testlet Variance	0.25			1.00			0.25			1.00		
Guessers %	1%	5%	15%	1%	5%	15%	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model												
Minimum	0.183	0.217	0.190	0.296	0.242	0.249	0.210	0.220	0.207	0.239	0.238	0.241
Mean	0.405	0.414	0.430	0.504	0.504	0.515	0.404	0.412	0.423	0.496	0.499	0.505
Maximum	1.090	1.122	1.178	1.285	1.296	1.451	1.071	1.145	1.268	1.247	1.290	1.504
Standard Deviation	0.087	0.109	0.138	0.113	0.132	0.161	0.085	0.101	0.131	0.112	0.124	0.149
Mixed-strategies Rasch model												
Minimum	0.189	0.214	0.188	0.283	0.238	0.247	0.207	0.216	0.205	0.236	0.244	0.251
Mean	0.404	0.414	0.429	0.503	0.504	0.517	0.403	0.411	0.422	0.496	0.498	0.506
Maximum	1.091	1.135	1.160	1.299	1.316	1.471	1.078	1.147	1.311	1.264	1.292	1.503
Standard Deviation	0.088	0.112	0.139	0.119	0.138	0.173	0.087	0.103	0.132	0.115	0.129	0.156
Rasch model												
Minimum	0.189	0.215	0.249	0.291	0.249	0.279	0.209	0.231	0.206	0.231	0.228	0.238
Mean	0.409	0.449	0.570	0.507	0.537	0.637	0.414	0.456	0.571	0.506	0.540	0.635
Maximum	2.218	3.147	3.443	2.345	2.959	3.468	3.466	4.110	3.951	3.625	4.036	4.011
Standard Deviation	0.120	0.261	0.442	0.136	0.254	0.427	0.172	0.294	0.453	0.178	0.280	0.432

Table 11

RMSE of Ability Parameter Estimates (continued)

Sample	5,000					
Testlet Variance	0.25			1.00		
Guessers %	1%	5%	15%	1%	5%	15%
Mixed-strategies Rasch testlet model						
Minimum	0.209	0.201	0.218	0.253	0.248	0.226
Mean	0.405	0.413	0.429	0.499	0.503	0.515
Maximum	1.146	1.253	1.303	1.421	1.405	1.764
Standard Deviation	0.088	0.104	0.134	0.116	0.125	0.151
Mixed-strategies Rasch model						
Minimum	0.208	0.198	0.215	0.253	0.237	0.214
Mean	0.405	0.412	0.429	0.498	0.503	0.517
Maximum	1.160	1.298	1.336	1.428	1.438	1.753
Standard Deviation	0.090	0.106	0.137	0.120	0.131	0.163
Rasch model						
Minimum	0.209	0.223	0.197	0.258	0.233	0.238
Mean	0.413	0.456	0.560	0.505	0.541	0.627
Maximum	2.735	3.103	3.712	2.710	3.143	3.722
Standard Deviation	0.145	0.282	0.414	0.156	0.268	0.393

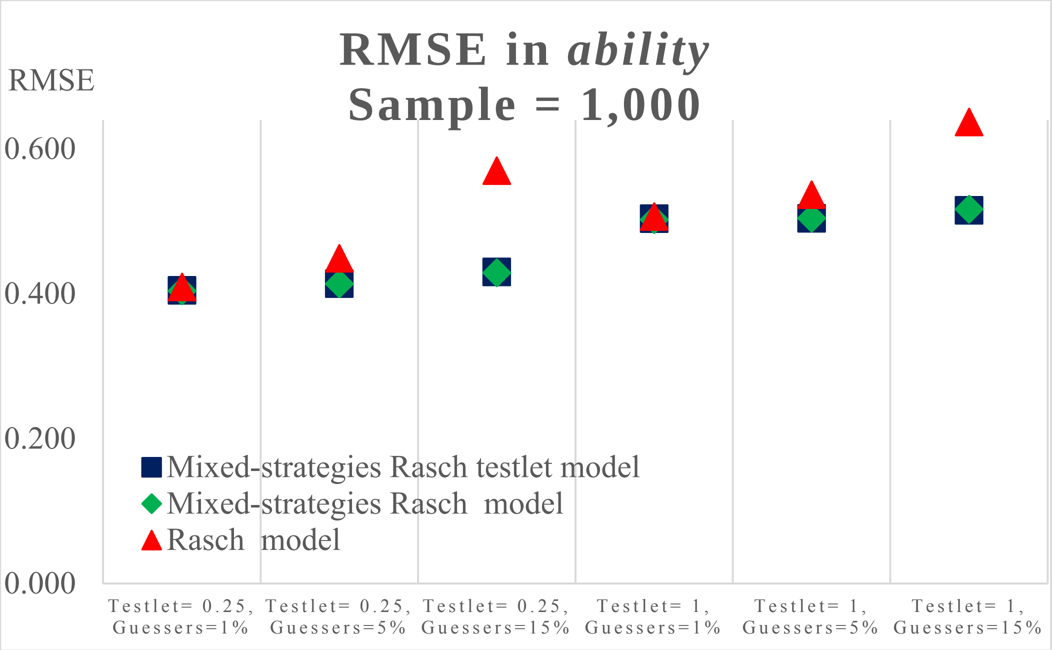

Figure 13. Plot of RMSE of Ability Parameter Estimates at a Sample = 1,000.

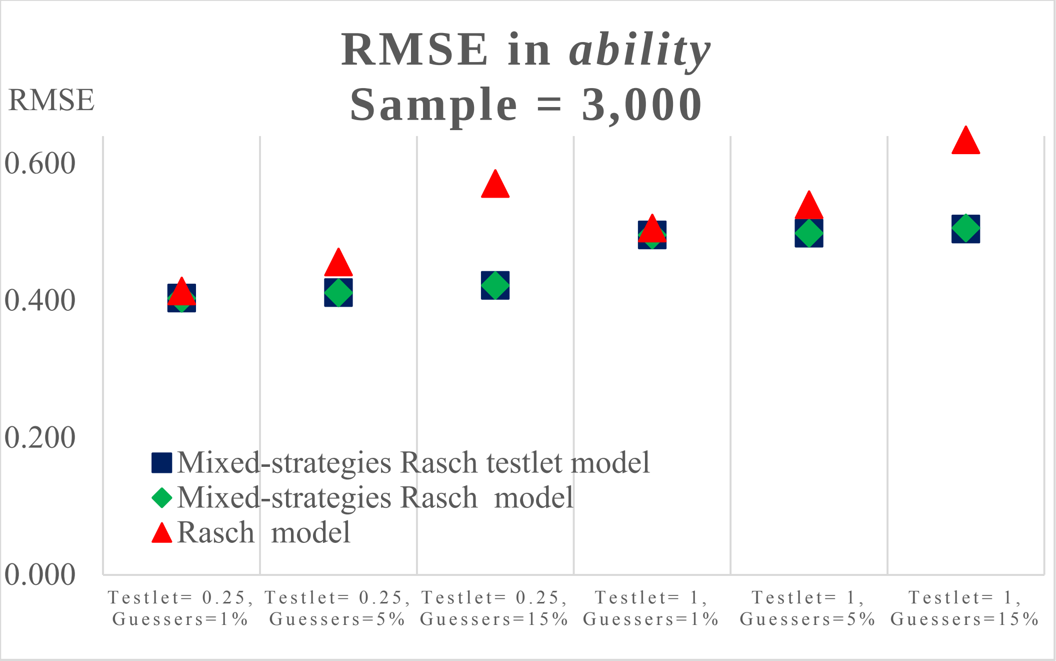

Figure 14. Plot of RMSE of Ability Parameter Estimates at a Sample = 3,000.

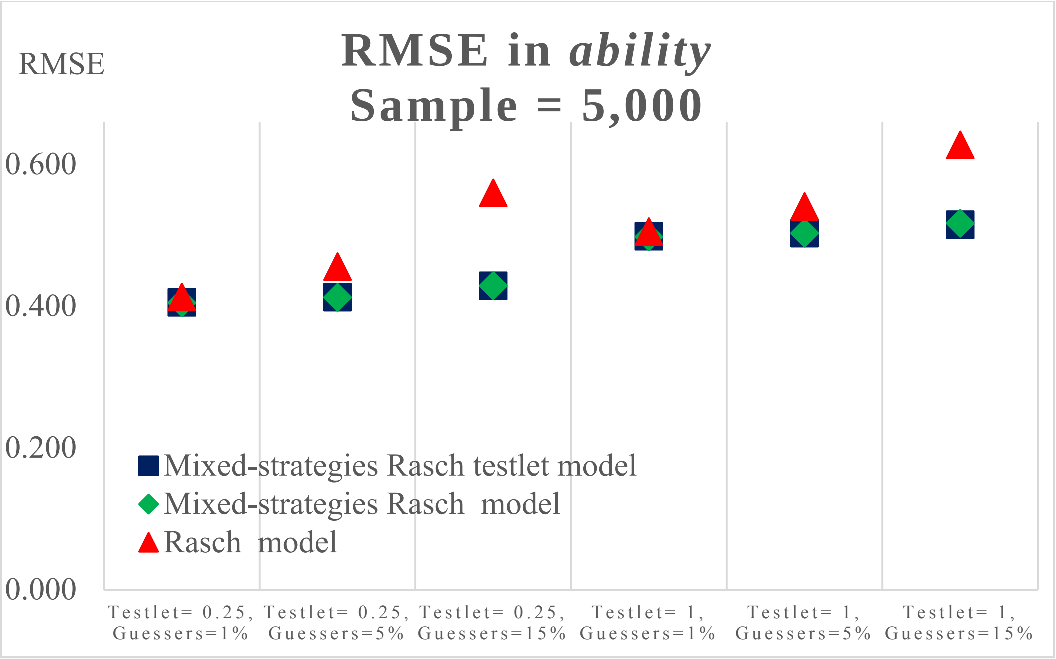

Figure 15. Plot of RMSE of Ability Parameter Estimates at a Sample = 5,000.

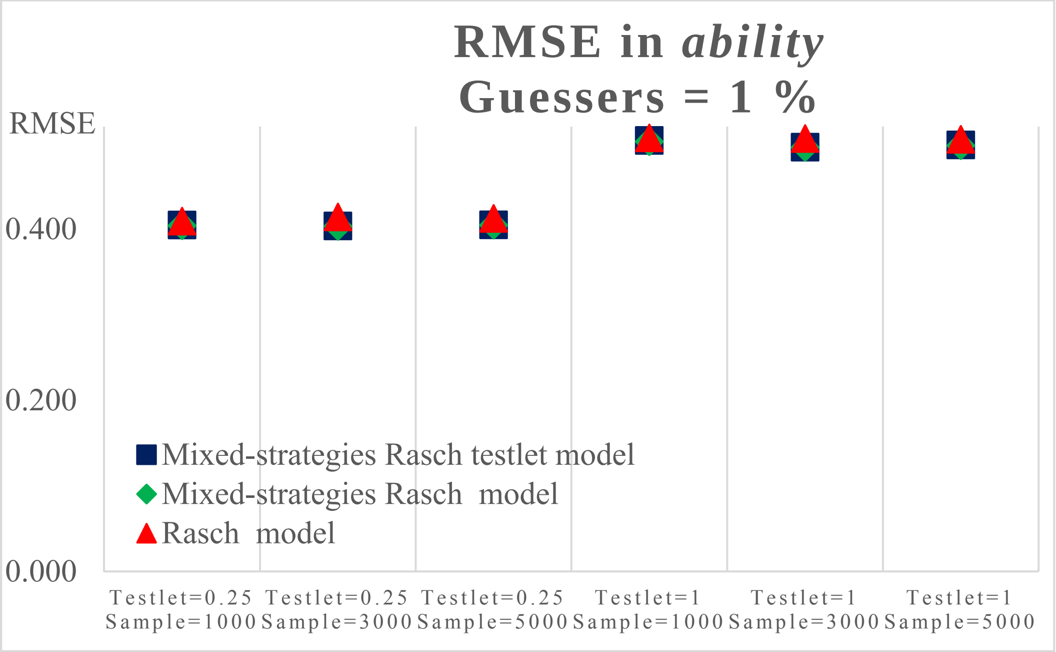

Figure 16. Plot of RMSE of Ability Parameter Estimates at Guessers = 1%.

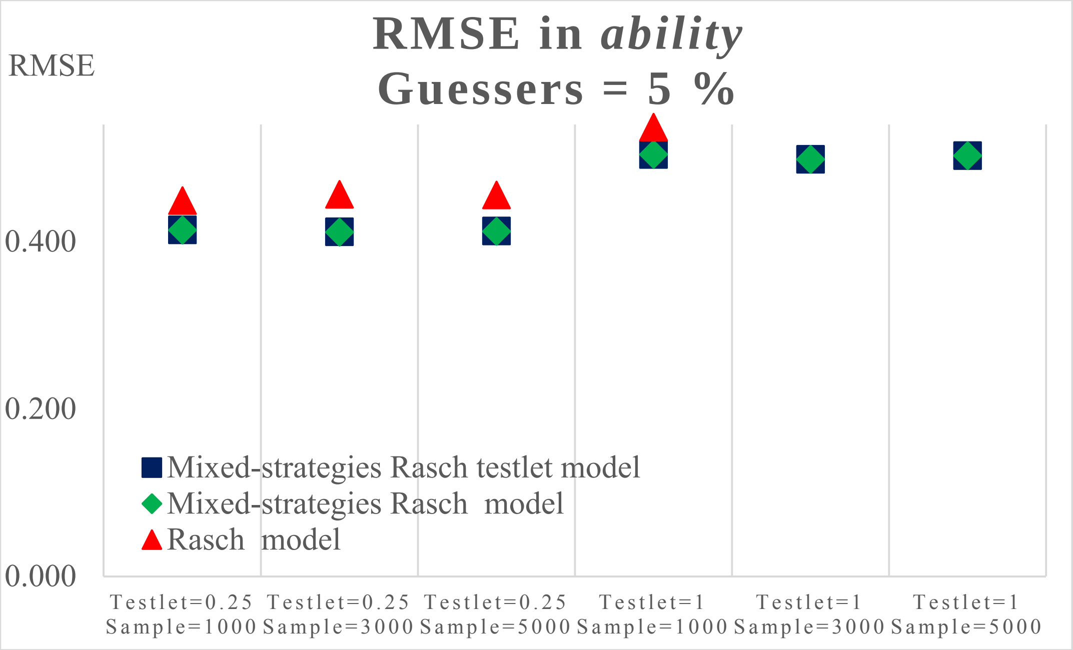

Figure 17. Plot of RMSE of Ability Parameter Estimates at Guessers = 5%.

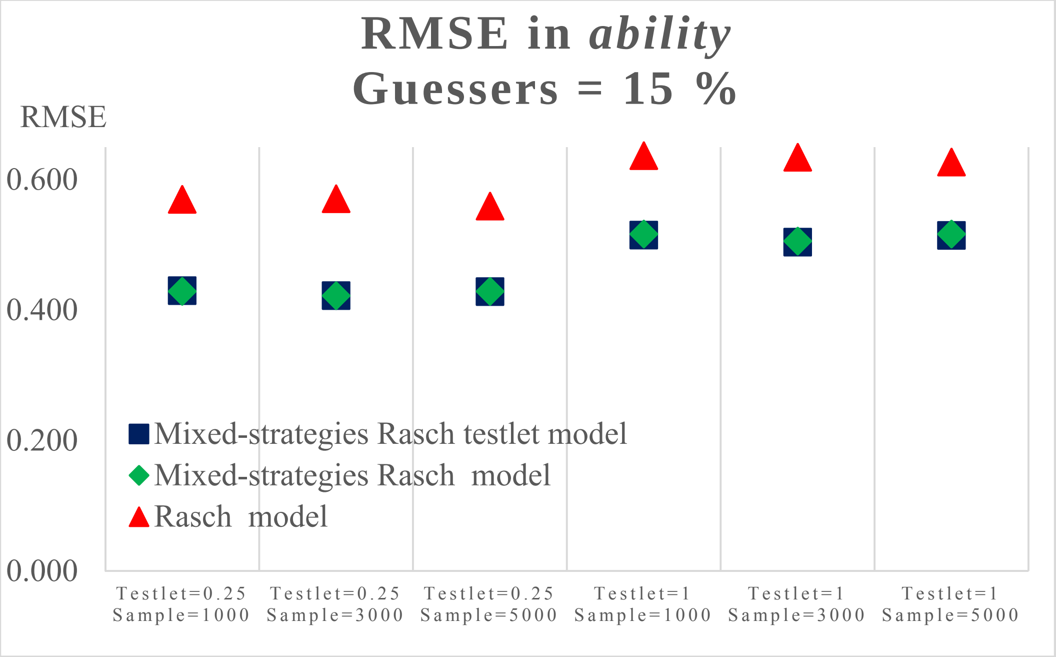

Figure 18. Plot of RMSE of Ability Parameter Estimates at Guessers = 15%.

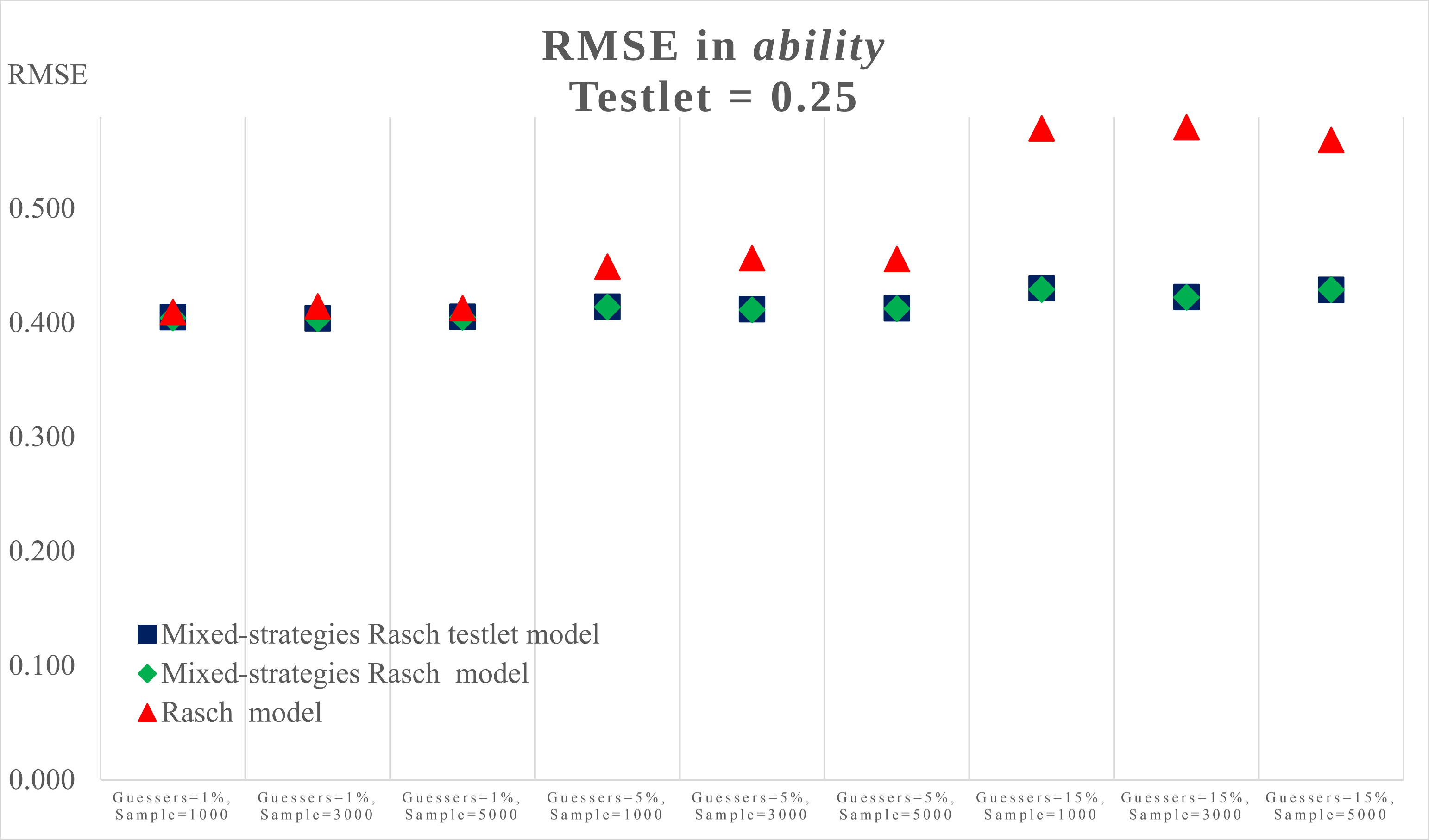

Figure 19. Plot of RMSE of Ability Parameter Estimates at a Testlet Variance = 0.25.

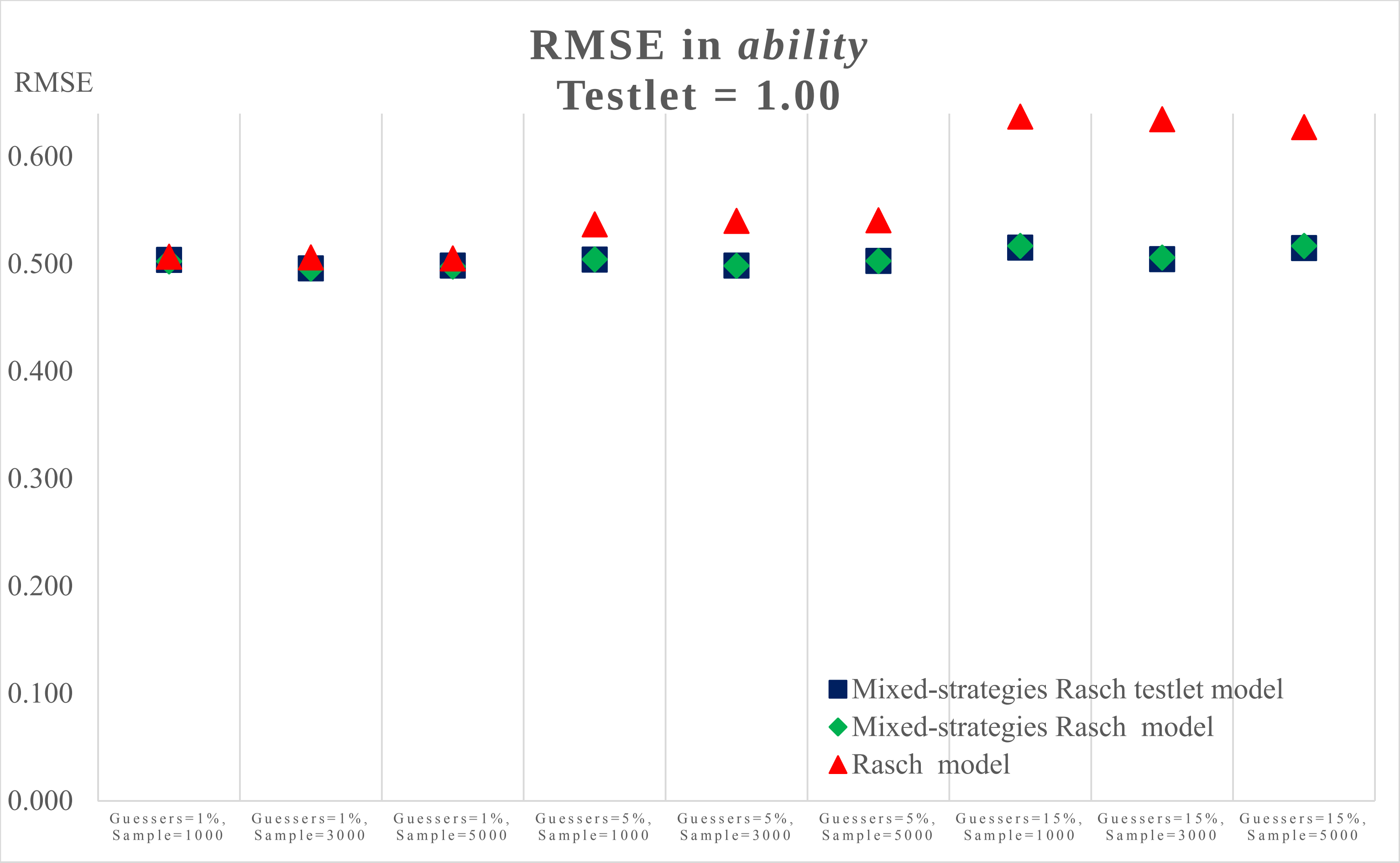

Figure 20. Plot of RMSE of Ability Parameter Estimates at a Testlet Variance = 1.00.

Recovery of ability and testlet variances. The estimated ability variances and testlet variances are summarized in Table 12. The estimates of ability variances were comparable among three estimation models across testing conditions. The ability variances were adequately recovered. The mixed-strategies Rasch testlet model also well recovered the testlet variances. The slight differences may be due to sample variations.

Table 12

Recovery of Ability and Testlet Variances

	Testlet Variance	0.25								
	Sample	1,000			3,000			5,000		
	Guessers %	1%	5%	15%	1%	5%	15%	1%	5%	15%
Rasch model	Ability	0.999	1.024	1.051	0.950	0.977	0.979	0.998	1.014	1.048
Mixed-strategies Rasch model	Ability	0.993	0.995	1.003	0.945	0.950	0.920	0.992	0.990	1.002
Mixed-strategies Rasch testlet model	Ability	1.048	1.049	1.060	0.992	0.998	0.965	1.043	1.040	1.054
	Testlet 1	0.294	0.283	0.306	0.261	0.277	0.263	0.261	0.263	0.262
	Testlet 2	0.279	0.282	0.318	0.270	0.265	0.265	0.266	0.260	0.272
	Testlet 3	0.297	0.295	0.290	0.255	0.273	0.273	0.260	0.265	0.254
	Testlet 4	0.289	0.286	0.310	0.270	0.269	0.270	0.263	0.255	0.268
	Testlet 5	0.298	0.315	0.301	0.269	0.268	0.265	0.253	0.256	0.261
	Testlet 6	0.314	0.294	0.288	0.259	0.273	0.264	0.261	0.267	0.252

Table 12

Recovery of Ability and Testlet Variances (continued)

	Testlet Variance					1.00				
	Sample		1,000			3,000			5,000	
	Guessers %	1%	5%	15%	1%	5%	15%	1%	5%	15%
Rasch model	Ability	0.918	0.922	0.969	0.851	0.888	0.904	0.889	0.917	0.962
Mixed-strategies Rasch model	Ability	0.910	0.889	0.896	0.845	0.857	0.830	0.884	0.887	0.900
Mixed-strategies Rasch testlet model	Ability	1.062	1.036	1.047	0.982	0.997	0.961	1.034	1.038	1.054
	Testlet 1	0.975	1.010	0.995	0.995	1.016	1.001	1.020	0.997	1.008
	Testlet 2	0.954	1.000	1.042	0.979	1.011	0.986	1.004	1.017	0.982
	Testlet 3	0.953	0.992	1.007	1.014	1.018	1.006	0.996	0.995	1.000
	Testlet 4	0.950	0.988	0.985	0.998	0.992	0.996	0.995	0.998	0.989
	Testlet 5	1.018	0.989	0.977	0.991	0.993	1.008	0.988	1.016	1.009
	Testlet 6	1.004	0.936	0.963	1.001	0.975	0.994	0.993	0.990	0.982

Chapter 5: Results of Empirical Study

This chapter presents the results of the empirical study. Section 5.1 presents the model selection and the descriptive statistics of model parameters; the analysis is designed to answer the research question 3. Section 5.2 explores the empirical factors that are potentially associated with the heterogeneity of test-taking motivation; a logistic regression analysis is conduced and summarized.

5.1 Real Data Application

The real dataset is fitted by the proposed model and the two comparison models. In case the real dataset may require more runs, the function "thin" in WinBUGS is used to reduce computer storage space before convergence is achieved; for example, the actual number of iterations carried out is 40,000 when 4,000 iterations are stored after thinning (thin = 10). When convergence is obtained, post-burn-in 5,000 iterations (without thinning) are run for drawing inferences (i.e., two Markov chains result in 10,000 iterations for inferences). Figure 21 shows example plots for convergence assessment. Generally, 15,000 iterations are needed for Markov chains to achieve convergence. The MC errors of the model parameters were less than 0.05, indicating convergence.

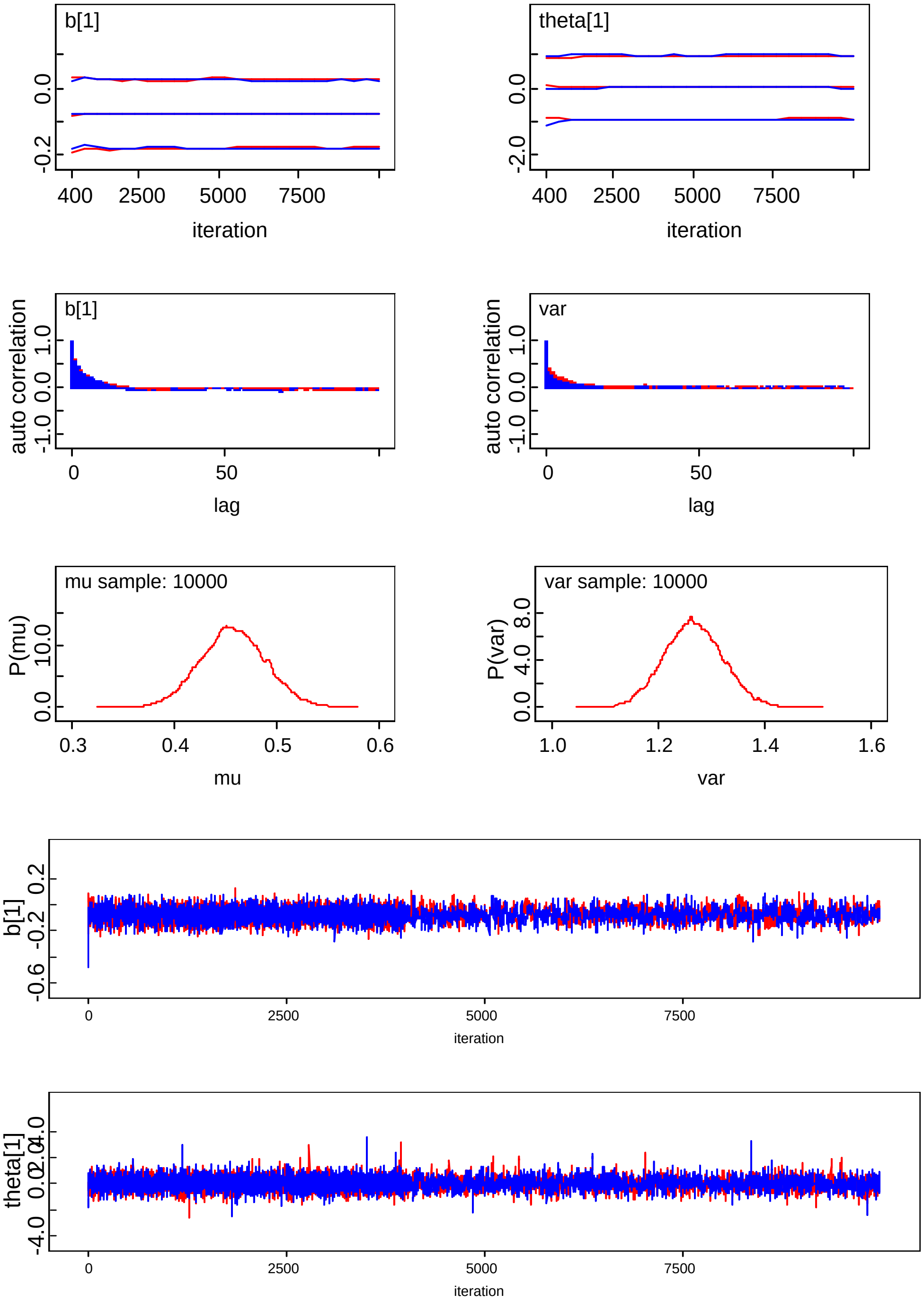

Figure 21. Examples of Plots of Convergence Assessment.

Results of model-fit statistics are shown in Table 13. All mode-fit indices pointed to the mixed-strategies Rasch testlet model as the best fitting model and the Rasch model as the worst. The results of model selection in the real data application agree with the findings in the simulation study: the model-fit statistics exhibited consistent and effective performance.

Table 13

Summary of Model Selection

	AIC	AIC_C	BIC	SABIC
Mixed-strategies Rasch testlet model	**51190**	**51200**	**51370**	**51270**
Mixed-strategies Rasch model	53000	53000	53130	53060
Rasch model	53150	53150	53280	53210

In the real dataset, of the 2,327 test takers, 46 (around 2.0%) were classified into the unmotivated class. Put it differently, the item response patterns of these 46 test takers were best characterized by the random guessing strategy model rather than by the item response theory model. Tables 14 and 15 present the descriptive statistics of the item and ability parameters, as well as the estimates of testlet variances in the mixed-strategies Rasch testlet model. The estimates of the item difficulty and ability parameters were of medium range (*b*: –1.129 to 2.314; *θ:* –2.244 to 2.817). The testlet variances ranged from 0.137 to 0.448, indicating the existence of item clusters in the real dataset.

Table 14

Descriptive Statistics of Item and Ability Parameters

Parameter	Minimum	Mean	Maximum	Standard Deviation
b	–1.129	0.000	2.314	0.814
θ	–2.244	0.458	2.817	0.934

Table 15

Estimates of Testlet Variances

Testlet Variance #	Estimate
Testlet variance 1	0.448
Testlet variance 2	0.440
Testlet variance 3	0.145
Testlet variance 4	0.301
Testlet variance 5	0.235
Testlet variance 6	0.137
Testlet variance 7	0.241

5.2 Potential Factors that Characterize Test-Taking Motivation Heterogeneity

To answer the research question 4, several variables are empirically explored in characterizing the latent classes of test-taking motivation heterogeneity. The logistic regression model with nine predictors and one binary dependent variable is fitted to the dataset. Unlike linear regression, the logistic regression does not require the linear, normality, and homogeneity assumptions (Lomax, 2007); yet the logistic regression requires that only minimal liner dependency occurs among a set of predictors. Prior to model fitting, the correlations among predictors are examined, and results indicate that all correlation coefficients were not exceeding |0.7|—a criterion suggested by Pallant (2007).

The Hosmer and Lemeshow test measures goodness of fit: a significant χ^2 indicates poor model fit. In this real data example, the logistic regression model passed the Hosmer and Lemeshow test, suggesting that the model fit the data well, $\chi^2 = 7.337$, $df = 8$, $p = .501$. The fitted logistic regression model is expressed by

$$\begin{aligned}\ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) &= 1.544 + 0.169(Gender) - 1.101(Language) - 0.013(ScienceProficiency) \\ &+0.318(ESCS) - 0.117(Enjoyment) - 0.086(Interest) - 0.034(SelfEfficacy) \\ &-0.014(SelfConcept) + 0.268(Motivation).\end{aligned} \tag{20}$$

The omnibus test of model coefficients was statistically significant, $\chi^2 = 58.572$, $df = 9$, $p = .000$, which implies that the subset of predictors jointly contributed to the heterogeneity of test-taking motivation. The Wald statistics in Table 16 indicate that all of the predictors, except for science proficiency, were unnecessary in the model; only science proficiency explained heterogeneity to a significant degree, Wald = 39.357, $p = .000$. The predicted probability of being an unmotivated respondent was higher at low scores of science proficiency ($\hat{\beta} = -0.013$). One unit decreased in science proficiency increased the predicted log odds by 0.013, holding all else constant. Put it differently, assuming that all else remained constant, for each unit decreased in science proficiency, there was a 1.3% increase in the odds of being an unmotivated respondent.

Table 16

Parameter Estimates for the Logistic Regression Model

	B	S.E.	Wald	*df*	Sig.	Exp(B)
Gender	0.169	0.346	0.239	1	0.625	1.184
Language	-1.101	1.033	1.138	1	0.286	0.332
Science proficiency	**-0.013**	**0.002**	**39.356**	**1**	**0.000**	**0.987**
ESCS	0.318	0.169	3.551	1	0.060	1.375
Enjoyment of science	-0.117	0.256	0.210	1	0.647	0.889
Interest of science	-0.086	0.195	0.196	1	0.658	0.917
Self-efficacy of science	-0.034	0.202	0.029	1	0.865	0.966
Self-concept of science	-0.014	0.236	0.004	1	0.952	0.986
Motivation to learn science	0.268	0.251	1.138	1	0.286	1.307
Constant	1.544	0.844	3.347	1	0.067	4.684

Chapter 6: Conclusion and Discussion

This chapter summarizes and interprets the findings of the simulation study and empirical application. It discusses how the research questions are answered as well as how the simulated design characteristics perform in terms of model parameter recovery. Implications, limitations, and directions for future research are also presented.

6.1 Summary and Discussion of Findings—Simulation Study

To answer the research question 1—the effect of overlooking test-taking motivation heterogeneity and testlet effects in low-stakes assessments—a variety of testing conditions are manipulated, and the performances of the proposed and comparison models are compared in a simulation study. Results show that neglecting test-taking motivation heterogeneity and testlet effects adversely affected model–data fit and model parameter estimation. The existence of these phenomena in data cannot be disregarded because these effects will, in turn, result in inaccurate targeted inferences.

To answer the research question 2—how well model parameters are recovered under test-taking motivation heterogeneity and testlet effects in low-stakes assessments—model–data fit, classification accuracy, and model parameter recovery are examined under simulated testing conditions. The results promisingly demonstrate the effectiveness of the proposed model and suggest its utility for low-stakes testlet-based assessments. More specifically, test-taking motivation heterogeneity and testlet effects were well controlled by the proposed model. The mixed-strategies Rasch testlet model outperformed the two comparison models, exhibiting superior model–data fit and satisfactory classification accuracy. The good fit between model and data indicates that observed item response patterns were in accordance with the expected item response patterns implied by the class-specific models. The high classification accuracy of the mixed-strategies Rasch testlet model is evidenced that test takers were responding in a manner (i.e., solution strategy or random guessing strategy) highly consistent with the hypothesized latent classes (i.e., a class of IRT model or a class of random guessing function). It is implied that the proposed model is capable of classifying examinees in terms of heterogeneous test-taking motivation. The mixed-strategies Rasch testlet model also demonstrated improved measures of model parameter estimates, making test results more accurate and reliable.

A close look at the item parameter recovery shows that the accuracy of the item parameter estimates was influenced by estimation model, magnitude of testlet effects, percentage of unmotivated respondents, and sample size. In the mixed-strategies Rasch testlet model, in which both test-taking motivation heterogeneity and testlet effects are accounted for, the item difficulty parameter estimates were very close to true values across testing conditions. In the mixed-strategies Rasch model, wherein only test-taking motivation heterogeneity is described, the accuracy of the item difficulty parameter estimates tended to decrease as the magnitude of testlet effects increased. Finally, in the Rasch model, which does not characterize test-taking motivation heterogeneity and testlet effects, the discrepancy between estimated and simulated parameters markedly expanded as testlet effects and/or the percentage of unmotivated test takers increased. Item parameter recovery generally improved as sample size increased, regardless of which model was fitted to the data. These findings have two implications. First, overlooking test-taking motivation heterogeneity and testlet effects exerted considerable influence on the precision of item parameter estimates, particularly when testlet effects and/or percentage of unmotivated respondents increased. This implication suggests the use of the proposed model given the need to carefully manage test-taking motivation heterogeneity and testlet effects. Second, item parameter recovery via the proposed model will benefit from a larger sample size, but a large sample size is not necessary. In the proposed model, item parameters recovered to a fairly satisfactory degree even under a small sample size (i.e., 1,000).

In the analyses of ability parameter recovery, the precision of ability parameter estimates differed in testlet effects, estimation model, and percentage of unmotivated respondents. The Rasch model generated the worst ability parameter recovery among three estimation models, especially when high testlet effects and/or numerous unmotivated respondents (i.e., 15%) were present in the data. The mixed-strategies Rasch testlet model and the mixed-strategies Rasch model offered fairly comparable recovery of ability parameters. Therefore, if one solely focuses on drawing inferences on the basis of ability parameters, no distinct difference is expected between the mixed-strategies Rasch testlet model and the mixed-strategies Rasch model. Based on the findings from this study, it is recommended the use of the mixed-strategies Rasch testlet model over the mixed-strategies Rasch model for two reasons: (1) the proposed model provided superior model–data fit and (2) facilitated the assessment of the existence and the magnitude of testlet effects, a task that the Rasch model or the mixed-strategies Rasch model is not able to accomplish.

In addition to the above-mentioned findings, two more findings were worthy of note in the simulation study. First, label switching of latent class did not occur in the proposed model. Mixture IRT and mixture Rasch modeling approaches frequently suffer from this problem (e.g., Cho et al., 2013; Dai, 2009; Li et al., 2009; Jiao et al., 2009). This drawback is attributed to the fact that in these mixture models, respondents from different latent classes correspond to the same form of item response functions. For example, the mixture Rasch model (Rost, 1990) involves at least two Rasch models, in which no constrains are deliberately imposed on a certain class (e.g., item difficulties are high for one class and low for other classes); therefore, the labels of latent classes can be switched across iterations within a single Markov chain or across replications. In the proposed model, test takers from distinct latent classes are characterized by either the random guessing function or the Rasch testlet model exclusively; thus, sufficient information is available for effectively classifying test takers into latent classes. The simulation results verify this hypothesis that the proposed model does not suffer from latent class label switching. The second noteworthy finding is that the model selection indices considered in this research worked equivalently well in classifying test takers into (true) simulated latent classes. The 100 % accuracy of model selection indicates that these model-fit indices effectively and consistently functioned across varied testing conditions and replications. On this basis, the four statistics examined—AIC, $\mathrm{AIC_C}$, BIC, and SABIC—are tremendously useful to researchers who apply the mixed-strategies Rasch testlet model. This study also recommends that researchers use accumulated evidence rather than a single index in determining model–data fit.

6.2 Summary and Discussion of Findings—Empirical Study

The proposed model allows for distinct subgroups of test-taking motivation be to modeled with different measurement functions. The empirical study is included to answer the research questions 3 (How does the proposed model perform in real low-stakes assessment data in terms of model–data fit? Are there unmotivated test takers and testlet effects identified?) and 4 (What are the potential factors that characterize heterogeneous test-taking motivation from empirical low-stakes assessment data?)

As stated in Chapter 3, the sample item response dataset extracted from the PISA 2006 science assessment is constructed under the Rasch model. Given that the PISA assessment is low stakes in nature (i.e., some unmotivated respondents are expected) and that the assessment is primarily comprised of testlet-based items (i.e., testlet effects), this study therefore hypothesizes that the proposed model will satisfactorily fit the item response dataset.

Three models are fitted to the real dataset: the Rasch model (i.e., the original calibration model for the dataset), the mixed-strategies Rasch model (i.e., the Rasch model that incorporates the heterogeneity of test-taking motivation), and the mixed-strategies Rasch testlet model (i.e., the Rasch model that manages testlet effects and test-taking motivation heterogeneity). All the model-fit indices exhibited preference for the mixed-strategies Rasch testlet model, suggesting that the real low-stakes assessment dataset was best fitted by the proposed model. Results of model–data fit verify the need for a more sophisticated model that integrates the IRT model and random guessing strategy model for low-stakes assessments. The consistency of model selection among indices echoes that indicated by the results of the simulation study.

Both test-taking motivation heterogeneity and testlet effects were identified in the extracted real dataset. A total of 2.0 % of test takers were classified under the unmotivated test-taking group, which appeared at a lower rate than has been reported in previous empirical studies (e.g., Brown & Gaxiola, 2010; Subedi, 2009; Sundre & Wise, 2003; Wise & DeMars, 2005, 2006). A possible explanation is that examinees who attended the PISA 2006 science assessment had high perceived value/expectancy of the test, thereby taking the test more seriously. The magnitudes of testlet variance were not negligible in the real dataset at a range of 0.137 to 0.448. These findings therefore support the study hypothesis on the occurrence of test-taking motivation heterogeneity and testlet effects in low-stakes assessments.

The follow-up exploratory study is intended to yield empirical evidence that demystifies the composition of unobservable latent class membership for whom test-taking motivation is distinct. Several variables are selected to characterize the heterogeneity of test-taking motivation in a logistic regression model. The findings from this empirical study are expected to help educators and practitioners identify potential sources that are associated with heterogeneity in real-world situations. Among the set of variables, science proficiency explained the heterogeneity of test-taking motivation at a statistically significant level. Low science proficiency was associated with a high likelihood of being an unmotivated respondent. More specifically, test takers with low science proficiency were more likely to have random guesses in the science assessment—a finding that agrees with Barry et al. (2010), Petridou and Williams (2007), and Wise et al. (2009). An important note is that this investigation is a methodological demonstration of exploring the potential characterizations of latent class membership after the proposed model identifies latent classes. This study has no intention to draw a definite conclusion regarding respondent motivation in taking the PISA science assessment. Furthermore, the findings from the follow-up exploratory study are based only on a single sample dataset and no strong evidence is derived as to cause–effect relationship. Identifying conclusive sources of test-taking motivation heterogeneity in real data necessitates future research and powerful support from both educational psychology theories and related empirical studies.

6.3 Limitations and Future Research Directions

Similar to the findings of any other studies, the interpretations in this study should be limited to the conditions considered. Several limitations and recommendations for future directions are addressed here.

Number of replications. The selection of the number of replications in this research is limited by practical considerations; that is, the heavy computational demand in MCMC estimation. Given that the importance of the number of replications in a simulation study is akin to that of the number of participants in an empirical study (Harwell et al., 1996), the small number of replications in this research could raise concerns on the generalizability of the findings. In addition, the number of replications could influence the sampling variance of the parameter estimates and the power with which effects are detected in a simulation study (Harwell et al., 1996). Ideally, researchers should perform as many replications as possible to ensure estimation precision. Such an approach will afford researchers more confidence in statistical inferences. This advantage particularly holds for more complex models, in which a higher number of parameters are estimated or convergence problems are more likely to occur. With the rapid growth of computer technology, the time required to run MCMC estimation can be substantially diminished in the near future. A larger number of replications or the inclusion of more simulation design factors in a simulation study will therefore be achievable.

Number of item characteristics. This study proposes the mixed-strategies Rasch testlet model for low-stakes assessments. In this model, the items are preselected on the basis of the one-parameter Rasch model. Thus, the proposed model is currently inapplicable to items that are calibrated under the 2PL model or 3PL model. Nevertheless, the promising findings obtained in this research permit the extension of the proposed model to low-stakes assessments that are calibrated under the 2PL model or 3PL model (i.e., a mixed-strategies two-parameter testlet model or a mixed-strategies three-parameter testlet model).

Type of response data. This research is interested in low-stakes dichotomously scored multiple-choice items, which means that the proposed model does not incorporate polytomously scored (e.g., a scoring range of 0 [no credit] to 3 [3 points]) or Likert scaled items (e.g., a scoring range of 1 [strongly disagree] to 4 [strongly agree]). Random guessing response patterns due to lack of test-taking motivation are expected in polytomously scored or Likert scaled items, making the extension of the idea in this research to a test with polytomously scored items a favorable endeavor.

Item format. This study investigates tests with fixed item format and testlet length. More specifically, all items in a test are testlet-based items, and the testlet lengths are constant across testlet units. In practice, some low-stakes assessments may contain a few items that are written individually rather than built upon testlets (e.g., Wainer & Wang, 2000; Wang & Wilson, 2005) and have varying testlet lengths in a test. Researchers can further extend the proposed model to a test with a mixture of testlet-based items and individual items by specifying zero testlet effect for individual items, as well as to a test with varying testlet lengths.

Null condition. This research includes a null model—the Rasch model—to address the effect of overlooking test-taking motivation heterogeneity and testlet effects upon model-data fit and model parameter recovery. This study, however, does not include “nothing to detect” conditions in the simulation study;

i.e., an absence of guessers or testlet effects. Therefore, little is known about the capability of the proposed model in recognizing an absence of guessers and testlet effects of data, as well as in determining how outcome statistics vary from baseline to simulated levels of targeted factors. Future studies should include null conditions of simulation factors to enable the comprehensive understanding of a targeted model's effectiveness.

Test-taking motivation. In this study, the heterogeneity of test-taking motivation stems from the low-stakes test results for individual test takers. Given no consequential effect on an individual test taker's academic records in low-stakes assessments, a proportion of test takers are assumed unmotivated to exert effort in taking a test and simply apply the random guessing strategy. On this basis, probability-based item response functions are created to represent the likelihood that a test taker in a given class will respond to items in a given manner—that is, adopt the solution strategy or random guessing strategy. Admittedly, test-taking motivation is a highly complicated psychological process that possibly drives the use of other test-taking strategies excluded in the current study. Furthermore, random guessing response patterns could result from other testing conditions. For future researchers, an interesting direction would be to investigate other types of heterogeneity in test-taking motivation or other patterns of random guessing responses under a particular testing scenario.

Predictors that characterize test-taking motivation heterogeneity. The empirical study in this dissertation provides an illustration of investigating whether science domain-specific predictors (e.g., enjoyment of science, interest of science) and latent classes are meaningfully related in the low-stakes science assessment. Such exploration is also worthwhile for other content domains, including mathematics and reading literacy, because they facilitate the empirical interpretation of domain-specific heterogeneous test-taking motivation. Furthermore, the random guessing strategy that test takers apply to items may vary as people age; e.g., the probability of being a random guesser in low-stakes assessments may be lower for younger test takers. An interesting initiative in educational psychology is the investigation of item response strategies across ages through longitudinal or cross-sectional studies. Such explorations can elicit useful findings on training programs or instructional courses that help test takers employ targeted response strategies in low-stakes assessment scenarios. In addition to person characteristics, item-related covariates (e.g., the type/level of skills required for solving an item) as well as the manner by which such features interact with person characteristics may be associated with test-taking motivation heterogeneity. Including item-related characteristics entails evaluations of cognitive levels (i.e., remembering, understanding, applying, analyzing, evaluating, & creating; in Bloom's revised taxonomy, Anderson & Krathwohl,

2001) or content areas assessed by the items. Therefore, the participation of domain-specific context experts is required.

6.4 Implications and Conclusion

IRT models implicitly assume that test takers are motivated to demonstrate proficiency and that no construct-irrelevant variances in item characteristics occur during assessments. In low-stakes assessments, some test takers do not put forth effort in performing well and testlet effects may be present as well, thereby limiting the utility of currently used IRT models. Measurement practitioners and professionals, therefore, needs to be cautious about such noise when developing, estimating, and interpreting test results from low-stakes assessments. They are responsible for maintaining the quality of estimation and the integrity of educational assessments.

In this research, the effects of overlooking test-taking motivation heterogeneity and testlet effects have been demonstrated to show negative influence on parameter estimation quality. The findings of this dissertation serve as substantive evidence of how construct-irrelevant variances adversely affect the precision of parameter estimates. This study highlights the psychological importance of the manner by which test takers heterogeneously respond to low-stakes assessments—the cognitive process underlying item responses. The proposed model is an evolution of a psychometric model combined with psychology. It enables the simultaneous modeling of test-taking motivation heterogeneity and testlet effects in low-stakes assessments. With the promising performance of the proposed model (good model–data fit, satisfactory classification accuracy, and well-recovered model parameters), assessment practitioners and professionals can confidently use it to improve estimation quality for low-stakes assessments. In addition to model estimation, another function of the proposed model is that it serves as a psychometric filtering tool for test-taking motivation and testlet effects; such filtering is based on item response patterns—an attribute that is truly helpful in verifying whether the studied datasets exhibit the examinee homogeneity and local item independence that are assumed in IRT models. Finally, the illustration in the empirical study serves as an example of explaining latent class membership. The incorporation of external variables in the follow-up exploratory study enables more practical and meaningful interpretations of the latent classes of test-taking motivation. To sum up, this dissertation provides empirical evidence related to the impact of test-taking motivation heterogeneity on model parameter estimation in testlet-based assessments. The findings of this study are anticipated to inspire more investigations into low-stakes assessments, on which educational policy and implications heavily rely.

Appendix A: Simulated Item Parameters

Item ID	Difficulty	Item ID	Difficulty
Item 1	0.137	Item 19	-0.341
Item 2	0.989	Item 20	-0.567
Item 3	1.257	Item 21	0.223
Item 4	-2.14	Item 22	-0.99
Item 5	-0.086	Item 23	-0.317
Item 6	0.823	Item 24	-0.273
Item 7	-0.968	Item 25	-0.047
Item 8	1.04	Item 26	0.075
Item 9	-1.677	Item 27	-0.779
Item 10	-0.711	Item 28	1.555
Item 11	-0.99	Item 29	0.176
Item 12	-0.558	Item 30	0.372
Item 13	0.012	Item 31	-2.029
Item 14	-0.262	Item 32	2.307
Item 15	-1.066	Item 33	0.474
Item 16	0.893	Item 34	-0.181
Item 17	-0.223	Item 35	-1.524
Item 18	0.936	Item 36	-0.944

Appendix B: WinBugs Codes for the Proposed Model

```
# J: the number of persons
# I: the number of items
# G: the label of latent classes
# b: item difficulty
# theta: ability

model
{
for (j in 1:J) {
for (i in 1:6)  {
p[j,i] <- (2-G[j])*1/(1+exp(1.0986))+(G[j]-1)*1/(1+exp(-(theta[j]+gam1[j]-b[i])))
       r[j,i] ~ dbern(p[j,i])
       }
for (i in 7:12)  {
p[j,i] <- (2-G[j])*1/(1+exp(1.0986))+(G[j]-1)*1/(1+exp(-(theta[j]+gam2[j]-b[i])))
       r[j,i] ~ dbern(p[j,i])
       }
for (i in 13:18)  {
p[j,i] <- (2-G[j])*1/(1+exp(1.0986))+(G[j]-1)*1/(1+exp(-(theta[j]+gam3[j]-b[i])))
       r[j,i] ~ dbern(p[j,i])
       }
for (i in 19:24)  {
p[j,i] <- (2-G[j])*1/(1+exp(1.0986))+(G[j]-1)*1/(1+exp(-(theta[j]+gam4[j]-b[i])))
       r[j,i] ~ dbern(p[j,i])
       }
for (i in 25:30)  {
p[j,i] <- (2-G[j])*1/(1+exp(1.0986))+(G[j]-1)*1/(1+exp(-(theta[j]+gam5[j]-b[i])))
       r[j,i] ~ dbern(p[j,i])
       }
for (i in 31:36)  {
p[j,i] <- (2-G[j])*1/(1+exp(1.0986))+(G[j]-1)*1/(1+exp(-(theta[j]+gam6[j]-b[i])))
       r[j,i] ~ dbern(p[j,i])
       }
gam1[j] ~ dnorm(0,taut1)
gam2[j] ~ dnorm(0,taut2)
gam3[j] ~ dnorm(0,taut3)
gam4[j] ~ dnorm(0,taut4)
gam5[j] ~ dnorm(0,taut5)
gam6[j] ~ dnorm(0,taut6)
G[j]~dcat(PI[])
pg[j]<- equals(G[j],1)
}

#priors
PI[1:2] ~ ddirich(alpha[])
mu ~ dnorm(0,1)
tau ~ dgamma(1,1)
var <- 1/tau
taut1 ~ dgamma(1,1)
taut2 ~ dgamma(1,1)
taut3 ~ dgamma(1,1)
taut4 ~ dgamma(1,1)
taut5 ~ dgamma(1,1)
taut6 ~ dgamma(1,1)
vart[1] <- 1/taut1
vart[2] <- 1/taut2
vart[3] <- 1/taut3
vart[4] <- 1/taut4
vart[5] <- 1/taut5
vart[6] <- 1/taut6

for (i in 1:I-1) {
   b[i] ~ dnorm(0, 1)}
b[I]<- -1*sum(b[1:I-1])
for (j in 1:J) {
   theta[j]~ dnorm(mu,tau)}

# Log Likelihood
for (j in 1:J) {
for (i in 1:I) {
lik[j,i]<-   log(p[j,i])*r[j,i]+log(1-p[j,i])*(1-r[j,i])}}
loglik <-  sum(lik[1:J,1:I])
AIC <-  -2*(loglik - np)
BIC <-  -2*loglik + np*log(J)
SABIC<- -2*loglik +np*log((J+2)/24)
AICC<-AIC+2*np*(np+1)/(J-np-1)
}
```

Appendix C: Simulation Procedure

What follows is a description of the simulation steps for data generation in MATLAB.

1. Generate 36 random beta values from a standard normal distribution.
2. Generate 1000 random theta values from a standard normal distribution.
3. Generate 3000 random theta values from a standard normal distribution.
4. Generate 5000 random theta values from a standard normal distribution.
5. Generate the first set of testing conditions: sample size = 1000; testlet variance = 0.25; guessers = 1%. For each testlet unit (testlet 1: items 1–6; testlet 2: items 7–12; etc.):

   5.1 Repeat theta values 6 times and create a matrix of theta values with dimensions = (1000, 6)

   5.2 For each testlet unit, repeat beta values 1000 times and create a matrix of difficulty values with dimensions = (1000, 6).

   5.3 Generate 1000 random gamma values from a normal distribution with a mean of 0 and a standard deviation of 0.5.

   5.4 Repeat gamma values 6 times and create a matrix of gamma values with dimensions = (1000, 6).

   5.5 For each testlet unit, compute a matrix of probabilities by using Equation (8) with $g = 1$.

   5.6 When the data generation for 6 testlet units are completed, a matrix of probabilities is created with dimensions = (1000, 36).
6. Introduce random guessing responses into the first set of testing conditions.

   6.1 Create a temporary matrix of probabilities of guessing responses by using Equation (8) with $g = 0$ ($\tau = -1.0986$). The dimensions of the matrix = (1000, 36).

   6.2 For the selected number of guessers (number 991 to 1000), their original matrix of probabilities (created in step 5.6) is replaced with the matrix of probabilities created in step 6.1.

   6.3 Create a matrix of item response data on the basis of the matrix of probabilities (created in step 6.2) by using binomial distribution with number of trials = 1 (i.e., Bernoulli distribution).
7. The design comprises 25 replications. For each replication, repeat steps 5 and 6 and save item response datasets.
8. Repeat steps 5 to 7 for the remaining 17 sets of testing conditions by varying sample sizes, theta values, testlet variances, and percentages of guessers in an examinee population.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alexeev, N., Templin, J., & Cohen, A. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313–332.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objective.* New York, NY: Longman.

Armstrong, R. D., & Shi, M. (2009). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*, 391–410.

Bao, H. (2007). *Investigating differential item function amplification and cancellation in application of item response testlet models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3283409)

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342–363.

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

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

Brown, J. M., & Gaxiola, C. A. (2010). Why would they try? Motivation and motivating in low-stakes information skill testing. *Journal of Information Literacy*, *4*, 23–36.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach* (2nd ed.). Berlin, Germany: Springer-Verlag.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavior Statistics*, *22*, 265–289.

Chen, Y.-F., Jiao, H., & van Davier, M. (2013). *Comparison of different approaches to dealing with testlet effects in mixture item response theory modeling*. Presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*, 336–370.

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2013). Markov Chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278–306.

Cho, Y., Jiao, H., & Macready, G. B. (2012). *Assessing the effects of different item parameter profiles in mixture Rasch models*. Paper presented at the meeting of the American Educational Research Association, Vancouver, Canada.

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

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

Cole, J. E. (2007). *Motivation to do well on low-stakes tests* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3322685)

Congdon, P. (2003). *Applied Bayesian modeling.* New York, NY: John Wiley.

Cook, K. F, Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement*, *3*, 1–20.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429–449.

Curtis, S. M. (2010). BUGS codes for item response theory. *Journal of Statistical Software*, *36*. Retrieved from http://www.jstatsoft.org/

Dai, Y. (2009). *A mixture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carlo estimation* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 304919186)

Dayton, C. M. (1999). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

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

de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*, 159–177.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory to testlet-based tests. *Journal of Educational Measurement*, *43*, 145–168.

Dodeen, H. & Darabi, M. (2009). Person-fit relationship with four personality tests in mathematics. *Research Papers in Education*, *24*, 115–126.

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 255–270). New York, NY: Springer.

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

Emons, W. H. M. (2008). Nonparametric person–fit analysis of polytomous item scores*. Applied Psychological Measurement*, *32*, 224–247.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, *39*, 1–35.

Finch, H., & Pierson, E. (2011). A mixture IRT analysis of risky youth behavior. *Frontiers in Psychology*, *2*:98. doi:10.3389/fpsyg.2011.00098

Frederickx, S., Frederickx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1996). *Bayesian data analysis*. London, UK: Chapman & Hall.

German, S., & German, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Glas, A. S. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*, 217–233.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159–180.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Dordrecht, Netherlands: Kluwer.

Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: Principles and applications.* Norwell, MA: Kluwer Academic Publishers.

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

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125.

Hasting, W. K. (1970). Monte Carlo sampling method using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*, 82–100.

Jiao, H., van Davier, M., & Wang, S. (2010). *Marginal maximum likelihood of the Rasch mixture testlet model*. Paper presented at the meeting of the American Educational Research Association. Denver, CO.

Jiao, H., von Davier, M., Kamata, A., & Chen, Y.-F. (2011). *A multilevel Rasch mixture testlet model*. Paper presented at the meeting of the American Educational Research Association. New Orleans, LA.

Jiao, H., Wang, S., & He, W. (2013). Estimation method for one-parameter testlet models. *Journal of Educational Measurement*, *50*, 186–203.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*, 311–321.

Jiao, H., Wang, S., & Lu, R. (2009). *Mixture Rasch model for dichotomously scored testlet-based assessments*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298.

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

Kiefer, V. E., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimation in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, *27*, 887–906.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*, 38–51.

Lau, A. (2009). *Using mixture IRT model to improve parameter estimates when some examinees are amotivated* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3366561)

Li, D. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3359398)

Li, F. (2009). *An information correction method for testlet-based test analysis: From the perspective of item response theory and generalizability theory* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3391265)

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

Lomax, R. G. (2007). Statistical concepts: A second course. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS–A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

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

McCutcheon, A. L. (1987). *Latent class analysis*. Beverly Hills, CA: Sage.

Meijer, R. R., & Sijtsma K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

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

Muthén, L., & Muthén, B. (1998–2010). Mplus user’s guide (6th ed.) [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Napoli, A. R. & Raymond, L. A. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in graded and ungraded conditions. *Research in Higher Education*, *45*, 921–929.

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

O'Neil, H. F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low–stakes tests*.* *Educational Assessment*, *3*, 185–208.

OECD (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006.* Retrieved from http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/pisa2006-publications.htm

OECD (2007a). *Database–PISA 2006* [Data file and codebook]. Retrieved from http://pisa2006.acer.edu.au/downloads.php

OECD (2007b). *PISA 2006 science competencies for tomorrow's world volume 1: Analysis* (Report No. 978-92-64-04000-7). OECD.

OECD (2009). *PISA 2006 technical report* (Report No. 978-92-64-04808-9). OECD.

Pallant, J. (2007). *SPSS Survival Manual*. New York, NY: Open University Press.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement*, *44*, 227–247.

Pintrich, P. R., & Schunk, D. H. (1996). *Motivation in education: Theory, research, and applications*. Englewood Cliffs, NJ: Merrill.

Preinerstorfer, D., & Forman, A. K. (2012). Parameter recovery and model selection in mixed Rasch model. *British Journal of Mathematical and Statistical Psychology*, *65*, 252–262.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3175148)

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

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

Sinharay, S. (2003). *Assessing convergence of Markov Chain Monte Carlo algorithm: A review* (Report No. PR-03-07). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*, 461–488.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *2*8, 237–247.

Smith, E. V., Jr., Ying, Y., & Brown, S. W. (2012). Using the mixed Rasch model to analyze data from the beliefs and attitudes about memory survey. *Journal of Applied Measurement*, *13*, 23–40.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & von der Linde, A. (2002). Bayesian measures of model complexity and fit. *Royal Statistical Society*, *64*, 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn. D. (2003). WinBUGS 1.4 user's manual. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/

Subedi, D. R. (2009). *Investigating unobserved heterogeneity using item response theory mixture models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3381368)

Sundre, D. L. (2007). The student opinion scale (SOS): A measure of examinee motivation. Retrieved from http://www.jmu.edu/assessment/resources/resource_files/sos_manual.pdf

Sundre, D. L., & Wise, S. L. (2003). *Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the meeting of the National Council on Measurement Education, Chicago, IL.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*, 162–188.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. Hancock & K. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, NC: Information Age Publishing.

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

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203–220.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). The Hague, Netherlands: Kluwer-Nijhoff.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109–128.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1-17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measurement of examinee motivation in computerized-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wise, S. L., Pastor, D. A., & Kong, X. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185–205.

Yamamoto, K. (1987). *A model that combines IRT and latent class models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8803242).

Yamamoto, K. (1989). *HYBRID model of IRT and latent class model* (Report RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen, R. J. Mislevy, I. I. Bejar. (Eds.), *Test theory for a new generation of tests* (pp. 275–295). Hillsdale, NJ: Lawrence Erlbaum Associates.

Yan, J. W. (1997). *Examining local item dependence effects in a large-scale science assessment by a Rasch partial credit model*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Yang, C.-C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computer Statistics and Data Analysis*, *50*, 1090–1104.

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

Yen, W. M., & Fitzpatrick A. R. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (pp. 111–153). Westport, CT: Praeger Publishers.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7*, 168–190.