ABSTRACT

Title of Document: IMPACTS OF LOCAL ITEM DEPENDENCE OF TESTLET ITEMS WITH THE MULTISTAGE TESTS FOR PASS-FAIL DECISIONS

Ru Lu, Doctor of Philosophy, 2010

Directed By: Professor Hong Jiao
Department of Measurement, Statistics and Evaluation

The primary purpose of this study is to investigate the impact of the local item dependence (LID) of testlet items on the performance of the multistage tests (MST) that make pass/fail decisions. In this study, LID is simulated in testlet items. Testlet items are those that physically share the same stimulus. In the MST design, the proportion of testlet items is a manipulated factor. Other studied factors include testlet item position, LID magnitude, and test length. The second purpose of this study is to use a testlet response model to account for LID in the context of MSTs. The possible gains of using a testlet model against a standard IRT model are evaluated. The results indicate that under the simulated conditions, the testlet item position has a very minimal effect on the precision of ability estimation and decision accuracy, while the item pool structure (the proportion of testlet items), the LID magnitude and test length have fairly substantial effects. Ignoring the LID effects and fitting a unidimensional 3PL model result in the loss of

ability estimation precision and decision accuracy. The ability estimation is adversely impacted by larger proportion of testlet items, the moderate and high LID levels and short test lengths. As the LID condition gets worse (large LID magnitude, or large proportion of testlet items), the decision accuracy rates decrease. Fitting a 3PL testlet response model does not reach the same level of ability estimation precision under all simulations conditions. In fact, it proves that ignoring LID and fitting the 3PL model provides inflated ability estimation precision and the accuracy of decision accuracies.

IMPACTS OF LOCAL ITEM DEPENDENCE OF TESTLET ITEMS WITH THE MULTISTAGE TESTS FOR PASS-FAIL DECISIONS

By

Ru Lu

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
2010

Advisory Committee:

Professor Hong Jiao, Chair
Professor Robert J. Mislevy
Professor Robert W. Lissitz
Professor Jeffrey R. Harring
Professor Robert G. Croninger

Dedication

To everyone who loves, encourages, motivates and inspires me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

In recent years, a clear trend in the testing field has emerged: computer-based testing (CBT). Many large-scale high-stakes test programs such as the Graduate Record Examinations (GRE), the Graduate Management Admission Test (GMAT), the Law School Admission Test (LAST), the Armed Service Vocational Aptitude Battery (ASVAB), and several certification or licensure tests such as the Uniform Certified Public Accountant (CPA) Examination are administered by computers now. Some other tests, for example, K-12 education and adult education tests are in the transition from traditional paper-pencil test (PPT) to CBT.

CBT, also known as Computer-Based Assessment (CBA), e-exam, computerized testing or computer-administered testing, is a method of administering tests in which examinees view and respond to test questions via a computer, and in some cases, via the Internet. The advantages of CBT over PPT include increased frequency of test delivery, administration and scoring efficiency, reduced costs associated with many aspects of testing such as test delivery and administration, improved security, consistency and reliability, instant scoring and faster decision making (Bergstrom & Lunz, 1999; Scheuermann & Bjornsson, 2009; TCExam, 2008). CBT also dramatically expands the realm of possibilities for innovations in assessment format (Parshall, Spray, Kalohn, & Davey, 2002).

CBTs may take many forms. Based on test designs, CBTs can be divided into linear tests or adaptive tests. Linear tests are those that use the computer only as means of

administering which is in most respects identical to a PPT. With adaptive CBTs, different tests are assembled for different examinees. Important distinctions can also be made within adaptive test designs. For example, computerized adaptive tests (CAT) select items individually, with a decision as to what to administer next being made following each response. Multistage tests (MST) select items in pre-assembled blocks or sets, with decisions made only after each group of items is administered.

In a CAT, items are selected for each examinee based on his or her responses to previous items in a way that targets and maximizes the precision of the examinee's estimated ability. Thus, one of the main advantages of CAT over PPT is that it offers a shorter test while still maintains an equivalent level of precision (Schnipke & Reese, 1997; Wainer, 2000; Weiss, 1982). Currently, many large-scale tests are delivered in the format of CAT. These include the ACCUPLACER postsecondary placement exams (College Board, 1993), the GRE exam (Eignor, Stocking, Way, & Steffen, 1993), and the ASVAB (Sands, Waters, & Mcbride, 1997).

The successful implementation of CAT requires a psychometric model, most often an item response theory (IRT) model. IRT describes the application of mathematical models to analyze response data collected during testing/survey situations whose main objective is to measure individual persons' latent trait, ability, or skill levels. By assuming these traits, abilities or skills on a continuous latent scale, the probability of a response of an item is modeled via a mathematical function of the student's trait parameters and the item parameters. The main advantage of IRT models is the invariance of the person and item parameters. It enables the administration of different

sets of items to different examinees while still being able to estimate their abilities on the same scale (Embretson & Reise, 2000).

One of the fundamental assumptions of IRT models is local independence, or conditional independence of item responses given item parameters and examinee proficiency parameters. It means that an examinee's performance on any item depends only on the examinee's ability and the item's characteristics, and that knowledge of the examinee's performance on other items does not add any further information (Hambleton & Swamnathan, 1985; Mislevy & Chang, 2000). However, in some situations, local independence assumption may be violated. For example, in a reading test, several items are associated with the same passage, or students become fatigued at the end of the test, which is known as *local item dependence* (LID). In these situations, using IRT can lead to inaccurate estimation of the examinee and item parameters and over-estimation of test reliability (Chen & Wang, 2007; Sireci, Thissen, & Wainer, 1991; Wainer & Thissen, 1996).

In a CAT, decisions about item choice are dependent upon the local item independence assumption. However, in real testing situations, this assumption may not hold. For instance, in a math test, items with highly similar content, such as two items about factoring, are administered in the same session. Or in another situation, a CAT starts the session with difficult items, fatigue may adversely affect the examinees performance on the items at the end of session. So, locally dependent items will not work in these situations with traditional IRT models.

In addtion to the LID problem , CAT has also been scrutinized with a number of non-psychometric problems. First, because CAT is administered "on demand" rather

than on a small number of fixed dates, items may be exposed at a faster rate than in conventional tests, which poses a serious test security problem (Schnipke & Scrams, 1999; Yi, Zhang, & Chang, 2006). Second, examinees taking a CAT are not permitted to review or skip items (Vispoel, 1998). Third, with CAT, it is possible to create millions of test forms from a single item pool, making it unfeasible for humans to review every test form in advance for quality assurance purposes (Luecht & Nungester, 1998).

Because of these administrative shortcomings of CAT, an alternative CBT known as MST has been proposed and implemented in several large-scale tests. Rather than adapting the test to the current ability estimation of examinees by item as in CAT, MST adapts by a group of items in stages. It is viewed as a hybrid or compromise between conventional PPT and CAT formats (Armstrong, 2002). Different names have been given to MST. For example, Luecht and Nungeter (1996, 1998, 2000) introduced it as *computer-adaptive sequential testing* or CAST. A similar design developed by Armstrong et al. (2004) is called *multiple form structure design* (MFSD).

Due to its capacity to eliminate some of the common criticisms of CAT, MST is gaining increased interest over the past few years. Several large-scale assessments have been implemented as MSTs. Examples include LSAT (LSAT), the Test of English as a Foreign Language (TOEFL), the National Council of Architectural Registry Board (NCARB), the National Assessment of Educational Progress (NAEP), the U.S. Medical Licensure Examination (USMLE) and the Uniform CPA Examination (Hendrickson, 2007; Luecht, Brumfield & Breithaupt, 2006).

The basic unit under the MST design is a module, which is composed of a group of items. The group of items could be a group of discrete items or a group of testlet items

(that share the same stimulus) or a combination of both discreate items and testlet items. The apparent reason to use testlet items is that it is more time efficient. To item writers, it is somewhat easier to write a series of related items around a common theme or stimulus than the same number of unrelated items or stand-alone items (Downing, 2006). Also if an examinee has to study a stimulus, it is more efficient to ask several related questions with the same stimulus. For example, in a medical certification exam, the common stimulus material could be a clinical situation, describing a patient's relevant history or presenting a problem in sufficient detail for examinees to respond to several questions, all related to the common stimulus, asking for diagnosis, decisions, lab tests and likely complications, and so on.

Testlets are known to be vulnerable to the problem of LID. Ignoring LID can lead to overestimates of reliability or information and underestimates of the standard error of the ability estimates (e.g., Chen & Wang, 2007; Jiao, Wang, & Kamata, 2005; Wainer & Thissen, 1996; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002) in the situation of PPTs. LID also can have substantial effect on CAT score precision (Pommerich & Segall, 2008). Due to the use of testlet items, LID is suspected to influence measurement precision as well as inferences based on it with MSTs. However, to date, there is no study investigating the effect of LID on MST with testlet items. Thus, this dissertation studies the impact of local item dependence on MST with testlet items for pass-fail decisions. To account for the LID due to testlets, this dissertation also explores using the testlet response model (TRT) in MSTs and comparing its performance with the 3PL model.

Objectives and Research Questions

There are two objectives for this study. The first objective is to investigate the impact of LID of testlet items on the performance of MSTs for pass-fail decisions. MSTs are manipulated to have different proportions of testlet items. Various degrees of LID are simulated with testlet items. Due to the adaptive nature of MSTs, testlets are put into different stages, thus the impact of the position of LID is also studied. Another studied variable is test length because it is seen as insufficient in current research literature (Stark & Chernyshenko, 2006). Such a close examination can help us understand the underlying impact of LID. The second objective is to compare the performance of the conventional three parameter logistic (3PL) IRT model and a 3PL testlet response model in MSTs. Rather than ignoring, the 3PL testlet response model can account for LID.

In sum, this study is intended to answer the following research questions.

1. If the 3PL model is the measurement model used in analysis, how are measurement precision and classification decisions impacted by the proportion of testlet items in an MST, the position of testlet items (which stage?), and the magnitude of LID? And if the LID exists and the 3PL model is the measurement model, how are the measurement precision and classification decisions impacted by the test length of MSTs?

2. Would the 3PL testlet model that can account for LID improve the overall measurement precision and classification decisions over the 3PL model?

To answer above questions, simulation studies are carried out. The 3PL testlet model is used to generate item responses. The 3PL model and the 3PL testlet model are used to calibrate, construct MSTs and score examinees respectively. The factors manipulated include the proportion of testlet items in the MST, the testlet /discrete item

position, the testlet item effect magnitude and the test length. Details about the simulation study are presented in Chapter Three.

Organization of the Study

This study is presented in five chapters. Chapter One addresses the background of this study, the research objectives and questions associated with this study. Chapter Two provides a literature review on the basics of IRT and MST. The review concentrates on four aspects: the basic components of IRT, CBT delivery models, the benefits of CAT, the components and construction of MST, and the problem of LID. Chapter Three describes the research design and the major steps of data preparation and analyses. Chapter Four provides a detailed report of the results for the analyses introduced in Chapter Three. Chapter Five summarizes the findings, discusses the implications of this study and provides some directions for future research.

Chapter 2: Literature Review

This chapter reviews related background information to the proposed study. It includes six major sections. The first section reviews the general framework of IRT, which is the theorectial background of many current operational tests. It includes IRT assumptions, models, and its characteristics. The second section reviews current CBT delivery models. As a predecessor of the MST, a brief review of CAT and its advantages and disadvantages is provided in the third section. The fourth section provides information about the MST framework, including its components, considerations in developing an MST, and the LID problem with current operational MSTs. The fifth section is about the problem of LID. Its causes, consequences and models that can account for LID are reviewed in this section. Finally, research questions are restated at the end of this chapter.

Item Response Theory

General Framework of IRT

Item response theory (IRT) is a family of statistical models used to analyze data from any tests or questionnaires. It has the unique property of *invariance*. With this property, it is possible to solve some important measurement problems that have been difficult to solve in classical test theory (CTT) framework, such as those encountered in test equating and CATs (Hambleton, Swaminathan, & Rogers, 1991). When used appropriately, it can increase the efficiency and flexibility of the testing process and improve the precision of item or ability estimates. Currently, IRT is used for many

measurement applications including item banking, test construction, adaptive tests, scaling, equating, standard setting, and score reporting.

The core of most IRT models is that it applies a particular mathematical function to describe the probability of a particular response, such as a correct response to an item, given item and person parameters. It is assumed that an examinee's responses to different items are conditionally independent.

Model Assumptions

There are three main assumptions underlying an IRT model. They include:

Dimensionality. IRT models use *examinee parameter(s)* (also called *person parameter(s), traits, abilities, or proficiencies*) to describe the dimension(s) on which there are important differences among examinees as measured by the test items. Models that use only one dimension to describe the examinees are called *unidimensional*. Thus, unidimensional IRT models assume that only a single trait is measured by the set of items, and it is commonly referred as *unidimensionality* (Hambleton & Swamnathan, 1985). Models that use two or more dimensions to describe the examinees are called *multidimensional*. However, in the majority of applications, unidimensional IRT models are used.

Local Item Independence. The local item independence assumption requires that given the person's ability, the response to one item is independent to the response to another item conditional on item and person parameters (Hambleton & Swamnathan, 1985). Under the assumption of the local independence, the probability of an examinee's response pattern (X_i) is equal to the product of probabilities of the examinee's responses to each of the J items given his/her ability θ_i . It is expressed mathematically as

$$P(X_i|\theta_i) = \prod_{j=1}^{J} P(X_{ij} = x_{ij}|\theta_i) \quad (1)$$

Namely, the probability of a correct response only depends on the item and person parameters (Lord, 1980; Mislevy & Chang, 2000). The assumption is that the true ability value is providing all the relavant information about the student's performance and that the contribution of each item to the test can be evaluated independently of all other items. The violation of this assumption is called *local item dependence* (LID). The basic idea underlying LID is that there are additional factors that may affect students' performance that are not accounted for by the IRT model. The causes and impacts of LID and models that can account for LID are reviewed more in the LID section.

Independence of Examinees. This assumption assumes that there is no relationship between the response patterns and the examinee subgroup memberships (such as gender, ethnicity, etc) after accounting for the differences of latent trait(s). With the assumption of examinee independence, the probability of observing all responses from all examinees is:

$$P(X|\theta) = \prod_{i=1}^{I} \prod_{j=1}^{J} P(X_{ij} = x_{ij}|\theta_i) \quad (2)$$

This probability is also known as the *likelihood for the data*. Either the item parameters or the person parameters or both can be estimated by solving Equation 2 using the maximum likelihood (ML) method.

Unidimensional Dichotomous IRT models

Unidimensional models differ in terms of the number of item parameters that are used to define each item's essential characteristics. In a dichotomous 3PL model, assuming that θ_i represents the ability level of person *i*, the probability of person *i* getting a correct response to item *j* can be modeled as (Hambleton & Swaminathan, 1985):

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]} \quad (3)$$

where $P_{ij}(\theta_i)$ defines the probability of a correct response to item *j* by person *i*, a_j is the *discrimination* parameter for item *j*, b_j is the *difficulty* parameter for item *j*, and c_j is lower asymptote or "*pseudo-guessing*" parameter for item *j*. Equation 3 is known *as item characteristic function* (ICF), or graphically, *item characteristic curve* (ICC) for the 3PL model.

The difficulty parameter b_j indicates the relative difficulty of item *j*. It increases in value as the item is more difficult. Theoretically, the range of the difficulty parameter is from $-\infty$ to $+\infty$. But most *b* values are typically between -3 and 3 on a logit scale. It is the point on theta scale where ICC has its maximum slope.

The discrimination parameter a_j identifies how well an item can distinguish between examinees in different regions on the latent continuum. The more discriminating an item is, the higher *a* value is. Theoretical values are positive and typically less than 2 (Hambleton, Swaminathan, & Rogers, 1991). In ICC, a_j equals to the slope value of ICC when $\theta=b_j$.

For items such as multiple-choice items, the parameter c_j indicates that the probability that an examinee with extremely low abilities can get the item correct and its value represents the lower asymptote of the ICC.

A *two-parameter logistic* (2PL) model can be obtained by setting the pseudo-guessing parameter at 0. For *one-parameter logistic* (1PL) model or the Rasch model (Rasch, 1960), it is further assumed that all items share the same discrimination parameter. These are the three commonly used IRT models in large-scale testing programs.

Information Functions and Standard Errors

An important concept in IRT is information. This term reflects the measurement precision at each ability level. The *item information function* (IIF) is defined as follow (Embretson & Reise, 2000):

$$I_j(\theta_i) = \frac{\left[P_j'(\theta_i)\right]^2}{P_j(\theta_i)\left[1-P_j(\theta_i)\right]} \quad (4)$$

where $I_j(\theta_i)$ defines the information provided by item *j* at ability level θ_i, $P_j(\theta_i)$ is the probability of a correct response on item *j* with ability θ_i as defined in Equation 3; and $P_j'(\theta_i)$ is the first derivative of $P_j(\theta_i)$ with respect to θ_i. For example, if the 3PL model is the measurement model, the item information function would be:

$$I_j(\theta_i) = a_j^2\left[\frac{1-P_j(\theta)}{P_j(\theta)}\right]\left[\frac{P_j(\theta)-c_j}{1-c_j}\right]^2 \quad (5)$$

The *test information function* (TIF) is the sum of the item information functions. It gives an overall impression of how much information a test could provide across all the items. Thus,

$$TI(\theta_i) = \sum_{j=1}^{J} I_j(\theta_i) \quad (6)$$

The *standard error of measurement* (SEM), also known as *conditional standard error of measurement,* is defined as the reciprocal of the square root of the test information at a given ability level, which is:

$$\mathrm{SEM}(\theta_i) = 1/\sqrt{\mathrm{TI}(\theta_i)} \qquad (7)$$

Thus, the more information a test provides, the smaller the measurement error will be. As we will see later, target information function is one of the major specifications in the current automated test assembly (ATA) procedures for test constructions.

IRT Model Parameter Estimation

Item Parameter Estimation. Two common item parameter estimation methods are the joint maximum likelihood (JML) and the marginal maximum likelihood (MML) method. Currently, the latter is more frequently implemented, such as in the BILOG-MG software (Zimowski, Muraki, & Mislevy, 2003). BILOG-MG is used to calibrate item parameters in this dissertation for the 3PL model, thus its algorithm is briefly described as follows.

Assuming that $g(\theta)$ is the probability density for θ, the marginal likelihood that is maximized is

$$\mathrm{P}(\mathrm{X}|\{a, b, c\}) = \int \prod_{i=1}^{I} \prod_{j=1}^{J} \mathrm{P}(x_{ij}\,|\theta_i) g(\theta) d_\theta \qquad (8)$$

The item parameters are estimated by finding the maximum of Equation 8. In most cases, a continuous $g(\theta)$ is replaced by a finite set of discrete θ values, called *quadrature points*. In addition, an iterative Expectation and Maximization (EM) algorithm is applied (Mislevy & Stocking, 1989). This algorithm iterates to convergence between a) estimating the numbers of theoretical examinees with a particular values of θ that are expected to give response *j* to item *i*, and b) finding the item parameters that maximize

the likelihood of observing those numbers of examinees with those responses (Yen, 2006).

Person Parameter Estimation. When the item parameters are known, there are typically two methods to estimate the person parameters: Maximum likelihood (ML) or Bayesian methods. With Bayesian methods, a prior distribution is assumed for the parameter being estimated. Usually, a normal distribution is assumed for the ability parameters. Along with the likelihood of the observed item scores given the measurement model, a posterior distribution of ability parameters is obtained. When the mode of the posterior distribution is taken as the ability estimate, it is called *maximum a posterior* (MAP); when the mean of the posterior distribution is taken as the ability estimate, it is called an *expected a posteriori* (EAP) estimate. EAP method is applied in this dissertation to score examinees, thus its estimate (Mislevy & Bock, 1982) is described below.

$$\hat{\theta} = \frac{\sum QL(Q)W(Q)}{\sum L(Q)W(Q)} \qquad (9)$$

where Q is a quadrature point in the ability scale, W(Q) is weight of the quadrature point. $L(Q)$ is the likelihood of a person's response pattern at Q quadrature point.

CBT Delivery Models

There are at least five categories of CBT delivery models: (1) computerized fixed tests (CFT); (2) linear-on-the-fly (LOFT) tests; (3) item-level CAT; (4) testlet-based CAT; and (5) MST.

CFT is a fixed-length test that pre-constructed, intact test forms that are administered by computers (Drasgow, Luechet, & Bennett, 2006). Different examinees may see different forms of the test; however, all examinees administered a given form see

exactly the same items, although the presentation sequence may be different during the administration. A CFT is directly analogous to having fixed-item PPT.

LOFT is a fixed-length test, with test items uniquely assembled for each examinee according to pre-defined content and statistical specifications (Drasgow, Luechet, & Bennett, 2006; Prometric, 2010). That is, this method adjusts the item selection routine to account for item exposure. The benefits of LOFT include all those associated with CFTs with the addition of more efficient item pool usage and reduced item exposure.

Item-level CAT is a form of CBT that adapts to the examinee's ability. The test length associated with CAT can either be fixed or variable. The core idea behind CAT is that each item presented to examinees is based on his/her ability estimation on previously administrated items.

Testlet-based CAT involves the adaptive administration of testlets to examinees, rather than single items. The primary adaptive unit is testlets, rather than items. Note that testlet-based CATS are only partially adaptive because items within a testlet are administrated in a linear fashion.

MST is also a partial adaptation of the test to individual examinees. Rather than adapting the test to each examinee item by item in CAT, it adapts to examinees in stages. Each stage is composed of pre-constructed items (discrete items or testlet items).

Computer Adaptive Testing

General Framework of CAT

CAT grew out of a motivation for more efficient and precise measurement of examinees across the entire distribution compared to that accomplished by linear tests

(Lord, 1980; Wainer, 1990). It is different from conventional linear testing in that examinees are not presented with the same set of items in a particular form. An item is selected based on the scoring of the most recent response as well as the cumulative scoring (pattern of responses). It is considered as a more efficient way of testing because instead of answering all the items, only those items that are near the examinee's ability level are selected. Items that are too easy or too difficult for a given examinee are not administered.

The successful implementation of CAT requires at least five important components: a) a large item pool, b) a starting rule, c) a continuing rule, d) a scoring rule, and e) a stopping rule. With these five components, it is administered as follows:

1. The first item is presented to the examinee.
2. Based on his/her previous response(s), the remaining items in the pool are searched for the next item according to the item selection rule.
3. The examinee responds to the next item.
4. The ability estimate is updated, based on his/her previous responses.
5. The termination criterion is checked. If the termination criterion is reached, then stop the test, otherwise steps 2-4 are repeated.

Usually, nothing is known about the examinee prior to the testing. The first item is often of easy to medium difficulty.

An IRT model is the most important element in each component of the CATs except the starting rule. It is used to calibrate item pools, to update examinee's ability estimation, to select the next item, and in some cases to terminate the test.

Benefits of CAT

First, CAT can provide uniformly precise scores for most test-takers (van der Linden & Glas, 2000). Secondly, CAT can typically reduce the test length to 50% and still maintain a higher level of precision than a fixed test (Weiss & Kingsbugy, 1984). This means time and cost saving for both examinees and test administers. Third, like any CBT, CAT allows testing on demand, that is, examinees may take the test whenever wherever they are ready. It also may show results immediately after testing. Finally, it may reduce the item exposure of some items because examinees typically receive different sets of items rather than the whole population being administered a single set. However, items that of medium difficulty may have a higher risk of over-exposure.

Criticisms of CAT

Hendrickson (2007) summarized six potential problems with item-level adaptive tests. They include: (1) potential violation of the IRT assumptions of local independence and unidimensionality, (2) lack of control over non-statistical properties such as item ordering and context effect, (3) lack of control over content balancing, (4) the need for item exposure control, (5) lack of review opportunities for examinees, and (6) large data management and computer processing demands.

The dimensionality and local item independence assumptions have been introduced in previous IRT section. The cause of the violation of the two assumptions is that there are other underlying traits/factors that influence examinees' performance on the test that is not accounted for by the selected IRT model. For example, in the context of a

science test, violation of the two assumptions may occur if performing well on the science test requires high reading ability.

The criticism of lack of control of non-statistical properties such as content balancing and possible item ordering effect and context effect is evident in that with the item-level CAT, millions of different test forms can be created with the same item bank, it is impossible for content experts to review each of these forms for quality assurance purposes (Luecht & Nungester, 1998) before test administration.

The criticism of lack of review opportunities for examinees is due to the algorithm of CAT. A typical CAT does not allow examinees to skip or review test items. This may force them to abandon some of their favorite test-taking strategies and makes them complain the most about not being able to skip, review, or revise.

Another criticism is the item exposure associated with CAT. This is one of the most serious problems with CAT. With CAT, items of medium difficulties have a higher risk of over-exposure. Highly exposed items can affect the accuracy and validity of test scores.

Some other concerns of CAT include the use of test data collected from the CAT administration. Due to the sparseness nature of these data, it is difficult to conduct equating, different item functioning analysis or recalibration of item parameters (Armstrong, Jones, Koppel, & Pashley, 2004; Ban, Hanson, Yi, & Harris, 2002; Mead, 2006; Stark & Chernyshenko, 2006).

For these practical reasons, MST is gaining popularity in both the research fields and practical testing situations. The next section introduces MST and reviews its main components, routing rules, and test assembly rules associated with MSTs.

Multistage Tests

Multistage tests (MST) are those in which pre-constructed sets of items are administered adaptively and scored as a unit (Hendrickson, 2007). They are very similar to CATs in that items are selected for each examinee based on their previous responses, but rather than selecting a single item, a set of items is selected which builds tests in stages. Thus it results in fewer adaptive points than an item-level CAT but is more adaptive than the traditional PPTs in which all examinees receive the same set of items. In an ideal situation, MST combines the advantage of both the adaptive and linear test forms (Berger, 1994).

The idea of MST is not new. A kind of non-computerized MST was developed (e.g., Cronbach & Bleser, 1965; Lord, 1971 &1980) prior to CAT and applied in some operational tests. However, MST research was eclipsed by CAT (Mead, 2006). The newly improved MST formalizes a set of statistical targets and other specifications into a template that can be used in conjunction with automated test assembly (ATA) to generate large-scale, adaptive tests with desired parallel statistical and content characteristics. In comparison to CAT, MST provides better quality assurance because test forms can be created ahead of test administrations. Also, it can reduce test security risks by creating multiple parallel test forms (Luecht, Brumfield, & Breithaupt, 2006).

Components of MST

Like CAT, the successful implementation of MST requires an item pool and a testing algorithm. Additionally it has several unique components. Using the terminology developed by Luecht & Nungester (1998), the new components include: *modules*, *stages*, *routing rules*, and *panels*.

Modules. Modules are sets of items that are preconfigured. The number of items within a module may range from several items to well over 100 items. They are also refered to as *item bundles* or *testlets* (Jodoin, Zenisky, & Hambleton, 2006; Hendrickson, 2007) or *bins* (Armstrong, Jones, Koppel, & Pashley, 2004 & 2006). A module may include discrete items or items that share a common stimulus. To avoid confusion, in this disseration, the term *testlet* means a set of items which share the same stimulus. Thus, a module is larger than a testlet. A module may contain several discrete items or one or more testlets. In MST, modules are designed for different ability groups. They are targeted to have specific statistical properties (e.g., a particular averaged item difficulty) and content balancing.

Stages. A test taker visits exactly one module at each stage of an MST. The modules are administered in sequence, one stage at a time. Each stage can have one or more modules.

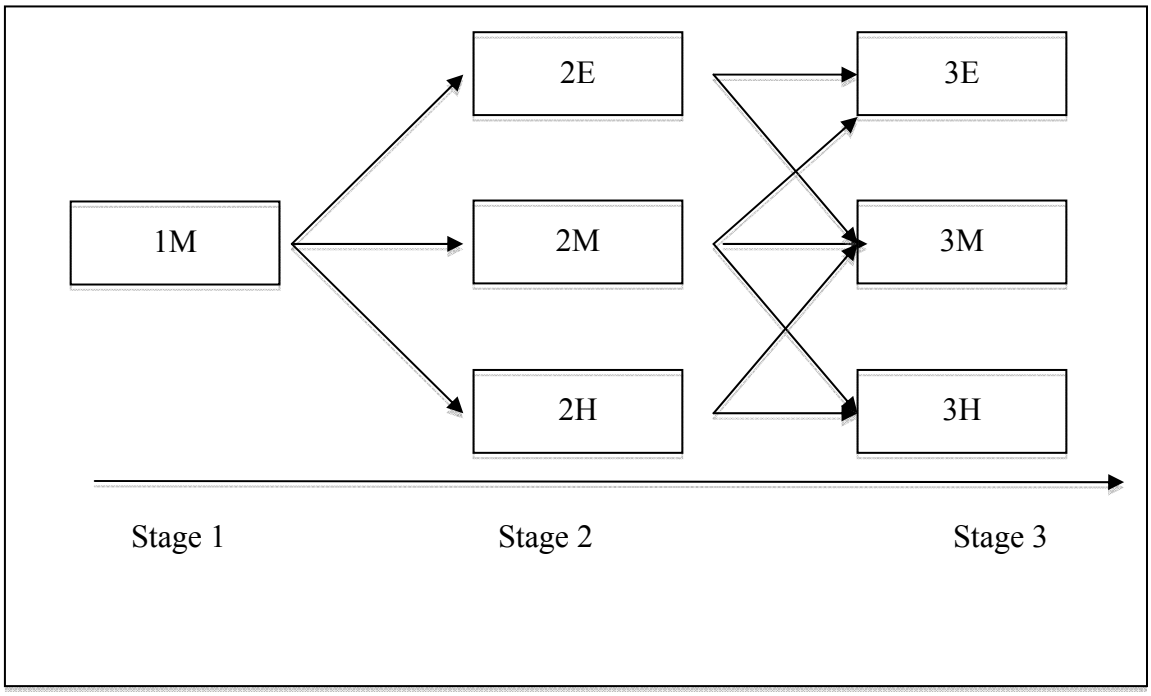
Routing Rules. After each module, a decision must be made as to which module an examinee should take in the next stage. The rules must be based on the examinee's recent ability estimation.

Panel. A panel is a particular combination of modules and routing rules. Each panel must meet the specified statistical targets, content areas, as well as other constraints. To control for the exposure of modules and items, multiple panels can be assembled and randomly assigned to examinees just like multiple test forms.

Figure 1 shows a panel of a three-stage MST in which there is one module on Stage 1, three modules on Stage 2 (2*E*, 2*M*, and 2*H*) and three modules on Stage 3 (3*E*, 3*M*, and 3*H*). It is labeled as a *1-3-3 panel* and has been used in several studies (e.g., Hambleton

& Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006; Keng, 2008; Luecht & Nungester, 1996 & 1998). The letters *E*, *M*, and *H* denote the average difficulty of the modules (E = relatively easy, M = moderately difficulty; and H = relatively hard). Each line in the figure represents a particular route. Routing from Stage 1 to Stage 2 is purely based on examinees' performance on Stage 1. Lower performance examinees are routed to the easy module on Stage 2 (2E); moderate performance examinees are routed to the moderately difficulty module on Stage 2 (2M), and top performing examinees are routed to the difficulty one on Stage 2 (2H). Routing decisions to Stage 3 is based on the examinee's cumulative performance on previous two stages. The specific criteria for determining which module an examinee should take on the next stage is built into the routing rules, which is one of the most important decisions to make in designing an MST.

Figure 1: A Panel of 1-3-3 MST



In Figure 1, There are seven possible routes indicated by the panel: 1-2E-2E; 1-2E-3M; 1-2M-3E; 1-2M-3M; 1-2M-3H; 1-2H-3M; and 1-2H-3H. The panel is explicitly constructed so that any of those pathways provides a content balanced test that meets all relevant test-level specifications (e.g., item counts, content balance, word counts, etc). Many panels like this can be constructed before the operational use.

Considerations in Developing MST

Creating an MST requires many of the same considerations as developing a linear PPT or an item-level CAT. Some basic considerations include: the purpose of the test, the type of the test (norm-referenced or criterion-referenced), the examinee population, and the decisions to make after the test. These considerations will help to guide the following decisions, such as the structure of the item pool, the number of stages, the number of

modules of each stage, and the number of items or testlets under each module, the routing rules, as well as the test assembly process. Details of these test design considerations as well as studies that have examed them are summarized as below.

Item Pool

An MST requires modules to be assembled to meet both the psychometric and content requirements. In MST design, modules of different difficulties must be created. In practical MST designs, parallel versions of modules are also needed. Thus, the item pool must support the assembly of an MST (Hendrickson, 2007). Available methods to assemble modules are reviewed in the next section. This section will further review several studies that have studied the impact of the item pool to MSTs.

Xing & Hambleton (2004) studied the effect of the item bank size and the item quality on an MST. In their study, the item bank size was set at two values: 240 and 480. 240 represented the size of an existing credential testing item pool; 480 was the size of a hypothesized item pool which doubles the current pool. Item quality was defined as the average of the discriminating values of the items in the bank. Three different levels of item quality were studied: .60, 1, and 1.4 representing poor, original and improved item bank quality respectively. Their results showed that the doubled bank size and the improved item quality can improve the final measurement precision of the MST design.

With a real item pool of 238 items, Jodoin, Zenisky, & Hambleton (2006) compared two traditional linear test forms with several variations of MSTs for making pass-fail decisions. They tried with three different passing scores. They found that all 60-item tests, regardless of the MST design or the passing scores, all test forms produced accurate ability estimates and acceptable decision consistency and decision accuracy.

However, they observed that the 60-item MSTs did not perform better than the original test form. The reason they explained was that the MSTs were held to a somewhat higher standard of content matching and that the test assembly software could not match the items in the current pool to those intended targets. That is, the current item bank was just not deep enough in quality and quantity. The conclusion is that MSTs will not be optimal unless there is an item bank rich enough to support the test design.

With a simulated relatively larger item bank (with 3222 items), which statistically reflects a current operational item bank (with 358 items), Zenisky & Hambleton (2004) studied the effects of target test information on a highly selective test (with a pass rate of 30%). They observed that as the amount of test information decreased, the levels of misclassification and inconsistent classification increased.

From above studies, one can see that not only the pool itself (pool size, item quality) has a direct impact on the measurement precision of the MST design, also the way of the pool used influence the accuracy of ability estimation and classification decisions made with the MST.

MST Structure

Before assembling an MST, several questions need to be answered first. As Hendrickson (2007) put it, these questions include: the number of stages, the number of modules at each stage, the number of items within a module or the total length of the test.

Number of Stages. Theoretically, the possible number of stages ranges from one to the total number of items. Most recent research and applications have used three or four stages (Hendrickson, 2007). More stages and more variety of difficulty of modules within the stages allows for greater adaptation and thus more flexibility. Patsula (1999) found

that increasing the number of stages from two to three increased the accuracy of ability estimates as well as the efficiency of the MST design relative to the PPT and the CAT at most ability levels (-.75 to 2.25). However, researchers cautioned that adding more stages to the test increases the complexity of the test assembly, without necessarily improve the measurement precision of the final test forms (Luecht, Nungester, & Hadadi, 1996; Luecht & Nungester, 1998).

Number of Modules in Each Stage. Many MST studies have used one module in the first stage and two or three modules in later stages. For example, Figure 1 presents a design of 1-3-3. Patsula (1999) found that increasing the number of modules in later stages from three to five increased the accuracy of ability estimation. Zenisky & Hambleton (2004) studied the design structure effects in which they compared four designs: 1-2-2, 1-3-3, 1-2-3, and 1-3-2. However, they did not find any design structure differences with respect to decision accuracy. In general, research indicates that a maximum of four modules is desirable at the last stage and that three levels may be adequate (Armstrong, 2002; Armstrong, et al., 2004).

Number of Items in a Module. Some recent studies (e.g., Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006) have implemented 20 items within each module. The length of the modules may also vary across the stages. Some tests have longer first stage module(s) and shorter modules in subsequent stages (e.g., Schnipke & Reese, 1997; Xing & Hambleton, 2004). Kim and Plate (1993) found that increasing the length of the first stage test was most important in reducing the size of the ability estimation errors. However, in their study, the total number of items in the test was not fixed. Patsula (1999) studied the effect of the distribution of item numbers in each stage. In his study,

the total test length was kept constant. He found that at most ability levels (-.75 to 2.25), varying the number of items per stage had little effect on the accuracy of the ability estimates. Luecht and Nungester (1998) discussed that using smaller modules in later stages allows test developers to better target the information provided by these latter-stage modules toward the extremes of the ability distribution.

From Equation 6, we can see that the total test information is directly associated with the number of items. Zenisky (2004) did a simulation study in which the total test information was held constant and the distribution of information over stages was a variable. She found that with limited amount of overall test information, it is better to capitalize more information on early stages so that examinees can be routed into more appropriate difficulty level of modules of later stages.

One caution is that for the reason of fairness, the module length on each stage should be kept constant for all examinees so that the total test length is consistent. An exception is that in some classification test situations, different examinees may receive different numbers of modules. This is because as any *variable-length* classification test, the main purpose of the test is to classify examinees into mutually exclusive categories rather than a precise estimate of examinees' abilities.

Routing Method

After an examinee finishes a module in one stage, a decision has to be made as to which module in the next stage to be presented to the examinee. Routing rules are used to make such decisions. Two kinds of routing rules exist. One is based on the cumulative number of correct responses. The other is based on the most recent ability estimation. Also, the routing can be designed either to classify the examinees into ability groups or to

maximize information. Maximizing information is a standard criterion in CAT. Thus, using maximizing information method or minimizing standard error can achieve better ability estimation. However, if the goal is to classify examinees to ability groups and the modules are created to match that purpose, the classification approach can also provide as much as information as the maximizing information method (Armstrong, 2002).

Luecht & Nungester (1998) empirically demonstrated that the cumulative number of correct scoring is sufficiently accurate for purposes of selecting modules. IRT scoring (maximum likelihood or Bayesian estimation) is also possible after each stage. Armstrong (2002) did a simulation study in which four routing rules are compared in the MST design in terms of ability estimation precision. The four rules he studied include: (1) route to maximize information, basing the decision on the ability estimate for the examinees at the time of routing; (2) route to classify the test taker to a percentile group defined by the design, basing the decision on the ability estimate at the time of routing; (3) route to maximize information, basing the decision on the number of correct responses; and (4) route to classify the examinees into a percentile group, basing the decision on the number of correct responses. Item pools for Logical Reasoning, Analytical Reasoning and Reading Comprehension sections of the LSAT were used to create the MSTs. For ability estimates, he used the EAP method. His results did not show any significant differences among any of the rules with regard to the correlation between the true ability and the ability estimate or the root mean square error (RMSE) of the scoring measure being compared. However, the first rule performed slightly better than the other rules and showed on average a 2% or 3% drop in RMSE.

Scoring and Ability Estimation Method

While the number correct responses could be used to score items for adaptation, a measurement model is still needed for calculating true scores and for final ability estimation. Recent research and applications of MST have often used the 3PL model (e.g. Jodoin, Zenisky, & Hambleton, 2006; Luecht & Nungester, 2000) or a polytomous IRT model (e. g. Davis & Dodd, 2003; Thissen, Steinberg, & Mooney, 1989). If the 3PL model is applied, it is assumed that the items between and within modules are conditionally independent from each other; if a polytomous model is applied to accumulate scores over individual items, it is only assumed that the items between modules are independent. The testlet response theory (TRT) model is an extension of the dichotomous IRT models, in which the LID within a testlet is modeled as part of the item characteristic function. Zenisky (2004) and Hendrickson (2007) suggested to use the TRT as the measurement model for MST. Keng (2008) first tried the use of TRT with MST design using a item pool for a statewide reading test. However, for each examinee, the testlet effect variable was assumed to be the same across different testlets in his study.

After an IRT model is selected, any method used in CAT to estimate final ability can be applied in an MST. These methods include the MLE (e.g. Jodoin, Zenisky, & Hambleton, 2006; Kim & Plake, 1993), EAP (e.g. Armstrong, 2002; Hambleton & Xing, 2006; Keng, 2008; Luecht, Brumfield, & Breithaupt, 2006) or MAP (e.g. Schnipke & Reese, 1999).

Test Assembly

Before a real multistage test can be administered, different panels must be assembled. The assembly process is very complicated because multiple panels must be

constructed to be parallel and subjected to both the psychometric properties and content constraints. The psychometric property constraints involve the use of target test information function (target TIF, Luecht & Nungester, 1998). Other constraints include test length, content balancing, item format, word count, and answer key positions etc. These constraints must be specified before the assembly.

Luecht and Nungester (1998) discussed two strategies for panel assembly—bottom up and top down. With the bottom up strategy, items are assembled into modules such that each module as a self-contained unit meets the requisite information, content, and item feature targets selected for the test. With this method, modules are interchangeable and can be mixed and matched to create multiple overlapping panels. The top down strategy requires only test level specifications of statistical and non-statistical targets. Modules are assembled in such a fashion that any path through the panel will result in a test of appropriate precision, content, and item type, although modules are not exchangeable either within or across panels. Examples of bottom-up (e.g. Jodoin et al., 2006; Luecht et al., 2006) and top-down (e.g. Davis & Dodd, 2003) strategies can also be found in the MST literature.

The *automated test assembly* (ATA) algorithm which uses optimization algorithms or heuristics, or both to select items from a bank can be used with both bottom up or top down strategies (e.g., Armstrong, Jones, Koppel, & Pashley, 2004; Luecht & Nungester, 1998; van der Linden, 1998 & 2005). In MST, the target TIF for individual modules or particular routes (i.e., module combinations) is expressed as *objective functions*; then this function is subject to other requirements and restrictions (i.e., the total test length, word count, etc). Once the model is defined, it is solved by mixed-integer

programming (MIP) methods in which algorithms iteratively assess every possible solution relative to the target until the optimal combination is reached. Some commercial computer softwares such as CPLEX 10.0 can help test developers to solve large and complex test assembly problems with discrete items.

Actually, the development of ATA is not unique to MST and it has occurred in more general contexts of optimal test design and assembly (Parshall, Spray, Kalohn, & Davey, 2002; Swanson & Stocking, 1993; van der Linden, 1998; van der Linden, 2005). Most of the recent MST studies have used the ATA method. However, when the constraints are relatively few, manual assembly of MST is also possible (e.g., Davis & Dodd, 2003; Keng, 2008).

Advantages of MST

Compared to traditional linear tests, MST allows for more efficient and precise measurement across the proficiency scale (Kim & Plake, 1993; Schnipke & Reese, 1997; Patsula, 1999). It can lead to reduced test length and testing and score reporting time. It has been shown to provide equal or higher predictive and concurrent validity of score inferences (Wainer, 1995; Weiss, 1982) .

Compared to the item-level CAT, advantages of using MST include: (1) with MST design, adaption happens between modules, thus content experts could review pre-assembled modules and make the quality assurance more feasible, (2) MST provides examinees with opportunities to skip, review or revise within a module without the concern of test integrity, (3) with MST, multiple parallel panels can be built and item exposure rate can be controlled at each modules, and (4) the data from MST are block sparse and thus more tractable to statistical analysis than CAT.

In summary, MSTs allow test developers to have more control over the test process and test quality assurance. Thus more appealing in operational use. The GRE revised General Test which is scheduled to launch in August 2011 chooses the MST design over the CAT design which is seen in current GRE test (ETS, 2010).

Local Item Dependence Problem with MST

As long as there are testlet items, MST is not exempt from the problem of LID. Hendrickson (2007) discussed that MSTs can better assure the local item independence assumption because items within a module can be treated as one polytomous item and thus independence of responses within the module is not required. While the use of polytomous IRT model have been shown to work well in some situations (e.g., Wainer, 1995), there are two circumstances where it falls short (Wainer, Bradlow, & Du, 2000). One situation occurs when more information such as item characteristics is needed. With a regular polytomous IRT model (e.g., Masters' partial credit model (Master, 1982); Samejima's graded response model (Samejima, 1969)) cannot differentiate the response pattern within a module and thus each item's characteristics under the same module are ignored. The second one is when ad hoc testlet construction is needed. For example, a stimulus has a total of twenty items and only ten items are needed to be presented to examinees along with the stimulus. A polytomous IRT model cannot help with such intelligent selection of the ten items. Thus far, there is no study evaluating the robustness of MST to the violation of the local item independence assumption. The next section will review more about the LID problem and studies that address the LID problem with other testing designs such as linear tests and CAT.

Local Item Dependence

Previous IRT section of this chapter reviews the local item independence assumption. The violation of this assumption is called LID. This section will review the LID and the causes of LID, consequences of LID, as well as models that can account for LID.

LID and Causes of LID

LID arises from the existence of an additional factor that consistently affects the performance of students on some items to a greater extent than on other items (Habing & Roussos, 2003). LID can be positive or negative (Yen, 1993). Positive LID between items means that, if a student perform better (or lower) than expectation on one item, he or she will perform higher (or lower) than expectation on the other. Negative LID between items means that if a student performs unusually well on one item, he or she probably will perform unusually poorly on the other.

Yen (1993) listed a variety of reasons that can cause LID. They include: external assistance or interference with some items, speededness, fatigue, practice, variation in response format (such as multiple-choice vs. constructed-response), a shared stimulus or passage, item chaining, items requiring explanation of a previous answer, cloze items (in which examinees fill in multiple blanks in one passage), scoring rubrics or raters, unique content knowledge or abilities, and different opportunity to learn. With the recent popularity of performance assessment, researchers also found that performance assessment tend to have items that are locally dependent due to common stimulus information or the requirement of explanation of previous response (e.g., Sireci, Thissen, & Wainer, 1991; Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999). In

short, as Yen & Fitzpatrick (2006, Page 141) stated: “the basic principle involved in producing LID is the existence of an additional factor that consistently affects the performance of some students on some items to a greater extent than on other items”.

Chen & Thissen (1997) divided LIDs into two categories: “underlying local dependence” (e.g., items share the same stimulus) and “surface local dependence” (e.g., item similarity or test speededness effect). This study will focus on the former LID that is caused by the shared stimulus.

Consequences of Ignoring LID

Ignoring LID can result in biased IRT person and item parameter estimation, overestimation of reliability, and equating errors (e.g., Chen & Thissen, 1997; Embretson, 2000; Hambleton & Swaminathan, 1995; Sireci, Thissen, & Wainer, 1991; Tuerlinckx & De Boeck, 2001). Ackerman (1987) reported that when LID exists, the calibrated dependent item discrimination parameters were over-estimated; difficulty estimates tend to become homogeneous; and ability estimates were affected by the degree of dependence increased. Thissen, Steinberg, and Mooney (1989) showed that the testlet information with the 3PL model was substantially larger than the testlet information with a polytomous model for passage related items. Yen (1993) identified artificially inflated information curves when LID items from language arts and mathematics performance assessments were treated as independent dichotomous items. Wainer & Wang (2000) found that if the 3PL model was applied to analyze testlet response data, the item difficulties were well estimated and the estimates for the item discrimination and pseudo-guessing parameters were biased, and that test information was substantially over-estimated. In a simulation study by Glas, Wainer, & Bradlow (2000), they compared the

performance of the 3PL and the 3PL testlet model when the data was simulated using the 3PL testlet model. They found that when the testlet effect was ignored, the mean absolute error of the estimates of the discrimination and diffiulty parameters were both worse using the 3PL model. Wainer, Bradlow, & Du (2000) did a similar study and got similar results in terms of the mean absolute error of the parameter estimation. They further showed that the inferences (such as classification decisions) made based on the 3PL model would be biased if the testlet effect was ignored. DeMars (2006) reported that when LID exists the 3PL model inflated reliablity for ability estimates. Zhang (2008) did a simulation study in which he compared the equating results using the 3PL model and a polytomous IRT model (GPC, generalized partical credit model) in the existence of LID, he found that the GPC method was more effective in equating. His results suggest that ignoring LID would lead to less precise parameter estimates.

In CAT, items are chosen adaptively to provide the most efficient measurement. Usually, the calibration of item parameters and the decision about the item choices are based on the assumption of local item independence. Reese (1999) first explored the impacts of LID on CATs. She pointed out that the impact of LID with PPTs is different from that with CAT in that the effects are equalized across examinees, since all examinees are asked to respond to the same set of items. By directly manipulating the correlation structures among test items, she found that only extreme level of LID is problematic in the CAT design. Pommerich & Segall (2008) found strong evidence for local dependence in a CAT of mathematics tests. They further did a simulation study to evaluate the impact of LID on the precision of test scores when the 3PL model is used for

item selection and scoring. Their results suggested that LID in examinees' responses had a fairly substantial effect on score precision, depending on the degree of LID present.

In sum, ignoring LID can negatively affect the IRT model parameter estimation as well as inferences based on the model estimation.

Models That Can Account for LID

To date, different models have been proposed to account for LID within a testlet. One general idea is that a random variable is added into the unidimensional model to account for the LID caused by the shared stimulus. The other idea is to model the LID as a second dimension.

Examples of adding a random effect variable into the model include Bayesian random-effects testlet models (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wang, Bradlow, & Wainer, 2002; Wainer, Bradlow, & Wang, 2007), one-parameter multilevel testlet model (Jiao, Wang, & Kamata, 2005), and Rasch Testlet models (Wang & Wilson, 2005). With the Bayesian random-effect testlet models, a random-effect parameter is added into the standard two- or three- parameter IRT models. Jiao et al. (2008) further proved that the one-parameter multilevel testlet model is algebraically equivalent to the Rasch testlet model, which is also a special case of Bayesian random-effect three-parameter testlet response model (3PL testlet, Wainer, Bradlow, & Wang, 2007). The 3PL testlet response model can be expressed as:

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j - \gamma_{id(j)})]}{1+\exp[a_j(\theta_i - b_j - \gamma_{id(j)})]} \qquad (10)$$

where θ_i is the person ability, a_j, b_j, and c_j denote the item difficulty, discrimination and pseudo-asymptote parameters respectively, $\gamma_{id(j)}$ is the testlet effect (interaction) of item j

with person i that is nested within testlet $d(j)$, and $P_{ij}(\theta_i)$ is the probability of a correct response at the ability level of θ_i to item j. Note that the testlet effect parameter $\gamma_{id(j)}$ is both a person and testlet parameter. This means that for a given testlet $d(j)$, the effect of the local dependency varies for the examinees. $\sigma^2_{\gamma_{id(j)}}$ is used to represent the magnitude of testlet effect. The larger the $\sigma^2_{\gamma_{id(j)}}$, the greater the proportion of total variance in test scores that is attributable to the given testlet.

As in IRT models, a two-parameter logistic testlet (2PL testlet) response model (Bradlow, Wainer, & Wang, 1999) can be obtained by assuming that the pseudo-guessing parameter in Equation 10 is 0. And by further assuming the items share the same discrimination parameter, a Rasch testlet model (Wang & Wilson, 2005) can be obtained.

Li, Bolt, & Fu (2006) pointed that the previous testlet response model (Equation 10) applies a common item discrimination parameter to both the general ability and testlet factor. They relaxed this assumption and included separated discrimination parameters for ability and the testlet effect. Their results suggested that their new model provided better fit to both simulated testlet response data and real data. Though they did not include a pseudo-guessing parameter in their model, based on their suggestions, an alternative testlet response model can be written as

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{\exp[a_{j1}\theta_i - t_j + a_{j2}\gamma_{id(j)}]}{1+\exp[a_{j1}\theta_i - t_j + a_{j2}\gamma_{id(j)}]} \qquad (11)$$

where t_j is a threshold parameter related to difficulty of the item; a_{j1} and a_{j2} indicates the discriminating power of an item with respect to θ and γ.

Above random effect testlet models can be estimated by the Bayesian estimation using Markov Chain Monte Carlo (MCMC) method. For detailed specifications of the

priors and hyper-prior of the parameters in the models, please refer to Wainer, Bradlow, & Wang (2007) and Li, Bolt, & Fu (2006).

Reckase's compensatory multidimensional extensions of the 3PL model (Reckase, 1997) also can be used to model the LID with testlet. His model can be expressed as:

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{\exp[\underline{a}_j'\underline{\theta}_i + d_j]}{1+\exp[\underline{a}_j'\underline{\theta}_i + d_j]} \quad (12)$$

where $\underline{a}_j'$ is the vector of item discriminations for item *j* in each of the dimensions; $\underline{\theta}_i$ is examinee *i*'s vector of abilities; and d_j is a scalar parameter that is affected by the difficulty of the item *i*. Note that d_j is added, rather than substracted as in the unidmensional IRT models. Thus, higher values of d_j indicates easier items. When the LID is modeled as the second dimension, Equation 12 can be re-written as:

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{\exp[a_{j1}\theta_1 + a_{j2}\theta_2 + d_j]}{1+\exp[a_{j1}\theta_1 + a_{j2}\theta_2 + d_j]}, \quad (13)$$

By treating the testlet effect as the secondary trait, DeMars (2006) used such a bi-factor model in Equation 13 to estimate LID with testlets. She further compared the results based on four different models: the bi-factor model, the 3PL model, the testlet-effects model as in Equation 10, and a polytomous IRT model. Her results suggests that parsimoniously the model choice should be the testlet response model expressed in Equation 10.

Rijmen (2009) reviewed three multidimensional IRT models that can account for LID: the bi-factor model, the testlet model, and the second-order model. In the second order model, the items only load on the testlet-specific factors. The correlations between the testlet-specific factors are modeled through a second-order factor. In his paper, it is

shown that the testlet model is equivalent to the second-order model with a dimension for each testlet. And the testlet model in turn is a bi-factor model.

Among all these reviewed models here, the random testlet effect model expressed in Equation 10 can explicitly model the LID within testlet without adding complexity. It also can easily facilitate the transformation from the convential IRT to TRT models mathematically and conceptually (Bradlow et al, 1999; Wainer et al. 2007). Thus, it is used in the current study to generate locally dependent response data.

Research Statement

In most research or operational use of the MST, a measurement model (e.g. the 3PL model) that assumes the local item independence is applied. When testlet items are included in MST, LID is suspected to influence the ability estimation as well as classification decisions. However, to date, there is no study investigating the effect of LID on the MST with testlet items yet. Thus, this dissertation evaluates the impact of local item dependence of testlet items with MSTs for pass-fail decisions. Also, to explore the possibility of accounting for the LIDs associated with testlet items in the MST, a 3PL testlet model is applied.

Specifically, this study has three purposes:

Firstly and mainly, to exam the impact of different LID conditions on MSTs with testlet items. The basic research design for this purpose follows the logic of previous LID impact studies on PPTs. Locally dependent data are first generated and calibrated with a unidimensional IRT model. Unidimensional item parameter estimates are used to construct the MST panels. The administration of panels are then simulated. During the administration, examinee responses are simulated to be locally dependent, while ability

estimation, assuming unidimensional data, is based on the unidimensional item parameter estimates. The estimated abilities are evaluated against "true" abilities to assess the impact of local dependence.

However, unlike previous studies, this study focuses on the potential capacity of MST to "control" the local dependence by prescribing the proportion of testlet items in an MST and the position of testlet /discrete items. The rationale for such a focus is explained as follows:

In this study, discrete items are those physically independent items. Testlet items refer to item sets that are physically clustered under common stimuli. In real tests, examinees may encounter different combination of discrete items and testlet items. For example, the two real data analysis reported in Wainer et al. (2000) include two different testlet item proportions: about 30% for the GRE-Verbal test and about 50% for the SAT-Verbal test, while the Analytical Reasoning section and Reading Comprehensive section of the LSAT have 100% of testlet items in their MST forms appeared in Armstrong, Jones, Koppel, & Pashley (2004). More testlet items imply more chances for LID. Thus, the proportion of testlet items is manipulated. Particularly, in this study, MSTs are built with the same proportion of testlet items as the studied item pool. Thus, the proportion of the testlet items in the item pool is then manipulated in this study.

The position of the testlet /discrete items is a new factor that is uniquely related with MST designs. The nature of MST is that it is an adaptive test: a module of items is administered to examinees and that the selection of next module is based on the performance of previous modules. If the testlet items appear in the early stages, the impact of LID would be complicated: a simple model such as the 3PL model gives an

imprecise temporary estimate of the examinee's ability; which in turn could route the examinee to an inappropriate module for the next stage. And inappropriate items within inappropriate modules would further lead to worse ability estimate and thus make more routing mistakes till the final ability estimated. On the contrary, if the testlet items appear in a later stage of MST, the impact of LID would be expected to be simpler and smaller than those in the early stages.

The second purpose of this study is to investigate the effect of another MST factor: test length. Test length is an important factor that can influence the measurement precision. In previous studies, for example, Jodoin, Zenisky, & Hambleton (2006) compared two-stage 40-item MST with three-stage 60-item MST for classification decisions; Zhang (2006) studied the multidimenionality issue with three-stage 60-item MSTs. Stark and Chernysheko (2006) commented that 40 or 60 items are too many for an MST. Current AICPA test has a structure of 1-2-2 with 25 or 30 items in total. This study will study two test length situations: one has 36 items as the long test; the other has 24 items as the short test.

The third purpose of this study is to explore the possible use of the 3PL testlet model to account for LID associated with testlet items in the MSTs. By adding a random effect variable to the 3PL model, the 3PL testlet model can account for LID with testlet items. And it has been proved useful in some certification tests (e.g., Wainer, et al., 2006).Thus, this study also attempts to apply the 3PL testlet model to account for the LID. Specifically, the 3PL testlet model is used to calibrate item parameters, to construct MSTs, and to get examinees' both interim and final ability estimates. The possible gains

on ability estimation precision and classification decisions using the more complicated 3PL testlet model against the use of the 3PL model are studied.

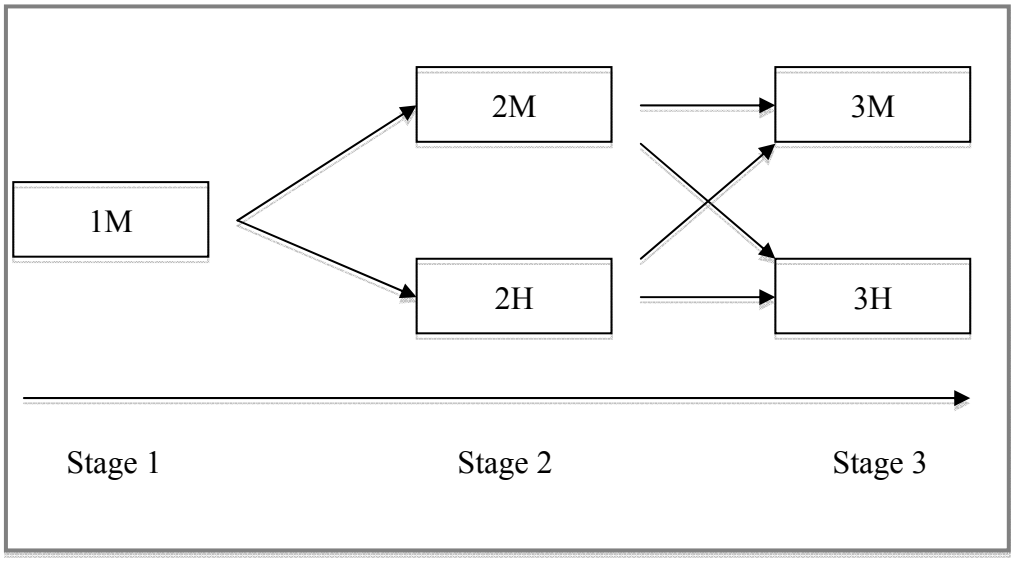
Chapter 3: Methodology

This chapter addresses the methodology applied in this study. It is divided into three main sections. The first section specifies the MST design simulated in this study. It is followed by a detailed description of the factors of investigation. In the second section, a step-by-step data generation method is provided, which includes the generation of item pools, the construction of MST panels and the administration and scoring of the MST tests. Finally, the evaluation methods of the results are presented.

Specification of the MST design

Chapter 2 introduced an MST with a structure of 1-3-3, where modules have three different levels: easy, moderate, and hard. The MST design used in this study, however, is of a structure of 1-2-2, which is presented in Figure 2. It is chosen to resemble the current certified public accountant credential test (Breithaupt & Hare, 2007). For a certification test like the CPA, it has two primary purposes: one is to provide accurate and consistent pass-fail decisions; the other is to provide diagnostic information for those who fail the test. This structure has also been studied in Zhang (2006) and van der Linden, Breithaupt, Chuah, & Zhang (2007).

Figure 2: The MST design used in this study



The MST presented in Figure 2 is a panel with five modules of multiple-choice items. Each module is targeted at a specific difficulty level: Stage 1 has one moderate module; Stage 2 and Stage 3 each have a moderate module and a hard one. During the administration, a panel is randomly selected first. An examinee encounters a moderate module in the first stage. According to the routing rule, the examinee is routed to one of the two modules in Stage 2. And at the end of stage 2, the routing rule is applied again to select another module in Stage 3. In total, each examinee is administered three modules, and selections are tailored at Stage 2 and Stage 3 to the ability of the examinee.

Following Breithaupt & Hare's (2007) example, the H modules in this study are specified so that test information would be maximized within the area of the passing score of $\theta = 1$; and the M modules are set to have most precision at $\theta = 0$. The midpoint $\theta = .5$ between the M and H modules is used as the routing cut score. That is, after Stage 1 or 2, if the ability estimate is less than or equal to .5, the examinee is routed to an M module; otherwise, the examinee is routed to an H module.

With this design, both interim and final abilities are estimated using the EAP method. To reduce estimation bias, a relatively weak normal prior with mean of zero and standard deviation of two is used for the target examinees. The EAP method is selected over the MLE method is because that though MLE has the desirable property of unbiased estimation, it is unstable for short tests and can be unbounded (Davey & Pitoniak, 2006). The application of EAP is also seen in other MST studies (e.g., Armstrong, 2002; Hambleton & Xing, 2006; Luecht, Brumfield, & Breithaupt, 2006).

No content control is implemented in this study. In real tests, the content control is implemented through the MST assembly procedure by specifying a set of constraints either at the module level or at the route level. A full treatment of content control is beyond the scope of this study. No content control is also seen in Edwards & Thissen (2007) and Pommerich & Segall (2008). However, the maximum item exposure rate is controlled at .25 in this study.

Specification of the Manipulated Factors

Measurement Models: the 3PL Model vs. the 3PL Testlet Model

With MST, a measurement model can be applied in three places. First, the measurement model is used to calibrate the item parameters in the pool. The calibrated item parameters are then used to assemble the panels. Secondly, the measurement model can be used to estimate the examinee's ability temporally to route examinees to the next stage. Third, the measurement model is applied to estimate the examinee's ability at the end of the test.

Two measurement models are considered in this study: the 3PL model and the 3PL testlet model. The 3PL model is the one that is the most applied measurement model

in operational tests and commonly seen in the literature about MSTs (e.g., Armstrong & Roussos, 2005; Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006; Xing & Hambleton, 2004). The 3PL testlet model is the one that can account for the LID within testlets. It has been applied in many PPTs. However, only Keng (2008) tried it with an MST.

Item Pool Structure

In this dissertation, the studied item pools are of dichotomously scored multiple-choice (MC) items. Two kinds of MC formats are seen in many operational tests: discrete items and testlet items. In this dissertation, discrete MC items are those that have a single stem and several options. They do not share the same stimulus and are conditionally independent from each other. Testlet items are defined as those items that share the same stimulus. With a fixed test length, more testlet items in an MST are expected to have more influence on the final inferences. It is also assumed that the final MSTs are constructed to have the same proportion of testlet items as the item pool. Four proportions will be studied: 0, .33, .67, and 1, correspondingly to represent no item, 33%, 67%, and 100% of the items in the pool are testlet items. They are presented in the second column of Table 1.

Table 1: Item Pools

Item Pool	Proportion of Testlet Items	Number of Items (Discrete/Testlet Items)	Magnitude of LID
1	0	1200/0	N/A
2	0.33	800/400	0.25
3	0.33	800/400	1
4	0.33	800/400	1.5
5	0.67	400/800	0.25
6	0.67	400/800	1
7	0.67	400/800	1.5
8	1	0/1200	0.25
9	1	0/1200	1
10	1	0/1200	1.5

According to Table 1, there are 1200 items in each item pool. The choice of 1200 items is based on the consideration of previous studied MST pool size and the needs of the current study. Different pool sizes have been used in previous MST studies. It varies from 238 to 3222. However, around 1000 is more commonly applied. For example, Ariel, van der Linden, & Veldkamp (2006) applied an item pool of 1066 items; Breithaupt & Hare (2007) reported the use of a pool of 1340 items; and Keng (2008) used an item pool of 1008 items. The reason that a pool size of 1200 used in this study is due to the combination of the discrete items and testlet items.

LID Conditions

The studied LID conditions include LID magnitudes and LID positions. LID magnitudes have been previously proved to have an impact on PPTs (Glas, Wainer, & Bradlow, 2000; Wainer, Bradlow, & Du, 2000). Three levels of LID magnitude are studied by setting $\sigma^2_{r_{id(j)}}$=.25, 1, and 1.5 or standard deviation of .5, 1, and $\sqrt{1.5}$

correspondingly to represent small, moderate and large effects. These magnitudes of testlet effects have been previously studied by Wainer, Bradlow, & Du (2000) or Wang &Wilson (2005). Empirical work (e.g., SAT, GRE, TSE and North Carolina Test of Computer Skills) has demonstrated that these are plausible values. Further assuming that testlets in the same pool have the same magnitude of LID, along with the item pool structure factor, a total of ten item pools are studied in this dissertation. They are listed in Table 1.

Another studied LID condition in this dissertation is the LID position factor. In the MST context, due to its adaptive nature, the position of the LID may also influence the final ability estimation. In an MST, modules are selected based on examinees' performance on previous stages. If LID exists in an early stage and the 3PL models applied, the impact of LID would be two-fold: one is on the routing decision that is based on the current ability estimation; the other is on the final ability estimation. On the contrary, if the LID exists in a later stage, even though it would impact the final ability estimation, it is expected that this impact would be less than those MSTs with LID in early stages. To fully understand the LID position effect, depending on the proportion of testlet items in the MST test, eight different LID position conditions are studied. They are listed in Table 2.

Table 2: LID position conditions

Proportion of Testlet Items	Stage1	Stage2	Stage3
0	dsct	dsct	dsct
0.33	dsct	dsct	tslt
0.33	dsct	tslt	dsct
0.33	tslt	dsct	dsct
0.67	dsct	tslt	tslt
0.67	tslt	dsct	tslt
0.67	tslt	tslt	dsct
1	tslt	tslt	tslt

Note. dsct: Discrete items; tslt: Testlet Items.

In Table 2, column 2 to column 4 represents the item property on each stage: either discrete items or testlet items. Particularly, if 33% items are testlet items, Item Pool 2, 3, and 4 are applied. If 67% items are testlet items, Item Pool 5, 6, and 7 are applied. Similarly, Item Pool 8, 9 and 10 are applied when all items are testlet items.

Test Length

Test length is an important consideration in test development. The primary reason for the developments of CAT or MST is the improvement of measurement efficiency. Previously MST studies have mainly set the test length as fixed (e.g., Jodoin, Zenisky, & Hambleton, 2006; Xing & Hambleton, 2004; Zenisky & Hambleton, 2004; Zhang, 2006). Only Jodoin (2003) and Keng (2008) studied the effect of test length of the performance of MST. They found that the overall test reliabiliy and conditional measurement precision as well as classification precision would increase as the test length increases.

Two test length conditions are simulated in this dissertation: long and short. Tests under the *long* test condition consist 36 items, with each module having 12 items. This test length has been seen in Armstrong, Jones, Koppel, & Pashley (2004) and Edwards &

Thissen (2007). Tests under the *short* test length condition consist 24 items, with each module having 8 items. An exam that is two thirds of the length could reduce exam costs for examinees and the testing agency, reduce testing time, lower item exposure levels, and possibly require smaller item banks.

Data Generation

A group of 2500 examinees are simulated to take each panel of the MST. They are defined as 100 θs from -3 to 3 in increments of .25. This flat distribution is used so that the precision of ability estimates across the entire ability range could be determined accurately.

From the initial item pool creation to the administration and scoring of MST, data preparation can be divided into three major steps. These steps are briefly described below as an overview, and more detailed descriptions of each step are provided later in this section.

Step1: Simulation of Items Pools. In this step, item parameters are generated instead of taken from a real test. To mimic the reality that true parameters are never known, item responses for item pool calibration are first generated and then used to calibrate the item parameters in the pool.

Step 2: Assembly of MSTs. In this major step, item information is calculated at two ability levels. Based on their information, items are selected and assembled into panels.

Step 3: Administration and scoring of MSTs. In this major step, target examinees' responses to the selected MST items are generated using true item and ability parameters. Then the administration and scoring of examinees of MST are simulated.

BILOG-MG and SCORIGHT are used to calibrate item parameters. All other steps are executed using SAS. Detailed procedure for each step is described as follows:

Step 1: Simulation of Item Pools

This major step includes the generation of item parameters, the generation of item responses for item pool calibration, as well as the calibration of item parameters in the pools.

Generation of Item Parameters

For any condition in Table 1 that has discrete items, the item parameters are generated according to the following specifications:

1. The item difficulties are drawn from a normal distribution with mean of 0 and standard deviation of 1, within the interval of (-2.5 to 2.5),
2. The item discrimination parameters are drawn from a log-normal distribution with mean of 0 and standard deviation of .5, with the range of (.4, 1.5).
3. The guessing parameters are set to have a random uniform distribution with range of (0, .3).

Items are simulated to have various degrees of discrimination, difficulty, and guessing. The draws of item parameters are made independently. The summary statistics of true item parameters are listed inTable 3. The same set of item parameters are used in each item pool.

Table 3: Summary statistics of true item parameters (N=1200)

Item Parameter	Min	Max	Mean	S. D
a	0.4004	1.4998	0.9201	0.2896
b	-2.4380	2.4334	-0.0398	0.9629
c	0.0002	0.2996	0.1462	0.0852

For any item pool in Table 1 that has testlet items, testlets are simulated instead of discrete items. Each testlet has 20 items. In the literature, the reported testlet size varies from 2 to 32 (e.g., Keng, 2008). The reason to choose 20 in this study is to enable and illustrate the testlet ad hoc construction property mentioned in Chapter 2. The item parameters within each testlet are randomly generated according to the same specifications as above. The LID parameters $r_{id(j)}$ are generated from $N(0, \sigma^2_{r_{id(j)}})$ with specified $\sigma^2_{r_{id(j)}}$ in Table 1.

Generation of Item Responses for Pool Calibration

For the calibration sample, 3000 known ability levels are drawn from a standard normal distribution. Response strings are generated for each simulee for each item in the pool. 3000 represents a large number of examinees that can help get stable parameter estimates. Particularly, if an item is a discrete item, the response is generated using the 3PL model; if an item is a testlet item, the response is generated using the 3PL testlet model.

In real testing situations, the calibration of such large item banks is not feasible as it is impossible for each examinee to take all the items in the item pool at one time. A common practice is to have a calibration scheme in which sparse data are collected to

ensure that all items can be jointly calibrated to a common scale (e.g., Ban, Hanson, Yi, & Harris, 2002; Chuah, Drasgow, & Luecht, 2006; Wainer & Mislevy, 2000).

The algorithm for generating dichotomous response data is as follows: for each simulee and each item, the probability of obtaining a correct response is computed using Equation 3 or Equation 10 depending on whether it is a discrete item or a testlet item. Then a random number from the uniform distribution between 0 and 1 is generated and compared with the probability. If the probability is larger than the random number, a score of 1 is assigned to the item; otherwise a score of 0 is produced.

Calibration of Item and Testlet Parameters

The 3000 simulees' response data created for each pool are used to calibrate item parameters under each measurement model. If the 3PL model is the measurement model, the item parameters are calibrated using the BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) software. If the 3PL testlet model is the measurement model, the item parameters are calibrated using SCORIGHT (Wang, Bradlow, & Wainer, 2005). SCORIGHT is a very general computer program for scoring tests. It models tests that are made up of dichotomously or polytomously rated items or any kind of combination of the two through the use of a generalized IRT formulation. The items can be presented independently or grouped into testlets or in any combination of the two. The estimation is accomplished within a fully Bayesian framework using MCMC procedures. In particular, the priors used in the Bayesian framework for TRT models include (Wainer, Bradlow, & Wang, 2007, p. 136):

$$\theta_i \sim N(0,1)$$

$$\log(a_j) \sim N(\mu_a, \sigma_a^2)$$

$$b_j \sim N(\mu_b, \sigma_b^2)$$

$$\gamma_{id(j)} \sim N(0, \sigma_{d(j)}^2)$$

$$\text{logit}\,(c_j) \sim N(\mu_c, \sigma_c^2)$$

For the data analysis here, the default hyperpriors for parameters in SCORIGHT are applied. SCORIGHT was run using two independent chains for an initial burn-in period of 8000 draws and 2000 iterations thereafter.

BILOG-MG uses MML procedures to estimate item parameters. To avoid possible confounding of model with estimation method, the 3PL model could have been estimated using SCORIGHT too. However, one of the purposes of this study is to compare these two models as they are used in practice. A direct comparison of estimation results from the MML method to those from the MCMC is also seen in DeMars (2006).

For each item pool and each measurement model, item parameter estimates ($\hat{a}$, $\hat{b}$, and $\hat{c}$) and testlet effect estimates $\hat{\sigma}_{r_{id(j)}}^2$ are stored and used in the following MST panel constructions and administrations.

Step 2: Assembly of MST Panels

The bottom-up strategy by Luecht & Nungester (1998) is applied in this study. Particularly, the assembly of MST is accomplished following Armstrong et al. (2004)'s two-phase method, though it is a simplified version by not considering content control or other practical constraints, such as word counts or answer key count. In the first phase, items are selected from the original item pool to appear in the panels of MST. In the second phase, parallel panels are created with the selected items.

With eight panels, the panel exposure rate for each test examinee group is expected to be 12.5%. Thus, the module exposure rate is expected to be no larger than 12.5%. However, discrete items and testlets can be shared among modules between panels. Thus, the actual item exposure rate might be higher than 12.5%. For example, if the same item in a testlet appears in two panels, the expected exposure rate would be 25%. Thus, to control for the item exposure rate to be less than 25%, each independent item and testlet is set to appear in no more than two modules in total.

Phase I

From an item pool with estimated item parameters, select items/testlets in the pools to be assembled. From Figure 2, we know that each panel has five modules, with three modules at moderate level, and two modules at hard level. Target information functions are introduced in Chapter 2. They are used to select discrete items or testlets to each module. In this study, the TIFs for the three moderate modules are set to peak at $\theta = 0$ [1]and the TIFs for the two hard modules are set to peak at $\theta = 1$. If the MSTs constructed with only discrete items, the MST is constructed as follows:

The item information of each individual item in the pool is calculated both at $\theta = 0$ and $\theta = 1$. The item information function for discrete items is given in Equation 5. Constrain that each item can only appear in one module and in one panel, a total of (8*5*module length) are needed. For example, if the module length is 12, then a total of 480 items are selected from the item pool, of which 288 (8*3*12) items that provide the most information at $\theta = 0$ and 192 (8*2*12) items that provide the most information at

[1] In operational ATA, the target information functions are typically computed for a vector of theta points along the proficiency scale.

$\theta = 1$. Note that some items in the pool may appear in both the 288 most informed items at $\theta = 0$ and the 192 most informative items at $\theta = 1$. Further assume that these items can only appear in one group and it is only selected for the location that they provide more information than the other. Thus, a total of 480 discrete items out of 1200 are selected to be assembled in the MSTs. If the studied module length is 8, then only 8*3*8=192 items are needed for moderate modules; and 8*2*8=128 items are needed for hard modules. They are selected using the same procedure as previously described.

For MSTs constructed with only testlet items, the item information of each item in each testlet is calculated both at $\theta = 0$ and $\theta = 1$. If the measurement model is the 3PL model, the item information is calculated by using Equation 5. If the measurement model is the 3PL testlet model, the calculation of item information within a testlet is elaborated as follows:

According to Wainer, Bradlow, & Du (2000), the expected Fisher information at θ_i, for a single item in a testlet is given by

$$I(\theta_i) = a_j^2 \left(\frac{\exp[a_j(\theta_i - b_j - r_{id(j)})]}{1+\exp[a_j(\theta_i - b_j - r_{id(j)})]}\right)^2 \frac{1-c_j}{c_j+\exp[a_j(\theta_i - b_j - r_{id(j)})]} \qquad (14)$$

To calculate item information, a_j, b_j, and c_j in Equation 14 can be replaced with estimated parameters $\hat{a}_j$, $\hat{b}_j$, and $\hat{c}_j$. Li (2009)'s method of treating unknown $r_{id(j)}$ is applied here. Since testlet effect parameter $r_{id(j)}$ is independent of ability parameter θ, it is appropriate to obtain the expected information by taking the integral over $r_{id(j)}$. Thus, Equation 14 can be rewritten as:

$$I(\theta_i) = \int \{\hat{a}_j^2 \left(\frac{\exp[\hat{a}_j(\theta_i - \hat{b}_j - r_{id(j)})]}{1+\exp[\hat{a}_j(\theta_i - \hat{b}_j - r_{id(j)})]}\right)^2 \frac{1-\hat{c}_j}{\hat{c}_j+\exp[\hat{a}_j(\theta_i - \hat{b}_j - r_{id(j)})]}\} \varphi(r_{id(j)}) d(r_{id(j)}) \qquad (15)$$

where $\varphi(r_{id(j)})$ represents the distribution of $r_{id(j)}$.

In the MCMC estimation, $r_{id(j)}$ is assumed to be drawn from a normal distribution with mean of 0 and variance of $\sigma^2_{\gamma_{id(j)}}$. $\sigma^2_{\gamma_{id(j)}}$ is estimated during item pool calibration process. Quadrature points can be used to approximate the continuous distribution. Lesaffre & Spiessens (2001) suggested that it is often sufficient to use 10 quadrature points. 15 equally spaced quadrature points from -4 to 4 are used in this study. 15 is also the default number of quadrature points used in BILOG-MG. Thus, item information within a testlet can be further written as:

$$I(\theta_i) \cong$$

$$\sum_{K=1}^{15}\{\hat{a}_j^2\left(\frac{\exp[\hat{a}_j(\theta_i-\hat{b}_j-P_k(r_{id(j)}))]}{1+\exp[\hat{a}_j(\theta_i-\hat{b}_j-P_k(r_{id(j)}))]}\right)^2\frac{1-\hat{c}_j}{\hat{c}_j+\exp[\hat{a}_j(\theta_i-\hat{b}_j-P_k(r_{id(j)}))]}w(P_k(r_{id(j)}))\} \quad (16)$$

where $P_k(r_{id(j)})$ is the *kth* quadrature point and $w(P_k(r_{id(j)}))$ is the corresponding weight. The quadrature point weights are calculated using SAS PROBNORM function, a practice that has been applied by Raîche & Blais (2006) and Li (2009).

For each testlet, the items are sorted by their information. For a module that requires a testlet with 12 items, the 12 most informed items at $\theta = 0$ and $\theta = 1$ make a candidate module. Each testlet provides two candidate modules. The sum of the information provided by the 12 most informed items is the module information. Then the module information is sorted in descending order both at $\theta = 0$ and at $\theta = 1$. In total, 24 moderate testlets and 16 hard testlets are selected. 12 items are selected within each testlets.

In short, in this step, items are first screened inside testlets; testlets are then judged with the information that the selected items within that testlet can provide. Note that the same testlet with different combination of items can be selected to be in two candidate modules if it provides more information at both $\theta = 1$and $\theta = 0$ than the remaining testlets. To avoid that examinees may encounter the same testlet twice during the administration, it is constrained that a testlet cannot appear twice or more in the same panel.

For any MST that contains both discrete items and testlet items, items are selected with similar methods: discrete items are selected based on their information at the desirable location; testlets are selected based on their testlet information provided by the most informative items associated with it.

Phase II

This phase assigns selected items/testlets into different modules of the panels. As mentioned in Phase I, items/testlets are already selected based on their statistical properties. According to the desired properties of each module in each panel, the assignments of items or testlets to panels are done as follows.

For the panel constructed with only discrete items, the assignment of the selected items to the eight panels is done through the following two steps:

[Step 1]: The selected items that provide the most information at $\theta = 0$ are assigned to moderate modules of each panel. Initially, the first 1/3 of the most informative items are randomly assigned to the eight first-stage modules. For example, if the module length is 12, then 288 items have been selected to provide more information at $\theta = 0$. The first 96 of the 288 items are randomly assigned to the eight first-stage

panels. Then, the next 1/3 of the most informative items are randomly assigned as the second stage moderate modules. Finally, the remaining 1/3 of the selected items are randomly assigned to a panel as the third-stage moderate modules. Putting most informative items in early stages is based on the results of Zenisky (2004). She found that with limited amount of overall test information, it is better to get more information on early stages such that examinees can be routed into more appropriate difficulty levels of the modules of later stages.

[Step 2]: The selected items that provide the most information at $\theta = 1$ are assigned to hard modules of each panel. The same algorithm as in Step 1 is applied here: assign the first half of the items that has the most information at $\theta = 1$ randomly to the 8 hard modules on Stage 2; the remaining half items are randomly assigned to the hard panels on Stage 3.

For MSTs constructed with only testlet items, the assignments of testlets to modules are done through the following steps:

[Step 1]: The 24 selected moderate testlets are assigned to panels. First, the first 8 most informative testlets are randomly assigned to each of the eight first stage modules. Then, the next 8 most informative testlets are randomly assigned as the second stage moderate modules. Next, the remaining 8 are randomly assigned to a panel as the third stage moderate modules.

[Step 2]: The 16 hard testlets are assigned to panels. The same algorithm as in Step 1 is applied here: assign the 8 most informative testlets randomly to the 8 hard modules on Stage 2; the next 8 are randomly assigned to the hard panels on Stage 3.

[Step 3]: Check additional constraints. The constraints applied in this study include: no testlet appears more than once in the same panel, and no testlet appears more than twice in all panels.

[Step 4]: If there is violation of the constraints, repeat step [1-3] until there is no violation.

For any MST that has both discrete items and testlet items, discrete items are selected using the same procedure as the MST having only discrete items, and testlet items are selected using the same methods as those MSTs with only testlet items. Testlet items are then put into different stages according to the specified position in Table 2. And the general rule is that more informative items are put in earlier stages.

Step 3: Administration and Scoring of MSTs

Generation of Candidate Examinees' Responses

In this step, candidate examinees' responses to all items in all eight panels are simulated. Their responses are generated using the same algorithm as the examinees' responses; only that MST examinees' abilities are generated from a flat distribution.

Simulation of MST Administration

Each examinee is assigned to take each of the eight panels. The module in Stage 1 is first presented to the examinees. After he/she finishes the module in Stage 1, the ability estimate ($\hat{\theta}$) is computed using the EAP. If $\hat{\theta}$ is greater or equal to .5, the examinee is routed to the hard module in Stage 2; otherwise, the examinee is routed to take the moderate module in Stage 2. Similarly, at the end of Stage 2, $\hat{\theta}$ is updated based on his/her responses to those items in Stage 1 and 2. Again, if the $\hat{\theta}$ is greater or equal to .5, the examinee is routed to the hard module in Stage 3; otherwise, the moderate one is

administered. After Stage 3, the final ability is estimated with the entire set of responses using EAP. Classification decisions are made based on $\hat{\theta}$. If $\hat{\theta}$ is larger than or equal to 1, then the examinee passes the exam; otherwise, he or she fails the test.

It should be noted that though at first examinees' responses to all items in a panel are simulated, only those responses to the items in the selected routes are included in the analysis.

Data Analysis

This study evaluates the ability estimation through MST administrations in two perspectives: ability estimation precision and accuracy of classification decisions.

Evaluation of Ability Estimation

The ability estimation precision are assessed by the degree of true ability parameters (θ) recovered by the estimation ($\hat{\theta}$). To accomplish this purpose, two different measures are used in this study.

The first evaluation criterion is the bias. For each individual, bias is calculated as the difference between the estimated ability parameters and the true ability parameters. A positive value indicates that the ability was overestimated; negative value indicates that ability was underestimated. The second evaluation criterion is the root mean square error (rmse). It shows the extent to which the estimated ability estimate matches the true ability.

For each ability level with each simulation condition, the bias is calculated as:

$$\text{bias} = \frac{\sum(\hat{\theta}-\theta)}{n} \qquad (17)$$

For each ability level with each simulation condition, the rmse is calculated as:

$$\text{rmse} = \sqrt{\frac{\sum(\hat{\theta}-\theta)^2}{n}} \quad (18)$$

where n equals to 100 because at each ability level, there are 100 simulated examinees.

Using the 2500 (100 θs from -3 to 3 with increment of .25) examinee's response data to approximate a standard normal distribution, the overall BIAS and RMSE for the standard normal distribution with each simulation condition is calculated as follows:

$$BIAS = \frac{\sum \text{bias}*\text{weight}}{\sum \text{weight}} \quad (19)$$

$$RMSE = \sqrt{\frac{\sum(\text{rmse}^2)*\text{weight}}{\sum \text{weight}}} \quad (20)$$

where the weight is calculated using the SAS PROBNORM, the same practice used in previous calculation of quadrature point weights.

Evaluation of Decision Accuracy

The decision accuracy can be determined by comparing each simulated examinee's true ability (θ) and estimated score ($\hat{\theta}$) to the established cut score 1. The four possible classification decision outcomes are depicted in Figure 3.

Figure 3: A two-by-two table of possible decision classifications

		Estimated Ability	
		$\hat{\theta} > \theta_{cut}$	$\hat{\theta} < \theta_{cut}$
True Ability	$\theta > \theta_{cut}$	Correct Pass	False Negative
	$\theta < \theta_{cut}$	False Positive	Correct Fail

Correct-pass decisions occur when both the examinee's true ability and estimated score are greater or equal to the cut score. Conversely, *correct-fail* decisions occur when both the true and estimated ability score are below the cut score. *False-negative* errors

result when examinees who should pass, based on their true ability, fail to attain the passing score on a particular test form. *False-positive* errors occur when examinees that should not pass the examinee, based on their true ability, pass an examination by chance.

For each ability level, the decision accuracy rate is calculated as:

$$\text{da} = \frac{\text{correct pass} + \text{correct fail}}{\text{n}} \text{ , where n=100.} \quad (21)$$

The overall DA rate for the standard normal distribution for each simulation condition is calculated as follows:

$$DA = \frac{\sum \text{da} * \text{weight}}{\sum \text{weight}} \quad (22)$$

The above MST panel construction, MST administration and scoring, and data analysis are replicated 30 times for each simulation condition. All statistics are averaged across the 30 replications. Moreover, a series of analysis of variance (ANOVA) are applied to investigate the effects of studied factors. Finally, multiple comparison procedures are carried out if necessary.

For each ANOVA test, the partial eta-squared effect size measure is used to assess the degree of relationship of the dependent variables with the predictors. Specifically, the partial eta-squared measure describes the proportion of variance explained in the dependent variable by a factor partialling out other factors from the total non-error variation (Pierce, Block, & Aguinis, 2004).

The formula for partial eta squared is as follow:

$$\text{Partial } \eta^2 = \frac{SS_{\text{factor}}}{SS_{\text{factor}} + SS_{\text{error}}} \quad (19)$$

where SS_{factor} is the variation attributable to the factor and SS_{error} is the error variation. The general accepted regression benchmark for effect size is applied in this study: small=.01; medium=.06 and large=.14 (Cohen, 1988; Lomax, 2007; Stevens, 1992).

Chapter 4: Results

This chapter describes the results of the simulation study discussed in Chapter 3. It is divided into two sections. Each section addresses a research question.

Research Question I

To answer research question 1 "If the 3PL model is the measurement model, how are the measurement precision and classification decisions impacted by the proportion of testlet items in an MST, the position of the testlet items (which stage?), as well as the magnitude of LID?", the 3PL model is used to calibrate the item pools, to construct MST panels and to estimate examinees' abilities. The results of locally independent data (Pool 1) are described first, followed by the results of locally dependent data (Pool 2-10).

Results under Locally Independent Data

Figure 4 shows the results of bias and rmse at each ability level for the locally independent data (Pool 1) with the long test length. This condition represents the ideal condition where there is no local dependence in the data and the MST is long. The results under this condition serve as a baseline to which the other manipulated conditions are compared. It shows that lower end of abilities are overestimated and the large end of the abilities are underestimated. Largest magnitude of bias and rmse is obtained at the two tails of the ability level. Smallest magnitude of bias and rmse is obtained around θ=0.5. This is because the modules are designed to provide most information either at $\theta = 0$ or $\theta = 1$. A combination of the modules can provide most information at $\theta = .5$.

Figure 4: The bias and rmse plots under item independent condition

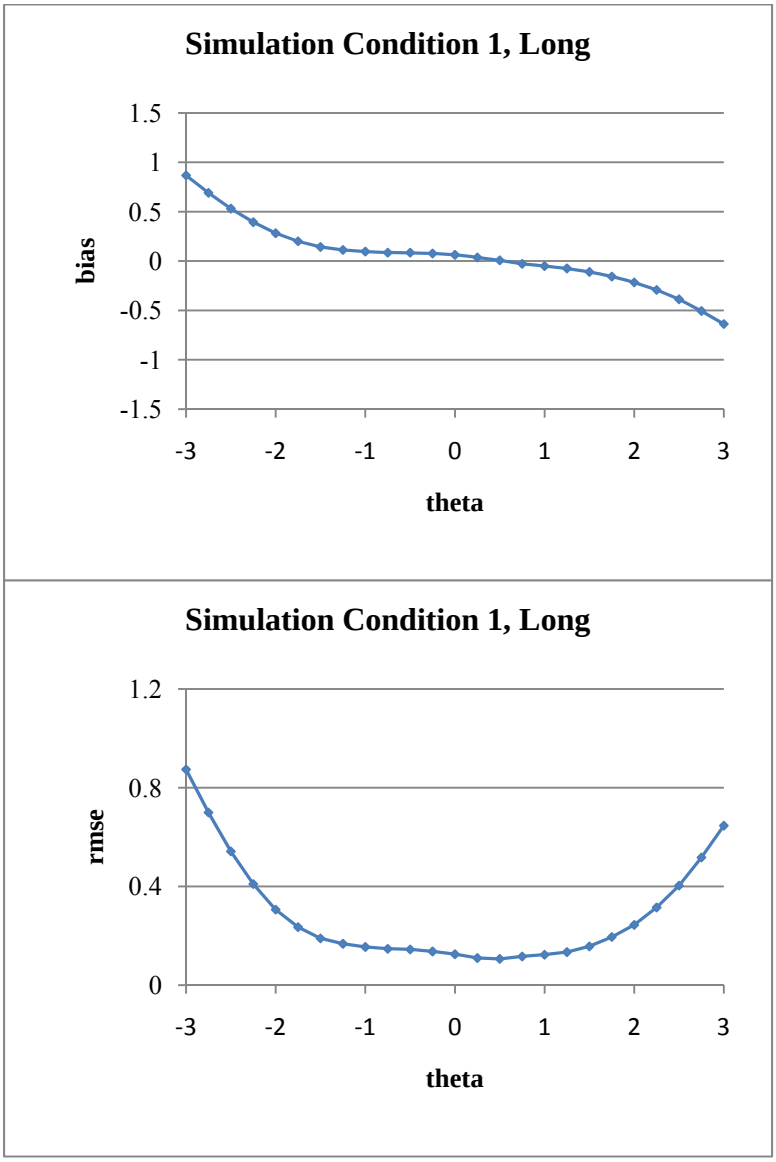


Table 4 contains the classification decisions for ability levels where examinees are misclassified. The frequencies in Table 4 are computed per 100 examinees for each ability level. Table 4(a) indicates false positive errors are made for examinees at ability level of $\theta = .75$. Table 4(b) indicates that false negative errors are made for examinees with ability level $\theta = 1$. Theoretically, the false negative errors and correct pass rates at the cut score of $\theta = 1$ should be 50% and 50%. Here a much lower false negative error

rate (32.8%) is observed because that EAP method is applied in this study to make ability estimation. EAP method intends to have estimates toward the mean of the prior distribution which is set at zero. False negative errors are observed at ability level of $\theta = 1.25$ (see Table 4(c)). The overall decision accuracy rate after correction for a standard normal distribution is 95.6%.

Table 4: Decision classifications under item local independence condition (Pool 1)

		Estimated Ability	
		$\hat{\theta} \geq \theta_{cut}$	$\hat{\theta} < \theta_{cut}$
True Ability $\theta = .75$	$\theta \geq \theta_{cut}$	0	0
	$\theta < \theta_{cut}$	.7	99.3

(a) True ability =.75

		Estimated Ability	
		$\hat{\theta} \geq \theta_{cut}$	$\hat{\theta} < \theta_{cut}$
True Ability $\theta = 1$	$\theta \geq \theta_{cut}$	32.8	68.2
	$\theta < \theta_{cut}$	0	0

(b) True ability =1

		Estimated Ability	
		$\hat{\theta} \geq \theta_{cut}$	$\hat{\theta} < \theta_{cut}$
True Ability $\theta = 1.25$	$\theta \geq \theta_{cut}$	94.8	5.2
	$\theta < \theta_{cut}$	0	0

(c) True ability =1.25

Results under Locally Dependent Data

Table 5 presents the averaged BIAS, RMSE, and DA rates over 30 replications under each simulation condition with the 3PL model. A total of 22 conditions are studied, in which the first simulation condition is the locally independent condition, and the remaining 21 are locally dependent data conditions. Item pool 1 has only discrete items. Item pool 8-10 has only testlet items. Thus the testlet/discrete item position factor is not investigated with these item pools. Depending on the proportion of testlet items in the MST, the testlet/discrete items are manipulated to appear on one of the three stages with item pool 2-7 according to Table 2. Table 5 also reports the evaluation criteria under two different test length conditions: long and short. Under the long condition, each module is composed of 12 items; under the short condition, each module has only 8 items.

BIAS evaluates the ability estimates against their true values. As shown in Table 5 the values of BIAS are all close to zero. This is probably because positive and negative biases cancel out each other between different ability levels and across the replications. There is no obvious trend of BIAS with different simulation conditions, except that the absolute values of BIAS under long test length conditions are smaller than those under short test length conditions which indicate that longer test yields better ability estimation.

RMSE represents the overall accuracy of ability estimates. Several trends can be observed from Table 5. First, under each simulation condition the long MST yields smaller RMSE. Second, with the same LID magnitude and the same testlet item proportion, the testlet/discrete item position effect appears to exist. For example, when the testlet item proportion equals to .33, with moderate or large LID magnitude, the RMSE decreases as the testlet-items' position changed from Stage 1 to Stage 3. However,

contrary to expectation, when the testlet item proportion equals to .67, with each studied LID magnitude, the RMSE also decreases as the discrete-items' position changed from Stage 1 to Stage 3. This phenomenon needs further investigation. Third, with the same proportion of testlet items, the RMSE increases as the LID magnitude increases. This result is consistent with those from other simulation studies of testlet effects (e.g., Bradlow, et. al., 1999; DeMars, 2006; Jiao & Wang, 2008; Wainer, et al. 2007). Fourth, with the same LID magnitude, the RMSE increases as the proportion of testlet items increases. It implies that the accuracy of ability estimates deteriorates as the level of model misspecification increases. The reason is that the 3PL model does not fit every item during the MST administration. As the proportion of testlet items increases, the level of model misfit increases.

The decision accuracy (DA) rates reported in Table 5 represent the proportion of examinees' true pass-fail status recovered through the MST administration under each simulation condition. The larger the DA rate, the better the recovery. Similar trends are observed with DA rates as the RMSE. First, with the same simulation condition, the long test yields higher DA rate. Second, with the same LID magnitude and the same proportion of testlet items, the effects of testlet/ discrete item position seem exist with decision accuracy. For example, when the testlet item proportion equals to .33, with each level of LID magnitude, the condition with testlet items positioned on Stage 1 produces the smallest DA; when the testlet item proportion equals to .67, with each level of LID magnitude, the condition with discrete items positioned on Stage 1 yields the largest DA. However, under each proportion of testlet items and each level of LID magnitudes, the DA rates with testlet/discrete items positioned on Stage 2 and Stage 3 are the same or

close with each other, which may imply that the position effect may not be significantly different between Stage 2 and Stage 3. Third, if the proportion of testlet items is kept the same, the DA rate decreases as the LID magnitude increases. Fourth, if the LID magnitude keeping the same, the DA rate decreases as the proportion of testlet items increases.

These results confirm our expectations that each of these studied factors may have an effect on the precision of the ability estimation and the accuracy of decision classifications with the MST design. Their effects are further studied with a series of ANOVA tests. The results are presented in the following sections.

Table 5: Evaluation criteria under the 3PL model

Simulation Condition	Item Pool	LID Magnitude	Testlet Item Proportion	Position	BIAS		RMSE		Decision Accuracy	
					Long	Short	Long	Short	Long	Short
1	1	0	0	N/A	0.039	0.038	0.167	0.205	0.956	0.948
2				tslt_s1	-0.014	-0.013	0.190	0.225	0.933	0.927
3				tslt_s2	-0.014	-0.013	0.185	0.223	0.939	0.933
4	2	0.25	0.33	tslt_s3	-0.013	-0.011	0.186	0.223	0.939	0.932
5				tslt_s1	-0.010	-0.006	0.209	0.246	0.930	0.922
6				tslt_s2	-0.011	-0.006	0.200	0.236	0.937	0.930
7	3	1	0.33	tslt_s3	-0.008	-0.004	0.199	0.236	0.938	0.931
8				tslt_s1	-0.002	0.000	0.216	0.253	0.930	0.923
9				tslt_s2	-0.007	-0.006	0.204	0.244	0.937	0.930
10	4	1.5	0.33	tslt_s3	-0.003	-0.001	0.203	0.243	0.938	0.929
11				dsct_s1	-0.011	-0.014	0.207	0.242	0.937	0.929
12				dsct_s2	-0.015	-0.016	0.205	0.240	0.932	0.926
13	5	0.25	0.67	dsct_s3	-0.015	-0.014	0.204	0.239	0.933	0.925
14				dsct_s1	-0.007	-0.005	0.240	0.279	0.934	0.925
15				dsct_s2	-0.006	-0.004	0.241	0.279	0.929	0.919
16	6	1	0.67	dsct_s3	-0.007	-0.003	0.238	0.277	0.929	0.919
17				dsct_s1	-0.005	-0.006	0.257	0.294	0.932	0.921
18				dsct_s2	-0.008	-0.008	0.261	0.298	0.923	0.914
19	7	1.5	0.67	dsct_s3	-0.008	-0.008	0.257	0.294	0.924	0.915
20	8	0.25	1	N/A	-0.018	-0.017	0.215	0.247	0.936	0.927
21	9	1	1	N/A	-0.014	-0.011	0.281	0.314	0.925	0.915
22	10	1.5	1	N/A	-0.011	-0.006	0.317	0.350	0.917	0.907

Note. tslt_s(*x*): testlet items positioned on Stage (x); dsct_s(x): discrete items positioned on Stage (x), *x*=1, 2, or 3.

Effect of Test Length

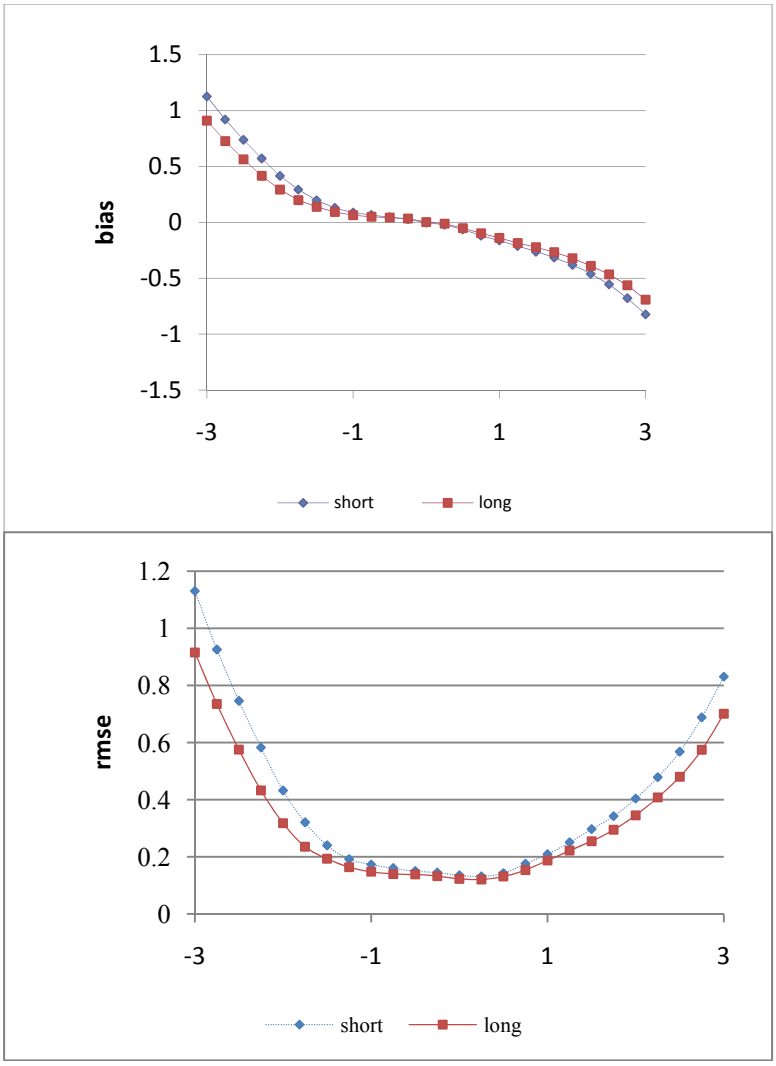
As discussed before, Table 5 shows that the BIAS and the RMSE under short test conditions are larger than those under long test conditions; the DA rates under short test conditions are lower than those under long test conditions. The 22 independent one-way ANOVA results are summarized in Table 6. The detailed ANOVA results for each simulation condition are presented in Appendix E-1. For all three evaluation criteria, the test length effects are found to be statistically significant under all simulation conditions (p-value$\leq$.05). This is what we expected that long tests produce high ability estimation precision. Reducing the test length would increase the standard error of measurement and thus decrease the decision accuracy. With the 22 simulation conditions, the averaged partial η^2 for BIAS is .549, the averaged partial η^2 for RMSE is .602; and the averaged partial η^2 for DAs is .581. According to the rule of thumb, the test length effect is large with all three evaluation criteria. Note that the experiment wise significanct level over 22 significance tests with a nominal .05 level is than than or equal to $.6765 = 1 - (1 - .05)^{22}$. Graphs are used to assist the interpretation of the ANOVA test results. The same caution are taken in the following multiple ANOVA tests.

Table 6: Summary statistics for the twenty-two ANOVA results for test length effect

Evaluation Criterion	Statistics	Sum of Squares	F-value	p-value	Partial eta squared
BIAS	min	0.000	0.550	0.000	0.009
	max	0.022	5420.458	0.461	0.989
	mean	0.007	1495.804	0.044	0.549
	std	0.010	2189.599	0.113	0.371
RMSE	min	0.000	0.000	0.000	0.000
	max	0.024	3924.519	0.985	0.985
	mean	0.007	885.472	0.094	0.602
	std	0.009	1288.030	0.248	0.378
Decision Accuracy	min	0.000	0.078	0.000	0.001
	max	0.024	5110.499	0.781	0.989
	mean	0.008	1360.944	0.067	0.581
	std	0.010	1848.775	0.193	0.391

Figure 5 compares the bias and the rmse curves between the two conditions of test length with simulation condition 2 along the ability distribution. In each graph, the X-axis is the ability scale; the Y-axis is the bias or the rmse. Higher magnitude of bias and rmse associated with short test length are observed at the two tales of the theta scale. This implies that the decreased overall ability estimation precision with the short test length is mainly attributable to the measurement of examinees with very high or very low abilities. The bias and the rmse curves for other locally dependent simulation conditions show similar pattern. They are presented in Appendix F-1.

Figure 5: Comparison of bias and rmse for test length effect using the 3PL model with simulation condition 2



Effect of Testlet/Discrete Item Position

From Table 1 we know that with item pool 2-4, testlet items are manipulated to appear on Stage 1, 2 or 3; with item pool 5-7, discrete items are manipulated to appear on Stage 1, 2 or 3. Since the test length is found to have effects on all three evaluation criteria, this section describes the results for the effects of testlet/discrete item position and possible interaction effects between test length and item positions.

The two-way ANOVA results are presented in Table 7 and Table 8. Based on the ANOVA outputs, the effect of testlet/discrete item position on each of the three evaluation criteria is found to be significant under all studied conditions except BIAS with item pool 2 and 6. Using the ANOVA outputs, it is reasonably to say that under small proportion of testlet items (item pool 2-4), the testlet item position has large effect on RMSE and DA. The partial η^2 for RMSE ranges from .361 to .840; and the partial η^2 for DA ranges from .467 to .616. The interaction between test length and testlet/discrete item position neither is significant (with p-value >.05) nor has large effect (with partial η^2 <.14) on each of the three evaluation criteria. Under moderate proportion of testlet items (item pool 5-7), the discrete item position has moderate effect on BIAS with small and large LID magnitudes (item pool 5 and 7) with partial η^2 values of .123 and .093 and small effect on BIAS with moderate magnitude (item pool 6) with the partial η^2 of .010. The discrete item position has large effect on RMSE with the partial η^2 ranges from .186 to .307, and large effect on DA with partial η^2 ranges from .280 to .563. The interaction between test length and testlet/discrete item position is neither significant (with p-value >.05) nor has large effect (with partial η^2 <.14) on each of the three evaluation criteria. In

summary, the position factor has large effect on RMSE and DA. There is no large effect caused by the interaction of test length and the position of testlet/discrete items.

Table 7: ANOVA results for testlet position, test length and their interaction effect with item pool 2-4

Item Pool	Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial Eta Squared
2	BIAS	Length	1	0.000	10.499	0.001	0.057
		Position	2	0.000	1.550	0.215	0.018
		Length*Position	2	0.000	0.363	0.696	0.004
	RMSE	Length	1	0.060	13037.115	0.000	0.987
		Position	2	0.000	49.074	0.000	0.361
		Length*Position	2	0.000	9.347	0.000	0.097
	Decision Accuracy	Length	1	0.002	202.210	0.000	0.537
		Position	2	0.001	76.129	0.000	0.467
		Length*Position	2	0.000	0.471	0.625	0.005
3	BIAS	Length	1	0.001	73.605	0.000	0.297
		Position	2	0.000	11.517	0.000	0.117
		Length*Position	2	0.000	1.902	0.152	0.021
	RMSE	Length	1	0.062	13690.309	0.000	0.987
		Position	2	0.004	427.482	0.000	0.831
		Length*Position	2	0.000	0.600	0.550	0.007
	Decision Accuracy	Length	1	0.002	255.138	0.000	0.595
		Position	2	0.003	139.375	0.000	0.616
		Length*Position	2	0.000	0.595	0.553	0.007
4	BIAS	Length	1	0.000	8.244	0.005	0.045
		Position	2	0.001	41.957	0.000	0.325
		Length*Position	2	0.000	0.162	0.850	0.002
	RMSE	Length	1	0.067	12423.334	0.000	0.986
		Position	2	0.005	457.733	0.000	0.840
		Length*Position	2	0.000	13.762	0.000	0.137
	Decision Accuracy	Length	1	0.003	336.220	0.000	0.659
		Position	2	0.002	119.061	0.000	0.578
		Length*Position	2	0.000	1.872	0.157	0.021

Table 8: ANOVA results for testlet position, test length and their interaction effect with item pool 5-7

Item Pool	Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial Eta Squared
5	BIAS	Length	1	0.000	1.650	0.201	0.009
		Position	2	0.000	12.230	0.000	0.123
		Length*Position	2	0.000	3.069	0.049	0.034
	RMSE	Length	1	0.057	8705.904	0.000	0.980
		Position	2	0.000	20.114	0.000	0.188
		Length*Position	2	0.000	0.153	0.858	0.002
	Decision Accuracy	Length	1	0.002	260.058	0.000	0.599
		Position	2	0.001	33.887	0.000	0.280
		Length*Position	2	0.000	2.981	0.053	0.033
6	BIAS	Length	1	0.000	26.720	0.000	0.133
		Position	2	0.000	0.858	0.426	0.010
		Length*Position	2	0.000	0.424	0.655	0.005
	RMSE	Length	1	0.067	11246.016	0.000	0.985
		Position	2	0.000	19.910	0.000	0.186
		Length*Position	2	0.000	4.567	0.012	0.050
	Decision Accuracy	Length	1	0.004	400.872	0.000	0.697
		Position	2	0.001	59.649	0.000	0.407
		Length*Position	2	0.000	0.181	0.835	0.002
7	BIAS	Length	1	0.000	0.024	0.878	0.000
		Position	2	0.000	8.900	0.000	0.093
		Length*Position	2	0.000	0.447	0.640	0.005
	RMSE	Length	1	0.063	8230.723	0.000	0.979
		Position	2	0.001	38.629	0.000	0.307
		Length*Position	2	0.000	0.371	0.691	0.004
	Decision Accuracy	Length	1	0.004	398.177	0.000	0.696
		Position	2	0.002	112.177	0.000	0.563
		Length*Position	2	0.000	1.503	0.225	0.017

Scheffe's multiple-comparison procedure is further applied to compare the mean differences of BIAS, RMSE, and DA with different testlet/discrete item positions. The results are summarized in Table 9 and Table 10, in which they suggest that the means of the interested criterion (BIAS, RMSE, or DA) on Stage 1 are significantly different from those on Stage 2 or Stage 3; and the mean differences may not be statistically significant between Stage 2 and Stage 3. For example, with large proportion of testlet items (item pool 5-7) the means of DA are not significantly different between discrete item positioned on Stage 2 and Stage 3. With small proportion of testlet items (item pool 2-4) the means of RMSE are not significantly different between testlet item positioned on Stage 2 and 3. In all studied conditions (item pool 2-7), the means of DA rates are not significantly different between testlet/discrete item positioned on Stage 2 and Stage 3.

Table 9: Group comparison results for testlet/discrete item position effect under long test length

Item Pool	Evaluation Criterion	Testlet Item Position		
		Stage 1	**Stage 2**	**Stage 3**
2	BIAS	A	A	A
	RMSE	A	B	B
	DA	A	B	B
3	BIAS	A	A	B
	RMSE	A	B	B
	DA	A	B	B
4	BIAS	A	B	A
	RMSE	A	B	B
	DA	A	B	B
Item Pool	**Evaluation Criterion**	**Discrete Item Position**		
		Stage 1	**Stage 2**	**Stage 3**
5	BIAS	A	B	B
	RMSE	A	A	B
	DA	A	B	B
6	BIAS	A	A	A
	RMSE	A	A	B
	DA	A	B	B
7	BIAS	A	B	B
	RMSE	A	B	A
	DA	A	B	B

*Note.*1. DA: Decision accuracy.

2. In the table, A, B, C are norminal values. They are only compared within each row.

3. Within each row, the same letter means that the means of an evaluation criterion between /among different groups are not significantly different; different letters indicate that there are significant mean differences between/among the comparison groups.

Table 10: Group comparison results for testlet/discrete item position effect under short test length

Item Pool	Evaluation Criterion	Testlet Item Position		
		Stage 1	**Stage 2**	**Stage 3**
2	BIAS	A	A	A
	RMSE	A	B	B
	DA	A	B	B
3	BIAS	A	B	B
	RMSE	A	B	B
	DA	A	B	B
4	BIAS	A	B	A
	RMSE	A	B	B
	DA	A	B	B
Item Pool	**Evaluation Criterion**	**Discrete Item Position**		
		Stage 1	**Stage 2**	**Stage 3**
5	BIAS	A	A	A
	RMSE	A	B	B
	DA	A	B	B
6	BIAS	A	A	A
	RMSE	A	B	B
	DA	A	B	B
7	BIAS	A	B	B
	RMSE	A	B	A
	DA	A	B	B

*Note.*1. DA: Decision accuracy.

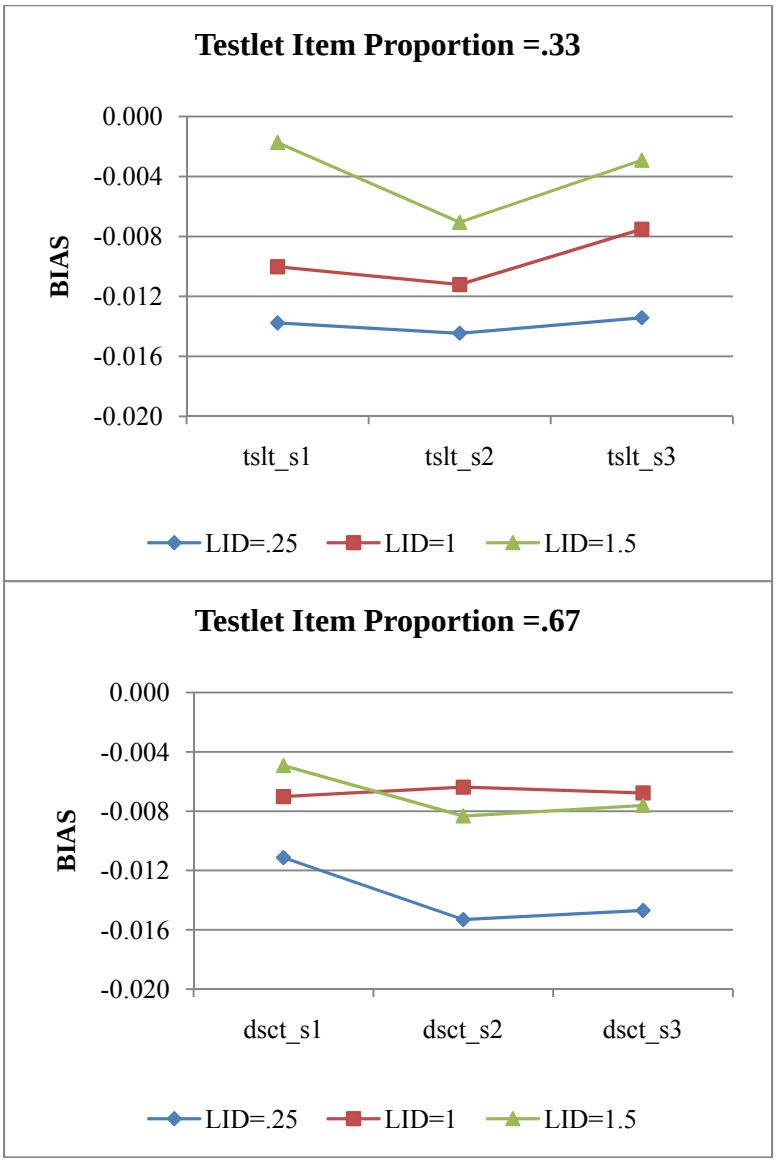
2. In the table, A, B, C are norminal values. They are only compared within each row.

3. Within each row, the same letter means that the means of an evaluation criterion between /among different groups are not significantly different; different letters indicate that there are significant mean differences between/among the comparison groups

The means of the three evaluation criteria with item pool 2-7 under both test length conditions are plotted in Figure 6 and Figure 7 respectively. In each figure, part (a) compares the BIAS differences; part (b) compares the RMSE differences; and part (c) compares the DA differences. In each part, two graphs which represent the proportion of

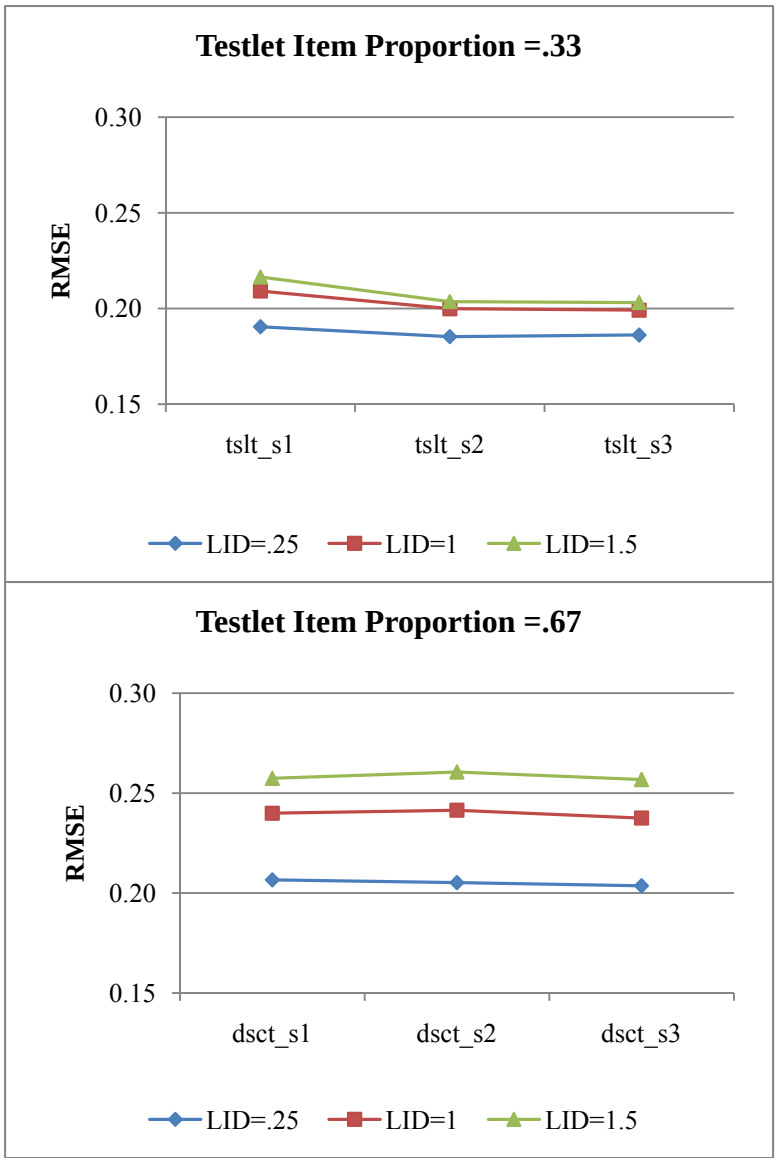
testlet items are presented side by side. In each graph, the X-axis specifies three levels of item positions; the Y-axis specifies an evaluation criterion; and each line represents a different LID magnitude which is shown in the legend. Figure 6(a) shows that under all conditions, the BIASes are smaller than zero. However, there is not a neat trend. For example, with testlet item proportion of .33, the absolute value of BIAS is largest when the testlet items are put on Stage 2. Figure 6(b) shows that with small proportion of testlet items, the RMSE changes slightly from the position of Stage 1 to the position of Stage 2; it almost remains the same as the position of testlet/discrete items changes from Stage 2 to Stage 3. It echoes with the previous finding that the mean differences of RMSE are not significant between positions of Stage 2 and 3. With moderate proportion of testlet items, the RMSE slightly decreases as the discrete items position changes from Stage 1 to Stage 2 and from Stage 2 to Stage 3. Figure 6(c) is based on the decision accuracy. With 1/3 of testlet items, the DA rate increases as the position of testlet items changes from Stage 1 to Stage 2; the DA rate stays almost at the same level as the position of testlet items changes from Stage 2 to Stage 3. With 2/3 of testlet items, the DA rate decreases as the position of discrete items changes from Stage 1 to Stage 2; the DA rate stays almost at the same level as the position of discrete items changes from Stage 2 to Stage 3.This result echoes with previous findings that the mean difference of DA is not significant between discrete items positioned on Stage 2 and Stage 3. Figure 7 shows the same trends as those observed in Figure 6. However, the differences of each evaluation criterion between positions of Stage 1 and Stage 2 in Figure 7 are smaller than those observed in Figure 6.

Figure 6: Comparison of testlet item positions with the 3PL model under the long test length condition



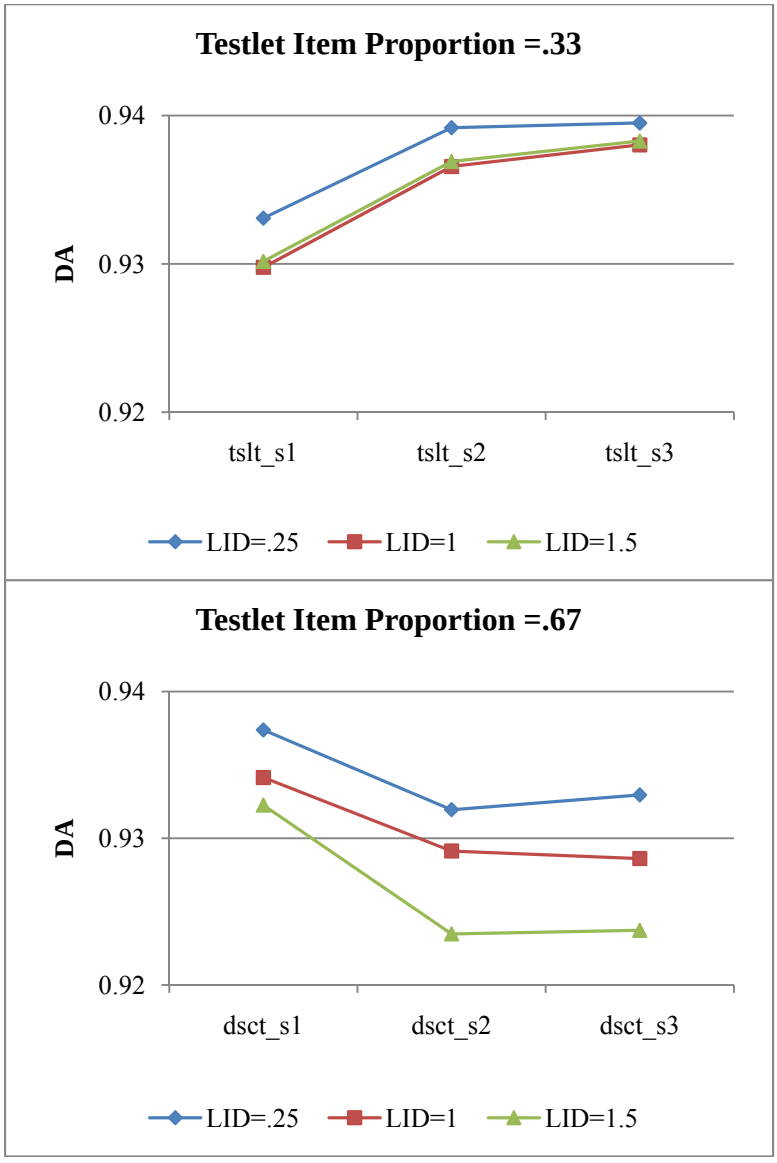
(a) BIAS

Figure 6, continued



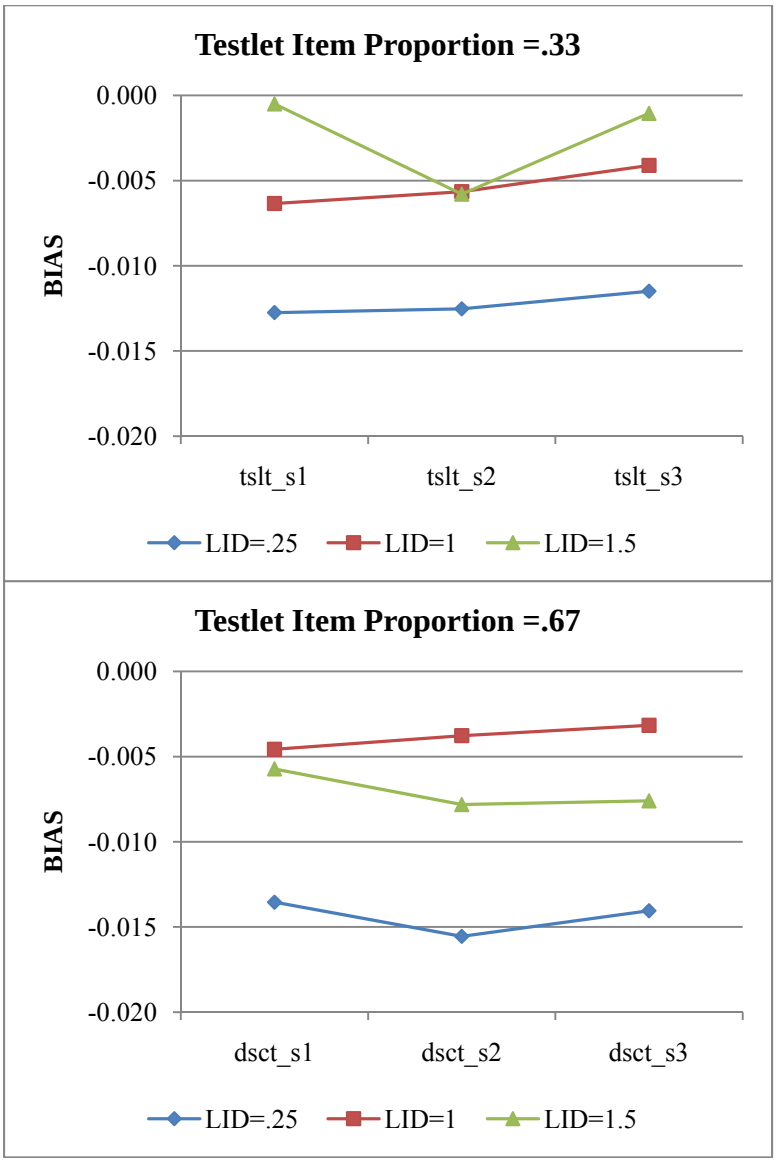
(b) RMSE

Figure 6, continued



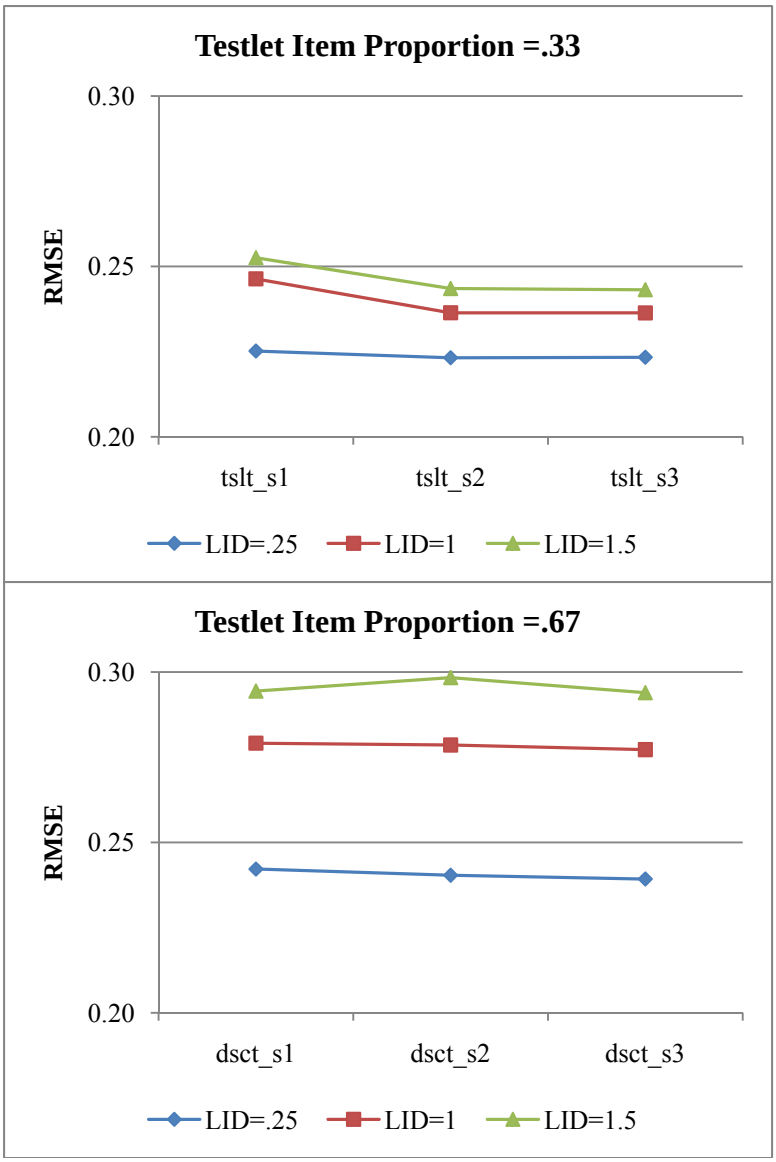
(c) Decision Accuracy

Figure 7: Comparison of testlet item positions under the 3PL model under the short test length condition



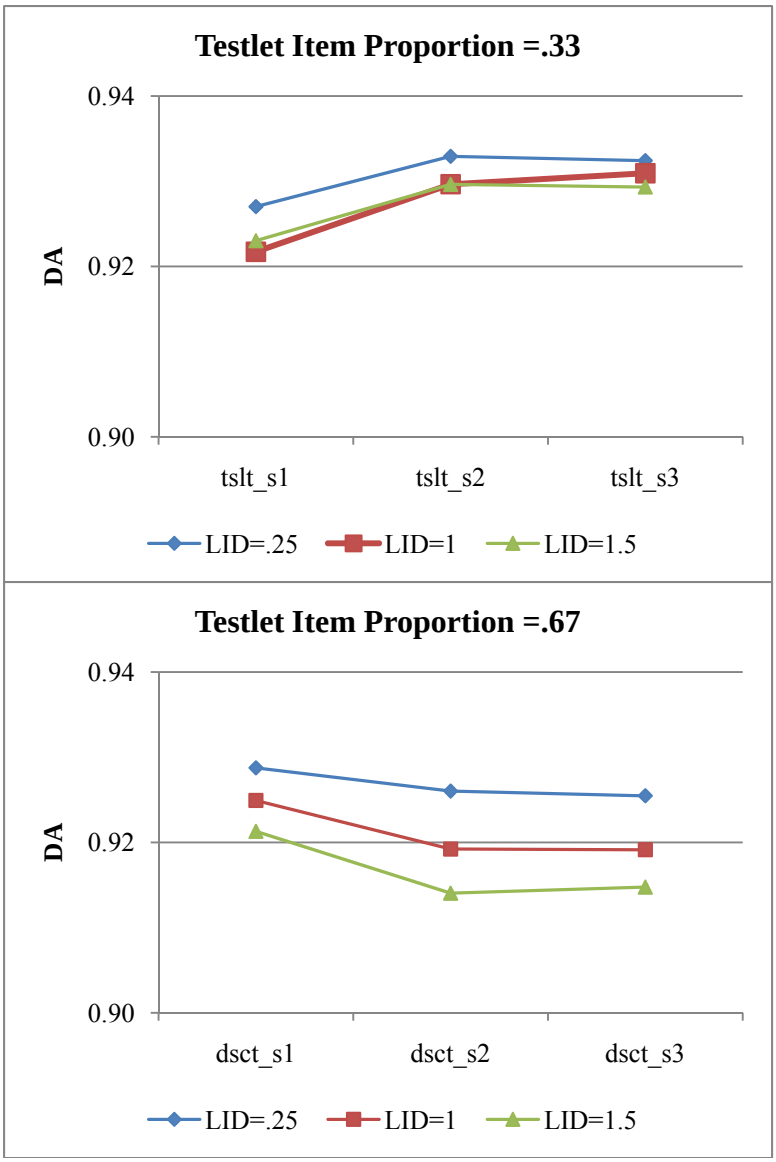
(a) BIAS

Figure 7, continued



(b) RMSE

Figure 7, continued

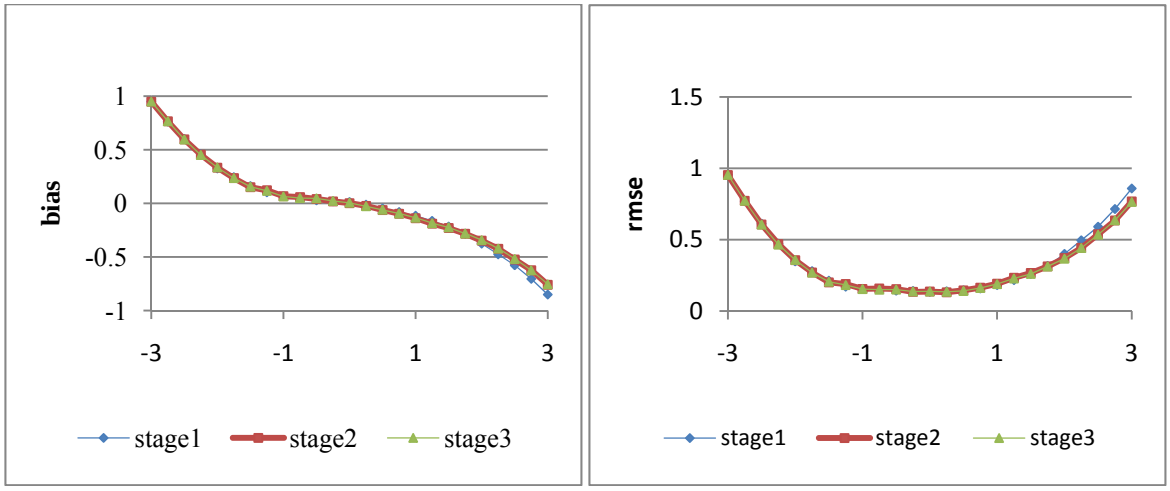


(c) Decision Accuracy

Figure 8 compares the bias and the rmse along the ability scale across the three discrete item positions with item pool 5 under the long test length condition. Each line in a graph represents a position of testlet/discrete items. In each graph, the three lines overlap with each other except at the tale of high abilities. At high ability levels, the

bias/rmse with the position of Stage 1 is quite different from those with the position of Stage 2 and 3, where the latter two stay close with each other. This indicates that the testlet/discrete item position mainly influences the estimation precision of high ability examinees. Since this study sets the cut score at 1 which is relatively high in the ability distribution, the accuracy of the final pass-fail decisions are influenced by the position of testlet/discrete items. Bias and rmse with other item pools (pool 2-4 and pool 6-7) show similar pattern. They are presented in Appendix F-2.

Figure 8: Comparison of bias and rmse for the item position effect with item pool 5



(a) bias

(b) rmse

Effect of Testlet Item Proportion

Figure 6 (b) and Figure 7(b) show that the RMSE under testlet item proportion of .33 are lower than those under testlet item proportion of .67. Figure 6(c) and Figure 7(c) indicate that the DAs under testlet item proportion of .33 are higher than those under testlet item proportion of .67. These results suggest that the testlet item proportion has effects on the precision of ability estimation and decision accuracy.

To study the effect of testlet item proportion and its possible interaction effect with test length on the MST design, the three evaluation criteria from item pool 2-7 are averaged across all testlet/discrete item positions with the same replication number. The averaged means for item pool 2-7 and means for item pool 1 and 8-10 under both test length conditions are reported in Table 11. These data are used in the following ANOVA analysis.

It appears that in Table 11 with the same test length condition and the same LID magnitude, the RMSE increases and the DA rate decreases as the proportion of testlet items increases. The BIAS does not have a clear trend.

Table 11: Averaged evaluation criteria with each item pool using the 3PL model

Item Pool	Testlet Item Proportion	LID Magnitude	BIAS		RMSE		Decision Accuracy	
			Long	Short	Long	Short	Long	Short
1	0	0	0.039	0.038	0.167	0.205	0.956	0.948
2	0.33	0.25	-0.014	-0.012	0.187	0.224	0.937	0.931
3	0.33	1	-0.010	-0.005	0.203	0.240	0.935	0.927
4	0.33	1.5	-0.004	-0.002	0.208	0.246	0.935	0.927
5	0.67	0.25	-0.014	-0.014	0.205	0.241	0.934	0.927
6	0.67	1	-0.007	-0.004	0.240	0.278	0.931	0.921
7	0.67	1.5	-0.007	-0.007	0.258	0.296	0.926	0.917
8	1	0.25	-0.018	-0.017	0.215	0.247	0.936	0.927
9	1	1	-0.014	-0.011	0.281	0.314	0.925	0.915
10	1	1.5	-0.011	-0.006	0.317	0.350	0.917	0.907

Several independent ANOVA tests are conducted to investigate whether there are significant mean differences with different proportion of testlet items. Based on the ANOVA outputs (see Table 12), it is reasonable to say that at each LID magnitude the proportion of testlet items has large effect on each of the three evaluation criteria (as their partial η^2 values are way above .14). The partial η^2 values for the interaction effect of test length and testlet item proportion on the RMSE range from .221 to .280, which indicates that the interaction effect on the precision of ability estimation is large. The partial η^2 values for the interaction of test length and the proportion of testlet items on the BIAS and DA are below .14 but larger than .01. These indicate that the interaction has at most moderate effect on BIAS and DA; and large effect on RMSE. In summary, the proportion of testlet items has large effect on RMSE and DA, and the interaction between test length and the proportion of testlet items has large effect on the RMSE, and at most moderate effect on BIAS and DA.

Table 12: ANOVA results for testlet item proportion, test length and their interaction effect

LID Magnitude	Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial Eta Squared
0.25	BIAS	Length	1	0.000	4.558	0.034	0.026
		Proportion	2	0.001	43.986	0.000	0.336
		Length*Proportion	2	0.000	3.489	0.033	0.039
	RMSE	Length	1	0.054	18049.265	0.000	0.990
		Proportion	2	0.020	3400.510	0.000	0.975
		Length*Proportion	2	0.000	33.311	0.000	0.277
	Decision Accuracy	Length	1	0.003	627.675	0.000	0.783
		Proportion	2	0.000	49.046	0.000	0.361
		Length*Proportion	2	0.000	8.204	0.000	0.086
1	BIAS	Length	1	0.001	66.211	0.000	0.276
		Proportion	2	0.002	108.088	0.000	0.554
		Length*Proportion	2	0.000	0.978	0.378	0.011
	RMSE	Length	1	0.059	15714.456	0.000	0.989
		Proportion	2	0.176	23424.953	0.000	0.996
		Length*Proportion	2	0.000	33.780	0.000	0.280
	Decision Accuracy	Length	1	0.003	772.367	0.000	0.816
		Proportion	2	0.004	406.357	0.000	0.824
		Length*Proportion	2	0.000	5.094	0.007	0.055
1.5	BIAS	Length	1	0.000	24.341	0.000	0.123
		Proportion	2	0.001	53.292	0.000	0.380
		Length*Proportion	2	0.000	12.699	0.000	0.127
	RMSE	Length	1	0.059	9401.560	0.000	0.982
		Proportion	2	0.341	27192.293	0.000	0.997
		Length*Proportion	2	0.000	24.739	0.000	0.221
	Decision Accuracy	Length	1	0.004	772.272	0.000	0.816
		Proportion	2	0.011	1145.551	0.000	0.929
		Length*Proportion	2	0.000	3.693	0.027	0.041

The group comparison results for the effect of the proportion of testlet items are presented in Table 13. It indicates that each of the evaluation criteria is significantly different under each proportion of the testlet items. The exceptions are the BIAS for the small LID magnitude under long test length condition and BIAS for moderate LID magnitude under short test length condition in which the BIAS may be statistically not significant different between two adjacent testlet item proportions. The DA rates under small LID magnitude under short test length condition are not significantly different between proportion of .67 and 1 neither.

Table 13: Group comparison results for testlet item proportion effect

Test length	LID Magnitude	Dependent Variable	Testlet Item Proportion		
			0.33	0.67	1
Long	0.25	BIAS	A	A	B
		RMSE	A	B	C
		DA	A	B	C
	1	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	C
	1.5	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	C
Short	0.25	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	B
	1	BIAS	A	B	B
		RMSE	A	B	C
		DA	A	B	C
	1.5	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	C

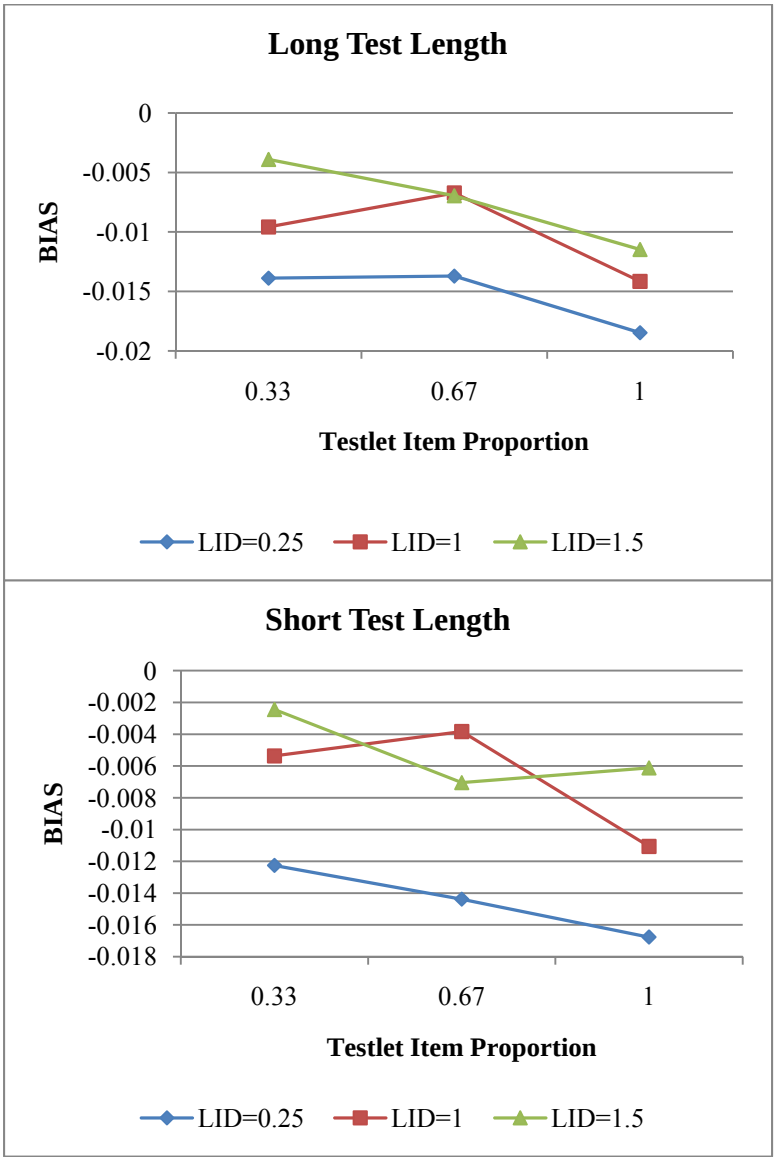
*Note.*1. DA: Decision accuracy.

2. In the table, A, B, C are norminal values. They are only compared within each row.

3. Within each row, the same letter means that the means of an evaluation criterion between /among different groups are not significantly different; different letters indicate that there are significant mean differences between/among the comparison groups

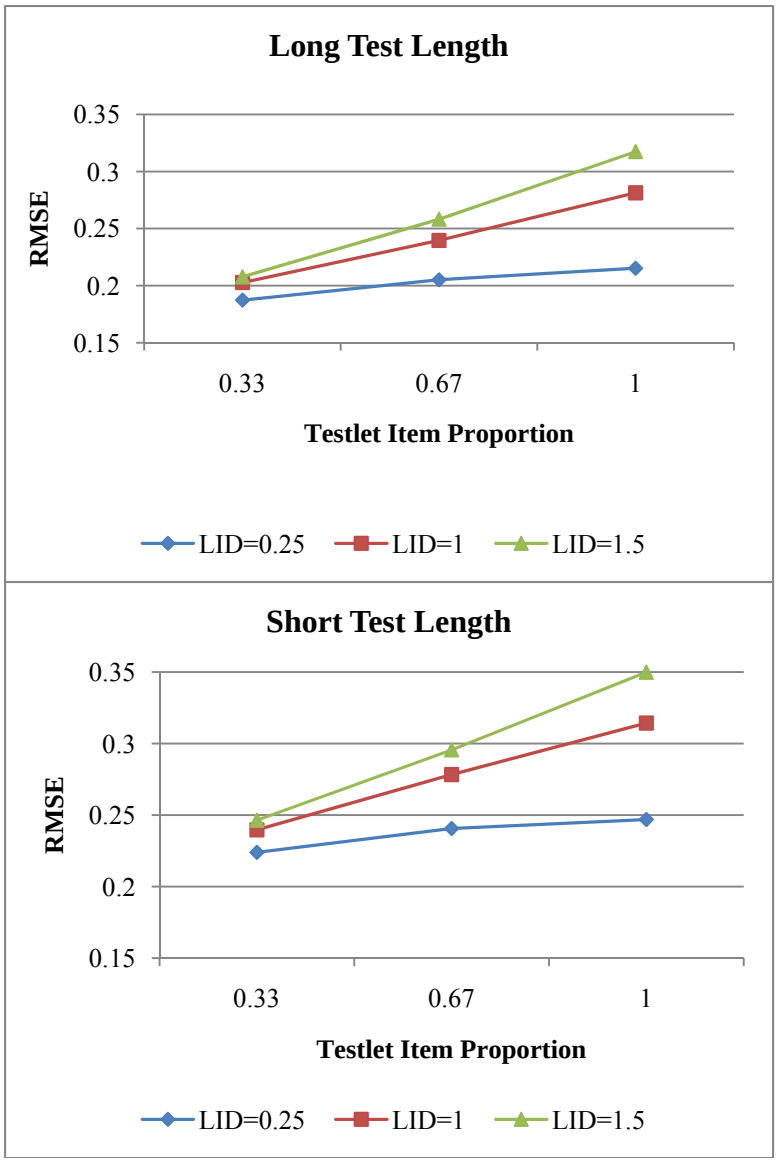
Figure 9 conveys the same message as that indicated in Table 5 and Table 11. With each level of the LID magnitude, there is no clear trend of BIAS as the proportion of testlet items increases; the RMSE increases and the DA decreases as the proportion of testlet items increases.

Figure 9: Evaluation criteria with different LID magnitude and testlet item proportion I



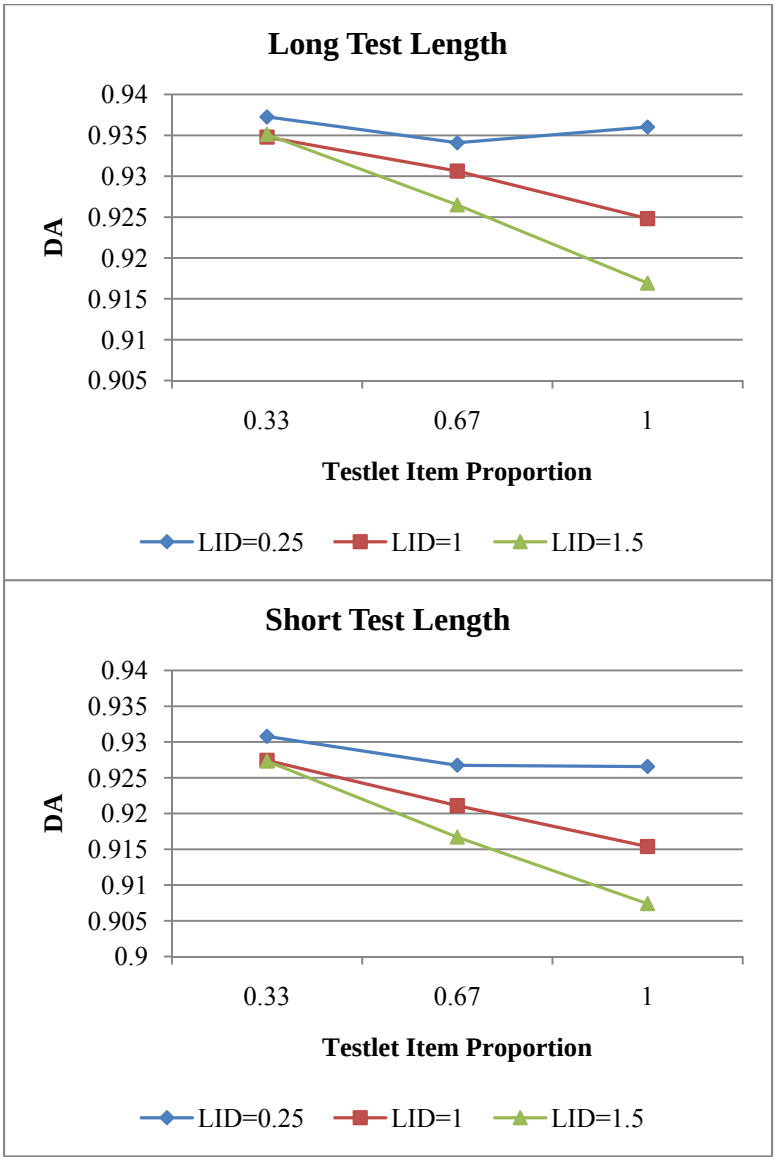
(a) BIAS

Figure 9, continued



(b) RMSE

Figure 9, continued

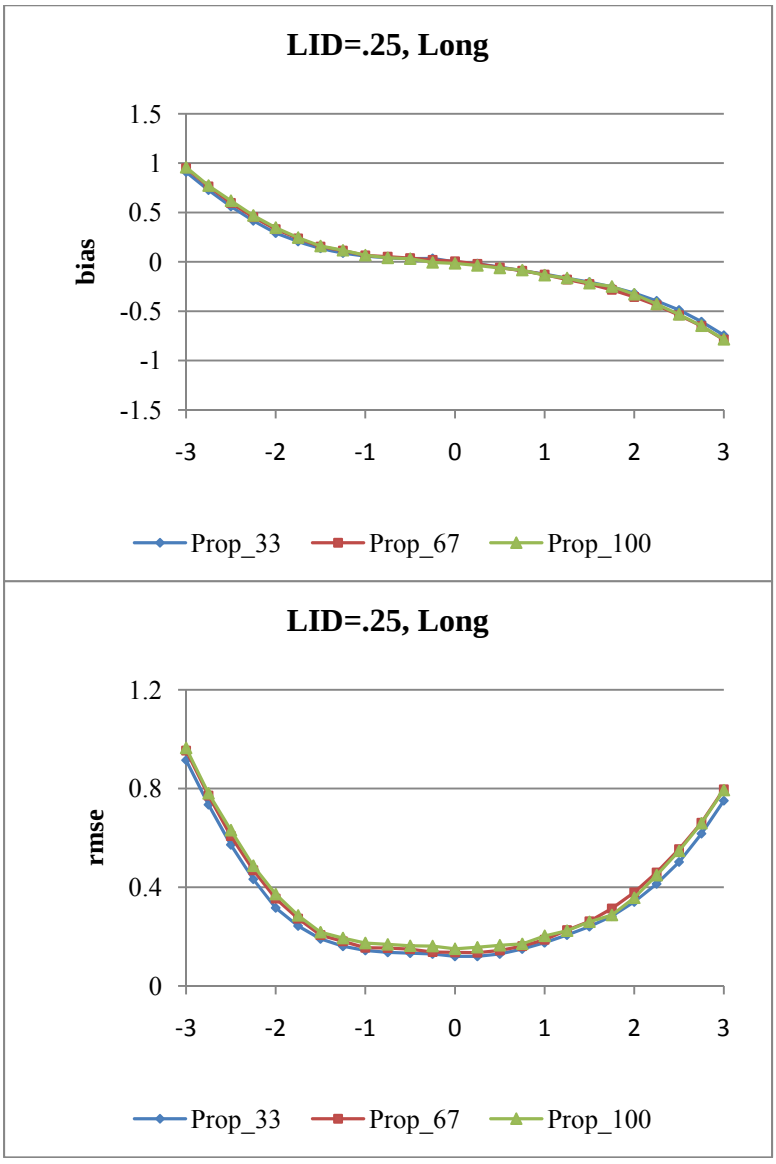


(c) Decision accuracy

Figure 10 compares the bias and the rmse across the three proportions of testlet items with the long test length. When the LID magnitude is small, Figure 10(a) shows very similar degrees of bias and rmse throughout the entire theta range across the three proportions, with smallest bias and rmse in the middle of ability distribution and increasing values as the ability become more extreme. When the LID magnitude is

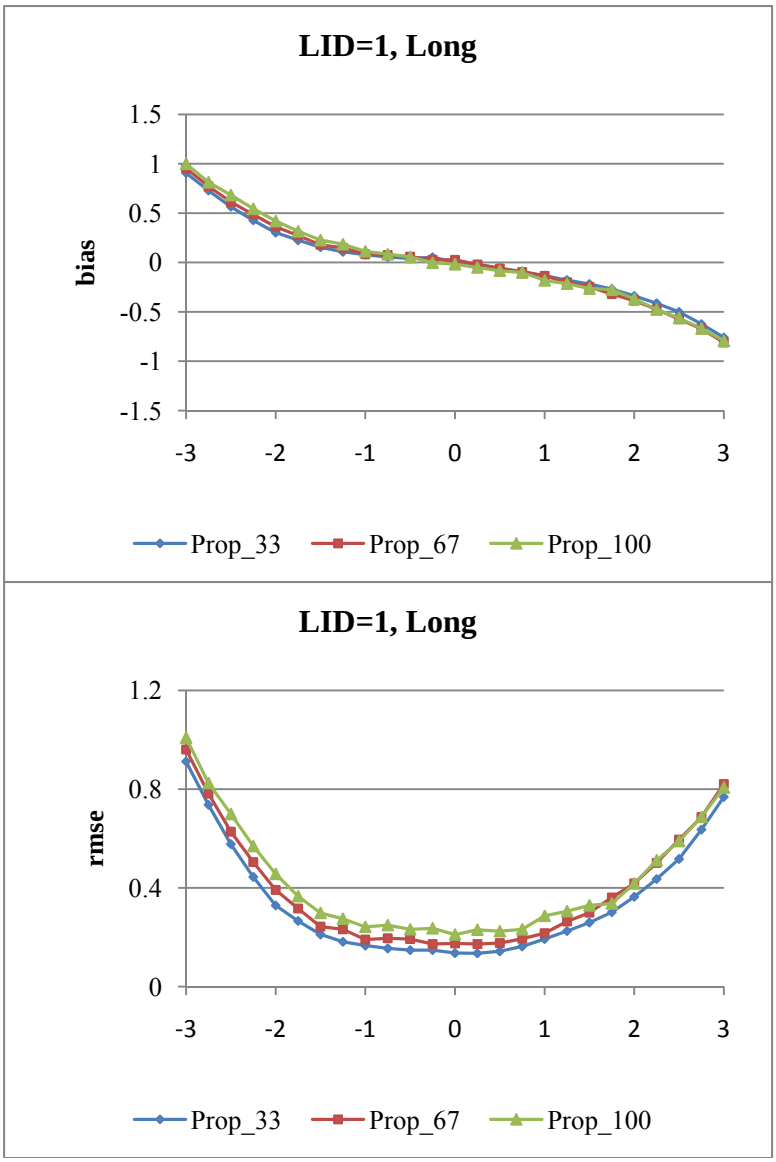
moderate or large, Figure 10(b) and Figure 10(c) show that the values of the bias and rmse also increases at each ability level as the proportion of testlet items increases. The pattern is consistent with those reported in Table 13. The bias and the rmse across the three proportions of testlet items with the short test length condition show similar pattern. They are presented in Appendix F-3.

Figure 10: Comparison of bias and rmse across different testlet item proportion levels with long test length condition



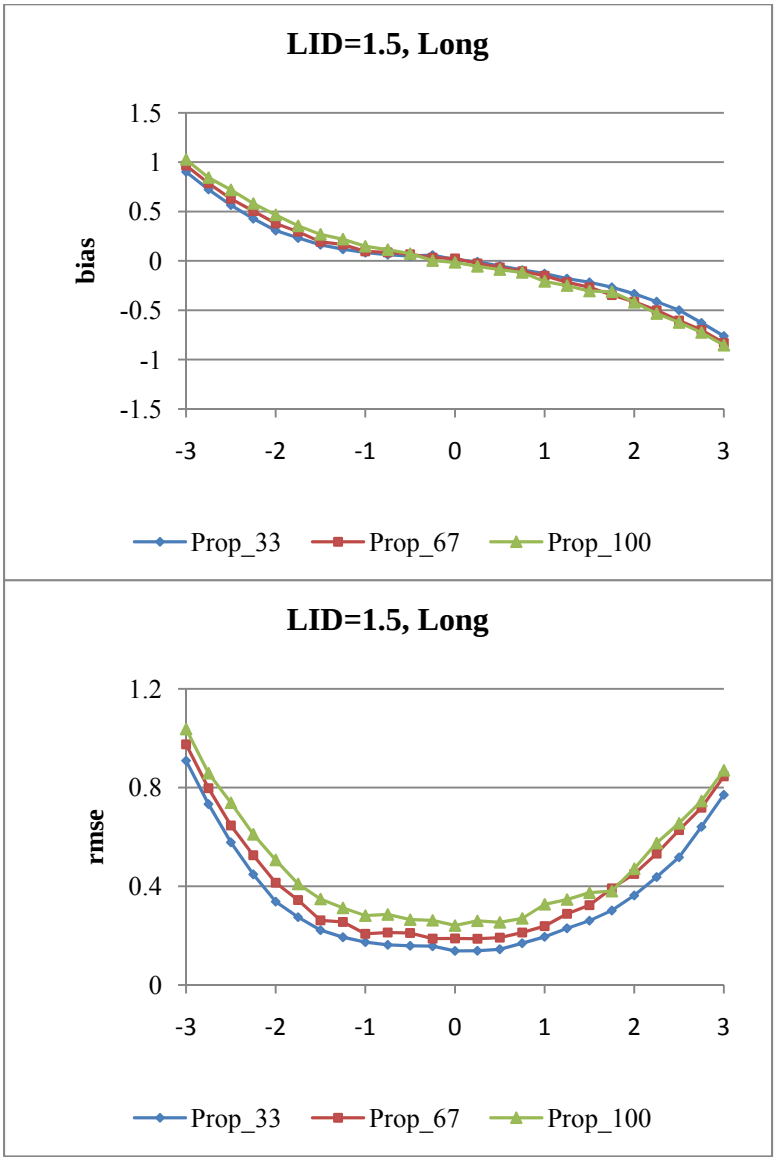
(a) Small LID magnitude, long test length

Figure 10, continued



(b) Moderate LID magnitude, long test length

Figure 10, continued



(c) Large LID magnitude, long test length

Effect of LID Magnitude

It appears in Table 5 and Table 11 that the RMSE increases and the DA decreases as the magnitude of LID increases under both test length conditions. In each graph of Figure 9, it shows that the RMSE of LID magnitude of .25 is lower than that of LID magnitude of 1 or 1.5 and the DA of LID magnitude of .25 is higher than that of LID

magnitude of 1 or 1.5. Figure 11 plots these trends. All these results suggest that the LID magnitude also has an effect on the precision of ability estimation and decision accuracy.

The results in previous sections indicate that the test length and the proportion of testlet items have effect on the precision of ability estimation and decision accuracy. Here a three-way ANOVA test is carried out to test the significance of the LID magnitude effect using the same dataset that is applied to test the effects of the proportion of testlet items.

The ANOVA results (See Table 14) suggest that the LID magnitude has large effect on each of the three evaluation criteria (with partial η^2 values of .624, .803 and .993 respectively). The interaction between the test length and the LID magnitude has small effect on each of the three evaluation criteria (with partial η^2 values of .032, .034 and .017 respectively). The partial η^2 value for the interaction between testlet item proportion and LID magnitude is .138 for BIAS which is very close to the threshold of large effect. The interaction effects between testlet item proportion and LID magnitude are large on RMSE and DA (with partial η^2 values of .978 and .638 respectively). The interaction effect among test length, testlet item proportion and LID magnitude is small on each of the three evaluation criteria (with partial η^2 values of .031, .011 and .021).

Table 14: Three-way ANOVA results for test length, testlet item proportion, LID magnitude and their interaction effect

Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial Eta Squared
BIAS	LEG	1	0.001	76.869	0.000	0.128
	PRO	2	0.003	161.150	0.000	0.382
	MAG	2	0.007	432.434	0.000	0.624
	LEG*PRO	2	0.000	9.932	0.000	0.037
	LEG*MAG	2	0.000	8.523	0.000	0.032
	PRO*MAG	4	0.001	20.839	0.000	0.138
	LEG*PRO*MAG	4	0.000	4.218	0.002	0.031
RMSE	LEG	1	0.172	39605.718	0.000	0.987
	PRO	2	0.435	50176.347	0.000	0.995
	MAG	2	0.328	37862.280	0.000	0.993
	LEG*PRO	2	0.001	82.251	0.000	0.240
	LEG*MAG	2	0.000	9.060	0.000	0.034
	PRO*MAG	4	0.102	5885.130	0.000	0.978
	LEG*PRO*MAG	4	0.000	2.849	0.023	0.021
DA	LEG	1	0.010	2169.765	0.000	0.806
	PRO	2	0.011	1197.180	0.000	0.821
	MAG	2	0.010	1063.647	0.000	0.803
	LEG*PRO	2	0.000	13.817	0.000	0.050
	LEG*MAG	2	0.000	4.550	0.011	0.017
	PRO*MAG	4	0.004	229.756	0.000	0.638
	LEG*PRO*MAG	4	0.000	1.479	0.207	0.011

Note. DA: Decision accuracy;
LEG: test length;
PRO: testlet item proportion;
MAG: LID magnitude.

The group comparison results are presented in Table 15. It indicates that the RMSE is strongly associated with the LID magnitude. With other factor's conditions keeping constant, the RMSE at one LID magnitude is significantly different from the RMSE at another LID magnitude. The DA rate also has a strong relationship with the LID magnitude. But when the testlet item proportion is small (.33), LID magnitude of 1 and 1.5 may produce the same level of DA rates. The relationship between the BIAS and the LID magnitude is relatively weak. Moderate and large LID magnitudes may produce the same level of BIAS, for example, with all testlet items under short test length condition. This is probably due to that the positive and negative BIAS cancel out each other during the replication and across different positions with item pool 2-7.

Table 15: Group comparison results for LID magnitude effect

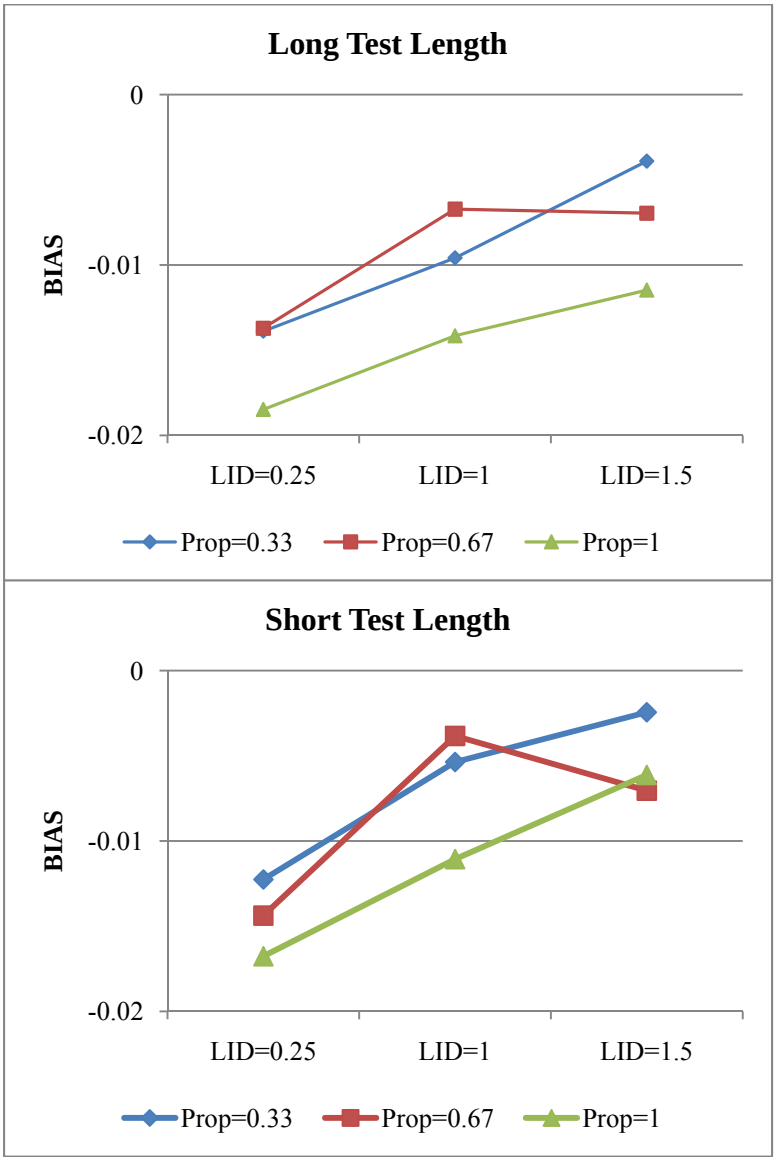
Test length	Testlet Item Proportion	Dependent Variable	LID Magnitude		
			0.25	1	1.5
Long	0.33	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	B
	0.67	BIAS	A	B	B
		RMSE	A	B	C
		DA	A	B	C
	1	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	C
Short	0.33	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	B
	0.67	BIAS	A	B	C
		RMSE	A	B	C
		DA	A	B	C
	1	BIAS	A	B	B
		RMSE	A	B	C
		DA	A	B	C

*Note.*1. DA: Decision accuracy.

2. In the table, A, B, C are norminal values. They are only compared within each row.

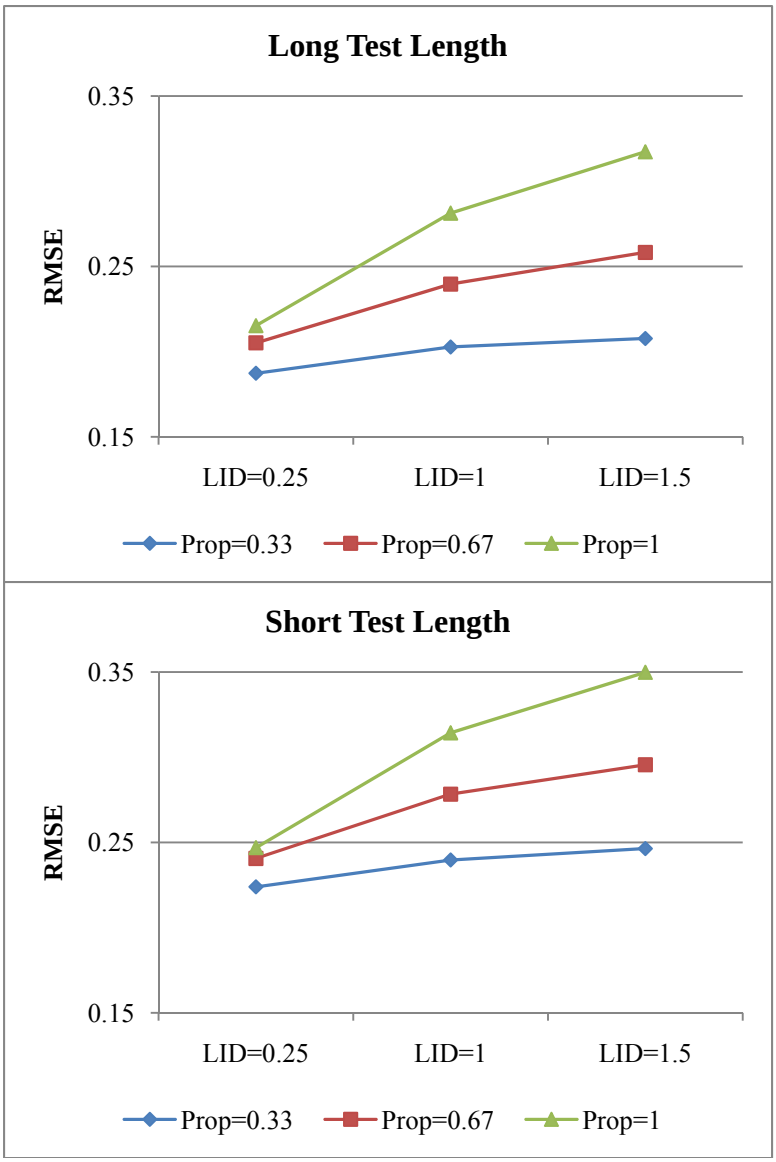
3. Within each row, the same letter means that the means of an evaluation criterion between /among different groups are not significantly different; different letters indicate that there are significant mean differences between/among the comparison groups.

Figure 11: Evaluation criteria under different LID and testlet item proportions II



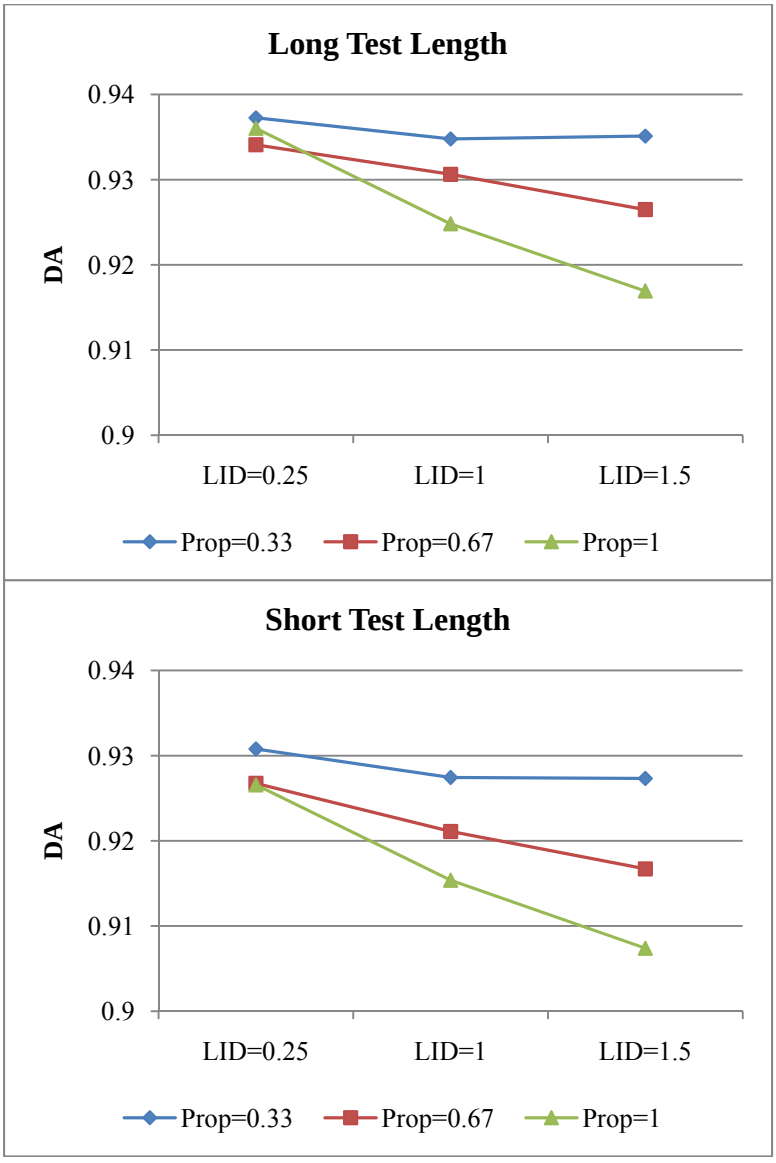
(a) BIAS

Figure 11, continued



(b) RMSE

Figure 11, continued



(c) Decision Accuracy

Previous results also suggest that the position of testlet/discrete items has effect on both ability estimation and decision accuracy. Thus, the effect of the LID magnitude, the position of testlet/discrete items, the test length, and their interaction effects are studied with item pool 2-4 and item pool 5-7 separately. The results are presented in Table 16 and 17.

Using the ANOVA output (Table 16) with the item pool 2-4 data, it is reasonable to say that with small proportion of testlet items, the interaction between the LID magnitude and the testlet item position has moderate effect on BIAS, large effect on RMSE and small effect on DA because their partial $\boldsymbol{\eta^2}$ values are .087, .357 and .025 respectively. The interaction effects among the LID magnitude, the position of testlet/discrete items, and the test length has no or small effect on each of the three evaluation criteria because their p-values are either larger than .05 or the partial $\boldsymbol{\eta^2}$ value smaller than .06. Using the ANOVA output (Table 17) for the item pool 5-7 data, it is reasonable to say that with moderate proportion of testlet items, the interaction between LID magnitude and testlet item position has small effect on BIAS, moderate effect on RMSE and small effect on DA because their partial $\boldsymbol{\eta^2}$ values are .041, .103 and .054 respectively. The interaction effects among the LID magnitude, the position of testlet/discrete items, and the test length has ignorable effect on each of three evaluation criteria because their p-values are larger than .05. In summary, the interaction between LID magnitude and the position of testlet items has effect on each of three evaluation criteria. However, these effects decrease as the proportion of testlet items increases. The interaction effect among LID magnitude, the position of testlet/discrete items and the test

length has no effect on BIAS and DA. And the interaction has small effect on RMSE. As the proportion of testlet items increases, the interaction effect on RMSE among LID magnitude, the position of testlet/discrete items and the test length becomes ignorable.

Table 16: ANOVA results for testlet/discrete item position, LID magnitude, test length, and their interaction effects with item pool 2-4 data

Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial eta square
BIAS	LEG	1	0.001	70.933	0.000	0.120
	POS	2	0.001	30.804	0.000	0.106
	MAG	2	0.009	392.716	0.000	0.601
	LEG*POS	2	0.000	0.886	0.413	0.003
	LEG*MAG	2	0.000	9.538	0.000	0.035
	POS*MAG	4	0.001	12.485	0.000	0.087
	LEG*POS*MAG	4	0.000	0.740	0.565	0.006
RMSE	LEG	1	0.189	39008.391	0.000	0.987
	POS	2	0.008	810.139	0.000	0.756
	MAG	2	0.044	4567.186	0.000	0.946
	LEG*POS	2	0.000	13.935	0.000	0.051
	LEG*MAG	2	0.000	11.304	0.000	0.042
	POS*MAG	4	0.001	72.495	0.000	0.357
	LEG*POS*MAG	4	0.000	5.464	0.000	0.040
DA	LEG	1	0.007	779.401	0.000	0.599
	POS	2	0.006	327.478	0.000	0.556
	MAG	2	0.001	54.486	0.000	0.173
	LEG*POS	2	0.000	1.036	0.356	0.004
	LEG*MAG	2	0.000	2.282	0.103	0.009
	POS*MAG	4	0.000	3.369	0.010	0.025
	LEG*POS*MAG	4	0.000	0.888	0.471	0.007

Note. DA: Decision Accuracy;
POS: Testlet/discrete item position;
LEG: test length;
MAG: LID magnitude.

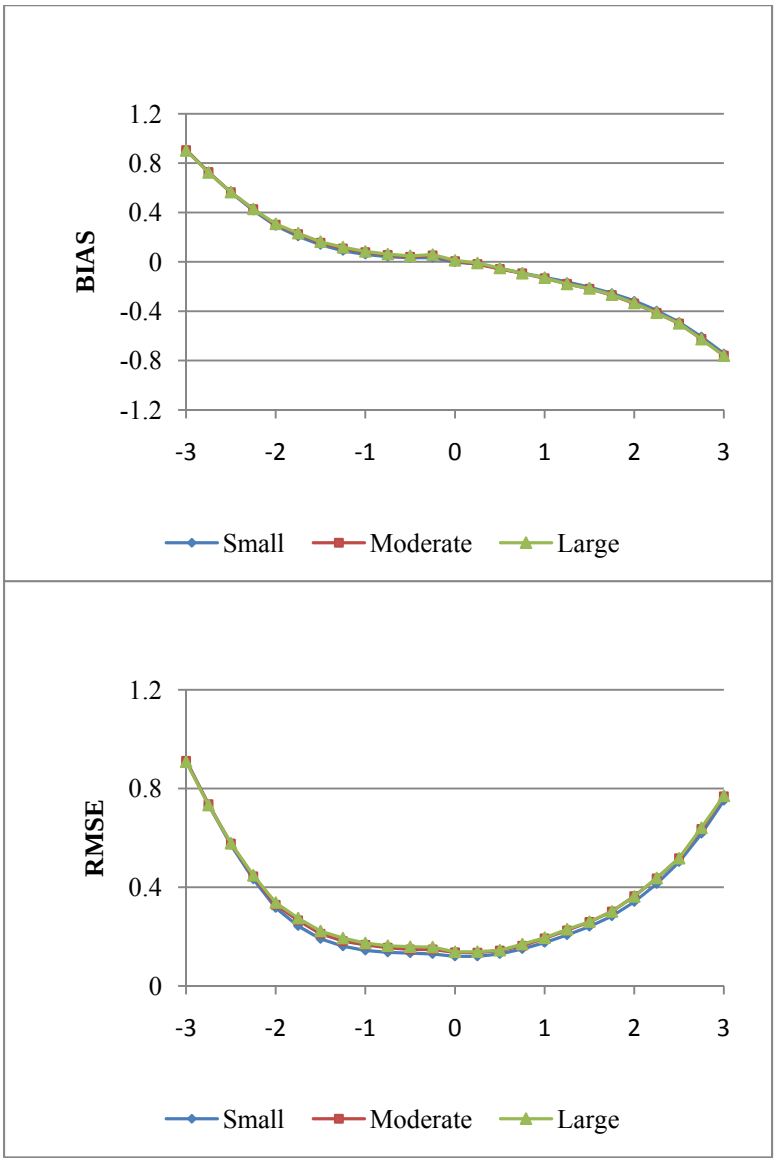
Table 17: ANOVA results for testlet/discrete item position, LID magnitude, test length, and their interaction effects with item pool 5-7 data

Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial eta square
BIAS	LEG	1	0.000	5.032	0.025	0.010
	POS	2	0.000	10.084	0.000	0.037
	MAG	2	0.008	286.040	0.000	0.523
	LEG*POS	2	0.000	2.495	0.083	0.009
	LEG*MAG	2	0.000	12.065	0.000	0.044
	POS*MAG	4	0.000	5.635	0.000	0.041
	LEG*POS*MAG	4	0.000	0.584	0.675	0.004
RMSE	LEG	1	0.187	27811.321	0.000	0.982
	POS	2	0.001	51.236	0.000	0.164
	MAG	2	0.273	20340.704	0.000	0.987
	LEG*POS	2	0.000	1.094	0.336	0.004
	LEG*MAG	2	0.000	17.424	0.000	0.063
	POS*MAG	4	0.000	14.985	0.000	0.103
	LEG*POS*MAG	4	0.000	1.778	0.132	0.013
DA	LEG	1	0.011	1053.817	0.000	0.669
	POS	2	0.004	196.400	0.000	0.429
	MAG	2	0.007	346.604	0.000	0.570
	LEG*POS	2	0.000	1.756	0.174	0.007
	LEG*MAG	2	0.000	8.018	0.000	0.030
	POS*MAG	4	0.000	7.494	0.000	0.054
	LEG*POS*MAG	4	0.000	1.392	0.235	0.011

Note. DA: Decision Accuracy
POS: Testlet/discrete item position;
LEG: test length;
MAG: LID magnitude.

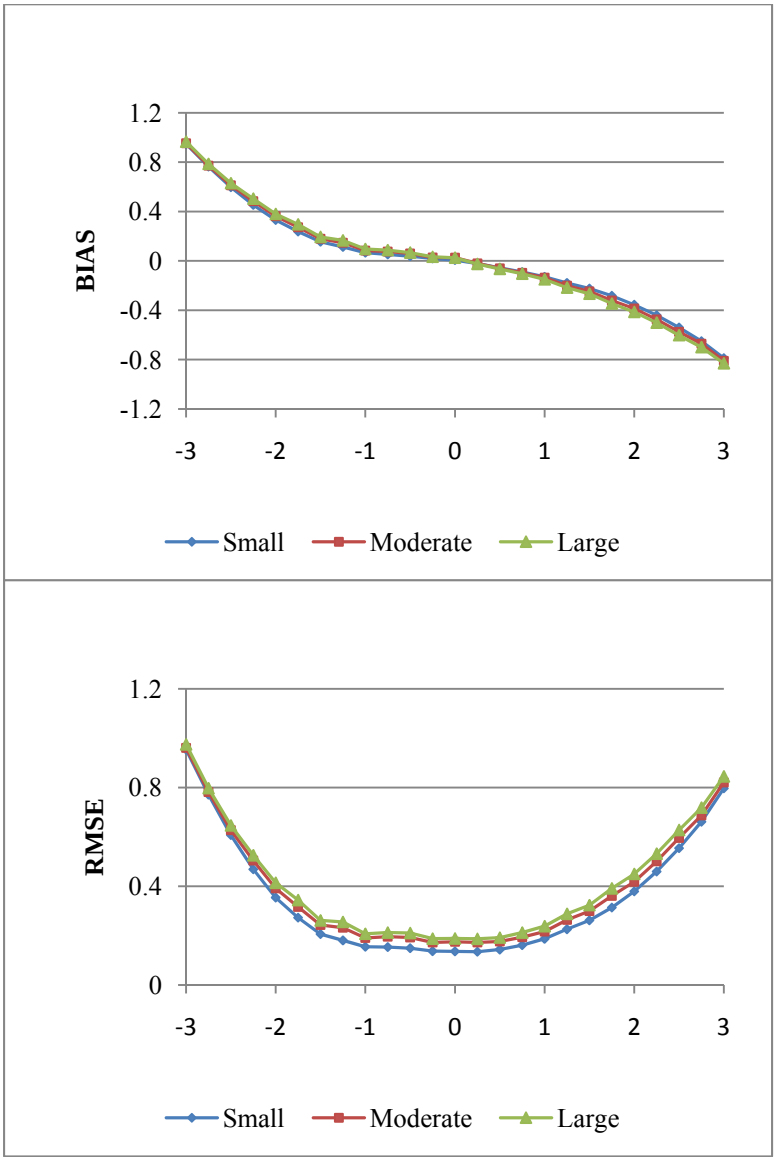
Figure 12 compares the bias and the rmse functions across the three LID magnitudes with the long test length condition. Figure 12 (a) shows that the bias and the rmse functions for the three LID magnitudes are virtually close to each other throughout the entire theta range when the proportion of testlet items is small. When the proportion of testlet items is moderate or large, Figure 12(b) and Figure 12(c) show that the bias and the rmse increases as the LID magnitude increases at each ability level. The differences of bias and rmse between any two LID magnitudes become large as the proportion of testlet items becomes large.

Figure 12: Comparison of bias and rmse across the three LID magnitude levels under long test length condition



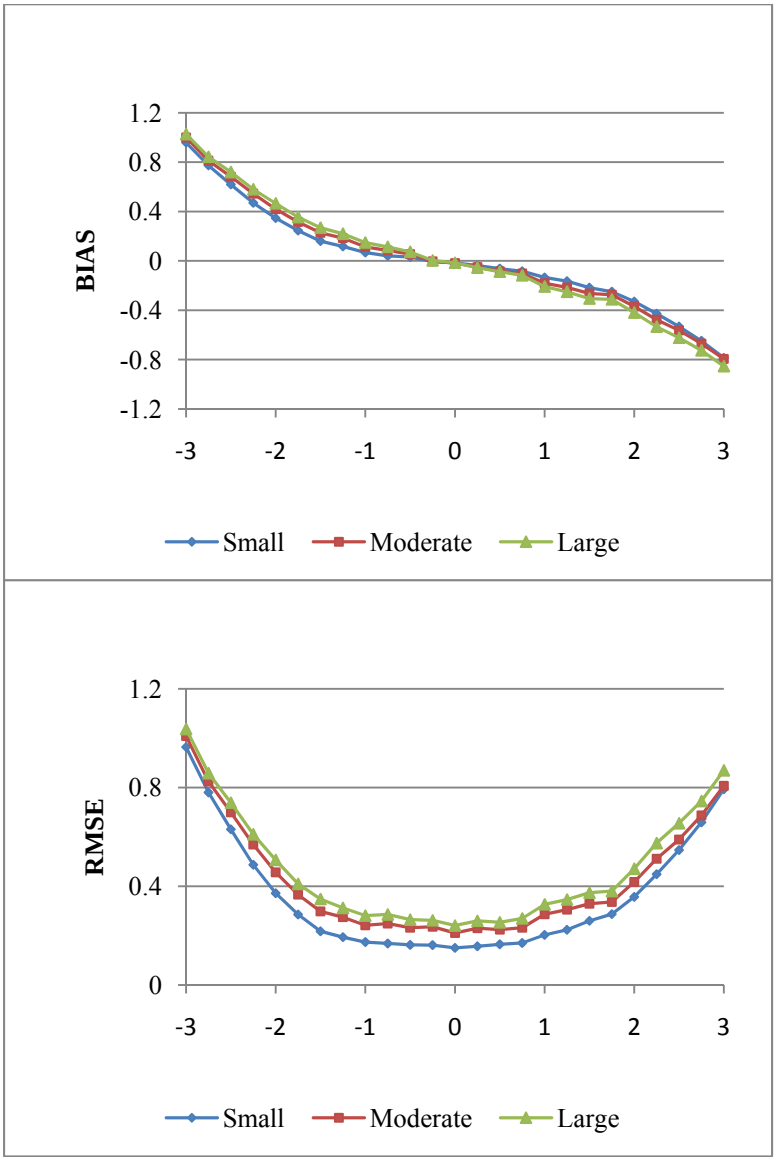
(a) The proportion of testlet items=.33; Long

Figure 12, continued



(b) The proportion of testlet items=.67; Long

Figure 12, continued



(c) The proportion of testlet items=1; Long

Summary

To answer research question 1, various locally dependent test data are simulated. The 3PL model is used to calibrate item parameters in the pool, to construct MST panels, and to get examinees' both interim and final ability estimates. The final ability estimate is further compared with the predefined cut score to make pass-fail decisions.

Each locally dependent test data is simulated according to a combination of specified factors. The studied factors include: the test length, the proportion of testlet items, the position of testlet/discrete items, and the LID magnitude associated with the testlets. The expectation is that each factor will contribute to the overall LID in the examinees response data. Ignoring LID and fitting a unidimensional 3PL model will result in the loss of ability estimation precision and decision accuracy. The following findings confirmed the expectations.

First, among all the simulated conditions, the panels of locally independent data with the long test length produce the lowest RMSE and highest DA. Comparing with the item locally independent data, all MSTs of locally dependent data yield larger RMSE and smaller DA.

Second, consistent with many test length studies, longer tests yield smaller magnitude of BIAS and RMSE, and higher DA. The test length of the MST design has large effect on each of the three evaluation criteria. The loss of ability estimation precision with short test length is probably mainly due to the poor measurement of examinees with high or low abilities.

Third, the position of testlet/discrete items in the MST design is found to have large effect on the overall decision accuracy. Its effect on the precision of ability estimation is ignorable or small. Its large effect on decision accuracy is mainly because of the estimation differences of high ability examinees between testlet/discrete items positioned on Stage 1 or on Stage 2 and Stage 3. In most situations, there are no significant mean differences of BIAS, RMSE or DA between position of Stage 2 and Stage3.

Fourth, the effects of the proportion of testlet items in the MST design are found to be statistically significant. When the LID magnitude is at least moderate, large proportions (.67 or 1) would produce large BIAS and RMSE, and small DA at each ability level; Small proportion of testlet items would produce small BIAS and RMSE, as well as large DAs at each ability level..

Fifth, the effects of the LID magnitude are found to be statistically significant in terms of ability precision and decision accuracy. Testlets with moderate and large LID magnitudes generate large magnitude of BIAS and RMSE and small DAs, and vice versa.

Sixth, the interaction effects exist at the four studied factors. Among all possible combination of interactions, the interaction between the LID magnitude and the proportion of testlet items have large effect on each of the three evaluation criteria. The other combinations may produce large effect on ability estimation but small to moderate effect on the decision classifications.

In short, each of the four studied factors would influence the precision of ability estimation and decision accuracy. Among all the studied factors, the testlet/discrete item positions are found to be less influential than the other factors. Among all the possible interactions, the interaction effects between the LID magnitude and the testlet item proportion and the interaction between the testlet item proportion and the test length are the most important. They are non-ignorable and may have large effect on the accuracy of ability estimation and decision accuracy.

Research Question II

To answer research question 2 "Would the 3PL testlet model that can account for LID improve the measurement precision, and decision accuracy over the 3PL model?", the 3PL testlet model is used to calibrate the item pools, to construct MST panels and to estimate examinees' abilities under the 21 item dependent conditions. The three evaluation criteria are then computed and compared with those under the 3PL model.

Model Effect

The means of BIAS, RMSE, and DA over 30 replications with the 3PL testlet model under both test length conditions are presented in Table 18. It appears that the overall BIAS are negative under all simulation conditions, indicating that examinees' true abilities are underestimated. Comparing with the results presented in Table 5, with the same simulation condition the absolute values of BIAS with the 3PL testlet model is larger than those with the 3PL model; the RMSEs with the 3PL testlet model are higher than those with the 3PL model; and the DA rates with the 3PL testlet model are lower than those with the 3PL model with a few exceptions under the short test length condition.

Table 18: Evaluation criteria with the 3PL testlet model

Simulation Condition	Item Pool	LID Magnitude	Testlet Item Proportion	Position	BIAS		RMSE		Decision Accuracy	
					Long	Short	Long	Short	Long	Short
2				tslt_s1	-0.039	-0.046	0.358	0.422	0.924	0.919
3				tslt_s2	-0.044	-0.045	0.357	0.420	0.933	0.934
4	2	0.25	0.33	tslt_s3	-0.043	-0.044	0.352	0.423	0.939	0.932
5				tslt_s1	-0.039	-0.053	0.377	0.439	0.920	0.913
6				tslt_s2	-0.047	-0.043	0.372	0.436	0.925	0.931
7	3	1	0.33	tslt_s3	-0.044	-0.048	0.365	0.444	0.933	0.930
8				tslt_s1	-0.030	-0.048	0.385	0.451	0.922	0.911
9				tslt_s2	-0.041	-0.034	0.378	0.446	0.926	0.930
10	4	1.5	0.33	tslt_s3	-0.039	-0.043	0.371	0.451	0.934	0.925
11				dsct_s1	-0.033	-0.043	0.400	0.456	0.913	0.920
12				dsct_s2	-0.021	-0.032	0.391	0.441	0.921	0.920
13	5	0.25	0.67	dsct_s3	-0.039	-0.041	0.400	0.443	0.922	0.916
14				dsct_s1	-0.040	-0.039	0.462	0.527	0.904	0.900
15				dsct_s2	-0.025	-0.040	0.463	0.517	0.906	0.906
16	6	1	0.67	dsct_s3	-0.037	-0.043	0.461	0.544	0.912	0.905
17				dsct_s1	-0.018	-0.025	0.480	0.565	0.901	0.893
18				dsct_s2	-0.019	-0.033	0.481	0.547	0.901	0.893
19	7	1.5	0.67	dsct_s3	-0.024	-0.043	0.479	0.544	0.908	0.905
20	8	0.25	1	N/A	-0.062	-0.064	0.472	0.489	0.919	0.908

21	9	1	1	N/A	-0.037	-0.040	0.639	0.641	0.881	0.879
22	10	1.5	1	N/A	-0.043	-0.047	0.721	0.728	0.858	0.858

Note. tslt_s(*x*): testlet items positioned on Stage (x); dsct_s(x): discrete items positioned on Stage (x), *x*=1, 2, or 3

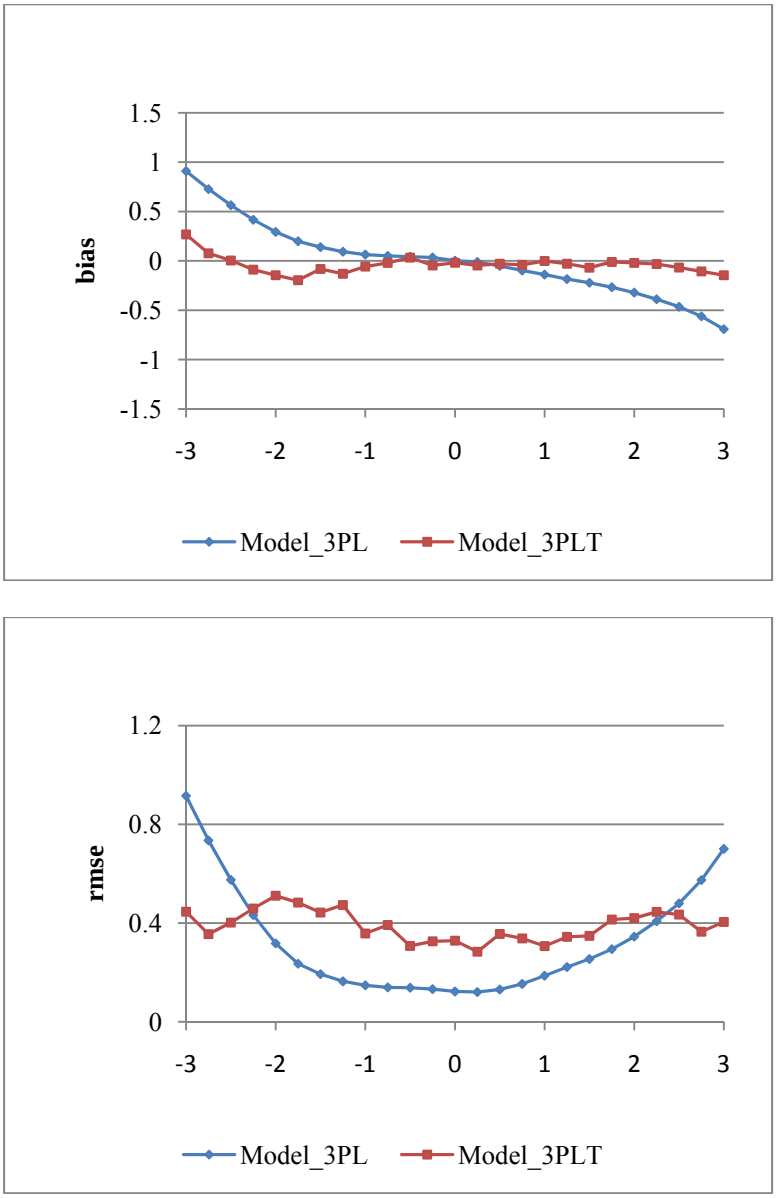
To study the model effect, a series of ANOVA tests are carried out. Since the simulation conditions of test length and measurement model are fully crossed, the interaction between measurement model and test length are also studied. Table 19 summarizes the twenty-one ANOVA results. Detailed ANOVA results are presented in Appendix E-2. The results suggest that on average the measurement model has large effect on each of the three evaluation criteria. On average the interaction effect between the test length and the measurement model also can have large effect on each of the three evaluation criteria.

Table 19: Summary statistics of ANOVA results for measurement model effect, test length and their interaction effect

Dependent Variable	Source	Sum of Squares	F-value	p-value	Partial eta squared
BIAS	Length	0.001	64.623	0.174	0.259
	Model	0.030	3557.569	0.000	0.953
	Length*Model	0.001	86.358	0.030	0.319
RMSE	Length	0.071	12645.684	0.000	0.985
	Model	1.568	280282.296	0.000	0.999
	Length*Model	0.006	1117.282	0.000	0.858
Decision Accuracy	Length	0.001	190.465	0.037	0.511
	Model	0.012	1914.186	0.004	0.709
	Length*Model	0.000	46.644	0.134	0.209

Figure 13 compares the BIAS and the RMSE between two models with simulation condition 2 of the long test length condition. Results for other simulation conditions and short test length conditions are similar (See Appendix F-5). Under the 3PL model, the BIAS and RMSE are smallest in the middle of ability distribution. The values of BIAS and RMSE increase as the ability becomes more extreme. Under the 3PL testlet model, the curves for BIAS and RMSE are more flat comparing to those with the 3PL model. They show less magnitude of BIAS and RMSE at the two tales of the ability distribution (large than 2 or smaller than -2), but higher values of BIAS and RMSE in the middle of ability distribution (between -2 and 2).

Figure 13: Comparison of bias and rmse under simulation condition 2 with long test length



Summary

To answer research question 2, with the twenty-one item local dependent simulation conditions, the 3PL testlet model is used to calibrate item parameters, to construct MST panels and to get examinees' both interim and final ability estimation. The final ability estimation is further compared with the predefined cut score to make pass-fail decisions. The results under the 3PL testlet model surprisingly do not "improve" the ability estimation under all simulation conditions. The ANOVA results suggest that the model effect is significantly large on all three evaluation criteria. A close examination of the bias and the rmse across the ability scale shows that comparing with the true 3PL testlet model, the 3PL models yields larger bias and rmse at the two tales of theta scale and smaller bias and rmse in the middle of ability distribution.

Examination of Item Parameter Calibration and Information Provided by MST Panels

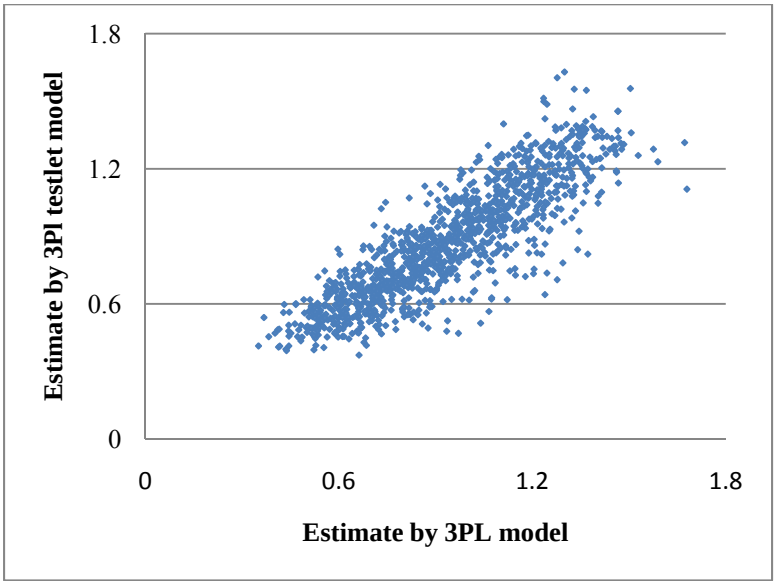
Since the 3PL testlet model does not provide the same level of accuracy of ability estimation and decision classifications, efforts are made to exam the performance of the 3PL testlet model in the first two steps of MST: the calibration of item parameters in the pool and the construction of MST panels.

Using the 3PL testlet model to calibrate items in the pools, it improves the item parameters estimation (as suggested in Appendix B), especially for the *a* parameters. Using the results presented in Table B-1, it appears that the *a*-parameter estimation is strongly associated with the properties of item pools. Item pools of larger proportions of testlet items and larger LID magnitudes of testlets will produce worse estimation of the *a* parameters as measured by the correlation between their true values and estimated values and the RMSE. The results presented on Table B-2 suggest that the estimation of the *a*

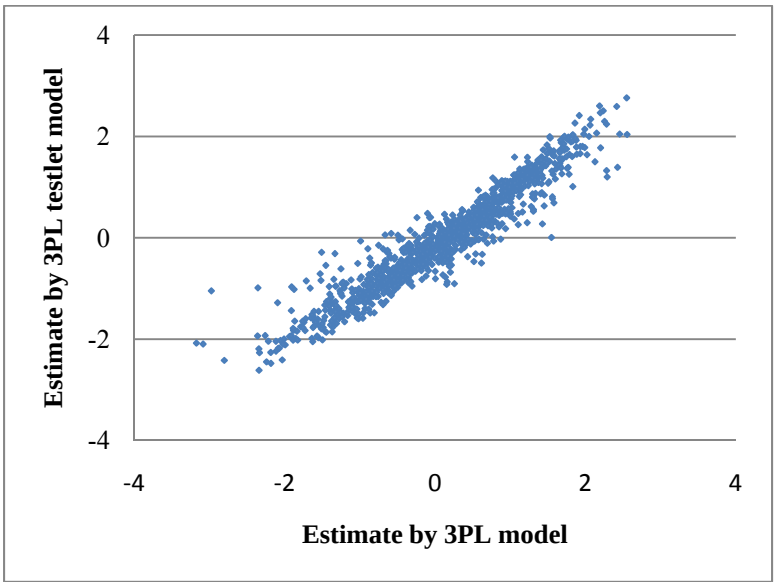
parameters are not influenced by the item pool structure or the magnitudes of testlets. No matter what the proportion of testlet items is and at what LID magnitude, using the 3PL testlet response model would generate the same level of estimation.

Figure 14 shows the comparison plots of the *a* and *b* parameters across the 3PL model and the 3PL testlet model with item pool 8. Figure 14(a) indicates larger *a* values for the 3PL model, while Figure 14(b) shows close alignment of the *b* parameters. These results are consistent with those from Acherman (1987) and Wainer et al. (2000) that when the LID exists the item discrimination parameters would be over-estimated using the IRT model, while the difficulty parameter are well estimated. This implies that the loss of the precision of ability estimation and decision accuracy is partially caused by the poor calibration of item parameters in the pool, especially the *a* parameters.

Figure 14: Comparison of the *a* and *b* parameters across the 3PL model and the 3PL testlet model



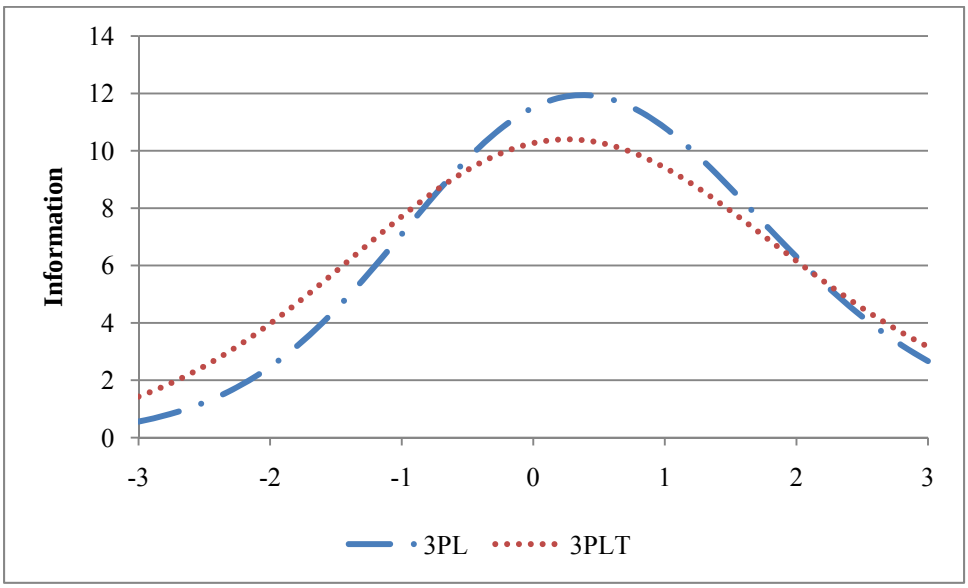
(a) *a* parameters



(b) *b* parameters

Figure 15 compares the information functions with a panel between the two models with item pool 8. Each information curve represents the information provided by a panel constructed under a model. Note that the two panels may be composed of different items or testlets. Figure 15 indicates that the panel constructed by the 3PL model provides more information in the middle of the ability distribution, while the 3PL testlet model has more information at the two tales of the ability distribution. This is different from the information curves listed in Yen (1993) in which the two information curves representing locally independent and LID never overlap with each other along the ability scale. This is probably because the information curved presented here includes the impact of the item selection and construction of MST panels, while in Yen (1993), the items are the same and the test is fixed. Figure 15 is also consistent with previous RMSE plot (Figure 13(b)) and implies that the inflated *a*-parameter estimation may cause inappropriate items or testlets to be included in the MST panels and produce inflated reliability for examinees that are in the middle of ability distribution.

Figure 15: Comparison of panel information functions across the two models



Chapter 5: Summary and Discussion

This chapter provides a summary and discussion of the study. It begins with a brief restatement of the research questions and a summary of the methodology used in this study. This is further followed by a discussion of the major findings of the study. Conclusions and practical applications are then described. Limitations of the study and directions for future research are given in the final section.

Restatement of Research Questions

As mentioned in Chapter 1, there are two objectives in this study. The first objective is to investigate the impact of LID of testlet items on the performance of MSTs for pass-fail decisions. The magnitude of LID is manipulated. Several MST design variables which could further influence the impact of LID are also studied. These variables include the proportion of testlet items, the testlet/discrete items positions, and the module/test length. The second objective is to apply the 3PL testlet model to account for LID, and to compare its performance with the 3PL model. In other words, this study tries to provide information in response to the following two research questions:

1. If the 3PL model is the measurement model, how are the measurement precision and classification accuracy impacted by the proportion of testlet items in an MST, the position of the testlet items (which stage?), the magnitude of LID, and the test length?

2. Would the 3PL testlet model that can account for LID improve the measurement precision and classification accuracy over the 3PL model?

Simulated data sets are used to investigate the objectives of this study. The selected MST design is of 1-2-2 structure. In this design, each panel has five modules.

Each module is targeted at a specific difficulty level. Stage 1 has one moderate module, Stage 2 and Stage 3 each has a moderate module and a hard one. The routing cut score is set at θ=.5. The final pass-fail cut score is set at θ=1. To control the item exposure rate less than .25, eight panels are constructed at each simualtion condition.

Four factors that are associated with the MST design is manipulated in this study. The first factor is the testlet item proportion in the item pool. The item pool is structured to have both discrete item and testlet items. It is assumed that the final MSTs are constructed to have the same proportion of testlet items as the item pool. Four proportions are studied: 0, 33%, 67%, and 100%, correspondingy to represent no items, 33%, 67%, and all items in the pool are testlet items. Each item pool is designed to have 1200 items. The second factor is the position of testlet/discrete items. Depending on the proportion of testlet items, the minority items could be placed on one of the three stages. The third factor is the LID magnitude. Three levels of LID magnitude are studied by setting $\sigma^2_{r_{id(j)}}$=.25, 1, and 1.5 or standard deviation of .5, 1, and $\sqrt{1.5}$ correspondingly to represent small, moderate and large effects. The last factor of interest is the module length. Two module lengths are considered in this study: 12 and 8.

The 3PL testlet model is used to generate item responses. The 3PL model and 3PL testlet model are used to calibrate, to construct MSTs and to score examinees seperately. A total of 88 simulated conditions are studied. Each simulation condition is replicated 30 times. A group of 2500 examinees are simulated to take each panel of MST. They are defined as 100 θs from -3 to 3 in increments of .25. The MSTs are constructed to put most informative items on early stages. During the administration, an examinee encounters a moderate module in the first stage. Accoring to the routing rule, the

examinee is routed to one of the two modules on Stage 2. At the end of Stage 2, the routing rule is applied again to select one of the two modules on Stage 3 for the examinee. In total, each examinee is administered three modules, and selections are tailored at Stage 2 and Stage 3 to the ability of examinees.

To evaluate ability estimation and classification results under various simulation conditions, the 2500 examinees true and estimated ability and pass-fail status are compared. Summary indices – BIAS, RMSE and DA over 30 replications are computed and used for final comparisons. ANOVA tests are conducted to identify the significant performance differences among the studied factors.

Discussion of Major Findings

Results of the above analysis are described in detail in Chapter Four. Here some highlights of the findings are summarized and discussed.

Research Question 1

First, the results of this study show that panels of locally independent data yield the smallest RMSE and the highest DA. Ignoring the testlet effects and fitting a unidimensional 3PL model result in the loss of ability estimation precision and decision accuracy. This finding is consistent with many of the previous studies of LID caused by testlet items on PPTs. This finding is also consistent with Glas & van der Linden (2003) and Pommerich & Segall (2008) in which the precision of ability estimation was negatively affected when the dependences between items were ignored. Their LID was caused by item cloning which belongs to another family of LID. The explanation is that ignoring LID in the response data and fitting a unidimensional IRT model is a case of

model misspecification, which generally leads to bias in parameter estimation and hence to the classification decisions.

Second, consistent with many test length studies, this study finds that MSTs composed of more items would produce smaller magnitude of BIAS and RMSE , and higher DA for a standard normal distribution. The test length has large effect on each of the three evaluation criteria. Comparing to other MST studies (e.g., Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006; Zhang, 2006), the long test length condition specified in this study is relatively short. But this study still produces comparable decision accuracy rates even with the worst LID conditions. This results proves Stark & Chernyshenko (2006)'s suspect that the test length specified in those studies might be too long to reveal the psychometric benefits of MST as compared to traditional static tests. As stated in Jodoin et al. (2006)'s paper, reducing test length would reduce exam costs for examinees, test developers as well as test administers with the benefit of reducing testing time, lowering item exposure levels, and requiring a smaller item pool. Of course, the test length is also determined by the needs of content coverage, the requirement of measurement precision as wells as other concerns of operational usage. Whether an even shorter test as used in this study is feasible in operation depends on the specific needs of testing agencies, the characteristics of examinee population and the inferences to be drawn from the examination.

Third, the position of testlet/discrete items in the MST design is found to have effect on the three evaluation criteria. Among the three studied testlet/discrete item positions, the DA associated with Stage 1 is significantly different from those associated with Stage 2 or Stage 3, where there are no significant differences between Stage 2 and

Stage 3. This is probably because the item type and the choice of measurement model for the first stage are more important than for the rest stages. For example, when the proportion of testlet items is .33, if the only testlet items are put on Stage 1, the misuse of the 3PL model would influence the ability estimation after Stage 1, thus the routing decisions by putting examinees into the wrong modules of Stage 2. If the only testlet items are put on Stage 2, the 3PL model is the correct model for Stage 1 items. Though it is still the wrong model for items on Stage 2, its impacts on the ability estimation is much smaller and may not influence the routing decisions because at this stage the number of items already doubled. If the only testlet items are put on Stage 3, the impact of using the 3PL model is only on the final ability estimation. In another words, the position effect might be mitigated by increasing the number of items that are used to making routing decisions in the MST design. Future studies might appropriately focus on simulation studies that aim to understand the misclassification of routing decisions and final classification decisions by having varying number of items on each stage.

Fourth, the effects of the proportion of testlet items are found to be significant. Large proportion of testlet items on the MST will produce large BIAS and RMSE, and small DAs at each ability level. The explanation for this finding is that the proportion of testlet items reflects the degree of model misspecification in an MST design. The large proportion of testlet items in an MST design is, the large the degree of model misspecification is. Thus the poor person parameter estimation is. As suggested in Yen (2006), while the effect of LID can be very large when estimating the amount of information that comes, the effects can be minimized by reducing the proportion of testlet items.

Fifth, this study finds that the ability estimation is adversely impacted by the moderate and high LID levels simulated here. Increasing the LID magnitude will lead to the increase in the BIAS and the RMSE and the decrease in the DA.

Sixth, this study finds that the interaction between the proportion of testlet items and the LID magnitude has large effect on each of the three evaluation criteria. As the overall LID conditions become severe (e.g., larger proportion of testlet items, bigger LID magnitude associated with each testlet), the ability estimation becomes worse and the classification errors spread to more ability levels. However, the 3PL model is robust with LID when the proportion of testlet items and the LID magnitudes associated with testlets are small.

Research Question 2

This study also uses the 3PL testlet model to calibrate item parameters in the pool, to construct MST panels and to score examinees. The results indicate that using the 3PL testlet model to calibrate items in the pools, it improves the item parameters estimation. The results also suggest that using the 3PL testlet model, the estimation of the *a* parameters are not influenced by the item pool structure or the magnitudes of testlets which contribute to the item parameter estimates using the 3PL model.

The panel constructed by the 3PL testlet model has more information at the two tales of the ability distribution, while the 3PL model provides more information in the middle of the ability distribution. This is different from the information curves listed in Yen (1993) in which the two information curves representing locally independent and LID never overlap with each other along the ability scale. This is probably because the information curved presented here includes the impact of the item selection and

construction of MST panels, while in Yen (1993), the items are the same and the test is fixed.

For ability estimation and decision classifications, the results indicate that the 3PL testlet model does not "improve" the overall ability estimation and the classification accuracy. Actually, the results prove that using the 3PL model to calibrate item pools, to construct MST panels and to score examinees is wrong when there is large proportion of testlet items and their LID magnitudes are large. As to the classification accuracy, the DA rates under the 3PL model are still higher than those using the 3PL testlet model. Whether test developers would want to use the 3PL testlet model in their MSTs would still be a question.

Practical Implications

MST is a new computerized test delivery technology aimed at enhancing the quality of credentialing exams. Many studies (e.g., Hambleton & Xing, 2006; Jodoin et al. 2006; Keng, 2008) have compared the MST design with PPT and CAT and concluded that though MST could not reach the same measurement precision as CAT; it has the potential to increase testing efficiency and decision accuracy comparing to traditional linear fixed length tests or computerized fixed tests. To test developers, MST provides better test security with a single item pool, as it is possible to create many panels that are parallel in content and information, and panels can be randomly assigned to examinees. It also allows greater control over test construction because subject experts have the opportunity to review all panels and conduct analyses examining dimensionality, adverse impact and differential test function before the publication of the test. Comparing to CAT, MST can better accommodate testlet items. Recently, there has been an increased

interest in applications of testlet items (Downing, 2006). Comparing to PPT, MST is adaptive in nature and is therefore more efficient than PPT. All these make the MST design appealing to many testing agencies. To test takers, comparing to CAT, MST provides greater flexibility. And examinees like the ability to review the items within modules and may apply their favorite testing strategy within a stage.

To design an MST test, there are a number of factors to consider. First consideration is the test specification which includes "the content covered by the test, proposed number of items, format(s) of the item, desired psychometric properties of items, and item and section arrangement" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Second important consideration is the item pools which include the pool size and pool composition. The third important consideration is about measurement. In other words, use which method (CTT or IRT) and which model if the method is IRT to describe the relationship between items, examinees abilities, and their responses. This study tries to address some of factors mentioned above, but not all. The addressed factors include test specification (the number of items in each module, the item types as well as their stage arrangements); the item pool structure (proportion of testlet items); and the measurement model (3PL model or 3PL testlet model). Thus, the findings from this study contribute to the expanding knowledge base in the field of research and provide practical guidelines to programs that are considering MST as the test delivery model.

First, this study evaluates the robustness of using the 3PL model with MSTs that are designed to make pass-fail decisions when there are testlet items which are causing

the LID problem. To the author's knowledge, that is no other literature addressing this issue yet. Each simulation generates data according to the 3PL testlet model. The evaluation determines how successfully the 3PL model would recover the examinees' true ability values and true pass-fail status despite the presence of LID. The results suggest that using the 3PL model with the MST design when there are testlet items is fairly robust to the violation of local independence assumption as long as the LID magnitude associated with the testlets is small and the proportion of testlet items is small. This warrants the test developers that if there are testlet items that have small LID magnitude, they still can be put on the MST panels. The conventional 3PL model may still be applicable to calibrate the item pools, to construct MST panels and to score examinees as long as the test length is sufficiently long.

Second, the study evaluates several factors that may contribute to the local item dependence in the examinees' response data through the MST administration and therefore affect the final ability estimation and decision accuracy. The factors include the proportion of testlet item in the MST panels (or in the pool), the position of testlet/discrete items, and the LID magnitude associated with each testlet. The results help us to understand each of the factors and their interaction effects on the final ability estimation and decision accuracy. To test developers, the simulation results may help them to decide the proportion of testlet items and their positions to appear in the MST panels. For example, if the test developers decide that the testlet item proportion is .33 and the test length is long, the testlet items appearing on Stage 2 or 3 do not have significant difference on decision accuracy. This may give test developers some flexibility in arranging their items without worrying about the loss of decision accuracy.

The simulation results also help them to decide what levels of LID magnitude are tolerable. Since the costs associated with item develop (item writing, pretesting) are high, test developers will be glad to retain testlet items in the MST panels knowing that they do not cost too much loss of the accuracy of ability estimation and classification decisions.

Third, the studied factors include the test length. Comparing to traditional PPT, MST has the benefits of efficiency. Efficiency is partially defined by the test length. Short MST panels with the same item pool means less item exposure rate and thus less test security problems. However, short test length is always accompanied with less test precision. The results in this study help test developers to see the lost ability estimation precision and decision accuracy, and help them to make a decision whether they would tolerate the lost precision and decision accuracy by shortening the test length. The results of this study also suggest that the effects of the above studied factors (e.g., testlet item proportion, testlet/discrete item position, and LID magnitude) may be large or small depending on the test length. Comparing to other MST studies, the short test length is 24 items (8 items per module), is much shorter than those on other MST studies. The results partially provide response to Stark & Chernyshenko (2006)'s question that shorter tests (15-20) of MST may provide increase in efficiency over traditional paper-pencil tests.

Fourth, the study demonstrates the use of 3PL testlet model with testlet items in constructing MST panels and scoring examinees. The testlet items are designed to appear in one module and with only one stimulus. Thus, in this study, the original number of items associated with a stimulus is high (twenty in the simulation design). Either twelve or eight items are selected during the construction of MST panels. The application of

polytomous response models to calibrate the item parameters and to select item intelligently may not be practical. This study also contributes to the research community with the computation of item information and EAP estimation of abilities with the 3PL testlet response model. All these demonstrate that the 3PL testlet model is a viable option for testing programs considering the MST design.

Finally, based on the results of this study, some suggestions can be made to minimize the effect of LID on the MSTs that are designed to make pass-fail decisions. First, LID and its effects can be minimized by constructing the tests with discrete items that are independent with each other. If testlet items are included in the MST design, testlet items can put on later stages rather than in the first stage. If the proportion of testlet items is large in the MSTs, the 3PL testlet model can be used as the measurement model appropriately to account for LID.

Limitations of This Study and Future Research Directions

In general, this study tries to mimic an MST from the calibration of item parameters in the pools, to construct MSTs and to administer and score examinees. Comparing to operational use, the simulation design has its limitations in the first two aspects.

The calibration of item parameters in the pool in this study uses a simple scheme in which 3000 examinees respond to all 1200 items in each pool. This is not achievable in reality. The item parameters are typically estimated by pilot- or field-testing of each item prior to its appearance on an operational test form. A common model for pretesting is to administer each examinee some number of pretest items (which do not contribute to scoring) alongside his or her operational test. Some equating methods are then used to put

item parameters in the same scale. Alternatively, testing programs can seek volunteers to take sets of new items. In either case, none of examinees respond to all of the pretest items. However, for testing agencies to carry out the calibration of new items, a lot of questions need to be answered before implementing it, such as what is the sample size requirement? What is the number of pretest items? What is the number of linking items to appear in each form? Where to put the new testlet items and how many of them in the case of test forms composed of both of testlet items and discrete item? Testing agencies want new items to be accurately and reliably calibrated before they can be used in the operational MST panels. This study does not seek to provide responses to those questions. Future studies can address these issues.

The construction of MST in this study only considers the psychometric property of items. There are several limitations. First, in practice, content control is a very important consideration during test construction. In operational tests, due to the limitation of item pools, sometimes the psychometric property has to be sacrificed in order to reach the goal of adequate content coverage. This study ignores that. Second, in the simulation, each testlet is set to have 20 items. 12 items are selected within a testlet to make up a module. Thus, a large proportion of items in a testlet would not be used. In reality, a module may consist both discrete and testlet items. Thus, the testlet itself does not have to have a large number of items. Third, the construction of MST in this study considers test information only at two ability levels: $\theta = 0$, and $\theta = 1$. There is no further assurance that in a wide range the test information provided by different MST panels are similar. van der Linden (2005) suggested to use linear programming approach to construct test forms. Luecht, Brumfield, & Breithaupt (2006) described some steps

necessary to construct MST panels via ATA. In both methods, complicated content specifications can be implemented. However, they only focus on discrete items. Future studies can explore the usage of those methods with testlet items in constructing MST panels.

As for any study that uses simulated data, the findings of this study are restricted by the prescribed LID data conditions and the levels of each studied factors.

As to the LID data condition, this study addresses the issues of LID with testlet items, in which LID could easily be described using the 3PL testlet model. There are other situations that may result LID. For examples, LID due to the relationship between item pairs, such as item clones, reverses, or alternatives. These LID items appear as discrete items. Future studies can evaluate the impact of LID caused by those factors on the MST design.

In this study, the proportion of testlet items and the position of testlet/discrete items are partially determined by the structure of the studied MST: 1-2-2. Other MST designs such as 1-3-3 or 1-2-3-4 might accommodate more levels of testlet item proportions. Or other small (e.g., .20) or large (e.g., .80) proportion of testlet items can be arranged with the current MST structure by specifying certain modules having a combination of both discrete and testlet items. The allocation of items across the stages (e.g., longer or shorter initial and final stages) can also be varied in future studies.

In the simulation, the LID magnitudes are kept constant (small, moderate and large) with each item pool. In reality, different testlet would probably exhibit varying levels of LID magnitude ranging from none to very large. Simulating a range of LID

magnitude rather than a constant value with the same item pool might be one option for future research.

The MSTs in this study are designed to provide pass-fail decisions. Only one cut score is specified in this study. And the hard module in the MST panels is designed to provide most information at the cut score. Other cut scores (e.g., .5, 0 or -.5) may produce different results with current simulation design.

In summary, this study investigates the impacts of using the 3PL model with testlet items in the MSTs where the local item independent assumption is violated and tries to solve the problem by using the 3PL testlet model. However, MST is a relatively new computer delivery model. From the initial item development to the final administration of MST panels, further investigation is still needed.

Appendix A: EAP Estimation with the 3PL Testlet Model

Under the 3PL model, the ability estimate is:

$$\hat{\theta} = \frac{\sum QL(Q)W(Q)}{\sum L(Q)W(Q)}$$

In which, Q is a quadrature point in the ability scale, and W(Q) is weight of the quadrature point. $L(Q)$ is the likelihood of a person's response pattern at Q quadrature point.

Under the 3PL testlet model,

$$\mathrm{P_{ij}}(\theta_\mathrm{i}) = \mathrm{c_j} + (1 - \mathrm{c_j})\frac{\exp[\mathrm{a_j}\left(\theta_\mathrm{i} - \mathrm{b_j} - \gamma_{\mathrm{id(j)}}\right)]}{1 + \exp[\mathrm{a_j}\left(\theta_\mathrm{i} - \mathrm{b_j} - \gamma_{\mathrm{id(j)}}\right)]}$$

Since $\gamma_{\mathrm{id(j)}}$ is unknown, the quadrature point idea is applied. Assume Q_θ represents quadrature point in the ability scale; Q_γ represents quadrature point in the testlet effect scale; W(Q_θ) is weight of the ability quadrature point; and W(Q_γ) is the weight of the testlet effect quadrature point, the EAP estimate is:

$$\hat{\theta} = \frac{\sum Q_\theta L(Q_\theta)W(Q_\theta)}{\sum L(Q_\theta)W(Q_\theta)},$$

Assuming that the testlet effect parameter is independent of theta, which is the standard assumption, one can get that $L(Q_\theta) = \sum L\left(Q_{\theta,r}\right)W(Q_r)$,

thus,

$$\hat{\theta} = \frac{\sum Q_\theta[\sum L\left(Q_{\theta,r}\right)W(Q_r)]W(Q_\theta)}{\sum[\sum L\left(Q_{\theta,r}\right)W(Q_r)]W(Q_\theta)}$$

In this study, for an MST that has testlet items on its all three stages, the ability estimate could be written as:

$$\hat{\theta} = \frac{\sum Q_\theta L(Q_\theta) W(Q_\theta)}{\sum [\sum L(Q_\theta) W(Q_\theta)}$$,

where $L(Q_\theta) = L_{tslt1}(Q_\theta) * L_{tslt2}(Q_\theta) * L_{tslt3}(Q_\theta)$, and

$L_{tslt(i)}(Q_\theta) = \sum L_{tslt(i)}(Q_{\theta,r}) W_{tslt(i)}(Q_r)$; i=1, 2, or 3.

For a MST with both individual and testlet items, $L(Q_\theta)$ can be seen as the combination of $L_{dsct}(Q_\theta)$ and $L_{tslt}(Q_\theta)$, where $L_{idsct}(Q_\theta)$ is the likelihood of the response pattern for discrete items and $L_{tslt}(Q_\theta)$ is the likelihood of the response pattern for testlet items. For example, if the first stage of an MST consists of testlet items and the second and third stages are discrete items, an examinee's ability can be obtained using the following formula:

$$\hat{\theta} = \frac{\sum Q_\theta L(Q_\theta) W(Q_\theta)}{\sum L(Q_\theta) W(Q_\theta)}$$,

In which $L(Q_\theta) = L_{tslt1}(Q_\theta) * L_{dsct2}(Q_\theta) * L_{dsct3}(Q_\theta)$, and $L_{tslt1}(Q_\theta)$ represents the likelihood of responses of stage 1 items; $L_{dsct2}(Q_\theta)$and $L_{dsct3}(Q_\theta)$ represent the likelihood of responses of stage 2 and stage 3 respectively.

Appendix B: Item Parameter Estimates

Table B- 1: Summary statistics for estimated item parameters by the 3PL model (N=1200)

Item Pool	Parameter Estimate	Min	Max	Mean	S.D	Correlation	RMSE
	a_est	0.389	1.949	0.991	0.296	0.935	0.128
	b_est	-2.966	2.672	0.167	0.942	0.967	0.321
1	c_est	0.012	0.400	0.200	0.071	0.688	0.083
	a_est	0.371	1.796	0.987	0.283	0.926	0.129
	b_est	-3.215	2.402	0.116	0.946	0.962	0.305
2	c_est	0.014	0.474	0.202	0.072	0.677	0.085
	a_est	0.346	1.764	0.953	0.271	0.879	0.143
	b_est	-2.606	2.351	0.123	0.937	0.959	0.318
3	c_est	0.019	0.479	0.202	0.072	0.661	0.086
	a_est	0.392	1.897	0.934	0.272	0.846	0.157
	b_est	-2.701	2.437	0.123	0.938	0.961	0.312
4	c_est	0.019	0.447	0.202	0.070	0.651	0.087
	a_est	0.385	1.764	0.975	0.276	0.914	0.130
	b_est	-2.858	2.635	0.116	0.943	0.960	0.312
5	c_est	0.016	0.419	0.202	0.071	0.656	0.086
	a_est	0.306	1.791	0.863	0.250	0.807	0.181
	b_est	-3.504	2.559	-0.009	0.986	0.926	0.376
6	c_est	0.000	0.399	0.165	0.090	0.386	0.099
	a_est	0.287	1.592	0.832	0.242	0.767	0.207
	b_est	-3.299	2.367	-0.004	0.991	0.918	0.397
7	c_est	0.000	0.474	0.165	0.092	0.354	0.103
	a_est	0.247	1.612	0.883	0.261	0.899	0.132
	b_est	-3.054	2.697	-0.018	1.030	0.936	0.363
8	c_est	0.001	0.405	0.165	0.089	0.409	0.097
	a_est	0.282	1.391	0.792	0.201	0.824	0.212
	b_est	-3.128	2.452	-0.010	1.022	0.918	0.406
9	c_est	0.001	0.500	0.167	0.093	0.363	0.103
	a_est	0.284	1.244	0.748	0.182	0.813	0.247
	b_est	-3.360	2.612	0.003	1.018	0.913	0.419
10	c_est	0.001	0.448	0.169	0.092	0.351	0.104

Table B- 2: Summary statistics for estimated item parameters by the 3PL Testlet model (N=1200)

Item Pool	Parameter Estimate	Min	Max	Mean	S.D	Correlation	RMSE
2	a_est	0.405	1.595	0.903	0.260	0.940	0.101
	b_est	-2.543	2.543	-0.034	0.947	0.962	0.264
	c_est	0.014	0.465	0.151	0.072	0.655	0.067
3	a_est	0.367	1.639	0.917	0.273	0.943	0.096
	b_est	-2.484	2.493	-0.064	0.935	0.966	0.249
	c_est	0.013	0.494	0.146	0.066	0.694	0.062
4	a_est	0.387	1.760	0.911	0.276	0.946	0.095
	b_est	-2.505	2.499	-0.059	0.928	0.967	0.247
	c_est	0.013	0.461	0.148	0.070	0.694	0.062
5	a_est	0.355	1.607	0.891	0.271	0.937	0.105
	b_est	-2.619	2.584	-0.080	0.962	0.964	0.260
	c_est	0.014	0.475	0.139	0.067	0.647	0.066
6	a_est	0.400	1.719	0.912	0.272	0.943	0.096
	b_est	-2.378	2.440	-0.048	0.924	0.958	0.278
	c_est	0.012	0.573	0.146	0.075	0.683	0.064
7	a_est	0.354	1.813	0.913	0.278	0.945	0.095
	b_est	-2.613	2.436	-0.057	0.948	0.970	0.236
	c_est	0.013	0.461	0.144	0.066	0.728	0.059
8	a_est	0.373	1.630	0.865	0.254	0.932	0.120
	b_est	-2.614	2.762	-0.011	0.969	0.955	0.290
	c_est	0.015	0.501	0.164	0.074	0.625	0.072
9	a_est	0.381	1.563	0.883	0.264	0.946	0.102
	b_est	-2.478	2.892	-0.057	0.969	0.968	0.245
	c_est	0.015	0.458	0.146	0.068	0.734	0.058
10	a_est	0.323	1.589	0.885	0.271	0.947	0.099
	b_est	-2.835	2.591	-0.073	0.970	0.966	0.255
	c_est	0.012	0.432	0.147	0.072	0.758	0.056

Table B- 3: Summary statistics for estimated testlet magnitude by the 3PL testlet model

Item Pool	True LID Magnitude	n	Min	Max	Mean	S.D
2	0.25	20	0.232	0.303	0.273	0.019
3	1.00	20	0.928	1.190	1.014	0.054
4	1.50	20	1.402	1.750	1.577	0.094
5	0.25	40	0.132	0.319	0.265	0.027
6	1.00	40	0.927	1.179	1.033	0.055
7	1.50	40	1.357	1.741	1.552	0.096
8	0.25	60	0.261	0.351	0.303	0.022
9	1.00	60	0.983	1.224	1.095	0.060
10	1.50	60	1.424	1.893	1.647	0.101

Appendix C: Example SAS codes to Assemble MST with Testlet Items and to Estimate Examinee Abilities with the 3PL Testlet Model

```
proc printto log='c:\dissertation\auto.log' new;
run;
Options notes;


%let pool=8;
%let Total_item=1200; * total number of items in the pool;
%let J=20; *number of total items with each testlet;
%let T=60;


libname one "c:\dissertation\pool&pool\";
libname two "c:\dissertation\pool&pool\scoright\step2\";
libname three "c:\dissertation\pool&pool\scoright\step2\step3\";

%let seed=1+round(1000*time());
%let Total_item=480; *total number of items in the panels;
%let length_module=12;

%let r=.25; *the magnitude of LID;

%let nqpt=15;
/* Set mean and variance for prior distribution */
%let mean = 0;
%let sigma = 4;
/* Set number of replications */
%let num_rpl=30;


/*calculate item information*/

data qp_tslt (keep=qp_tslt1-qp_tslt&nqpt);
    array qp_tslt{&nqpt} qp_tslt1-qp_tslt&nqpt;
    qp_tslt1=-4;
    do i=2 to &nqpt;
       qp_tslt{i}=qp_tslt{i-1}+2*4/(&nqpt-1);
    end;
run;
/*read in estimated testlet effect*/

filename tslt "c:\dissertation\pool&pool\scoright\res\testlet.est";
```

```
data tslt;
     infile tslt firstobs=4;
          input tslt $ a $ est se_est;
run;

data tslt_item;
    set one.item_3plt;
run;
proc iml;

 /*calculate testlet item information */
    use tslt_item;
              read all var {b_est1} into B;
              read all var {a_est1} into A;
              read all var {c_est1} into C;
           close tslt_item;

           w=j(1,15,1);
           b=b*w; a=a*w; c=c*w;

           *calculate item information at theta=0;
           use qp_tslt;
              read all into testlt;
           close qp_tslt;
           x=j(1200,1,1);
           testlt=x*testlt;


    D11=0-B-testlt;D12=A#D11;
          D13=EXP(D12);  D14=1+D13;
    D15=D13/D14;  D16=(D15#D15)#(A#A);
    D17=(1-C)/(C+D13);
          info0_tslt1=d16#d17;

    D11=1-B-testlt;D12=A#D11;
          D13=EXP(D12);  D14=1+D13;
    D15=D13/D14;  D16=(D15#D15)#(A#A);
    D17=(1-C)/(C+D13);
          info1_tslt1=d16#d17;

          use qp_tslt;
      read all into qp;
     close qp_tslt;

    use tslt;
             read all var {est} into tslt_1;
```

```
    close tslt;

    do i=1 to &T; /*do it for each items*/
      tslt_var=tslt_1[i,];
            qpw=probnorm((qp+2*4/((&nqpt-1)*2))/sqrt(tslt_var))
           -probnorm((qp-2*4/((&nqpt-1)*2))/sqrt(tslt_var));
                  qpw=t(qpw);

                  info0_tslt=info0_tslt1[(i-1)*20+1:20*i,]*qpw;
                  info1_tslt=info1_tslt1[(i-1)*20+1:20*i,]*qpw;
      infotsl0=infotsl0//info0_tslt;
      infotsl1=infotsl1//info1_tslt;
          end;

            infotslt=infotsl0||infotsl1;
            info=info_indi//infotslt;

                  create item_info from info;
                  append from info;

quit;

data a;
    do item_id=1 to 1200;
          output;
          end;
run;
data two.item_info;
    set a;
    set item_info;
          rename col1=I0;
    rename col2=I1;
run;


/*selection of tslt items*/
%macro select_items_tslt (level=);


data item_infoT;
     set two.item_info ;
run;

data a;
    do T=1 to 60;  /*the item pool has 60 testlets*/
            do j=1 to 20;
```

```
                output;
                end;
        end;
        keep T;
run;

data item_infoT;
    set a;
        set item_infoT;
run;


/*select the most 12 informed items at specified level;*/
data item_info&level;
    set item_infoT;
        keep  T item_id I&level;
run;

proc sort data=item_info&level;
    by T descending I&level;
run;

data top12at&level(drop=count);
  set item_info&level;
  by T descending I&level;
  if first.T then count=0;
  count+1;
  if count le &length_module then output;
run;

data top12at&level;
    set top12at&level;
        diff=&level;
run;

/*calculate the total information provided by 12 selected items within each testlet;*/

data info_eachTat&level (keep=T T_info&level);
    set top12at&level;
        by T;
        if first.T then T_info&level=0;
        T_info&level+I&level;
        if last.T then output;
run;

proc sort data=info_eachTat&level out=three.info_eachTat&level;
```

```
    by descending T_info&level;
run;

%mend select_items_tslt;


%macro assemble_module_tslt ;

    data one;
        set three.info_eachTat0 (obs=8);
          run;
          data two;
        panel=0;
        do i=1 to 8;
              x=ranuni(&seed);
                    panel=panel+1;
                    stage=1;
                    diff=0;
                    output;
              end;
    run;
    proc sort data=two out=two1 (keep=panel stage diff);
        by x;
    run;

          data stage1m_T;
        set one;
              set two1;
                    rename T_info0=T_info;
    run;

          data module_info_tslt1;
              retain panel stage diff module_info;
              set stage1m_T (rename=(T_info=module_info));
                    drop T;
          run;



          proc sort data=stage1m_T;
        by T;
    run;

          data stage1m;
        merge top12at0 stage1m_T ;
              by T;
```

```
        if panel ne .;
            drop T T_info;
            rename I0=info;
run;



data one;
set three.info_eachTat1 (obs=8);
run;
    data two;
  panel=0;
  do i=1 to 8;
        x=ranuni(&seed);
            panel=panel+1;
            stage=2;
            diff=1;
            output;
        end;
run;
proc sort data=two out=two1 (keep=panel stage diff);
  by x;
run;

data stage2h_T;
  set one;
        set two1;
            rename T_info1=T_info;
run;
    data module_info_tslt2;
        retain panel stage diff module_info;
        set stage2h_T (rename=(T_info=module_info));
            drop T;
    run;



    proc sort data=stage2h_T;
  by T;
run;

    data stage2h;
  merge top12at1 stage2h_T ;
        by T;
        if panel ne .;
            drop T T_info;
```

```
                rename I1=info;
 run;


 data one;
      set three.info_eachTat0 (firstobs=17 obs=24);
 run;
 data two;
    panel=0;
    do i=1 to 8;
         x=ranuni(&seed);
              panel=panel+1;
              stage=3;
              diff=0;
              output;
         end;
 run;
 proc sort data=two out=two1 (keep=panel stage diff);
    by x;
 run;

data stage3m_T;
         set one;
         set two1;
        rename T_info0=T_info;
 run;

      data module_info_tslt5;
         retain panel stage diff module_info;
         set stage3m_T (rename=(T_info=module_info));
              drop T;
      run;



      proc sort data=stage3m_T;
    by T;
 run;

      data stage3m;
    merge top12at0 stage3m_T ;
         by T;
         if panel ne .;
              drop T T_info;
              rename I0=info;
 run;
```

```
      data one;
    set three.info_eachTat0 (firstobs=9 obs=16);
      run;
      data two;
    panel=0;
    do i=1 to 8;
          x=ranuni(&seed);
                panel=panel+1;
                stage=2;
                diff=0;
                output;
          end;
run;
proc sort data=two out=two1 (keep=panel stage diff);
    by x;
run;

      data stage2m_T;
    set one;
          set two1;
                rename T_info0=T_info;
run;

      data module_info_tslt3;
          retain panel stage diff module_info;
          set stage2m_T (rename=(T_info=module_info));
                drop T;
      run;



      proc sort data=stage2m_T;
    by T;
run;

      data stage2m;
    merge top12at0 stage2m_T ;
          by T;
          if panel ne .;
                drop T T_info;
                rename I0=info;
run;

data one;
```

```
set three.info_eachTat1 (firstobs=9 obs=16);
run;
      data two;
    panel=0;
    do i=1 to 8;
          x=ranuni(&seed);
                panel=panel+1;
                stage=3;
                diff=1;
                output;
          end;
run;
proc sort data=two out=two1 (keep=panel stage diff);
    by x;
run;

data stage3h_T;
    set one;
          set two1;
                rename T_info1=T_info;
run;
      data module_info_tslt4;
          retain panel stage diff module_info;
          set stage3h_T (rename=(T_info=module_info));
                drop T;
      run;



      proc sort data=stage3h_T;
    by T;
run;

      data stage3h;
    merge top12at1 stage3h_T ;
          by T;
          if panel ne .;
                drop T T_info;
                rename I1=info;
run;


data module_info_tslt;
          set module_info_tslt1 module_info_tslt2
        module_info_tslt3 module_info_tslt4
        module_info_tslt5;
```

```
        run;

%mend assemble_module_tslt;

%macro ass;

%select_items_tslt (level=0);
%select_items_tslt (level=1);
%assemble_module_tslt ;

%mend ass;



%macro assemble_mst;

/*put moderate moduels together*/

data moderate;
    set stage1m stage2m stage3m;

run;

proc sort data=moderate;
     by panel stage;
run;

/*put hard moduels together*/

data hard;
    set stage2h stage3h;
run;


proc sort data=hard;
     by panel;
run;

/*put moderate and hard modules together */

data all_module;
     retain panel stage diff item_id Info;
     set moderate hard;
run;

proc sort data=all_module;
```

```
    by panel stage diff;
run;
%mend assemble_mst;


%macro mstfinal;

proc sort data=all_module;
     by item_id;
run;

data one;
     merge one.item_3plt all_module;
     by item_id;
     if panel ne .;
run;

proc sort data=one;
     by panel stage diff;
run;


data three.mst_items;
     retain panel stage diff T item_id a b c a_est1 b_est1 c_est1;
           set one;
run;

data three.module_info;
    set module_info module_info_tslt;
run;


%mend mstfinal;

%macro dupcheck;

data moderate_T;
    set stage1m_T stage2m_T stage3m_T;

run;

proc sort data=moderate_T;
     by panel stage;
run;

/*put hard moduels together*/
```

```
data hard_T;
    set stage2h_T stage3h_T;

run;


proc sort data=hard_T;
     by panel;
run;

/*put moderate and hard modules together */

data all_module_tslt;
     set moderate_T hard_T;
run;

proc sort data=all_module_tslt;
    by panel stage diff;
run;

data check;
     set all_module_tslt;
            keep T panel;
run;

proc sort data=check;
    by panel T;
run;


data dups nodups ;
  set check ;
  by panel T ;
/*
    Compare the values of the FIRST.CLASS and LAST.CLASS variables.
    Write an observation to NODUPS or DUPS, depending on the outcome
    of the comparison.
*/
  if first.T and last.T then output nodups ;
  else output dups ;
run;


data _null_;
if 0 then set dups nobs=count;
```

```
call symput("no",left(put(count,8.)));
stop;
run;

%mend dupcheck;

%macro frecheck;
/*check whether each panel has the same item appear more than once*/

proc freq data=all_module;
     tables item_id*panel /nocol norow nopercent out = item_freq (where = (count>1))
noprint;
run;

data _null_;
if 0 then set item_freq nobs=count;
call symput("no1",left(put(count,8.)));
stop;
run;
%mend frecheck;


%macro assemble;

%if (&no1 EQ 0) %then %do;
   %mstfinal
         %end;
 %else %do;
      %do %until  (&no1 EQ 0);

           %assemble_module_tslt ;
                          %assemble_mst;
                          %dupcheck;
                          %frecheck;
            %end;
      %mstfinal
         %end;
%mend assemble;


%macro initial;

 data module_info;
    retain panel stage diff module_info;
run;
```

```
%mend initial;



%macro step2;
%initial;
%ass;
%assemble_mst;
%frecheck;
%assemble ;

%mend step2;


%macro test_examinee_par ;

data three.examinees;
      retain i j id theta class_true;
                  id=0; theta=-3.25;
      do i=1 to 25;
            do j= 1 to 100;
                     theta=theta+.25;
                   if theta <1 then class_true='0';
                    else class_true='1';
                     id=id+1;
          output;
        end;
                  end;
                  drop q;
run;

%mend test_examinee_par;

%test_examinee_par


filename tslt "c:\dissertation\pool&pool\scoright\res\testlet.est";

data tslt;
    infile tslt firstobs=4;
          input tslt $ a $ est se_est;
run;

data a;
   do tslt_id=1 to &T;
           output;
```

```
        end;
run;

data two.tslt;
    set a;
        set tslt;
        keep tslt_id est;
        rename est=tslt_var;
run;

%macro item_pars;
/*read in estimated tslt var information*/
data items;
    set three.mst_items;
run;

data itempar_tslt;
    set items;
        do i=1 to &T;
          m=(i-1)*20;
          n=+20*i;
              if (item_id > m) & (item_id <= n) then tslt_id=i;

        end;
run;

proc sort data=itempar_tslt;
    by tslt_id;
run;


data itempar_tslt1;
    merge itempar_tslt two.tslt ;
          by tslt_id;
          if panel ne .;
run;

proc sort data=itempar_tslt1 out=itempar;
    by panel stage diff;
run;
%mend;




/*EAP estimation of 3PL testlet model*/
```

```
%macro qp;
data qp (keep=qp1-qp&nqpt) qpw (keep=qpw1-qpw&nqpt);

    array qp{&nqpt} qp1-qp&nqpt;
    array qpw{&nqpt} qpw1-qpw&nqpt;

  qp1=-1*(&thetamax);
    do i=2 to &nqpt;
       qp{i}=qp{i-1}+2*(&thetamax)/(&nqpt-1);
    end;

  *Determine weights of normal distribution at quadrature points;
  sum=0;
  do j=1 to &nqpt;
    qpw{j}=probnorm((qp{j}+2*(&thetamax)/((&nqpt-1)*2))/sqrt(&sigma))
           -probnorm((qp{j}-2*(&thetamax)/((&nqpt-1)*2))/sqrt(&sigma));
          /*the same procedure has been used in SIMCAT1.0 by Gilles Raiche &Jean-Guy
Blais*/
  end;

run;
%mend qp;

%qp



%macro qp_tslt (stage=);

   %if &stage=1 %then %do;
   data a;
            set itempar10 (firstobs=1);
  run; %end;

         %else %do;
            data a;
             set itempar&stage&diff (firstobs=1);
         run;
         %end;

   data _null_;
            set a;
                  if _n_=1 then
            call symputx ("sigma_tslt", tslt_var);
```

```
                    else stop;
    run;

data qp_tslt (keep=qp_tslt1-qp_tslt&nqpt) qpw_tslt(keep=qpw_tslt1-qpw_tslt&nqpt);

    array qp_tslt{&nqpt} qp_tslt1-qp_tslt&nqpt;
    array qpw_tslt{&nqpt} qpw_tslt1-qpw_tslt&nqpt;

  qp_tslt1=-1*(&thetamax);
    do i=2 to &nqpt;
       qp_tslt{i}=qp_tslt{i-1}+2*(&thetamax)/(&nqpt-1);
    end;

  *Determine weights of normal distribution at quadrature points;
  do j=1 to &nqpt;
    qpw_tslt{j}=probnorm((qp_tslt{j}+2*(&thetamax)/((&nqpt-1)*2))/sqrt(&sigma_tslt))
          -probnorm((qp_tslt{j}-2*(&thetamax)/((&nqpt-1)*2))/sqrt(&sigma_tslt));
          /*the same procedure has been used in SIMCAT1.0 by Gilles Raiche &Jean-Guy
Blais*/
  end;

run;

%mend qp_tslt;

%macro response (person=, item=);

Proc IML;

    use &person;
    read all var {theta} into theta; /* Theta matrix is person parameter vector*/

          N=nrow(theta); /*get the number of examinees */
          call symputx ("N_person", N);  /*generate macro variable N_person to use later*/

          Use &item;
          read all var {b} into B; B=J(N,1,1)*t(B);
          read all var {a} into A; A=J(N,1,1)*t(A);
          read all var {c} into C; C=J(N,1,1)*t(C);
          J=ncol(B); /*get the number of items */  call symputx ("N_item", J);
          x = J(1,J,1);    /*CREATES A 1 X J with all ones MATRIX */

          Theta = theta * x;

            tslt10=rannor(j(2500,1,345))*sqrt(&r);
```

```
                tslt20=rannor(j(2500,1,12345))*sqrt(&r);
                tslt21=rannor(j(2500,1,793451))*sqrt(&r);
                tslt30=rannor(j(2500,1,345789))*sqrt(&r);
                tslt31=rannor(j(2500,1,7934531))*sqrt(&r);

          tslt10=tslt10*j(1,12,1);
      tslt20=tslt20*j(1,12,1);
                tslt21=tslt21*j(1,12,1);
                tslt30=tslt30*j(1,12,1);
                tslt31=tslt31*j(1,12,1);


          tslt=tslt10||tslt20||tslt21||tslt30||tslt31;


          *compute probablity;
                D11=THETA-B-TSLT;D12=A#D11;
                D13=EXP(D12);  D14=1+D13;
      D15=D13/D14;  P=C+(1-C)#D15;

          *generate random variable;
      R=ranuni(J(N,J,&seed));

          *generate response;
                X=(P>=R);

          CREATE response from X;
          APPEND from X;

    Close &person;
    Close &item;

Quit;

%mend response;

%macro lkhd (stage=);  /*caculate likelihood at each quadrature point of theta on stage 1
or stage 3)*/

Proc IML;
    use qp;
    read all into theta;
          theta=t(theta);
          close qp;

          N=nrow(theta); /*get the number of examinees */
```

```
        call symputx ("N_person", N);  /*generate macro variable N_person to use later*/


        Use itempar&stage&diff;
                read all var {b_est1} into B; B=J(N,1,1)*t(B);
                read all var {a_est1} into A; A=J(N,1,1)*t(A);
                read all var {c_est1} into C; C=J(N,1,1)*t(C);
                  close itempar&stage&diff;


             use resp&stage&diff;
                read all into res;
             close resp&stage&diff;

        J=ncol(B); /*get the number of items */  call symputx ("N_item", J);
              x = J(1,J,1);       /*CREATES A 1 X J with all ones MATRIX */


        Theta = theta * x;


   *compute probablity;
                D11=THETA-B;D12=A#D11;
                D13=EXP(D12);  D14=1+D13;
                D15=D13/D14;  P=C+(1-C)#D15;

                x=j(15, 12,0); lk=j(15,1,1);
                     do  j=1 to 12;
                          if res[,j]=1 then x[,j]=p[,j];
                            else if res[,j]=0 then x[,j]=1-p[,j];
                                 else x[,j]=1;
                           lk[,1]=lk[,1]#x[,j];
                      end;
                     lkhd=t(lk);

        create lkhd&stage from lkhd;
        append from lkhd;
Quit;

%mend lkhd;

%macro lkhd_tslt (stage=);


Proc IML;
```

```
w=j(1,&nqpt,1);

      use qpw;/*read in quadrature point of theta*/
read all into qpw; /* Theta matrix is person parameter vector*/
      close qpw;

/*prepare item parameters*/
      Use itempar&stage&diff;
      read all var {b_est1} into B;  b1=b; *print b;
      read all var {a_est1} into A;  a1=a;
      read all var {c_est1} into C;  c1=c;
      close itempar&stage&diff;



      N_item=nrow(B); /*get the number of items */

do i=2 to &nqpt; /*do it for each item and each person */
            b=b//b1; a=a//a1; c=c//c1;
      end;
      b=b*w; a=a*w; c=c*w;

use qp;/*read in quadrature point of theta*/
read all into theta; /* Theta matrix is person parameter vector*/
      read all into qp;
      theta=J(&nqpt*N_item,1,1)*theta;  /*   (15*40)    */
      close qp;

      /*prepare tslt*/
        use qp_tslt;
           read all into tslt; tslt=t(tslt);
        close qp_tslt;
        use qpw_tslt;
                  read all into qpw_tslt;
        close qpw_tslt;
                  x=J(N_item,1,1); tslt2=x*tslt[1,];
        do i=1 to &nqpt; /*each quadrature point repeated for  times*/
          tslt2=x*tslt[i,];
          tslt1=tslt1//tslt2;
        end;
          tslt=tslt1*w;


      /*compute probability*/

            D11=THETA-B-TSLT;D12=A#D11;
```

```
D13=EXP(D12);  D14=1+D13;
D15=D13/D14;  P=C+(1-C)#D15;


/*prepare response data*/
Use resp&stage&diff;
read all var _num_ into response;
close resp&stage&diff;

/*nth person's response*/

  resp=response ;  /* nth person's response pattern 1*40  */
  res=t(resp);   /* nth person's response pattern 40*1  */
  res=res*w;

      res1=res;
      do i=2 to &nqpt;
        res=res//res1;
      end;

  /*compute p or q*/
    xx=j(&nqpt*N_item,&nqpt,0);
    do i=1 to &nqpt*N_item;
      do j=1 to &nqpt;
          if res[i,j]=1 then xx[i,j]=p[i,j];
            else xx[i,j]=1-p[i,j];
          end;
    end;

  /*compute likelihood at each quadrature point of tslt and each quadreture point
of theta*/

  lkhd=j(&nqpt+1,1,1); /*likelihood at all quadreutre points of theta*/

 do qp_theta= 1 to &nqpt; /*for each quadrature point of theta*/
lkhd_v=j(1,1,1);/*likelihood at  all quadrature point of tslt*/
    do i=1 to &nqpt;  /*for each quadrature point of testlet*/
        x=N_item*(i-1)+1; y=N_item*i;
            xxqp=xx[x:y,qp_theta];  /*xx for each quadrature point of testlet*/
      lk=1;
            do a= 1 to N_item; /*for each item*/
       xx_qp_qp=xxqp[a,];
                      lk=lk*xx_qp_qp;
                  end;
            lkhd_v=lkhd_v//lk;
              end;
```

```
                    lkhd=lkhd||lkhd_v;
                end;
                lkhd=lkhd[2:&nqpt+1,2:&nqpt+1];


    /* compute quadature point of tslt times the likelihood at each quadrature point of
tslt*/
        lkhd_qp_theta=qpw_tslt*lkhd;

                create lkhd&stage from lkhd_qp_theta;
                append from lkhd_qp_theta;
quit;

%mend lkhd_tslt;


%macro eap_est (stage=);
proc iml;

    if &stage=1 then do;
            Use lkhd&stage;
            read all var _num_ into lkhd&stage;
            close lkhd&stage;
        lkhd_qp_theta=lkhd&stage;
           end;
      else if &stage=2 then do;
                        use lkhd1;
                        read all into lkhd1;
                        close lkhd1;

          Use lkhd2;
                read all var _num_ into lkhd2;
                close lkhd2;
                lkhd_qp_theta=lkhd1#lkhd2;
                      end;

                    else if &stage=3 then do;
                        use lkhd1;
                           read all into lkhd1;
                        close lkhd1;

          Use lkhd2;
                   read all var _num_ into lkhd2;
                close lkhd2;

                        use lkhd3;
```

```
                                read all var _num_ into lkhd3;
            close lkhd3;

            lkhd_qp_theta=lkhd1#lkhd2#lkhd3;
                  end;

    use qpw;/*read in quadrature point of theta*/
    read all into qpw; /* Theta matrix is person parameter vector*/
          close qpw;

          use qp;/*read in quadrature point of theta*/
          read all into qp;
   /* compute EAP estimate for each person */
                a=t(qpw);
                denomi=lkhd_qp_theta*a;
                b=t(qp#qpw);
                numera=lkhd_qp_theta*b;
          eap=numera/denomi;

         /*write estimate to SAS data file*/
     CREATE est&stage from eap /*[colname=¡¯eap_theta¡¯]*/;;
     APPEND from eap;

quit;
%mend eap_est ;

%macro route (stage=);
   data est&stage;
       set est&stage;
            if col1≤.5 then diff=0;
               else diff=1;
            call symputx("diff", diff);
   run;
        %put _user_;
%mend route;

%macro itempar;


data itempar10;
    set items_panel ;
          if stage=1;
run;
data itempar20;
    set items_panel ;
          if stage=2 and diff=0;
```

```
run;
data itempar21;
    set items_panel ;
          if stage=2 and diff=1;
run;
data itempar30;
    set items_panel ;
          if stage=3 and diff=0;
run;

data itempar31;
    set items_panel ;
          if stage=3 and diff=1;
run;

%mend itempar;



%macro eap_all (panel=);

data est;
run;

%do n=1 %to 2500;

    data resp10(keep=col1-col12)
        resp20 (keep=col13-col24)
        resp21 (keep=col25-col36)
        resp30 (keep=col37-col48)
                    resp31 (keep=col49-col60)
            ;
            set response (firstobs=%eval(&n) obs=%eval(&n));
    run;


    data _null_;
              a=0;
                    if _n_=1 then
              call symputx ("diff", a);
                     else stop;
    run;
          %qp_tslt (stage=1);
       %lkhd_tslt (stage=1);
          %eap_est (stage=1);
          %route (stage=1);
```

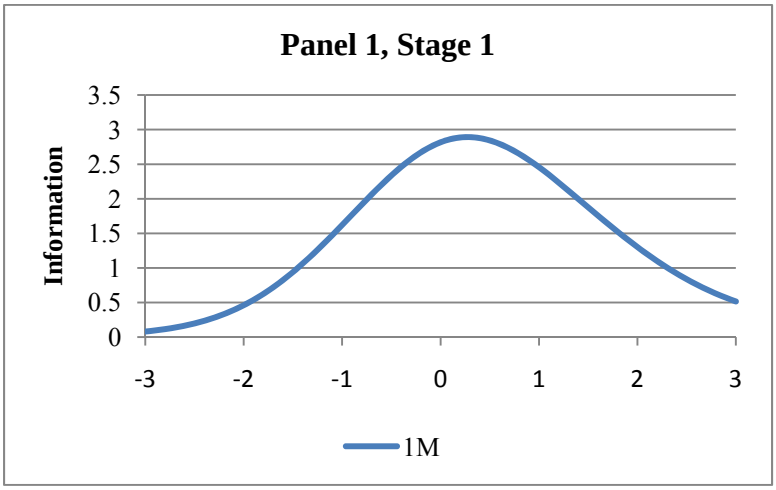
```
        %qp_tslt (stage=2)
   %lkhd_tslt (stage=2);
   %eap_est (stage=2)
   %route (stage=2);
   %lkhd_tslt (stage=3);
   %eap_est (stage=3);

data est;
          set est est3;
run;

%end;
data three.est&panel;
    set est (firstobs=2);
run;

%mend eap_all;

%macro result;
  %do panel=1 %to 8;
        data result;
            set result;
            set three.est&panel (rename=(col1=est&panel));
         run;
                %end;
%mend;



%macro step3;

%item_pars;

%do panel=1 %to 8;
   data items_panel;
       set itempar;
            if panel=&panel;
   run;

         %response (person=three.examinees, item=items_panel);
         %itempar;
         %eap_all (panel=&panel);

%end;

data result;
```

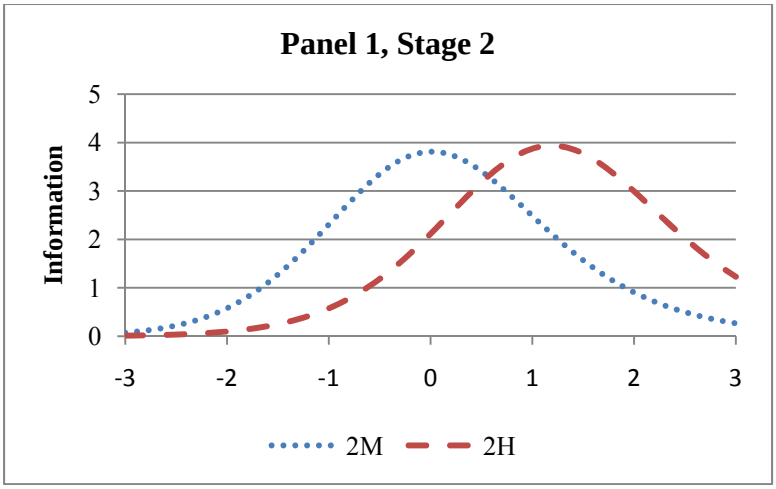
```
    set three.examinees;
run;
%result;


  /*calculate estimate*/

data estimate;
    set result;
          estimate=(est1+est2+est3+est4+est5+est6+est7+est8)/8;
          if estimate<1 then class='0';
            else class='1';
          bias=estimate-theta; bias_sq=bias*bias;
          *keep theta estimate class bias bias_sq;
          if (class='1') and (class_true='1') then cp=1; else cp=0;
    if (class='0') and (class_true='0') then cf=1; else cf=0;
          if (class='1') and (class_true='0') then fp=1; else fp=0;
    if (class='0') and (class_true='1') then fn=1; else fn=0;

run;

proc sort data=estimate;
    by theta;
run;

proc means noprint data=estimate;
    by theta;
          var estimate bias bias_sq;
          output out=estimate1 mean=estimate bias bias_sq;
run;

data three.bias&rep;
    set estimate1;
          drop _type_ _freq_;
          rmse=sqrt(bias_sq);
          drop bias_sq;
run;


/*evaluate pass-fail decisions*/
proc means noprint data=estimate mean;
    var cp cf fp fn;
          by i;
          output out=pass_fail mean= cp cf fp fn;
run;
```

```
data three.decisions&rep;
    set pass_fail;
          drop _type_ _freq_;
          deci_corr=cp+cf;
          deci_incorr=fp+fn;
run;


%mend step3;


%macro replication;
    %do rep=1 %to &num_rpl;
             %initial
             %step2
            %step3;
          %end;
%mend replication;

%replication;


proc printto log=log;
run;
```

Appendix D: Examples of MST Test Information Curves

Figure D-1: One of the panels with item pool 2 using the 3PL model





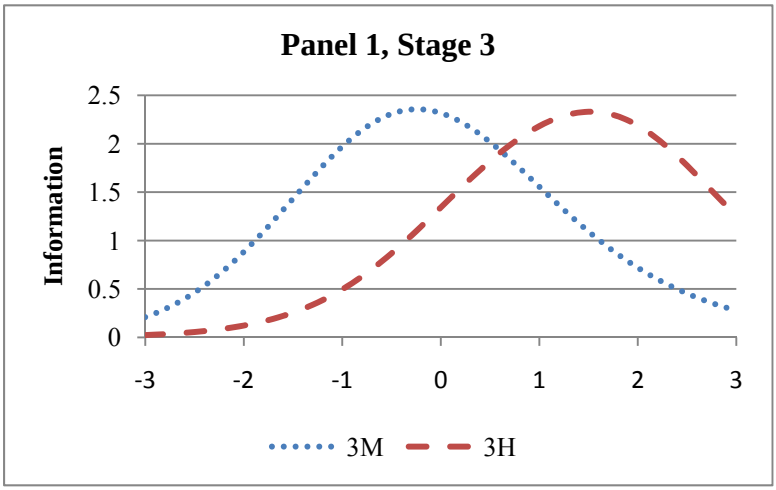


Figure D-2: One of the modules across the panels with item pool 2 using the 3PL model

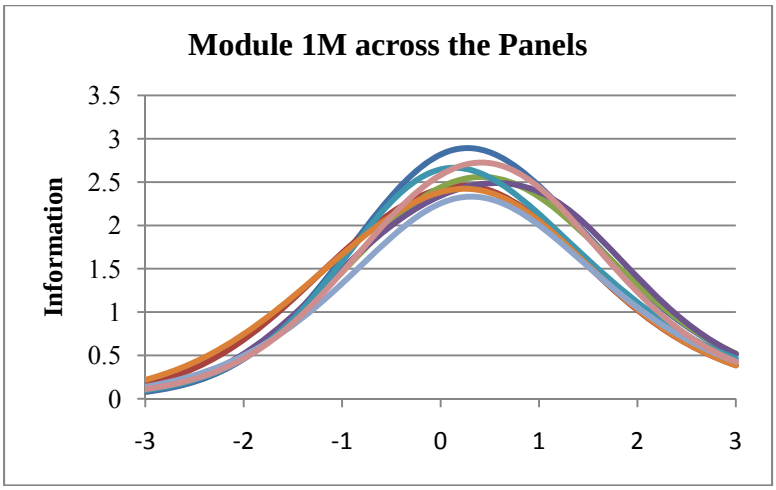


Figure D-3: One of the panels with item pool 8 using the 3PL testlet model

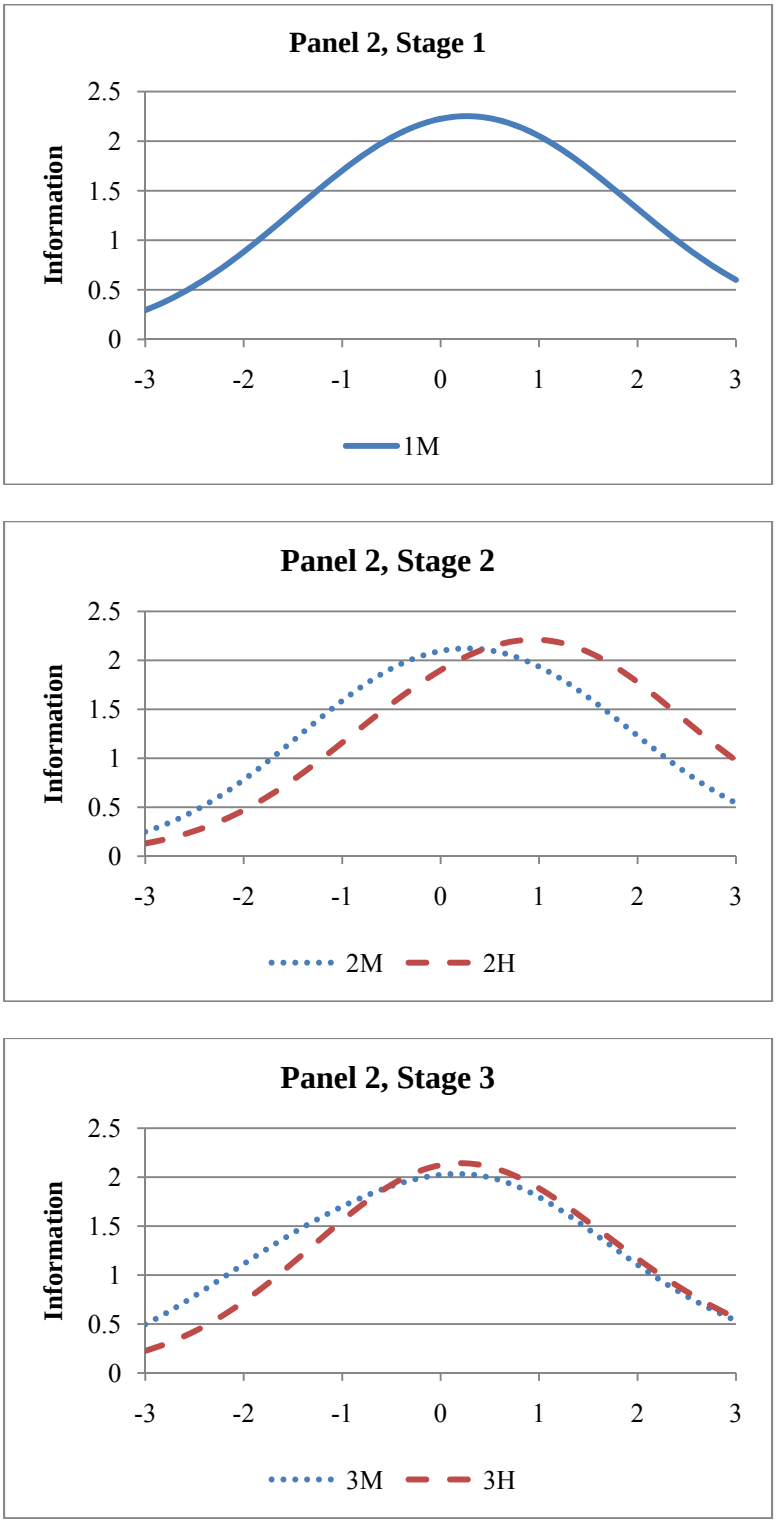
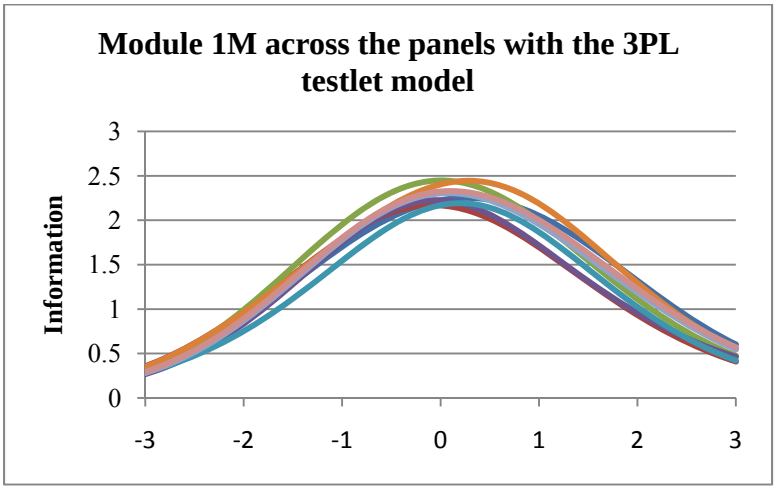


Figure D-4: One of the panels with item pool 8 using the 3PL testlet model



Appendix E: ANOVA Analysis Results

Table E- 1: ANOVA results for test length effect

Condition	Dependent Variable	df	Sum of Squares	F-value	p-value	Partial Eta Squared
	BIAS	1	0.000	0.550	0.461	0.009
	RMSE	1	0.021	5420.458	0.000	0.989
1	DA	1	0.001	118.397	0.000	0.671
	BIAS	1	0.000	1.548	0.218	0.026
	RMSE	1	0.018	3857.085	0.000	0.985
2	DA	1	0.001	60.703	0.000	0.511
	BIAS	1	0.000	5.037	0.029	0.080
	RMSE	1	0.022	4388.489	0.000	0.987
3	DA	1	0.001	65.696	0.000	0.531
	BIAS	1	0.000	4.351	0.041	0.070
	RMSE	1	0.021	4851.487	0.000	0.988
4	DA	1	0.001	75.976	0.000	0.567
	BIAS	1	0.000	19.772	0.000	0.254
	RMSE	1	0.021	4704.740	0.000	0.988
5	DA	1	0.001	124.330	0.000	0.682
	BIAS	1	0.000	46.280	0.000	0.444
	RMSE	1	0.020	4267.946	0.000	0.987
6	DA	1	0.001	70.437	0.000	0.548
	BIAS	1	0.000	14.150	0.000	0.196
	RMSE	1	0.021	4737.559	0.000	0.988
7	DA	1	0.001	71.120	0.000	0.551
	BIAS	1	0.000	1.582	0.214	0.027
	RMSE	1	0.020	3723.154	0.000	0.985
8	DA	1	0.001	117.038	0.000	0.669
	BIAS	1	0.000	2.327	0.133	0.039
	RMSE	1	0.024	3965.819	0.000	0.986
9	DA	1	0.001	73.778	0.000	0.560
	BIAS	1	0.000	5.296	0.025	0.084
	RMSE	1	0.024	4822.561	0.000	0.988
10	DA	1	0.001	170.095	0.000	0.746
11	BIAS	1	0.000	7.473	0.008	0.114

	RMSE	1	0.019	3037.227	0.000	0.981
	DA	1	0.001	104.027	0.000	0.642
	BIAS	1	0.000	0.078	0.781	0.001
	RMSE	1	0.019	3096.720	0.000	0.982
12	DA	1	0.001	91.236	0.000	0.611
	BIAS	1	0.000	0.460	0.500	0.008
	RMSE	1	0.019	2624.966	0.000	0.978
13	DA	1	0.001	73.020	0.000	0.557
	BIAS	1	0.000	7.237	0.009	0.111
	RMSE	1	0.023	5110.499	0.000	0.989
14	DA	1	0.001	112.652	0.000	0.660
	BIAS	1	0.000	7.476	0.008	0.114
	RMSE	1	0.021	2787.624	0.000	0.980
15	DA	1	0.001	143.074	0.000	0.712
	BIAS	1	0.000	12.150	0.001	0.173
	RMSE	1	0.024	3924.519	0.000	0.985
16	DA	1	0.001	149.568	0.000	0.721
	BIAS	1	0.000	0.675	0.415	0.011
	RMSE	1	0.021	3243.123	0.000	0.982
17	DA	1	0.002	158.465	0.000	0.732
	BIAS	1	0.000	0.307	0.581	0.005
	RMSE	1	0.021	2373.047	0.000	0.976
18	DA	1	0.001	170.560	0.000	0.746
	BIAS	1	0.000	0.000	0.985	0.000
	RMSE	1	0.021	2769.236	0.000	0.979
19	DA	1	0.001	90.817	0.000	0.610
	BIAS	1	0.000	2.914	0.093	0.048
	RMSE	1	0.015	2684.296	0.000	0.979
20	DA	1	0.001	206.915	0.000	0.781
	BIAS	1	0.000	9.845	0.003	0.145
	RMSE	1	0.016	2034.272	0.000	0.972
21	DA	1	0.001	188.294	0.000	0.765
	BIAS	1	0.000	20.094	0.000	0.257
	RMSE	1	0.016	1148.344	0.000	0.952
22	DA	1	0.001	149.873	0.000	0.721

Table E-2: ANOVA analysis results for test length, model and their interaction effect

Simulation Condition	Dependent Variable	Source	df	Sum of Squares	F-value	p-value	Partial eta square
2	BIAS	Length	1	0.000	35.122	0.000	0.232
		Model	1	0.025	3350.616	0.000	0.967
		Length*Model	1	0.000	63.627	0.000	0.354
	RMSE	Length	1	0.072	14831.005	0.000	0.992
		Model	1	0.996	205855.343	0.000	0.999
		Length*Model	1	0.006	1252.595	0.000	0.915
	DA	Length	1	0.001	129.829	0.000	0.528
		Model	1	0.002	323.586	0.000	0.736
		Length*Model	1	0.000	1.058	0.306	0.009
3	BIAS	Length	1	0.000	1.188	0.278	0.010
		Model	1	0.028	3003.511	0.000	0.963
		Length*Model	1	0.000	5.549	0.020	0.046
	RMSE	Length	1	0.075	13616.701	0.000	0.992
		Model	1	1.018	184119.217	0.000	0.999
		Length*Model	1	0.004	803.253	0.000	0.874
	DA	Length	1	0.000	24.794	0.000	0.176
		Model	1	0.000	20.270	0.000	0.149
		Length*Model	1	0.000	58.788	0.000	0.336
4	BIAS	Length	1	0.000	0.694	0.407	0.006
		Model	1	0.028	3723.139	0.000	0.970
		Length*Model	1	0.000	9.048	0.003	0.072
	RMSE	Length	1	0.089	23417.215	0.000	0.995
		Model	1	1.001	263863.827	0.000	1.000
		Length*Model	1	0.009	2343.294	0.000	0.953
	DA	Length	1	0.002	240.959	0.000	0.675
		Model	1	0.000	3.081	0.082	0.026
		Length*Model	1	0.000	0.051	0.821	0.000
5	BIAS	Length	1	0.001	98.619	0.000	0.460
		Model	1	0.044	5542.320	0.000	0.979
		Length*Model	1	0.002	292.875	0.000	0.716
	RMSE	Length	1	0.073	9922.876	0.000	0.988
		Model	1	0.978	132266.298	0.000	0.999

		Length*Model	1	0.004	607.051	0.000	0.840
		Length	1	0.002	217.633	0.000	0.652
		Model	1	0.003	342.600	0.000	0.747
	DA	Length*Model	1	0.000	2.438	0.121	0.021
		Length	1	0.001	93.472	0.000	0.446
		Model	1	0.040	4925.493	0.000	0.977
	BIAS	Length*Model	1	0.000	0.941	0.334	0.008
		Length	1	0.076	16524.941	0.000	0.993
		Model	1	1.040	226987.839	0.000	0.999
	RMSE	Length*Model	1	0.006	1235.844	0.000	0.914
		Length	1	0.000	0.209	0.648	0.002
		Model	1	0.001	92.938	0.000	0.445
6	DA	Length*Model	1	0.001	158.904	0.000	0.578
		Length	1	0.000	0.206	0.651	0.002
		Model	1	0.049	5800.673	0.000	0.980
	BIAS	Length*Model	1	0.000	47.785	0.000	0.292
		Length	1	0.102	21629.505	0.000	0.995
		Model	1	1.044	222023.062	0.000	0.999
	RMSE	Length*Model	1	0.013	2811.915	0.000	0.960
		Length	1	0.001	114.683	0.000	0.497
		Model	1	0.000	35.246	0.000	0.233
7	DA	Length*Model	1	0.000	11.554	0.001	0.091
		Length	1	0.002	198.784	0.000	0.631
		Model	1	0.043	4269.352	0.000	0.974
	BIAS	Length*Model	1	0.003	263.429	0.000	0.694
		Length	1	0.077	13161.189	0.000	0.991
		Model	1	1.009	171582.234	0.000	0.999
	RMSE	Length*Model	1	0.006	1101.372	0.000	0.905
		Length	1	0.002	401.376	0.000	0.776
		Model	1	0.003	540.166	0.000	0.823
8	DA	Length*Model	1	0.000	17.995	0.000	0.134
		Length	1	0.000	53.048	0.000	0.314
		Model	1	0.029	3554.243	0.000	0.968
	BIAS	Length*Model	1	0.000	23.645	0.000	0.169
		Length	1	0.088	17546.867	0.000	0.993
		Model	1	1.063	211819.642	0.000	0.999
9	RMSE	Length*Model	1	0.006	1221.214	0.000	0.913

		Length	1	0.000	5.141	0.025	0.042
		Model	1	0.001	94.811	0.000	0.450
	DA	Length*Model	1	0.001	124.390	0.000	0.517
		Length	1	0.000	9.066	0.003	0.072
		Model	1	0.046	6927.549	0.000	0.984
	BIAS	Length*Model	1	0.000	48.865	0.000	0.296
		Length	1	0.108	22655.125	0.000	0.995
		Model	1	1.061	221948.566	0.000	0.999
	RMSE	Length*Model	1	0.012	2517.824	0.000	0.956
		Length	1	0.002	417.523	0.000	0.783
		Model	1	0.001	90.206	0.000	0.437
10	DA	Length*Model	1	0.000	0.023	0.880	0.000
		Length	1	0.001	160.146	0.000	0.580
		Model	1	0.019	2537.829	0.000	0.956
	BIAS	Length*Model	1	0.000	62.127	0.000	0.349
		Length	1	0.062	10387.721	0.000	0.989
		Model	1	1.247	207430.108	0.000	0.999
	RMSE	Length*Model	1	0.003	501.856	0.000	0.812
		Length	1	0.000	2.638	0.107	0.022
		Model	1	0.009	1042.494	0.000	0.900
11	DA	Length*Model	1	0.002	221.106	0.000	0.656
		Length	1	0.001	112.348	0.000	0.492
		Model	1	0.004	501.202	0.000	0.812
	BIAS	Length*Model	1	0.001	102.282	0.000	0.469
		Length	1	0.055	12365.106	0.000	0.991
		Model	1	1.119	252398.045	0.000	1.000
	RMSE	Length*Model	1	0.002	387.111	0.000	0.769
		Length	1	0.000	64.634	0.000	0.358
		Model	1	0.002	395.326	0.000	0.773
12	DA	Length*Model	1	0.000	30.627	0.000	0.209
		Length	1	0.000	0.190	0.664	0.002
		Model	1	0.020	2560.812	0.000	0.957
	BIAS	Length*Model	1	0.000	2.960	0.088	0.025
		Length	1	0.046	9924.570	0.000	0.988
		Model	1	1.203	259441.378	0.000	1.000
	RMSE	Length*Model	1	0.000	79.510	0.000	0.407
13	DA	Length	1	0.001	189.722	0.000	0.621

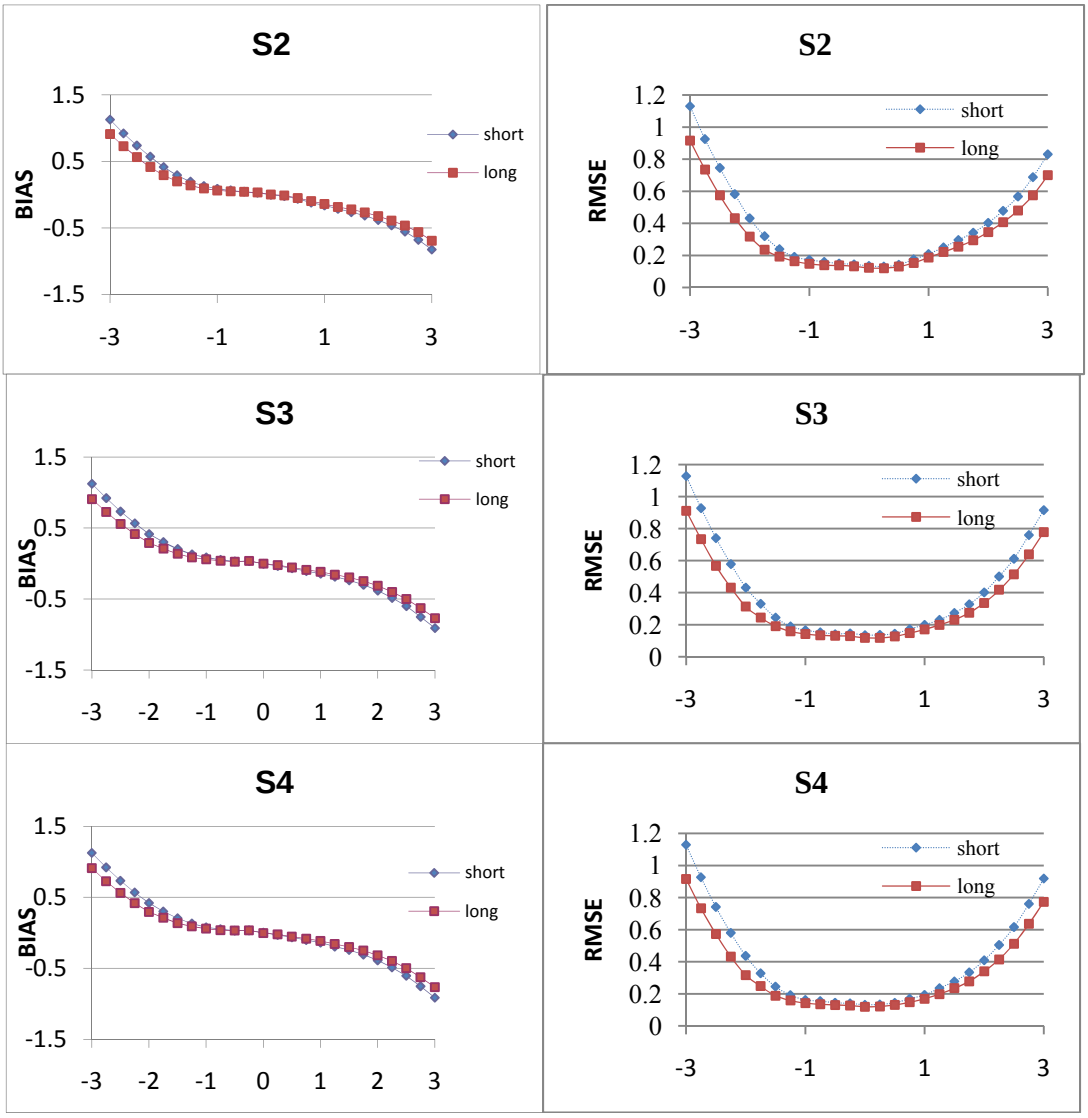
		Model	1	0.003	431.901	0.000	0.788
		Length*Model	1	0.000	1.422	0.236	0.012
		Length	1	0.000	11.183	0.001	0.088
		Model	1	0.033	4054.366	0.000	0.972
	BIAS	Length*Model	1	0.000	1.791	0.183	0.015
		Length	1	0.082	16518.971	0.000	0.993
		Model	1	1.653	333674.618	0.000	1.000
	RMSE	Length*Model	1	0.005	1031.274	0.000	0.899
		Length	1	0.001	128.943	0.000	0.526
		Model	1	0.022	2304.296	0.000	0.952
14	DA	Length*Model	1	0.000	23.317	0.000	0.167
		Length	1	0.001	123.409	0.000	0.515
		Model	1	0.022	2665.156	0.000	0.958
	BIAS	Length*Model	1	0.002	257.163	0.000	0.689
		Length	1	0.062	10353.238	0.000	0.989
		Model	1	1.587	266961.573	0.000	1.000
	RMSE	Length*Model	1	0.002	333.230	0.000	0.742
		Length	1	0.001	81.135	0.000	0.412
		Model	1	0.010	1107.118	0.000	0.905
15	DA	Length*Model	1	0.001	86.898	0.000	0.428
		Length	1	0.000	5.470	0.021	0.045
		Model	1	0.037	3642.278	0.000	0.969
	BIAS	Length*Model	1	0.001	73.409	0.000	0.388
		Length	1	0.111	6597.188	0.000	0.983
		Model	1	1.802	106777.697	0.000	0.999
	RMSE	Length*Model	1	0.013	798.307	0.000	0.873
		Length	1	0.002	246.412	0.000	0.680
		Model	1	0.007	873.187	0.000	0.883
16	DA	Length*Model	1	0.000	5.205	0.024	0.043
		Length	1	0.000	44.889	0.000	0.279
		Model	1	0.008	839.382	0.000	0.879
	BIAS	Length*Model	1	0.000	28.155	0.000	0.195
		Length	1	0.111	12399.554	0.000	0.991
		Model	1	1.819	203216.288	0.000	0.999
	RMSE	Length*Model	1	0.017	1900.034	0.000	0.942
		Length	1	0.003	317.438	0.000	0.732
17	DA	Model	1	0.027	3216.005	0.000	0.965

		Length*Model	1	0.000	7.456	0.007	0.060
		Length	1	0.001	131.297	0.000	0.531
		Model	1	0.010	1029.090	0.000	0.899
	BIAS	Length*Model	1	0.001	152.990	0.000	0.569
		Length	1	0.082	12398.162	0.000	0.991
		Model	1	1.649	249853.613	0.000	1.000
	RMSE	Length*Model	1	0.006	941.298	0.000	0.890
		Length	1	0.002	318.170	0.000	0.733
		Model	1	0.014	1912.356	0.000	0.943
18	DA	Length*Model	1	0.000	0.878	0.351	0.008
		Length	1	0.003	275.145	0.000	0.703
		Model	1	0.020	2031.291	0.000	0.946
	BIAS	Length*Model	1	0.003	276.270	0.000	0.704
		Length	1	0.078	12131.528	0.000	0.991
		Model	1	1.667	259210.674	0.000	1.000
	RMSE	Length*Model	1	0.006	891.762	0.000	0.885
		Length	1	0.001	122.846	0.000	0.514
		Model	1	0.005	506.528	0.000	0.814
19	DA	Length*Model	1	0.000	23.139	0.000	0.166
		Length	1	0.000	0.278	0.599	0.002
		Model	1	0.062	7767.534	0.000	0.985
	BIAS	Length*Model	1	0.000	14.720	0.000	0.113
		Length	1	0.018	5780.463	0.000	0.980
		Model	1	1.865	613153.501	0.000	1.000
	RMSE	Length*Model	1	0.002	553.189	0.000	0.827
		Length	1	0.003	650.983	0.000	0.849
		Model	1	0.009	1912.941	0.000	0.943
20	DA	Length*Model	1	0.000	3.598	0.060	0.030
		Length	1	0.000	0.010	0.922	0.000
		Model	1	0.020	2577.276	0.000	0.957
	BIAS	Length*Model	1	0.000	36.049	0.000	0.237
		Length	1	0.010	2228.398	0.000	0.951
		Model	1	3.514	823936.565	0.000	1.000
	RMSE	Length*Model	1	0.007	1629.035	0.000	0.934
		Length	1	0.001	208.756	0.000	0.643
		Model	1	0.048	10007.879	0.000	0.989
21	DA	Length*Model	1	0.000	81.445	0.000	0.412

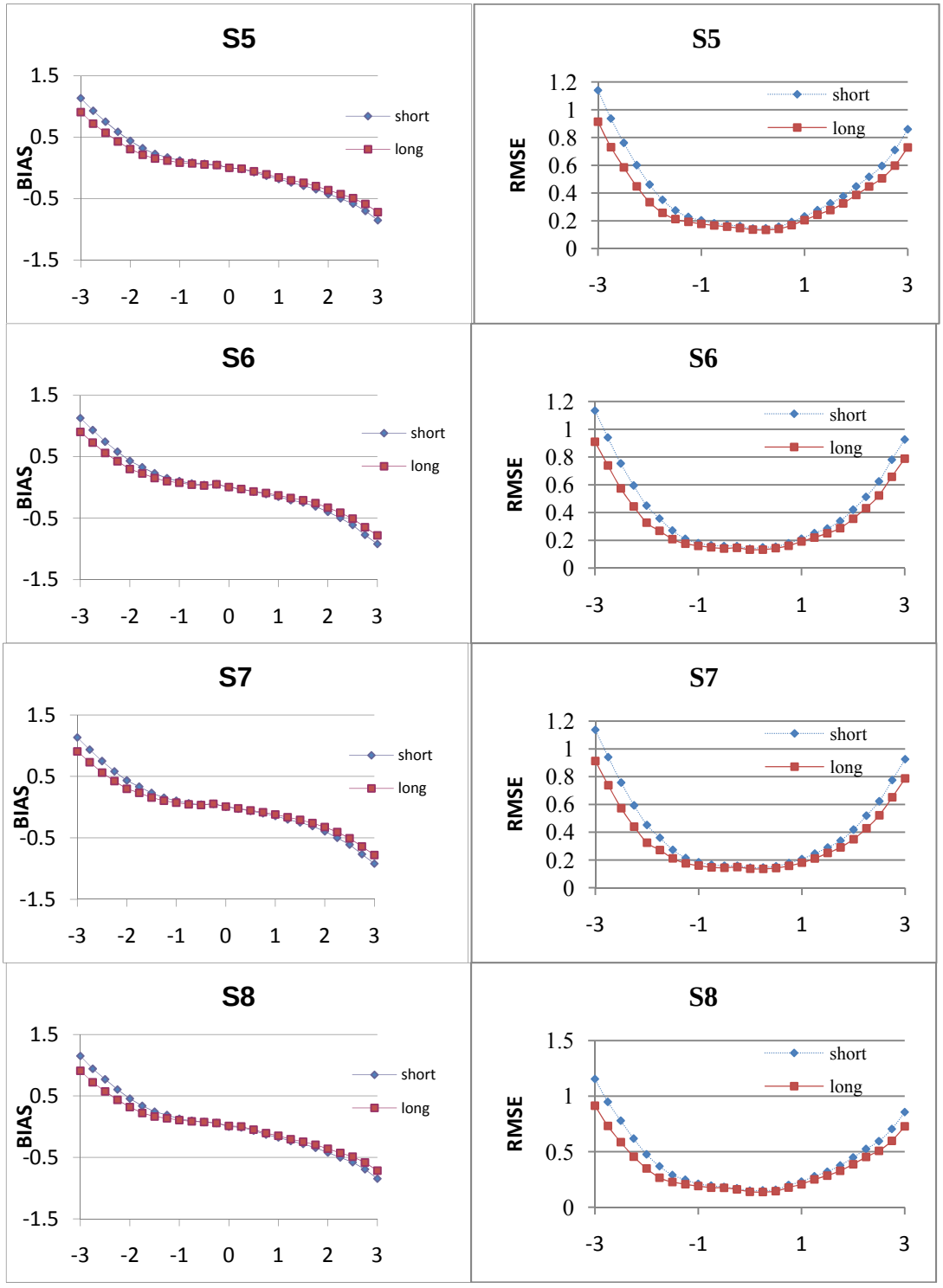
		Length	1	0.000	2.522	0.115	0.021
		Model	1	0.039	3405.834	0.000	0.967
	BIAS	Length*Model	1	0.001	49.836	0.000	0.301
		Length	1	0.011	1169.044	0.000	0.910
		Model	1	4.590	469408.138	0.000	1.000
	RMSE	Length*Model	1	0.005	521.961	0.000	0.818
		Length	1	0.001	115.937	0.000	0.500
		Model	1	0.087	14944.972	0.000	0.992
22	DA	Length*Model	1	0.001	119.223	0.000	0.507

Appendix F: Comparison of BIAS and RMSE under different simulation conditions

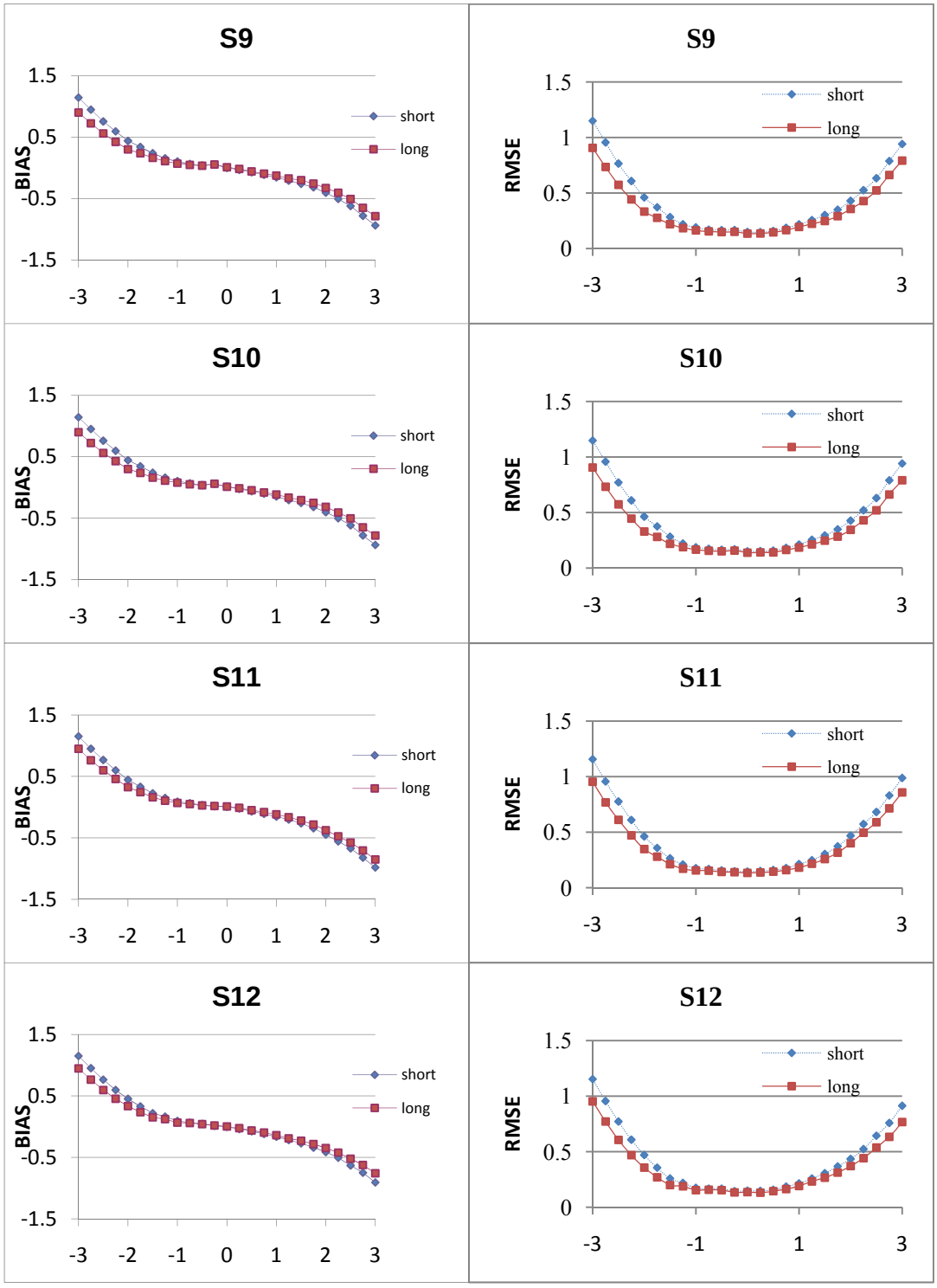
Figure F-1: Comparison of BIAS and RMSE for test length effect



F-1: Comparison of BIAS and RMSE for test length effect, continued

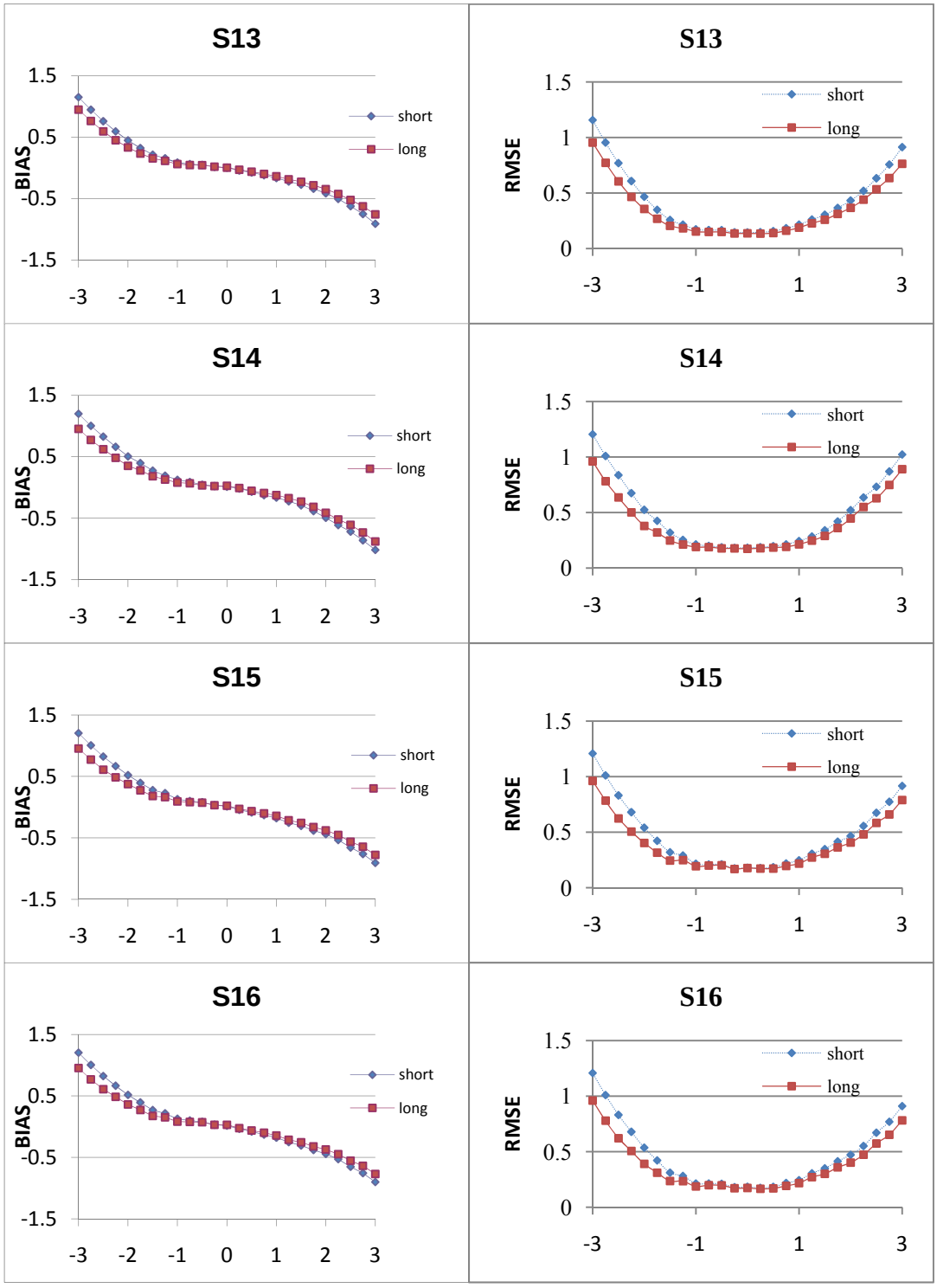


F-1: Comparison of BIAS and RMSE for test length effect, continued



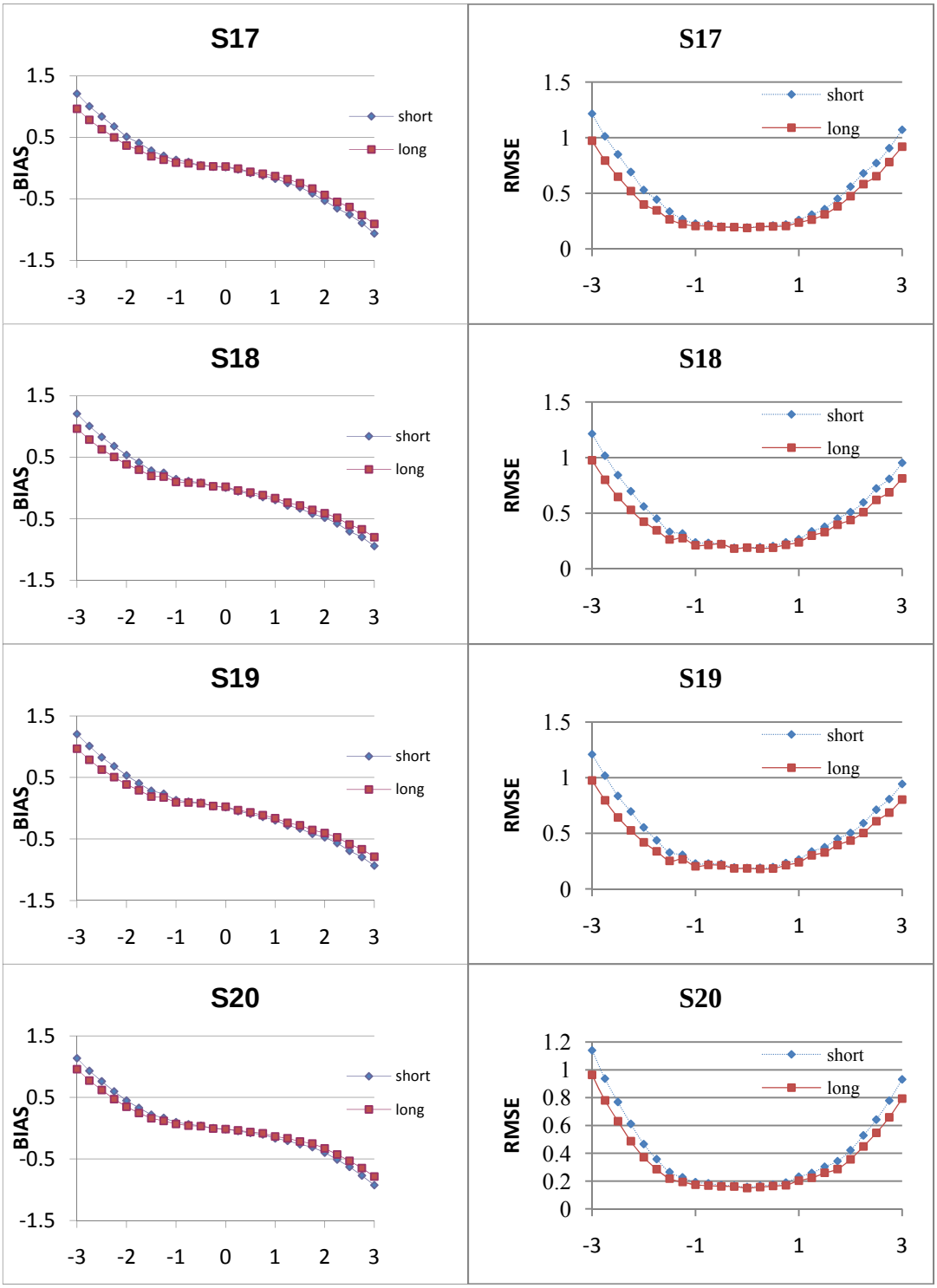
F-1:

F-1: Comparison of BIAS and RMSE for test length effect, continued



F-1:

F-1: Comparison of BIAS and RMSE for test length effect, continued



F-1: Comparison of BIAS and RMSE for test length effect, continued

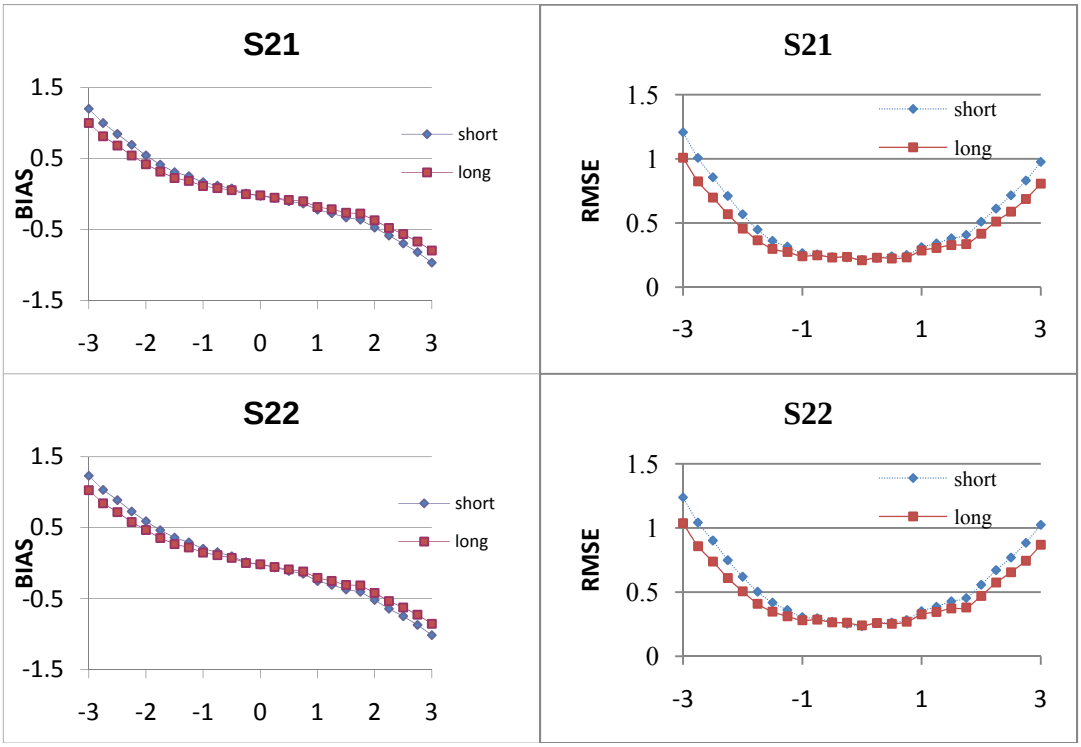
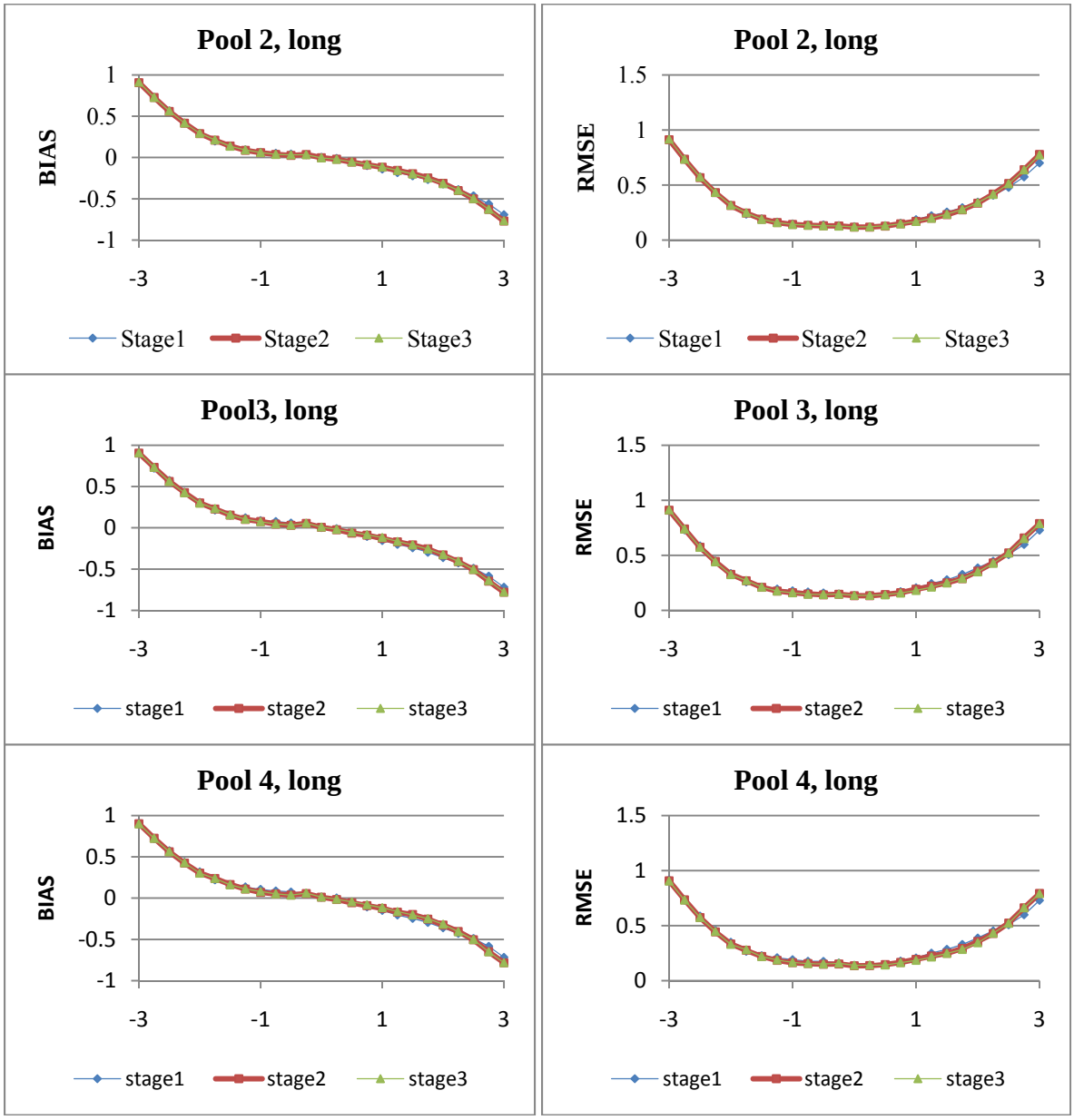
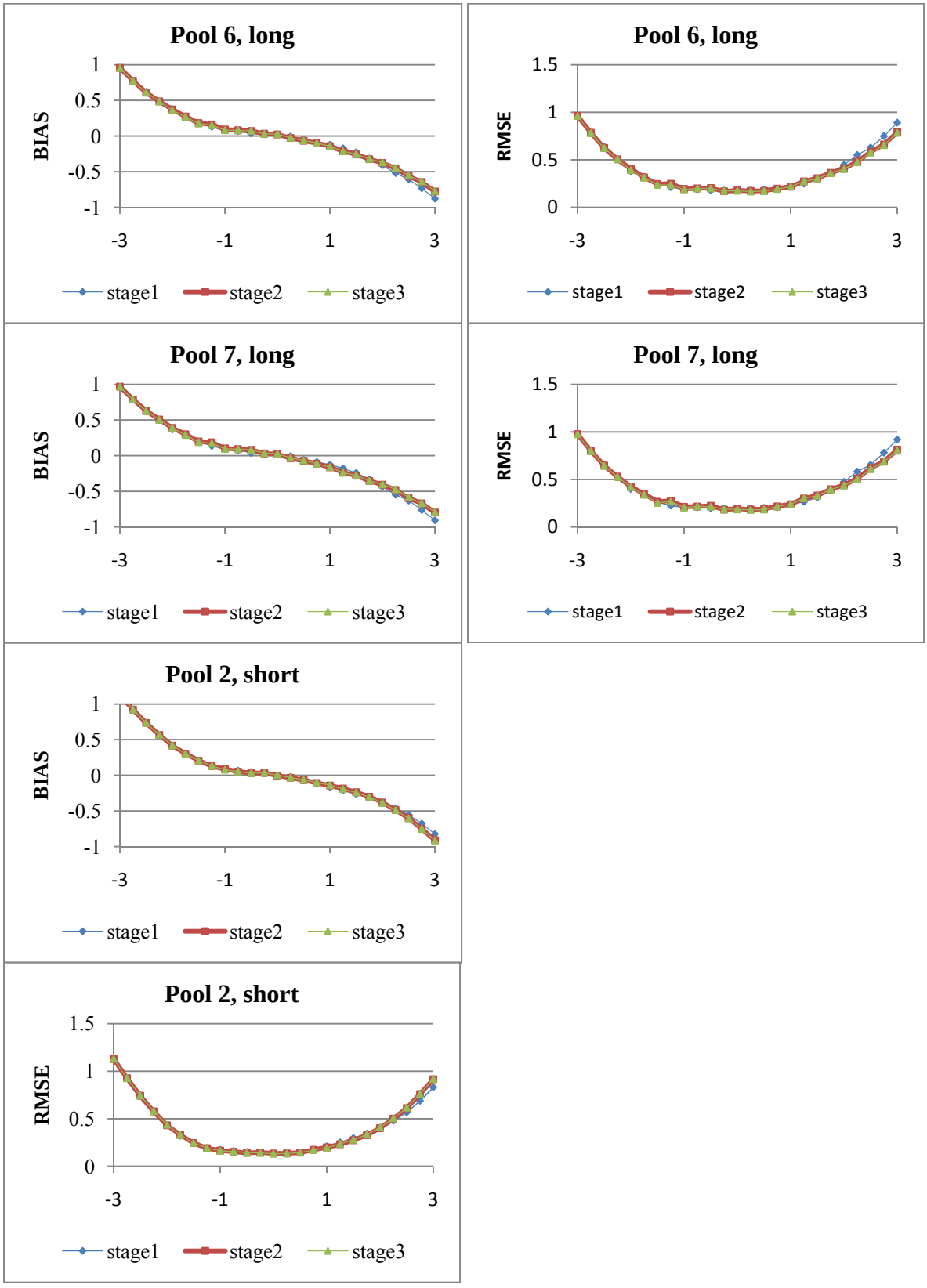
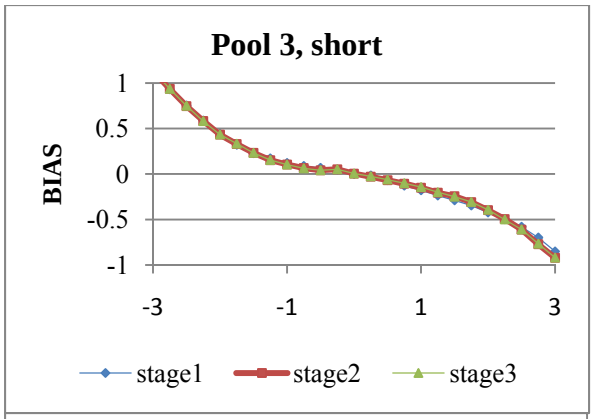


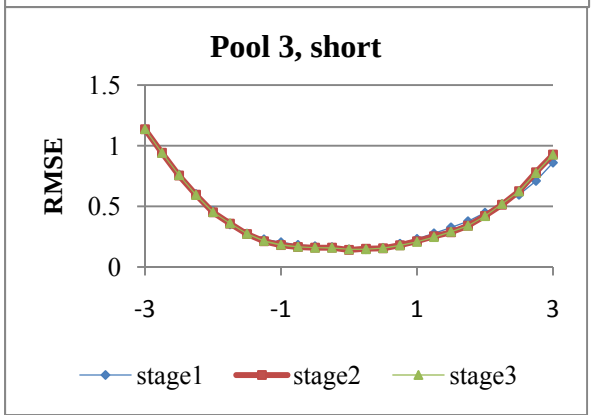
Figure F-2: Comparison of BIAS and RMSE for testlet/discrete item position effect



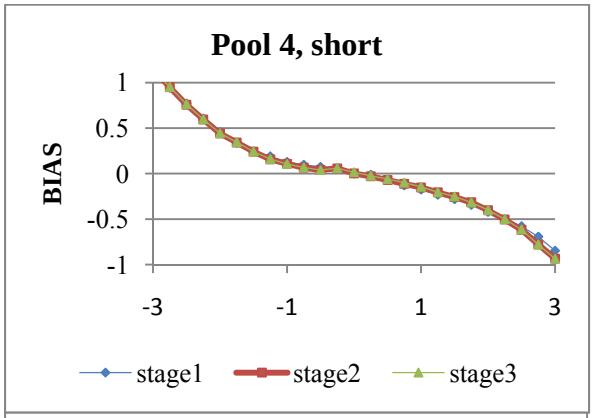
F-2: Comparison of BIAS and RMSE for testlet /discrete item position effect, continued

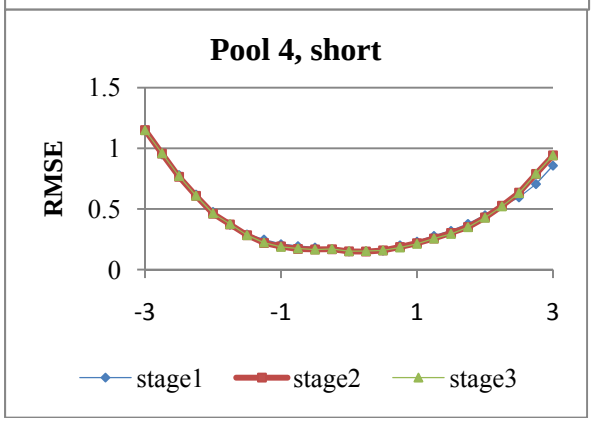


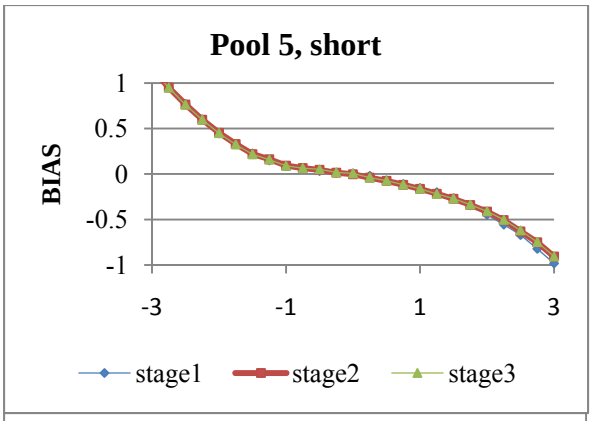


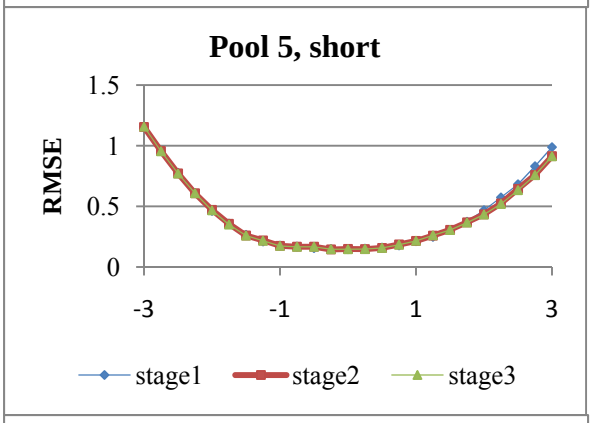


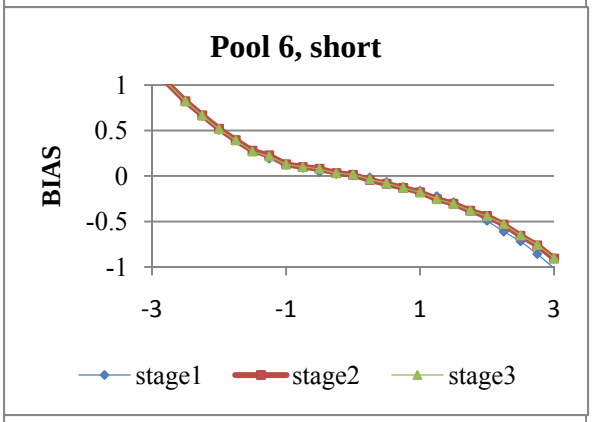
F-2: Comparison of BIAS and RMSE for testlet /discrete item position effect, continued

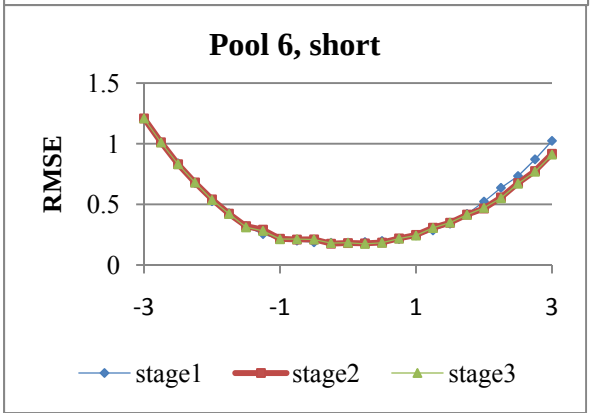





Pool 5, short
BIAS
1
0.5
0
-0.5
-1
-3
-1
1
3
stage1
stage2
stage3


Pool 5, short
RMSE
1.5
1
0.5
0
-3
-1
1
3
stage1
stage2
stage3


Pool 6, short
BIAS
1
0.5
0
-0.5
-1
-3
-1
1
3
stage1
stage2
stage3


Pool 6, short
RMSE
1.5
1
0.5
0
-3
-1
1
3
stage1
stage2
stage3

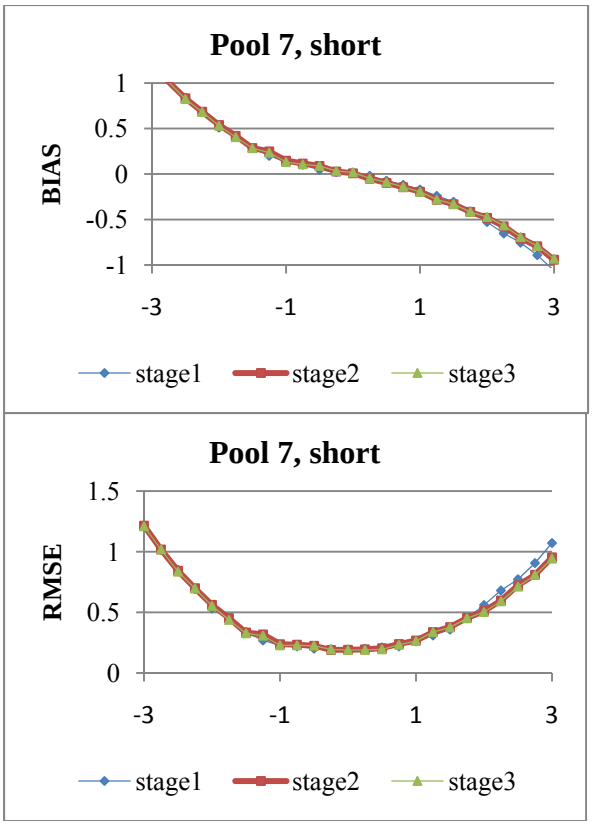
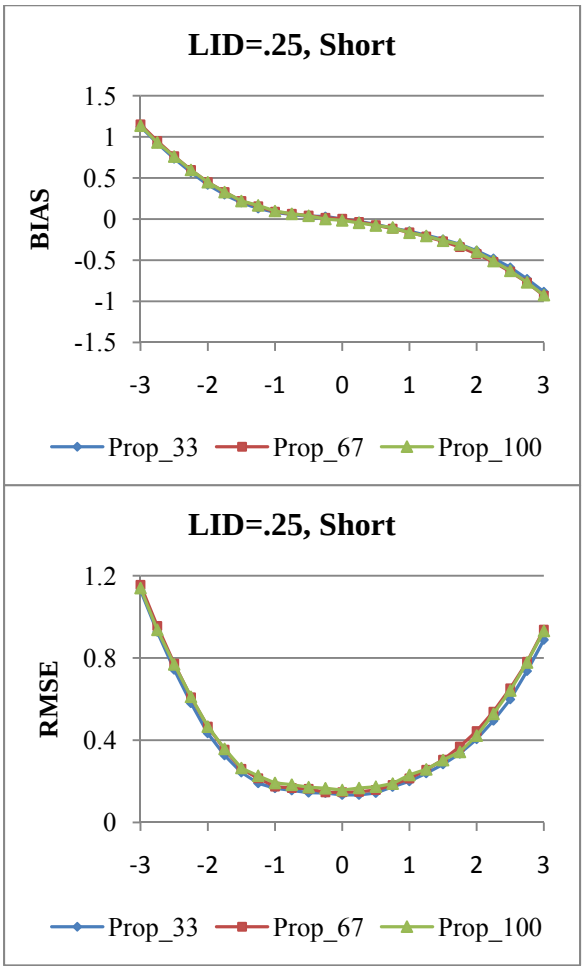
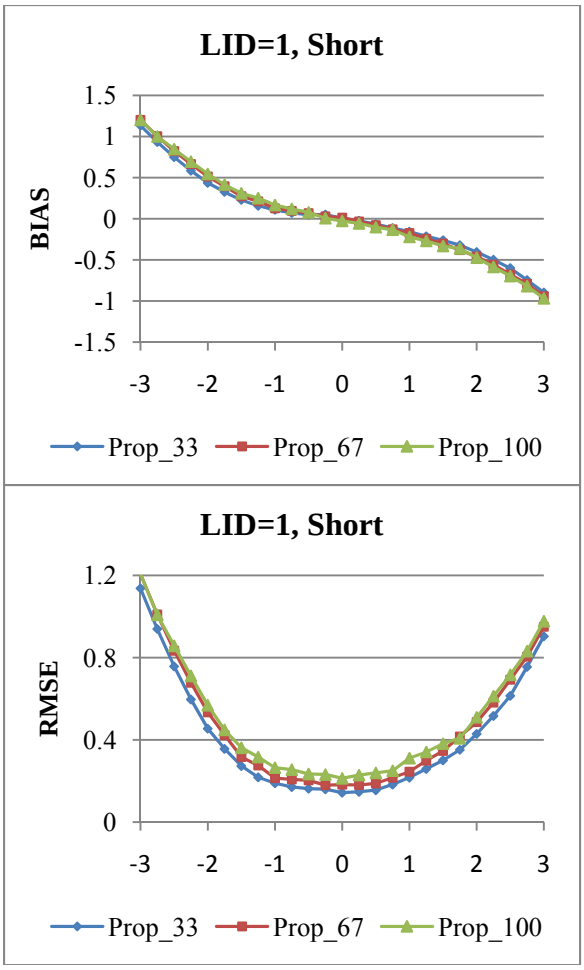

Pool 7, short
BIAS
1
0.5
0
-0.5
-1
-3
-1
1
3
stage1
stage2
stage3
Pool 7, short
RMSE
1.5
1
0.5
0
-3
-1
1
3
stage1
stage2
stage3

Figure F-3: Comparison of BIAS and RMSE for testlet /discrete item proportion effect under short test length conditions




LID=1, Short
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Prop_33
Prop_67
Prop_100
LID=1, Short
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Prop_33
Prop_67
Prop_100

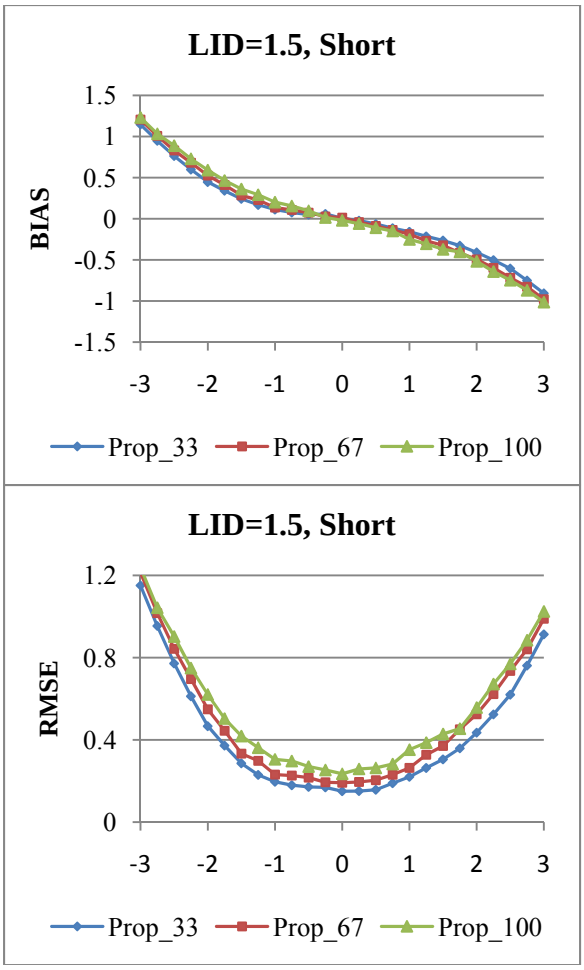
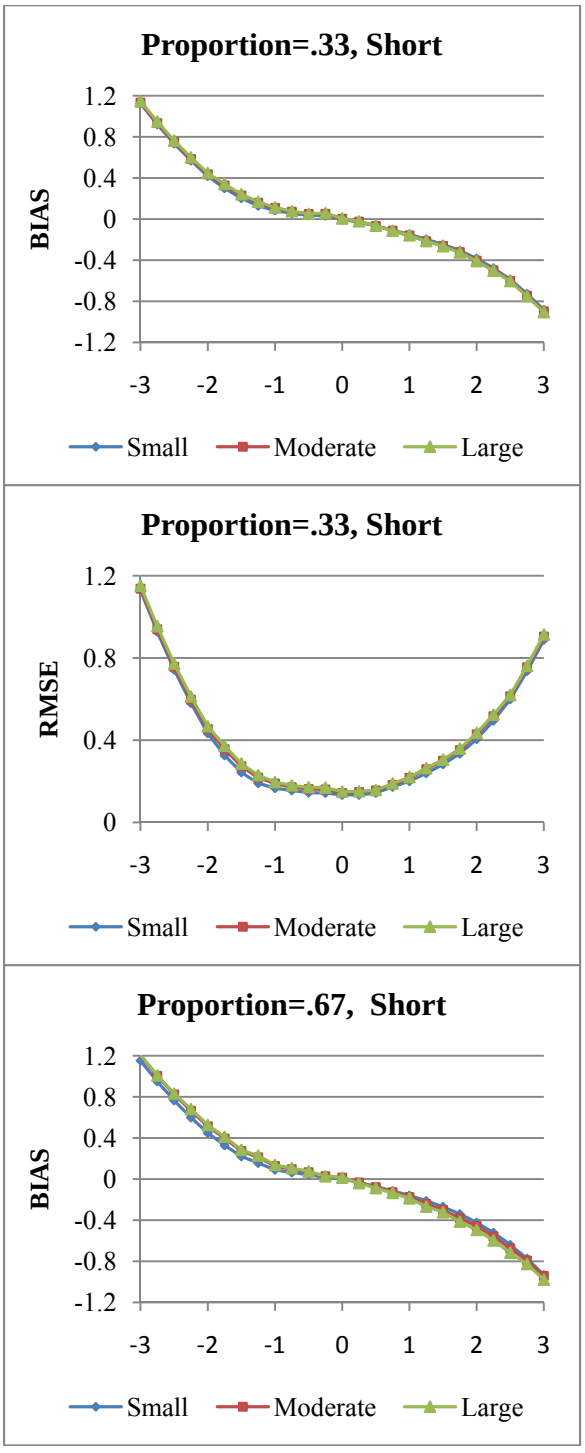

LID=1.5, Short
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Prop_33
Prop_67
Prop_100
LID=1.5, Short
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Prop_33
Prop_67
Prop_100

Figure F-4: Comparison of BIAS and RMSE for LID magnitude effect under short test length conditions



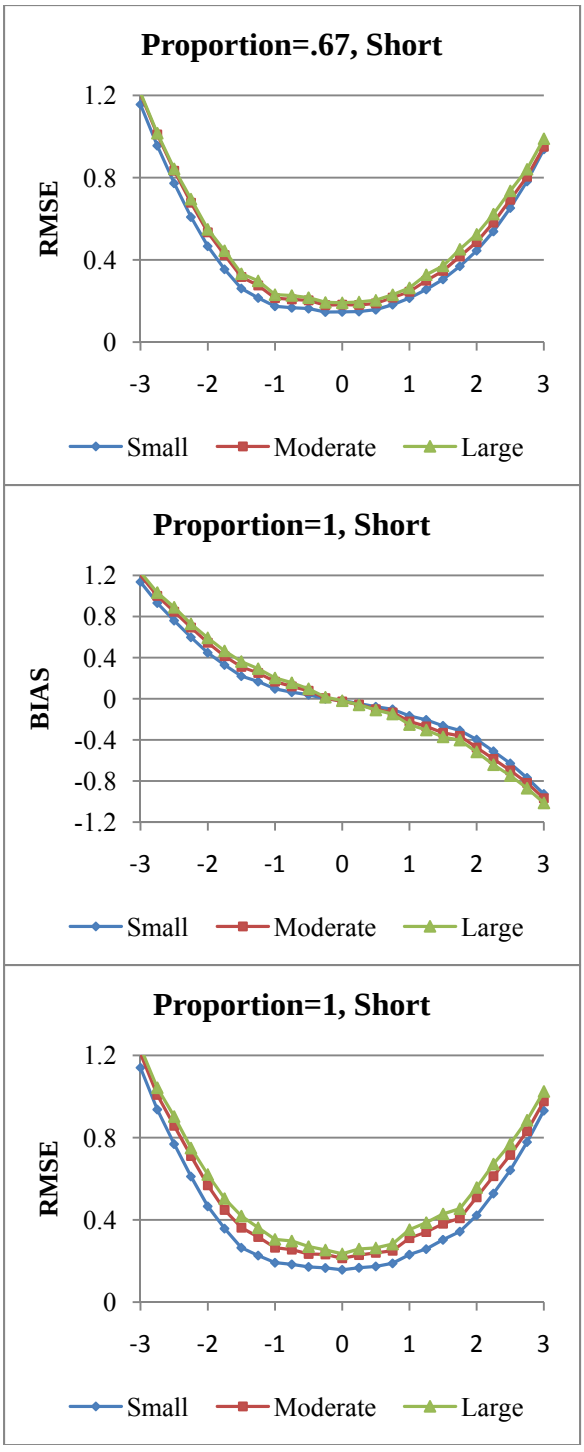
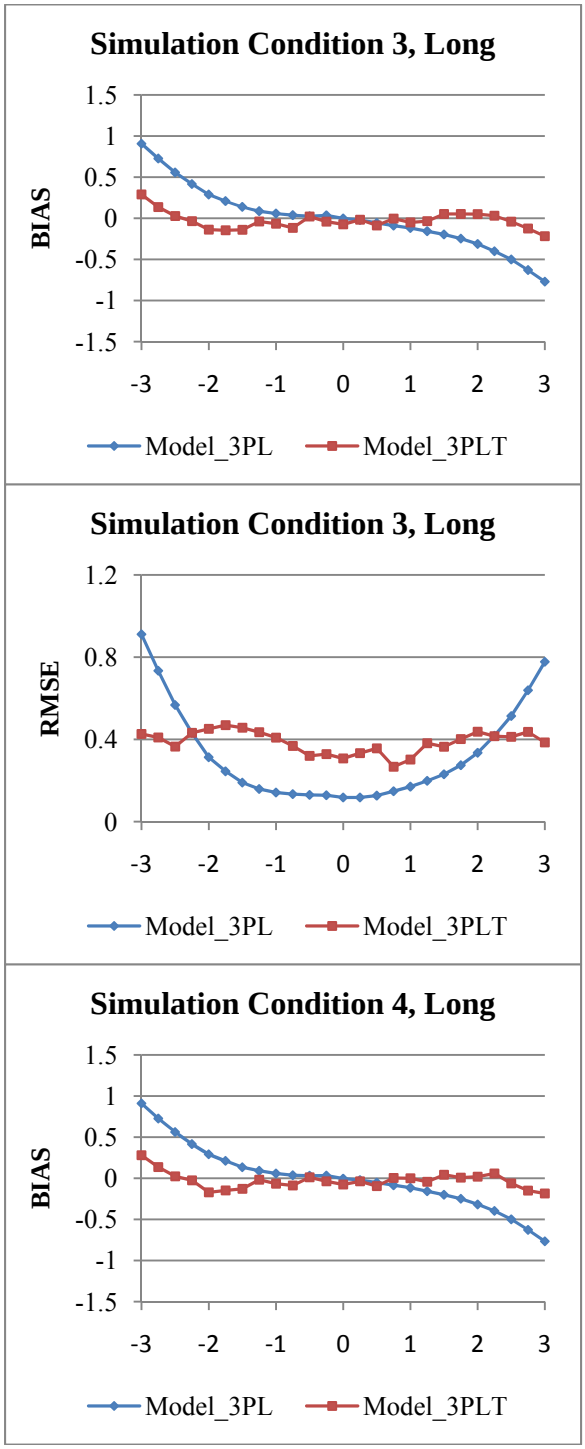
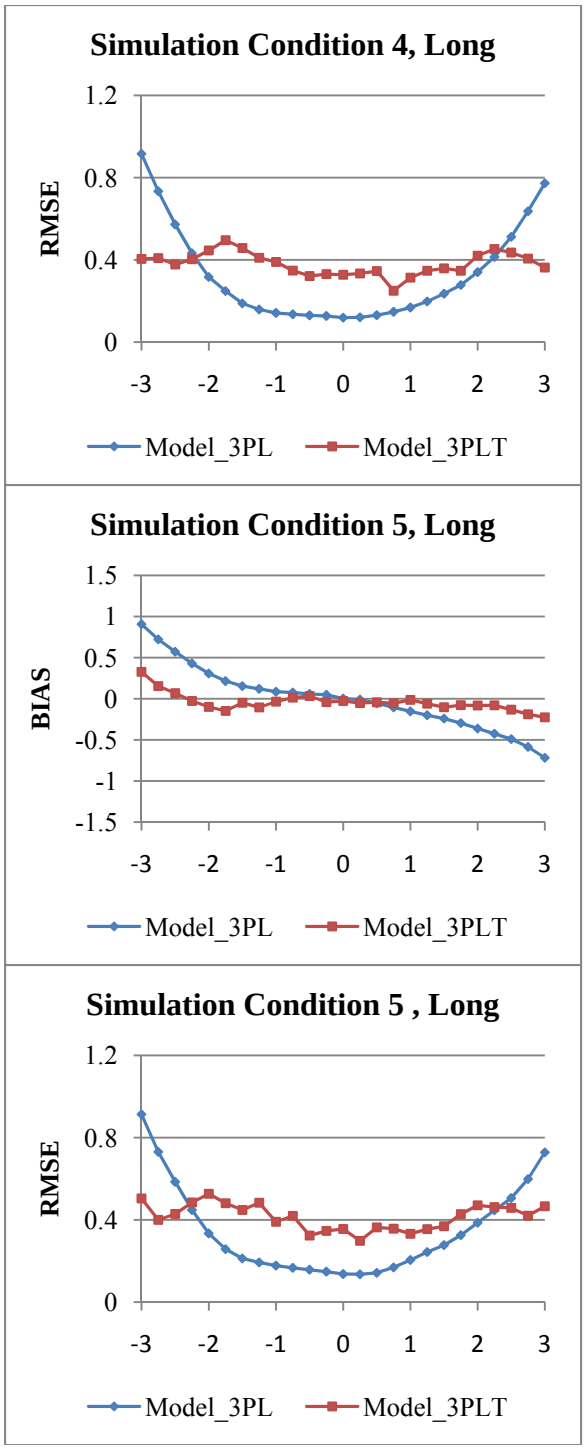
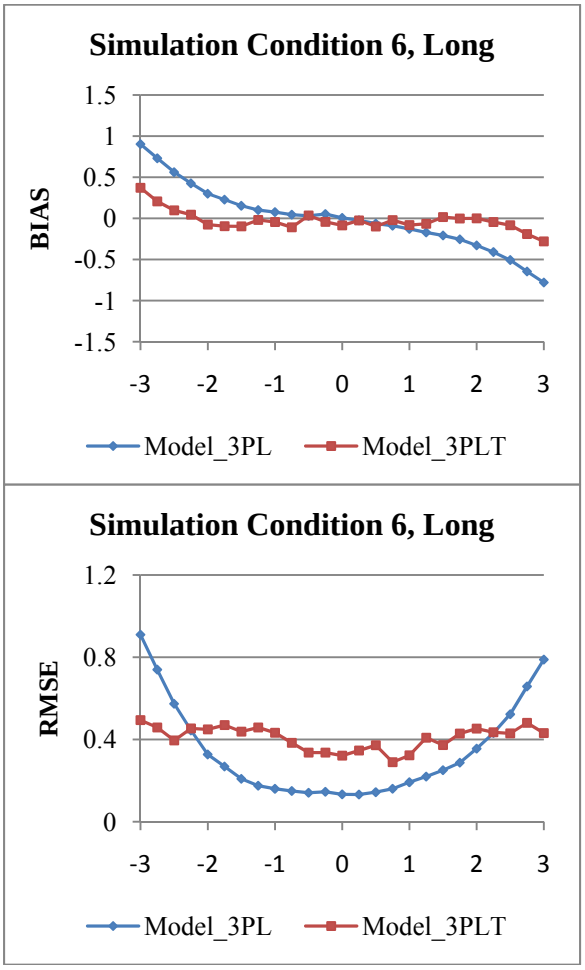

Proportion=.67, Short
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Small
Moderate
Large
Proportion=1, Short
BIAS
1.2
0.8
0.4
0
-0.4
-0.8
-1.2
-3
-2
-1
0
1
2
3
Small
Moderate
Large
Proportion=1, Short
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Small
Moderate
Large

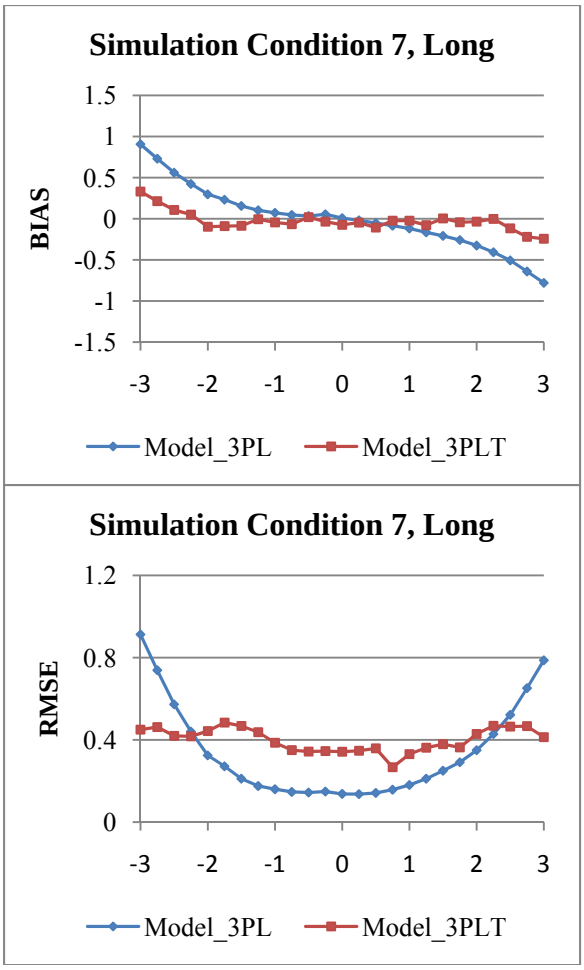
Figure F-5: Comparison of BIAS and RMSE for model effect



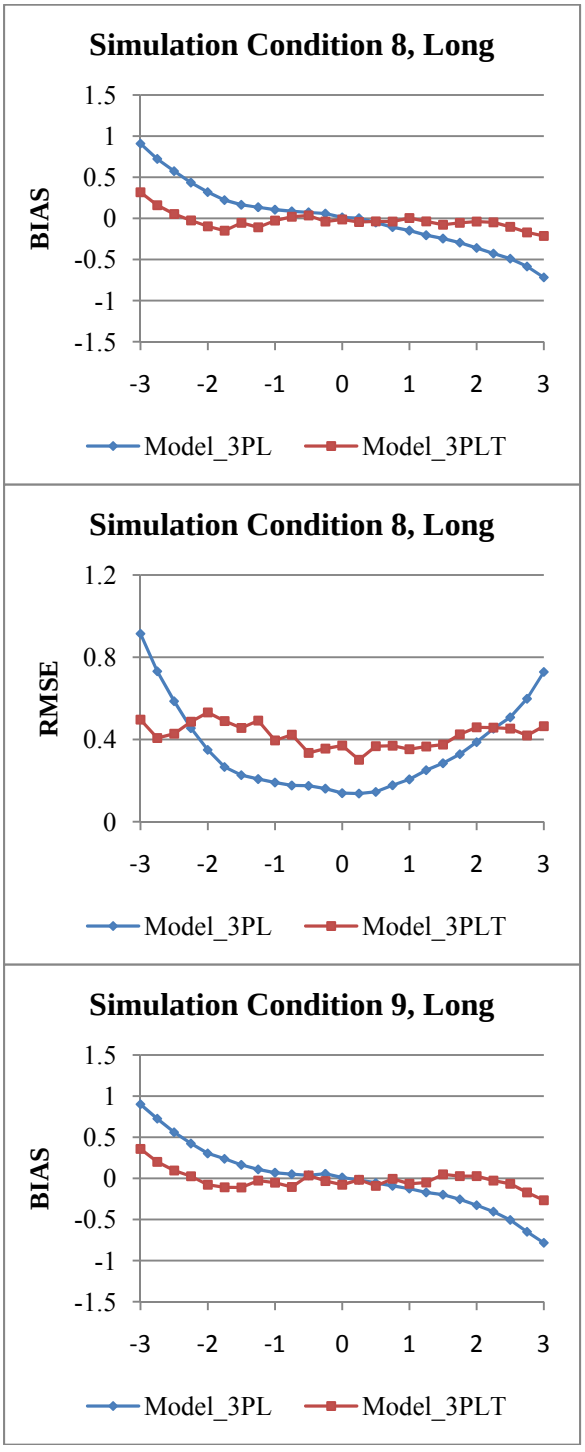

Simulation Condition 4, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 5, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 5 , Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

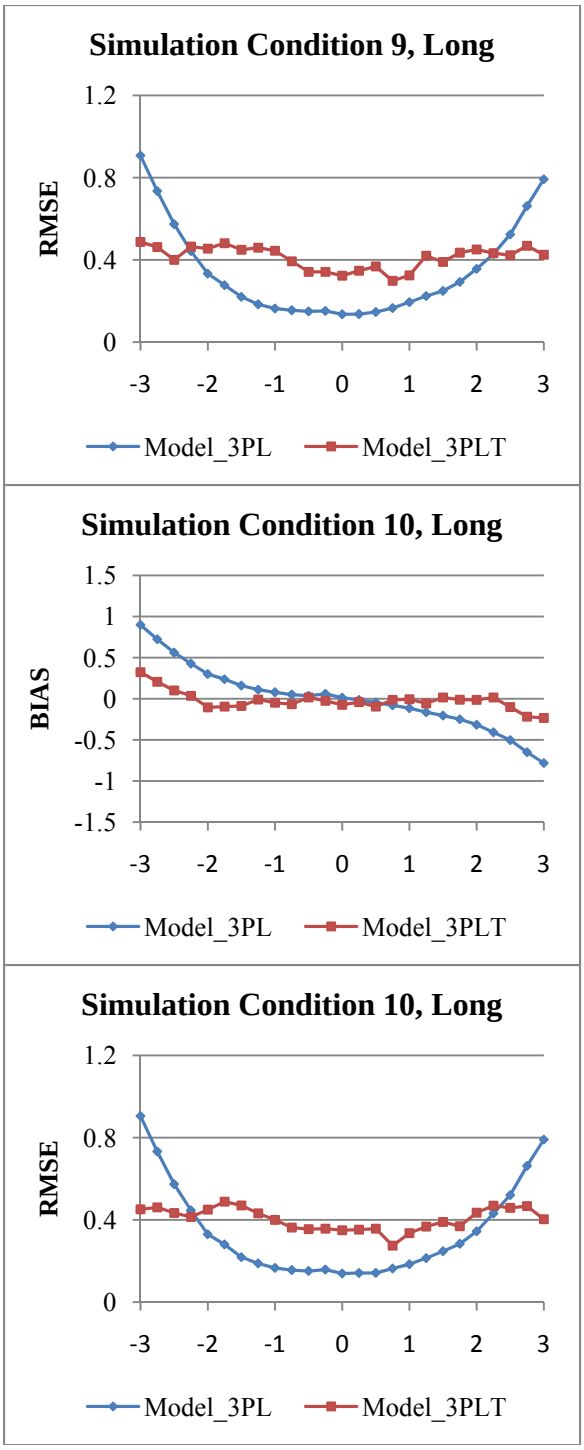
F-5: Comparison of BIAS and RMSE for model effect, continued



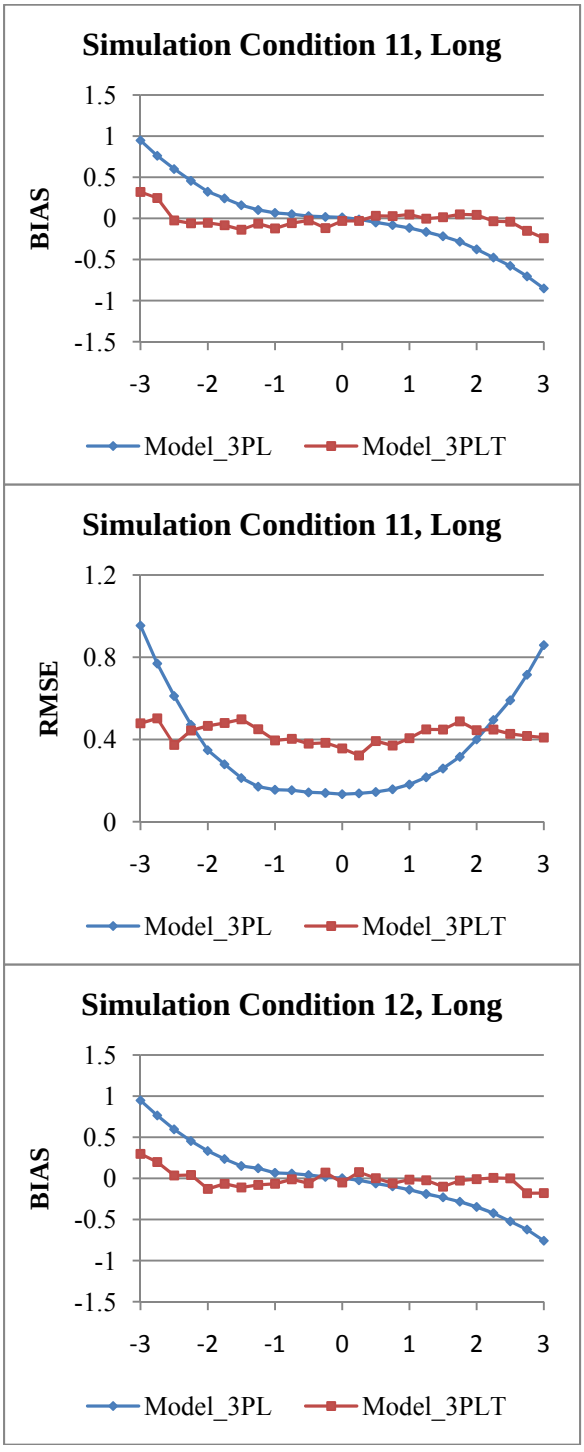

Simulation Condition 7, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 7, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

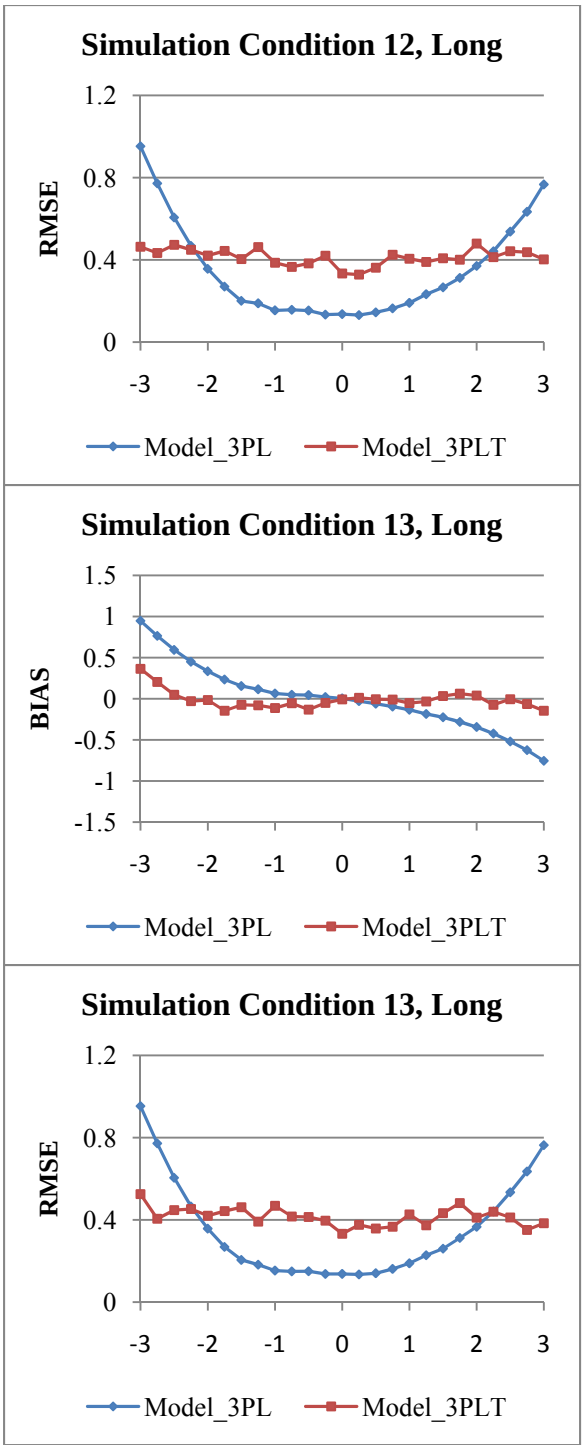
F-5: Comparison of BIAS and RMSE for model effect, continued



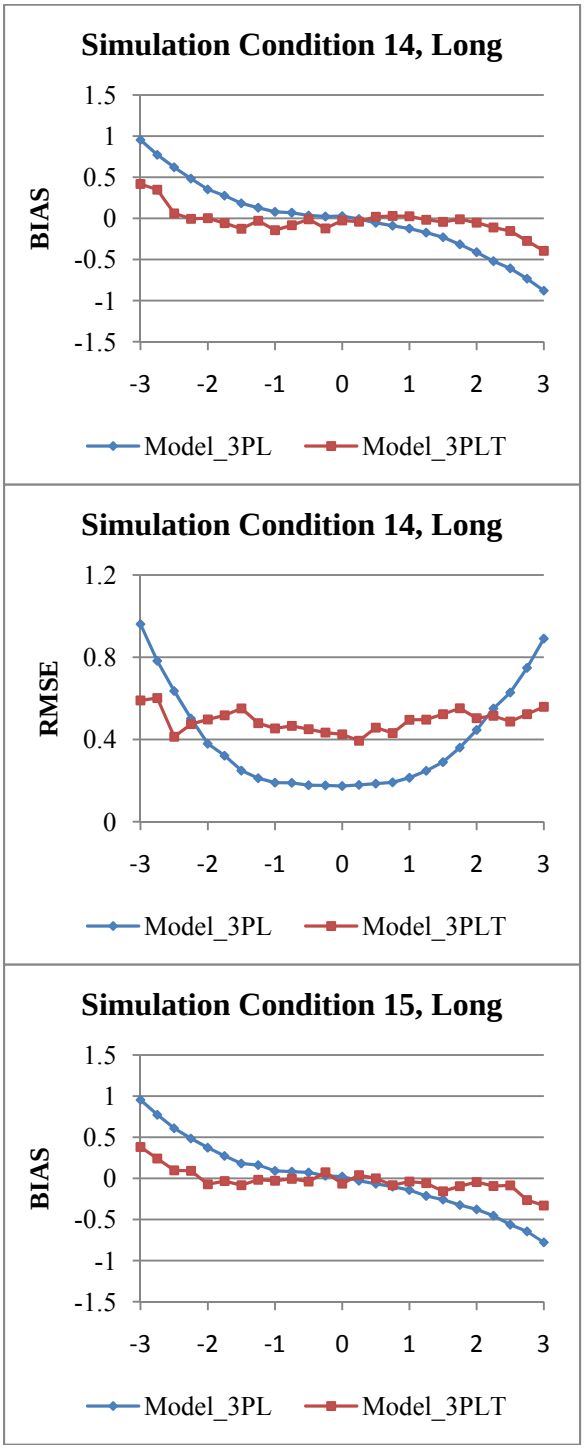

Simulation Condition 9, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 10, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 10, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

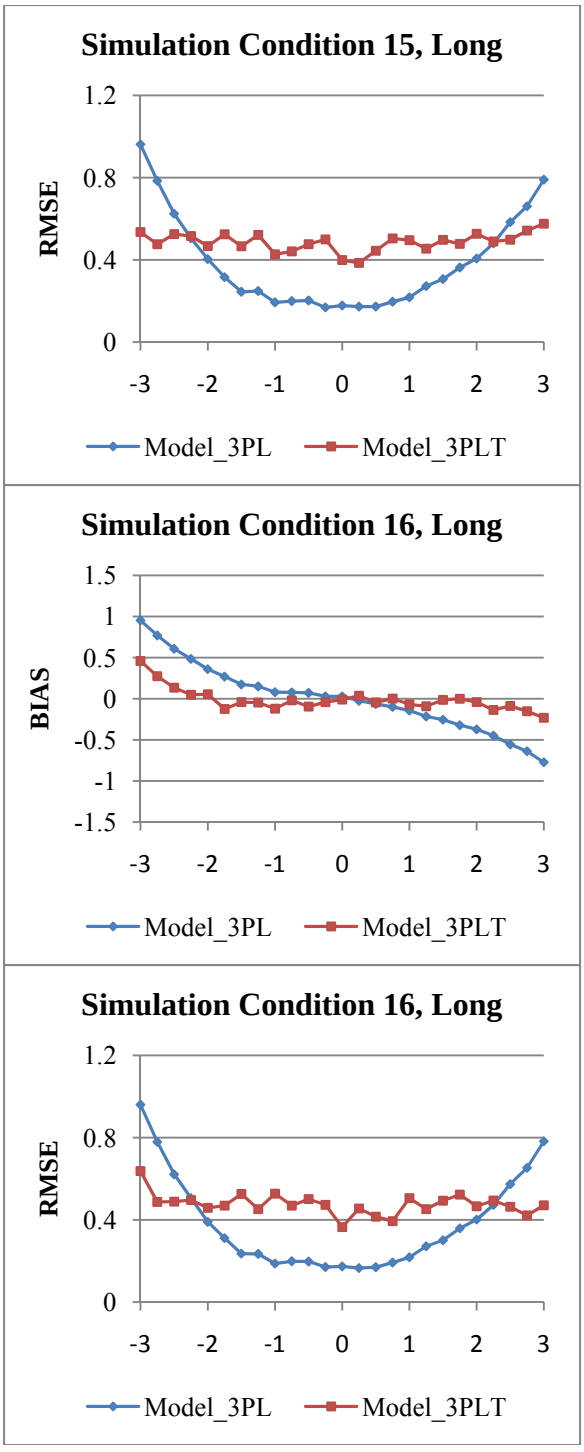
F-5: Comparison of BIAS and RMSE for model effect, continued



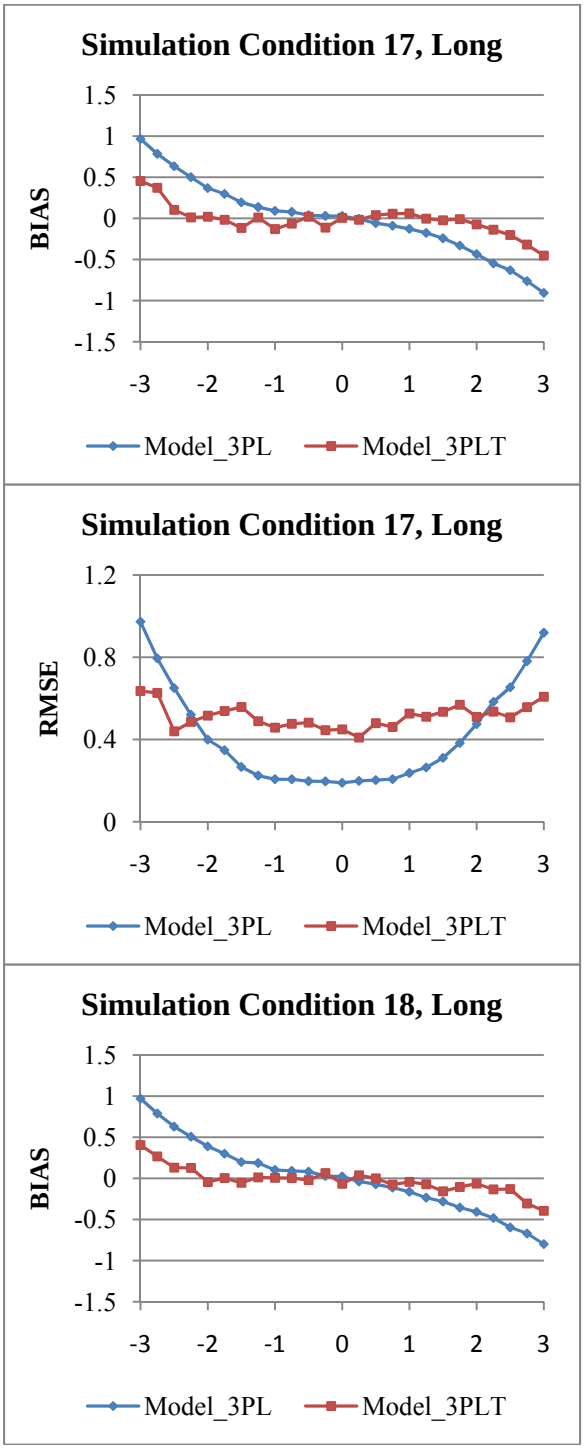

Simulation Condition 12, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 13, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 13, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

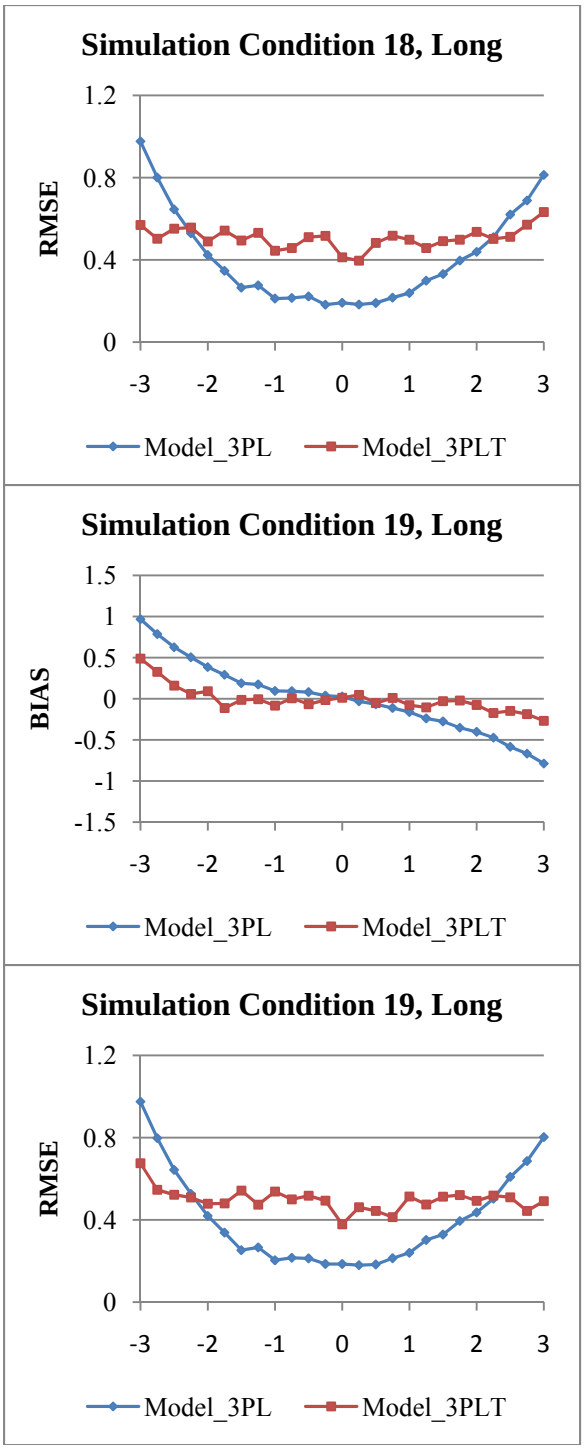
F-5: Comparison of BIAS and RMSE for model effect, continued



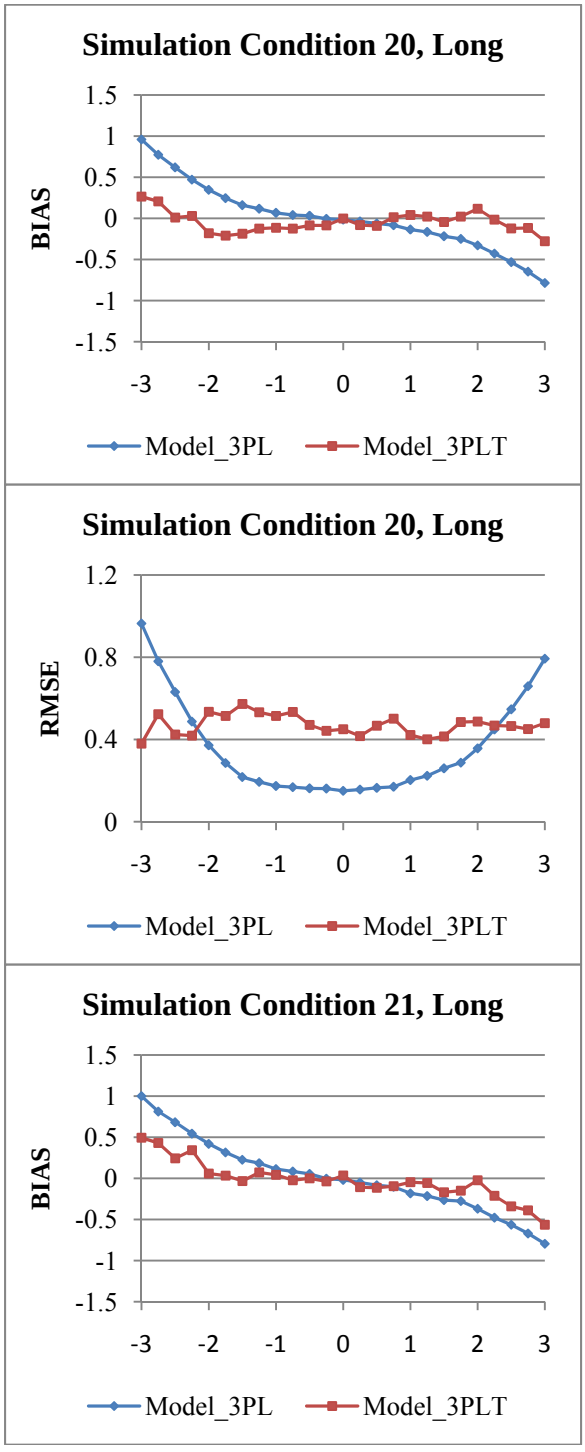

Simulation Condition 15, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 16, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 16, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

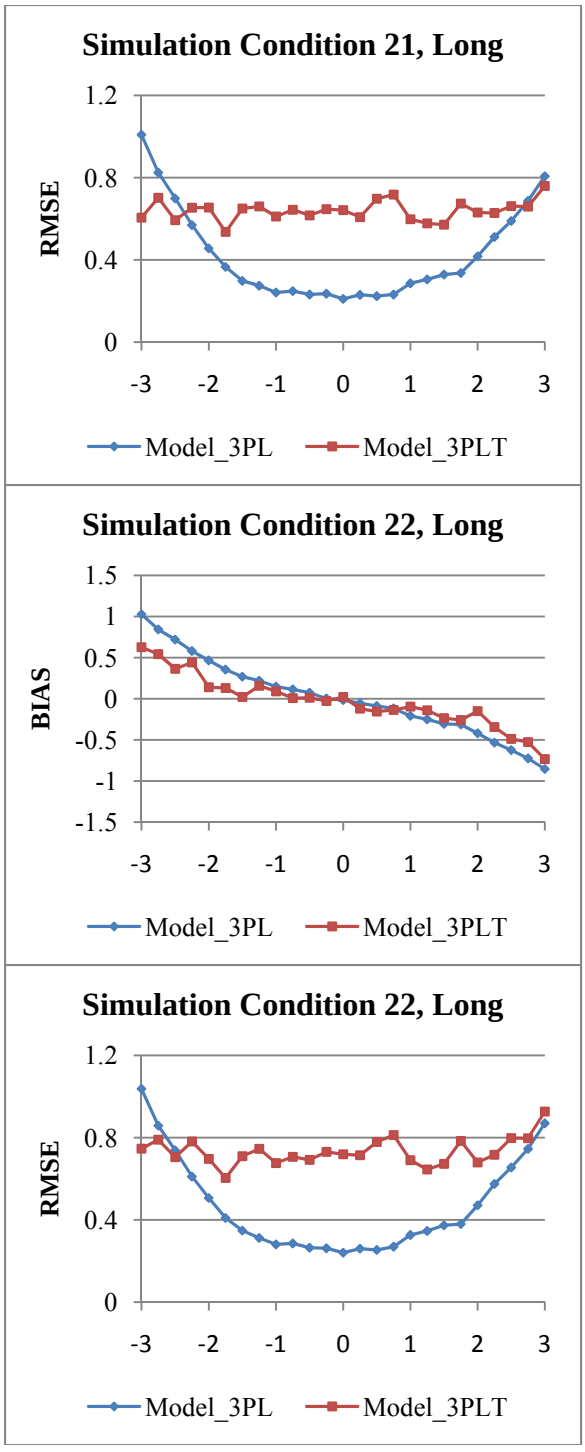
F-5: Comparison of BIAS and RMSE for model effect, continued




Simulation Condition 18, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 19, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 19, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

F-5: Comparison of BIAS and RMSE for model effect, continued




Simulation Condition 21, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 22, Long
BIAS
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT
Simulation Condition 22, Long
RMSE
1.2
0.8
0.4
0
-3
-2
-1
0
1
2
3
Model_3PL
Model_3PLT

Bibliography

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* ACT Research Report Series, 87-14, ACT, Iowa City.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika , 43* (4), 561-573.

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement , 43* (2), 85-96.

Armstrong, R. D. (2002). *Routing rules for multiple-form structures.* Law School Admission Council, Computerized Testing Report 02-08.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement , 28* (3), 147-164.

Armstrong, R., & Roussos, L. (2005). *A method to determine targets for multi-stage adaptive tests.* Law School Admission Council, Computerized Testing Report 02-07.

Baker, F. B., & Kim, S.-H. (2004). Estimation of item parameters of mixed models. In F. B. Baker, & S.-H. Kim, *Item Response Theory: Parameter estimation techniques* (pp. 283-293). New York: Marcel Dekker, Inc.

Ban, J.-C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement , 39* (3), 207-218.

Belov, D. I., & Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement , 29* (4), 239-261.

Berger, M. P. (1994). A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Applied Psychological Measurement , 18*, 141-153.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow, & J. B. Olsen-Bunchanan, *Innovations in computerized assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Birnbaun, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more norminal categories. *Psychometrika , 37* (1), 29-51.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT 4.0. Lincolnwood, IL: Scientific Software International.

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

Breithaupt, K., & Hare, D. R. (2007). Automated simutaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement , 67* (1), 5-20.

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing , 5* (3), 319-330.

Camilli, G. (2006). Test fairness. In R. L. Brennan, *Educational Measurement* (pp. 221-256). Westport, CT: Praeger Publishers.

Chang, H. H. (2004). Understanding computerized adaptive testing: from Robbins-Monro to Lord and beyond. In D. Kaplan, *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). SAGE Publications.

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement , 23* (3), 211-222.

Chen, C.-T., & Wang, W.-C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting item. *Applied Psychological Measurement , 31* (5), 388-411.

Chen, S.-K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and Psychological Measurement , 57* (3), 422-439.

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

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST. *Applied Measurement in Education , 19* (3), 241-255.

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

Cohen, J. (1992). Statistics a power primer. *Psychology Bulletin* (112), 155-159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

College Board. (1993). *ACCUPLACER: computerized placement tests: Technical data supplement.* New York.

Cranbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. Potenza, J. Fremer, & W. C. Ward, *Computer-based testing: Building the foundations for future assessments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum Associate.

Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. In S. M. Downing, & T. M. Haladyna, *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Measurement in Education , 27* (5), 335-356.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement , 43* (2), 145-168.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement , 19* (1), 5-22.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing, & T. M. Haladyna, *Handbook of Test Development* (pp. 287-301). Mahwah, NJ: Routledge.

Drasgow, F., Luechet, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan, *Educational Measurement* (pp. 471-515). Westport, CT: Praeger Publishers.

Edwards, M. C., & Thissen, D. (2007). Exploring potential designs for multi-form structure computerized adaptive tests with uniform item exposure. *In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*

Eggen, T. J., & Verschoor, A. J. (2006). Optional testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement , 30* (5), 379-393.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation.* Educational Testing Service, Princeton, NJ.

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

ETS. (2010). *The GRE® revised General Test is coming in August 2011*. Retrieved September 16, 2010, from http://www.ets.org/gre/revised_general/know

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large scale performance assessment. *Applied Measurement in Education , 10* (2), 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement , 36* (2), 119-140.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). ***Bayesian Data Analysis.*** Boca Raton, FL: Chapman & Hall/CRC.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement , 6* (1), 109-127.

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement , 27* (4), 247-261.

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

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: estimation procedures, population distribution, and the item pool characteristics. *Applied Psychological Measurement , 29* (6), 433-456.

Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika , 68*, 435-451.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice , 11* (1), 11-25.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education , 19* (3), 221-239.

Hambleton, R. K., & Swamnathan, H. (1985). ***Item response theory: Principles and applications.*** Boston, MA: Kluwer-Nijhoff.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice , 26* (2), 44-52.

Hol, A. M., Vorst, H. c., & Mellengergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: administration mode effects and a comparison with short forms. *Applied Psychological Measurement , 31* (5), 412-429.

Howard Wainer, L. M., Bradlow, E. T., Wang, X., Skorupski, W. P., Boulet, J., & Mislevy, R. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar,

Automated Scoring of Complex Tasks in Computer-Based Testing (pp. 169-200). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement , 6* (3), 311-321.

Jiao, H., Wang, S., & Wei, H. (2008). Parameter estimation of one-parameter testlet model. *Paper presented at the annual meeting of the National Council on Measurement in Education.* New York.

Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pools.* Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education , 19* (3), 203-220.

Keng, L. (2008). A comparison of the performance of testlet-based computer adaptive tests and multistage tests. Unpublished Doctorate Dissertation, University of Texas at Austin.

Keppel, G., & Wickens, T. (2004). *Design and analysis: A researcher's handbook.* Upper Saddle River, NJ: Person Prentice Hall.

Kim, H., & Plake, B. S. (1993). Monte carlo simulation comparison of two-stage testing and computerized adaptive testing. *Paper presented at the annual meeting of the National Council on Measurement in Education.* Atlanta, GA.

Lee, G. (2000). Estimating conditional standard errors of measurement for tests composed of testlets . *Applied Measurement in Education , 13* (2), 161-180.

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society. Series C (Applied Statistics) , 50* (3), 325-335.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computer mastery test. *Applied Psychological Measurement , 14* (4), 367-386.

Li, D. (2009). *Developing a common scale for testlet model parameter estiamtes under the common item non-equivalent groups design.* Unpublished Doctorate Dissertation, University of Maryland, College Park, MD.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement , 30* (1), 3-21.

Lomax, R. G. (2007). *Statistical Concepts* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika , 36* (3), 227-242.

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

Lu, R. (2008). *An Exploration of the multistage-CATs for a listening test.* Comprehensive Adult Student Assessment Systems, San Diego, CA.

Luecht, R. M. (2006). Designing tests for pass-fail decisions using item response theory. In S. M. Downing, & T. M. Haladyna, *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Luecht, R. M., & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linder, & C. A. Glas, *Computerized adaptive testing: theory and practice* (pp. 117-128). Dordrecht, Netherlands: Kluwer Academin Publishers.

Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996). Heuristic-based CAT: Balancing item information, content and exposure. *Paper presented at the annual meeting of the National Council of Measurement in Education.* New York.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests . *Applied Measurement in Education , 19* (3), 189-202.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika , 47* (2), 149-172.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education , 19* (3), 185-187.

Mislevy, R. J., & Bock, R. D. (1982). Adaptive EAP estimation fo ability in a microcomputer environment. *Applied Psychological Measurement , 6* (4), 431-444.

Mislevy, R. J., & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika , 65* (2), 149-165.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, , 13* (1), 57-75.

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

Olsen, J. B., & Bunderson, C. V. (2007). Validity and decision issues in selecting a CAT measurement model. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing.* Unpublished Doctoral Dissertation, University of Massachusetts Amherst, .

Pierce, C. A., Block, R. a., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement , 64* (6), 916-924.

Pommerich, M., & Segall, D. O. (2008). Local dependence in an operational CAT: diagnosis and implications. *Journal of Educational Measurement , 45* (3), 201-220.

Prometric. (2010). *Linear on the fly testing: Risk-free testing success.* Retrieved Oct 20, 2010, from Prometric: http://www.prometric.com/NR/rdonlyres/eepj7funjofq6gj3rqvgtopmkio4h3a7s5dvyhorczr2jznoddnhzqpmvyoghuluc72ke7jcgpwnaosfaujwvcecogf/CaseStudySOA.pdf

Raîche, G., & Blais, J.-G. (2006). SIMCAT 1.0: A SAS computer program for simulating computer adaptive testing. *Applied Psychological Measurement , 30*, 60-61.

Rasch, G. (1960). *Probabilisic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Linden, & R. K. Hambleton, *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer-Verlag.

Reese, L. M. (1999). *Impact of local item dependence item response theory scoring in CAT.* Computerized Testing Report 98-08. Law School Admission Council.

Reese, L. M., & Schinipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing.* Compuerized Testing Report 96-04, Law School Admission Council.

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison.* ETS RR-09-37. Princeton, NJ: ETS.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika , 53* (3), 349-359.

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-pencil tests.* Educational Test Service, RR-07-04, Princeton, NJ.

Samejima, F. (1969). Estimatin of latent ability using a response pattern of graded scores. *Psychometric Monograph , 17*.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph , 17*.

Sands, W. A., Waters, B. K., & Mcbride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Scheuermann, F., & Bjornsson, J. (2009). *The transition to computer-based assessment*. Retrieved May 20, 2009, from http://crell.jrc.it/RP/reporttransition.pdf

Schnipke, D. L., & Reese, L. M. (1997). A comparison testlet-based test designs for computerized adaptive testing. *Paper presented at the annual meeting of the American Educational Research Association,.* Chicago, IL.

Schnipke, D. L., & Scrams, D. J. (1999). *Item theft in a continuous testing environment: what is the extent of the danger?* Computerized Testing Report 98-01, Law School Admission Council.

Sireci, S. C., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement , 28* (3), 237-247.

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing: Widely or narrowly applicable? *Applied Measurement in Education , 19* (3), 257-260.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* Educational Testing Service, Princeton, NJ.

Stocking, M. L., & Lewis, C. (1995). *Controlling item exposure conditional on ability in computerized adaptive testing.* Research Report 95-24, Educational Testing Service, Princeton, NJ.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement , 17* (3), 277-292.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao, *Handbook of Statistics 26: Psychometrics* (Vol. 26, pp. 683-718). ELSEVIER.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement , 17* (2), 151-155.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military testing association* (pp. 973-977). San Diego, CA: San Diego: Navy Personnel Research and Development Center.

TCExam. (2008, 11 13). Retrieved from TCExam: http://www.tecnick.com/public/code/cp_dpage.php?aiocp_dp=tcexam

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of muti-category-response models. *Journal of Educational Measurement , 26* (3), 247-260.

Tuerlinckx, F., & De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods , 6*, 181-195.

Vale, C. D. (2006). Computerized item banking. In S. M. Downing, & T. M. Haladyna, *Handbook of Test Development* (pp. 261-285). New York: Routledge.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York: Springer.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement , 22*, 195-211.

van der Linden, W. J., & Glas, C. A. (2000). *Computerized Adaptive Testing: Theory and Practive.* Netherlands: Kluwer Academic Publishers.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement , 22* (3), 259-270.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement , 44* (2), 117-130.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement , 35* (4), 328-345.

Vos, H. J., & Glas, C. A. (2000). Testlet-based adaptive mastery testing. In W. J. van der Linden, & C. A. Glas, *Computerized adaptive testing: Theory and practice.* Dordrecht, Netherlands: Kluwer Academic Publishers.

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer* (Second ed.). Mahwah, NJ: Lawrence Erlbaum Associates,Inc.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education , 8* (2), 157-186.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement , 24* (3), 185-201.

Wainer, H., & Mislevy, R. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorens, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, et al., *Computerized Adaptive Testing: A Primer* (2nd Edition ed., pp. 61-99). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Wainer, H., & Wang, X. (2000). Use a new statistical model for testlet to score TOEFL. *Journal of Educational Measurement , 37* (3), 203-220.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York, NY: Cambridge University Press.

Wainer, H., Bradlow, L. M., Bradlow, E. T., Wang, X., Skorupski, W. P., Boulet, J., et al. (2006). An application of testlet response theory in the scoring a complex certification exam. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar, *Automated Scoring of Complex Tasks in Computer-based Testing* (pp. 169-199). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement , 35* (2), 109-135.

Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement , 65* (1), 5-27.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement , 29* (2), 126-149.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *A user's guide for SCORIGHT (version 3.0): A computer program built for scoring test unit of testlets including a module for covariate analysis.* Princeton, NJ: Educational Testing Service.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement , 6*, 473-492.

Weiss, D. J., & Kingsbugy, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement , 21*, 361-375.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika , 60* (2), 181-198.

Xing, D., & Hambleton, R. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurment , 64* (1), 5-21.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement , 8* (2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managaing local item dependence. *Journal of Educational Measurement , 30* (3), 293-313.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan, *Educational Measurement.* Westport, CT: American Council on Education and Praeger Publishers.

Yi, Q., Zhang, J., & Chang, H.-H. (2006). *Severity of organized item theft in computerized adaptive testing: An empirical study.* RR-06-22, Educational Testing Service.

Zenisky, A. L., & Hambleton, R. K. (2004). Effects of selected multi-stage test design alternatives on credentialing examination outcomes. *Paper presented at the annual meeting of NCME, San Diego, CA.*

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement , 39* (4), 291-309.

Zhang, J. (2008). *Dichotomous or polytomous model? Equating of testlet-based tests in light of conditional item-pair correlations.* Doctoral Dissertation, The University of Iowa.

Zhang, Y. (2006). *Impacts of multidimensionality and content misclassification on ability estimation in computerized adaptive sequential testing.* Unpublished Doctorate Dissertation, University of Delaware.

Zimowski, M., Muraki, E., & Mislevy, R. J. (2003). BILOG-MG 3: Multiple-group IRT analysis and test maintenance for binary items. Chicago, IL: Science Software International Inc.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement , 32* (4), 341-363.