ABSTRACT

Title of dissertation: INVESTIGATING ITEM PARAMETER DRIFT (IPD) AMPLIFICATION AND CANCELLATION AT THE TESTLET-LEVEL ON MODEL PARAMETER ESTIMATION IN A MULTISTAGE TEST

Rosalyn Bryant, Doctor of Philosphy, 2018

Dissertation directed by: Professor Hong Jiao
Department of Education

A test fairness goal is to design testing systems where performance is measured at acceptable degrees of accuracy across a wide range of test taker ability levels and across subgroups of student population.

Research has shown that with adaptive test, a high degree of test score accuracy is realized. Furthermore, standard item response theory (IRT) models, are the predominately used measurement models in educational testing for computerized multistage adaptive tests (MST). Moreover, item sets, or testlets, are widely used items types in MSTs.

Fitting standard IRT models to response data comes with item invariance and local independence assumptions. In practice, unexpected shifts in parameter values, or item parameter drift (IPD), across test administrations have been reported. Moreover, testlet items have been known to exhibit local item dependence (LID) due to interactions between the test taker and the common testlet stimulus. When IPD and/or LID are exhibited, these are likely violations of standard IRT assumptions threatening ability estimate accuracy and test score validity. Furthermore, a conjecture in this study is that the accumulation of insignificant IPD may be significant at the testlet level due to amplification or become insignificant at the testlet level due to cancellation. To date, no studies have investigated the combined impact of IPD amplification or cancellation at the testlet level with LID on ability estimation accuracy in an MST system.

In this study, MST ability estimates generated under the two-parameter logistic (2PL) IRT and 2PL testlet response theory (TRT) models are compared to determine if there are significant differences when the amplification and cancellation of IPD to the testlet-level and LID are exhibited and when LID is not exhibited. Further, this study examines the combined impact of amplification and cancellation of IPD to the testlet-level and/or LID on MST system routing performance.

This study reveals that ability estimation, routing, and decision accuracy are not significantly impacted by combined amplification, cancellation, and/or LID effects. However routing accuracy is impacted by module difference, routing error stage, or testlet effects. Finally, moderate ability test takers are found to be more likely misclassified than low or high ability test takers.

INVESTIGATING ITEM PARAMETER DRIFT (IPD) AMPLIFICATION AND CANCELLATION AT THE TESTLET-LEVEL ON MODEL PARAMETER ESTIMATION IN A MULTISTAGE TEST

by

Rosalyn Bryant

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 Philosphy
2018

Advisory Committee:
Professor Hong Jiao, Chair
Professor Emeritis Robert Lissitz
Professor Emeritis George Macready
Professor Tracy Sweet
Professor Steven Ross

Dedication

To all that motivates and drives me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Background

1.1 Introduction

When the parameters of a test item exhibit item parameter drift (IPD), the probability of responding correctly to the item may be influenced by unanticipated interactions between test takers and the affected item. This unanticipated influence may be of practical significance, adversely impacting the accuracy of final ability estimates and consequently the validity of inferences derived from the final estimates. Hence, results from a variety of IPD studies from differing perspectives work to inform test developers of testing situations that may be vulnerable to IPD effects. Some IPD studies focus on cause identification or detection procedures (Cook, Eignor, & Taft, 1988; De Mars, 2004; Donoghue & Isham, 1996), others on the indirect impact of IPD effects on the validity of test score inferences (Han, Wells & Sireci, 2012; Li & Zumbo, 2009; Wollack, Sung, & Kang, 2006), and still others on the direct impact of IPD effects on final test score accuracy and/or testing system performance (Han, Wells, & Sireci, 2012; Melican & Deng, 2009; Wei, 2013; Wells, Subkoviak, & Serlin, 2002). This present study is of the latter type, focusing on investigating the impact of cumulative patterns of IPD to the testlet-level on the final ability estimation accuracy and computerized multistage testing (MST) routing performance. As an outcome measure closely associated with ability estimation accuracy and routing performance, MST decision classification accuracy rates are also reported.

1.2 Multistage Test

To date, research regarding the impact on final test score accuracy due to the influence of IPD has been primarily implemented under testing conditions where statistically independent test items are administered. These item types are primarily used

in IPD investigations even though testlets, or bundled item sets, have become popular for use as test items (Hendrickson, 2007; Lissitz & Jiao, 2012; Wainer, Bradlow, & Wang, 2007; Yan, von Davier, & Lewis, 2014). Because prior research has shown that it is possible for items within a testlet to be statistically dependent (Wainer, Bradlow, & Wang, 2007), generalization of findings from statistically independent item IPD research to testlet-based testing situations is limited.

In fact, testlets are regularly used as the testing unit when MST system components are constructed (Hendrickson, 2007). Although studies have investigated the impact of IPD effects on final scores within a testlet based MST system (Wei, 2013), no studies have examined IPD effects in an MST system when the testlet items are assumed statistically dependent. To gain insight into potential MST system vulnerabilities to IPD effects when the testing units are testlets constructed using statistically dependent items, a brief introduction to MST system components is presented next.

MST systems consist of a number of interdependent components that are named and described in detail in Luecht and Nungester (1988) but summarized only briefly here. In an MST, groups of items that are scored as a unit are referred to as modules. The process of allocating items to modules is typically guided by statistical and non-statistical specification goals of test developers.

Once items are allocated to modules, modules are assigned to an MST test stage. Several modules can be assigned to a test stage and an MST system can consist of multiple test stages. Modules assigned to the same test stage are commonly distinguishable by the average difficulty of all items allotted to the module. Test takers

are administered specific modules, within one test stage to the next, based on module selection routines and routing decision rules.

The collection of modules assigned to test stages that are administered to test takers as a single test form an MST pathway. In other words, pathway administered items are scored and reported as a test score for a test taker. An MST may consist of multiple overlapping test pathways. That is, multiple pathways may share some of the same modules, but the complete composition of modules by pathway is unique.

Pathways are assigned to an MST panel. An MST system can have multiple panels that can be individually assigned to test takers in a manner deemed acceptable to test developers. The number of modules and stages within a single MST panel is determined by its design specification. A simple numerical abbreviation system is an approach that is commonly used to distinguish MST panel designs. For instance, a 1-2-2 MST panel design has three test stages with one module assigned to test stage one and two modules assigned to test stages two and three. The product of these numbers indicates how many possible pathways are within the panel. Hence for a 1-2-2 MST panel design, there are four possible pathways. The number of permissible pathways that are actually available to test takers is determined by the routing decision rule and may not be equal to the total number of possible pathways. Thus, a single MST panel contains all assigned modules, test stages, pathways, module selection routines, and routing decision rules needed for successful test administration.

When deciding which next test stage module along a pathway to route a test taker, MST routing rules take into account prior item performance, module level statistics, and interim measures of ability. Hence, a computerized MST is a computer-based testing

(CBT) system that can be more specifically classified as adaptive. Non-adaptive CBT systems sequentially administer items, always using the same set of test items for all test takers regardless of prior performance information. To the contrary, available performance information is used by adaptive CBT systems to adapt or make customized changes to the composition of items to be administered to test takers during the test. These real-time item adaptations adjust the test difficulty level for a test taker based on available prior item answering performance information. Thus, on the same testing occasion each test taker can be administered their own tailored test by an adaptive CBT. The test adaptation functionality within an MST is controlled by the MST routing decision rule procedure. Therefore, ensuring that acceptable levels of routing efficiency holds under testing conditions, including the test item types to be used, are important considerations for MST developers.

Dallas (2014) notes that within an MST system "while all testlets can be considered modules, not all modules could be considered as testlets" Dallas (2014). That is, any combination of testlets and individual items can collectively be allocated to a single module. Yet in practice, there are applications where the MST system module consists of only a single testlet, for example, the MST modules constructed for the reading section of the revised Graduate Record Examination evaluated in Davey and Lee (2011). It follows that within an MST system that utilizes this module construction approach, the module-level properties and statistics are identical to that of the single assigned testlet. Therefore, testlet related design considerations, such as item statistical dependence, become MST module related design considerations as well.

1.3 Benefits of a Multistage Test

MSTs are not the only commonly used CBTs. Unlike an MST, a computer adaptive test (CAT) system design adapts the difficulty of a test after selecting, administering, and scoring single items (Brossman & Guille, 2014; Hendrickson, 2007; Wise & Kingsbury, 2000; Yan, von Davier, & Lewis, 2014). Yet, the MST system design has been shown to overcome some known challenges associated with CAT system administration. For instance, since CAT algorithms assemble test forms item-by-item during test administration, additional algorithms that may potentially add complexity to the base CAT algorithm are often needed to meet desired non-statistical specifications (Weiss & Kingsbury, 1984, 2000) such as content balancing or item exposure control. MST systems help to diminish the need for this added complexity, often to the benefit to test developers, test takers, and test performance monitors.

To the benefit of test developers, all MST components can be constructed and assembled prior to test administration, unlike a CAT. This provides MST developers with a level of design flexibility and control to meet the desired non-statistical specifications of a test form (Zwick & Bridgeman, 2014). MST developers can pre-plan and implement test design control without having to introduce additional algorithmic complexity as may be needed to administer a CAT with similar non-statistical specifications.

Furthermore, and to the benefit of test takers, because MST testing units are item sets scored as a unit, items within a testing unit can be viewed or answered in any order. In addition, within the testing unit, test takers can change any item response prior to submitting the unit for scoring. Comparable item and response review features are not supported in CAT systems since the CAT item-by-item selection algorithm requires

individual item responses to be submitted and scored prior to it selecting and administering the next test item.

Finally, and to the benefit of testing system performance monitors, MST grouped item sets and particularly testlets are known to help constrain to the item set potentially adverse influences on ability estimation accuracy. For instance, Wainer, Bradlow, and Wang (2007) present a discussion on how the use of testlets can help constrain item positioning context effects. Once constrained to the testlet, the context effects are more easily detectable by performance monitors. Whereas entire testlets in an MST system can simply be allocated to modules prior to testing to reap the constraint benefits, implementing similar levels of context effect control and constraint within a CAT system may require algorithmic add-ons (Wise & Kingsbury, 2000).

It follows that when items within a testlet are affected by IPD, the idea of constrained adverse IPD effects within the testlet could also apply. That is, the cumulative effect of IPD within testlet items may make the IPD more easily detectable to performance monitors. Thus, the results of investigations reporting on how test score accuracy and MST system performance are impacted by cumulative patterns of IPD constrained to the testlet level may be of interest to testlet-based MST developers.

The motivation to consider the influence of constrained cumulative IPD effects on MST systems is bolstered due to the widespread use of testlets as MST testing units. Hence, this dissertation examines the impact of cumulative IPD effects, constrained at the testlet level, on test taker final ability estimates, MST routing system performance, and decision classification accuracy. It seeks to provide the educational testing field with empirical evidence related to the impacts of cumulative insignificant magnitudes of IPD,

accumulating or cancelling to the testlet level, within a testlet-based MST system. This study aims to serve as an additional reference to MST test developers attempting to determine whether cumulative IPD should be investigated at the testlet level, in addition to conducting routing IPD evaluations typically performed at the individual item level, when evaluating the performance of a testlet-based MST system.

1.4 Amplification and Cancellation of IPD Effects

Evaluation of cumulative item parameter shifts existing as amplification or cancellation has been readily implemented in studies related to differential item functioning (DIF). When items are affected by DIF, expected response probabilities may differ by demographic subgroup for test takers that have similar levels of the ability being measured by the test (Bock, 1997). Thus, the presence of significant magnitudes of DIF is a test fairness issue.

Testlet-level DIF amplification refers to a testing situation in which the cumulative effect of insignificant magnitudes of DIF in the same direction may be unacceptably high at the testlet level even though the DIF magnitude may not be detected as significant for individual testlet items (Nandakumar, 1993). Testlet-level DIF cancellation at the testlet level results when the magnitude of DIF for individual testlet items is bi-directional, resulting in an overall magnitude decrease in DIF at the testlet level (Bao, Dayton, & Hendrickson, 2009; Roznowski, 1988).

Although IPD is a form of DIF, IPD amplification and cancellation studies have not yet received comparable attention in the literature. Since it is not uncommon for DIF investigative approaches to be applied to IPD studies (Donoghue & Isham, 1996), an overview of related DIF amplification and cancellation research is presented next. The

IPD amplification and cancellation investigation reported in this dissertation builds upon this research.

According to Bao, Dayton, and Hendrickson (2009), DIF amplification or cancellation may be present at the testlet level if an interaction between the test taker and the testlet stem content occurs. Research has shown that this interaction could occur when testlet items are grouped in relation to a common stimulus (Wainer, Bradlow, & Wang, 2007). In practice, examples of such item grouping types in test development include passage-based reading comprehension or mathematics testlets where items are associated with a graph or table of values.

For some test takers, their previous exposure to or lack of knowledge of the testlet stem content may influence how successful they are when responding to the item (Bradlow, Wainer, & Wang, 1999; Wainer & Kiely, 1987; Yen, 1993). If present, this test taker and testlet stem interaction results in a type of local item dependence (LID) where the manner in which a test taker responds to a testlet item may affect their responses to other items within the testlet. Tuerlinckx and de Boeck (2001) distinguish between two types of LID. One type of LID results when more than one ability influences test takers' item responses. The other is due to item interactions with test takers, as may occur with testlet use. When this latter type of LID exists, the items within the testlet may exhibit statistical dependence and the potential for DIF amplification or cancellation at the testlet level may be present (Bao, Dayton, & Hendrickson, 2009). This dissertation proposes that, similar to testlet-level DIF, cumulative patterns of IPD may also exist as amplification or cancellation at the testlet level when within testlet items are statistically dependent.

Parameters of measurement models, such as those derived under Item Response Theory (IRT), are used to quantify item characteristics and other influences like LID. Standard IRT models have a local item independence usage assumption and are therefore not formulated to account for the presence of LID in testlets. Even as testlets are widely used as testing units in an MST, MST test form construction, administration, test scoring, and statistical analysis of results is primarily carried out based on standard IRT measurement model usage assumptions (Yan, von Davier, & Lewis, 2014). Research has shown that neglecting the presence of LID (by fitting testlet response data using standard IRT models that ignore the presence of LID) results in inflated measures of test reliability, ability estimate accuracy, and item difficulty parameter estimation accuracy (Li, Bolt & Fu, 2006; Lu, 2010; Wainer, Bradlow, & Wang, 2007; Wainer & Lukhele, 1997). Due to their widespread use for ability estimations in testlet-based MST systems, a brief overview of standard IRT models is presented next.

1.5 Standard IRT Models

Standard IRT model use is limited to representing response data where local item independence is assumed, a single dominant ability is measured by the test, and items are scored dichotomously. For MST testlets where the items are associated with a common stem, the standard IRT model local item independence assumption may not hold if a significant magnitude of LID is present (Wainer & Kiely, 1987).

Each standard IRT model represents a mathematical relationship between test taker ability levels and the probability of a correct response to an item. The three standard IRT models include the one parameter logistic (1PL) (Rasch, 1960; Wright, 1977), the two parameter logistic (2PL, Birnbaum, 1968), and the three parameter logistic (3PL;

Birnbaum, 1968) IRT models. The formulation for and a brief discussion of each are presented next.

The 1PL IRT model:

$$\mathrm{P}_{ni}(\mathrm{x}=1|\theta_i)=\frac{1}{1+\mathrm{e}^{-(\theta_i-b_n)}} \quad (1)$$

The 2PL IRT model:

$$\mathrm{P}_{\mathrm{ni}}(\mathrm{x}=1|\theta_i)=\frac{1}{1+\mathrm{e}^{-\mathrm{a_n}(\theta_i-\mathrm{b_n})}} \quad (2)$$

The 3PL IRT model:

$$\mathrm{P}_{\mathrm{ni}}(\mathrm{x}=1|\theta_i)=(1+\mathrm{c}_n)+\mathrm{c}_n\frac{1}{1+\mathrm{e}^{-\mathrm{a_n}(\theta_i-\mathrm{b_n})}} \quad (3)$$

Each standard IRT model formulation represents the probability of test taker i providing a correct response to item n. The calculated probability is conditional on test taker of ability level θ_i being administered the n^{th} item that has IRT item difficulty parameter b_n , and/or discrimination parameter a_n, and/or guessing parameter c_n.

The standard IRT models are distinguishable based on the number and type of item parameters. The 1PL IRT model is the simplest of the three standard IRT models. It has only one item parameter which detects and accounts for differences in item difficulty. The popular IRT Rasch model (Rasch, 1960) is the same as the 1PL IRT model with the value of the discrimination parameter constrained to 1.0 for all items.

The 2PL IRT model is an extension of the 1PL IRT model, permitting each item to have both a difficulty and a discrimination parameter. Test score results generated

using items that have discrimination parameters allow test developers to better distinguish between high and low performing test takers than do test score results generated without discriminating items.

Finally, the 3PL IRT model is an extension to the 2PL IRT model, additionally accounting for an influence on response probabilities resulting from test taker use of guessing to obtain the correct response to an item.

Approaches have been suggested that are aimed at helping test developers diminish the impact of IPD and LID effects in MST systems where the use of standard IRT items are proposed. Suggestions for diminishing IPD effects are presented first followed by a discussion on approaches to diminish the impact of or account for LID effects.

1.6 Addressing IPD Within an IRT-based MST

To diminish IPD effects, item parameters should be calibrated as accurately as possible prior to testing. Four approaches that may work to improve item parameter calibration accuracy are (1) using a large enough random calibration sample, (2) selecting the best-fitting IRT model, (3) field testing new items to gather empirical evidence regarding how they function in practice, and (4) selecting an optimal calibration sample based on consideration of the items, the intended test taker population, and the intended test purpose (Ban, Hanson, Wang, Yi, & Harris, 2001; Buyske, 1998; de Ayala, 2009; Hambleton, Jones, & Rogers, 1993; Kingsbury, 2009).

Another approach to help minimize IPD effects and consequently the impact on test scores is monitoring item parameter magnitude changes across testing occasions. Significant shifts in the magnitude of item parameters could signal the presence of IPD or

some other unanticipated influence in need of investigation in order to determine an acceptable responsive action.

A commonly used graphical approach to standard IRT model item parameter monitoring makes use of item, testlet, or test level mathematical curves across testing occasions to detect changes to item parameter values. To illustrate this approach, the graph of 2PL IRT item characteristic curves (ICC), also called item response functions (IRF), for the same item over two separate testing occasions are displayed in Figure 1.

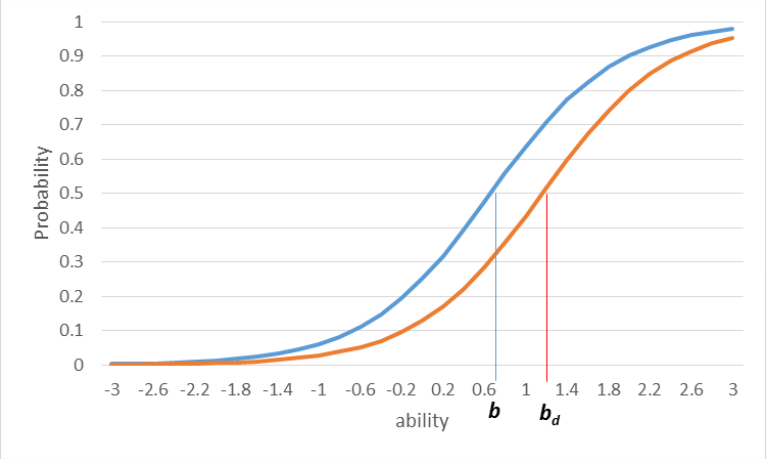

Figure 1. Graph of 2PL IRT ICCs

Test and testlet-level characteristic curves can be generated by summing the ICCs for all the items in the test or within the testlet respectively. For a single item and its parameters, Figure 1 shows that the vertical axis of an ICC graph represents the probability of a correct response to the item. The continuum of test taker ability levels is represented along the horizontal axis.

The blue curve in Figure 1 illustrates an item ICC at time of calibration. The red curve illustrates an ICC of the same item generated after a later test administration. The two graphs are not the same for both testing occasions due to a shift in the value of the

difficulty parameter from b to b_d. In theory, from the time of calibration, the ICC of an appropriately functioning item should not change significantly in shape across repeated test administrations. Changes in the shape of an ICC could be due to IPD effects or to some other unanticipated and thus un-modeled influence. If ICC changes are monitored, shifts in item parameter value can be detected. Then, test developers can perform investigations to determine how best to proceed regarding the significance of the change and consequently what future action should be taken regarding the affected item.

Two popular control chart approaches for detecting and monitoring IPD over time in MST systems are discussed in Lee, Lewis, and von Davier (2014**).** They include the Shewhart (1931) and the cumulative sum control **(**CUSUM; Veerkamp & Glas, 2000) approaches. The Shewhart control chart flags an affected IPD item when the monitored means of its item parameters fall outside of pre-determined bounds. The CUSUM statistic can be calculated as a cumulative sum using detected item parameter value changes over successive recalibrations. If a pre-determined value of the CUSUM statistic is reached, it is an indication that a significant change in the value of an item parameter has occurred. Lee, Lewis and, von Davier **(**2014) also mention use of a Lagrange multiplier statistic that is discussed in detail in Glas (2010), for detecting IPD in MST response data. This approach also makes use of control charts not to monitor item parameter value changes but to monitor changes in test score means across test administrations.

1.7 Testlet Response Theory for Addressing LID

To account for the influence of LID in testlet response data due to test taker and item interactions there are several model-based approaches, used in place of standard IRT models, that have been reported in the research literature. First, use of Testlet Response

Theory (TRT; Wainer, Bradlow, & Wang, 2007) models are suggested as extensions to standard IRT models. TRT model formulations include a random parameter, referred to as a testlet effect parameter, which captures and accounts for the influence of LID on final scores resulting from test taker and testlet stem interactions. Under TRT, the variance of the testlet effect parameter serves as an indicator of the magnitude of the LID effect. Unlike standard IRT models, TRT models do not require that the local item independence assumption hold among items within the testlet (Wainer, Bradlow, & Wang, 2007). Even so, TRT models do assume that local item independence exists between testlets (Rosenbaum, 1988).

When used in place of standard IRT models, TRT models have been shown to increase final ability estimate accuracy when LID resulting from test taker and testlet stem interactions exists (Bradlow, Wainer, & Wang, 1999). TRT has been regularly used in large-scale assessment research to model testlet response data when investigators want to examine LID effects. For example, Eckes (2013) applied the two-parameter logistic TRT (2PL TRT) model to the listening section of the Test of German as a Foreign Language. Moreover, the three parameter logistic TRT (3PL TRT) model was applied by Wainer and Wang (2000) to the listening and reading comprehension testlets in the Test of English as a Foreign Language and also by Wang, Bradlow, & Wainer (2002) to the of the North Carolina Test of Computer Skills as well as the Educational Testing Service's Test of Spoken English.

The bifactor model developed by Gibbons and Hedeker (1992) is another model that has been used in research (e.g., De Mars, 2006) to account for LID in testlet response data. The bifactor model is a multidimensional model that can be used to model item

response data that exhibit a bifactor, or two measured abilities, structure. Thus, in a bifactor model, each item response is modeled as a function of a primary ability and a secondary ability. To that end, for passage-based testlets where LID due to item interactions is assumed to exist, under the bifactor model the primary ability is the ability assumed to be measured by the items on the test and the secondary ability is the testlet effect.

Distinguishing it from the TRT model, the bifactor model includes an additional discrimination parameter associated with differing levels of the testlet effect parameter. In other words, when modeling testlet response data using the bifactor model, items can vary in discriminating levels for the primary ability measured in a manner that is dissimilar from discriminating levels that are associated with the testlet effect (De Mars, 2006). In testing situations where test takers may be influenced by more than on factor associated with the testlet stem, the additional discrimination parameter in the bifactor model formulation is derived to account for these differing effects.

De Mars (2006) found that when item response data were generated with either the bifactor or TRT model, the true item difficulty parameters were recovered well for both models. However, a slightly higher root mean square error (RMSE) was observed for the bifactor model, implying that when using the bifactor model, recovered item parameters tended not to match the true parameters as well as the item parameter matches observed using the TRT model. Therefore, the more parsimonious TRT model over the bifactor model is a feasible solution when both models are found to be a suitable fit given the testlet response data. That is, TRT models can explicitly model the LID in testlet

response data with less added model complexity (Lu, 2010) when compared to bifactor model use.

Other measurement models that have been proposed to account for LID when calculating ability estimates includes item-based multilevel testlet response theory models, such as the three-level testlet response theory model discussed in Jiao, Wang, & Kamata (2005). In addition to accounting for LID, the multilevel testlet response theory models also take local person dependence (LPD) into account when calculating ability estimates. LPD can occur when test takers clustered within a hierarchy tend to perform more similarly to one another than to test takers outside of the cluster. For example, test takers attending the same school in a district respond more similarly to test items than test takers who attend another school within the same district.

Multilevel testlet response theory models have been applied in educational research including to DIF detection and item generation studies (Beretvas & Walker, 2012; Glas, & van der Linden, 2003; Ravand, 2015). However, to date they have not been used when constructing MST components nor when investigating MST system performance. Hence, discussions involving use of multilevel testlet response theory models to model LID in MST systems is beyond the scope of this study.

However, TRT models have been shown to model LID well due to test taker and item interactions and are more parsimonious than the bifactor and the multilevel testlet response theory models. TRT models have also been used in previous studies to model LID in MST testlet response data (Lu, 2010) as well as to construct an MST system (Keng, 2008). Furthermore, the 2PL TRT model in particular has been used in DIF amplification studies that have compared its performance under LID conditions to that of

the 2PL IRT model (Bao, Dayton, & Hendrickson, 2009). The 2PL TRT model has also been used in MST studies that do not include investigation of cumulative item parameter shift effects (Keng, 2008). Finally, the 2PL TRT model, the first TRT model proposed and investigated by Bradlow, Wainer, and Wang (1999), has also been shown to recover item parameters better than the 2PL IRT when LID is present. For these reasons, the performance of a testlet-based MST system under 2PL TRT and 2PL IRT assumptions is compared in this present study.

1.7.1 The 2PL TRT Model

The 2PL TRT model formulation is embedded in a Bayesian framework (Wainer, Bradlow, & Wang, 2007). It is through the use of probability distributions, called hyperpriors, that a hierarchical Bayesian structure of the 2PL TRT model is imposed. The hyperpriors govern the priors of the model parameters and are derived from the estimates of the posterior distribution of the item and the testlet effect parameters. In the 2PL TRT Bayesian framework, priors are imposed on the item parameters, testlet effect parameter, and the ability θ distribution. To identify the 2PL TRT model, the mean and variance of the ability θ distribution are fixed at 0 and 1.0 respectively.

To model the probability of providing the correct response for item n within a testlet conditional on test taker i with ability level θ_i being administered the n^{th} item, the 2PL TRT model is formulated as follows:

$$P_{ni}(x = 1|\theta_i) = \frac{e^{(a_n(\theta_i - b_n - \gamma_{id(n)})}}{1 + e^{(a_n(\theta_i - b_n - \gamma_{id(n)})}} \qquad (4)$$

$d(n)$ maps item n to the testlet of which it belongs. For the testlet $d(n)$, $\gamma_{id(n)}$ represents the value of the testlet effect for test taker i. Furthermore, responses to items n and $n^{'}$ have the same $\gamma_{id(n)}$ value if items n and $n^{'}$ are in the same testlet administered to test taker i. For items modeled under the 2PL TRT, the value of the testlet effect for each item is the same for a single test taker within the same testlet. However, the magnitude of the testlet effect for an item can change in value from one test taker to the next.

When using the 2PL TRT model, test developers are not required to construct tests where every item belongs to a testlet or exhibits some magnitude of LID (Wang & Wilson, 2005). In fact, when $\gamma_{id(n)} = 0$ or when the variance of the testlet effect parameter is equal to 0, then the standard IRT model local item independence assumption has not been violated. Hence, the presence of no testlet effect in the response data across all testlets reduces the 2PL TRT to the standard 2PL IRT model with Bayesian priors for both the item parameters and the ability distributions (Wainer & Kiely, 1987).

1.8 The Need for More IPD Testlet Based MST Research

It has been discussed in the above sections how the use of testlets as testing units in an MST system may create testing situations where IPD amplification or cancellation at the testlet level and/or LID may exist due to test taker and testlet stem interactions. It follows that the influence of these combined effects on ability estimation and system performance may be particularly troublesome given the in practice approach of regularly constructing, administering, and scoring testlet-based MSTs under standard IRT model assumptions.

Investigations reporting on testlet-based MST ability estimate and system performance impact due to the combined presence of testlet-level IPD amplification or

cancellation and/or LID effects have not yet been presented in the literature. Only by conducting investigations under these conditions in testlet-based MST settings can these combined influences be evaluated to determine if the impact on final ability estimates are negligible and can be ignored, or if there are practical implications to test fairness that should be addressed by test developers.

Furthermore, such investigations could also provide insight into potential MST system vulnerabilities, namely system operations that depend on the stability of item parameter values across testing occasions. For example, testlet-level statistics used in MST test form construction and module-level routing procedures are calculated using item parameter values. Thus, when testlet-level IPD exists in the presence of LID, it is important to consider to what extent and at what magnitudes testlet-based MST system operations that depend on the accuracy of module-level statistics are most vulnerable to these effects. To gain insight into potential vulnerabilities, a brief overview of MST test form construction is presented next, followed by a discussion on MST routing operations.

MST test form construction depends on item parameter stability in the calculation of module-level statistics. For instance, Luecht (2014) discusses an approach to MST test form construction that uses module-level statistical targets for selecting items into modules to ensure desired levels of measurement accuracy. Once the modules have been constructed they are used to assemble MST test forms. Thus, the level of measurement accuracy expected upon repeat administration of the assembled test forms depends on the estimated accuracy and stability of item parameter values.

MST routing procedures may also depend on module-level statistics. When MST modules consist of a single testlet, testlet-level properties are essentially module level

properties. One testlet-level statistic used in MST routing operations that may be impacted by the presence of cumulative IPD is the average item difficulty for the module. When designing the MST, test developers select testlets for module inclusion that reflect the intended average item difficulty desired for that module. Bryant and Jiao (2016) showed that for a testlet-based linear test, insignificant levels of IPD existing as positive amplification to the testlet-level may increase the average item difficulty levels from what is expected. Under similar unanticipated increases in average module difficulty within an MST, a test taker may potentially be adaptively routed to a next stage testlet that is not suitable for their ability level.

1.9 MST Routing and the Problem of IPD

Weissman (2014) describes MST routing procedures in detail; however, only a brief overview of MST routing procedure is presented here. To initiate the MST routing procedure, first an initial measure of ability is assigned to a test taker. The assignment can be done randomly or by making use of test taker performance information that is available prior to testing. Next, test takers are administered one module at the initial test stage. This module is often called a routing test. The routing test is typically constructed using items that are moderately difficult. Once the routing test has been administered and scored, a decision is made regarding which second stage module to route the test taker to. This process of calculating module-level scores for use in routing decision making is repeated until the final module in an MST pathway has been administered.

MST routing rules used for module-level scoring can be classified as either static or dynamic (Weissman, 2014). Under standard IRT assumptions, static MST routing rules use standard IRT item parameters available prior to testing to calculate number

correct (NC) scores. Then, NC lookup tables are constructed and used to make routing decisions during testing. NC scores calculated under IRT assumptions are IRT true scores derived using the sum of the ICCs of the test items. That is, the ICCs can be summed to produce a test level IRT curve called either a test characteristic curve (TCC) or a test response function (TRF). The general formulation for the TCC under the 2PL IRT model is as follows.

$$TCC = \sum_{n=1}^{L} \mathrm{P_n}(\mathrm{x} = 1|\theta, a_n, b_n) \qquad (5)$$

where across all L test items administered and at each test taker ability level θ, a sum is calculated for the probability of providing a correct response to each item n.

In Figure 2, the vertical axis of the TCC represents the expected true score that a test taker, on average, would obtain if the test was theoretically administered an infinite number of times. Along this axis, true scores range from 0 to the highest NC score possible for the test.

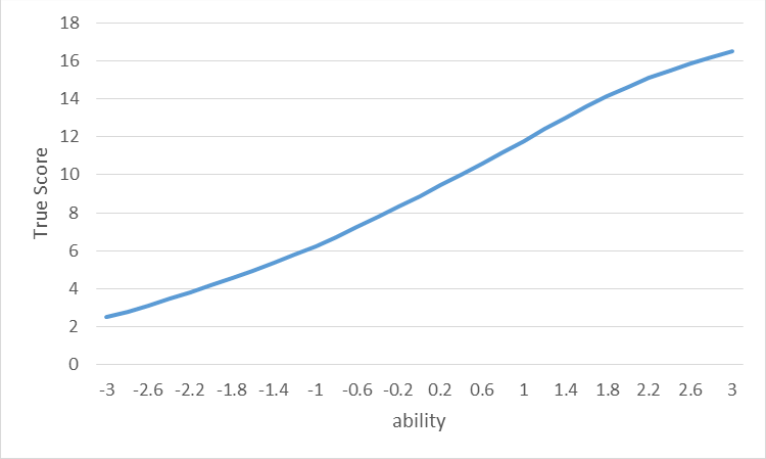

Figure 2. Graph of 2PL IRT TCC

NC routing uses the IRT true scores that are associated with the location of pre-determined ability level cut scores. The NC score calculation procedure used in this study

is presented in Chapter 3. Once NC scores are calculated, NC lookup tables are used by the routing module selection method to inform routing decision making during test administration. No IRT true scores nor IRT based interim ability estimation procedures are used during MST test administration under NC routing. Hence when static MST routing procedures such as NC routing are used, if item parameters significantly shift from expected values, there are no opportunities during test administration to update the NC lookup tables to reflect the item parameter value changes.

To the contrary, dynamic routing rules do make use of IRT item parameters during operational testing to calculate interim ability estimates to inform test taker routing decisions. Thus, Weismann (2014) notes that dynamic routing rules are more efficient at incorporating currently available item pattern of response information into routing decisions than NC static routing rules are. However, in practice, due to its ease in implementation, NC routing rules are preferable to dynamic routing rules (Weissman, 2014). Dallas (2014) points out that the use of NC routing scoring methods provides relief to testing centers as they would no longer have to perform IRT calculations onsite. Research however has found IRT based and NC routing rules to be feasible alternatives to each other (Armstrong, 2002; Dallas, 2014; Zenisky, 2004). Research also shows that under both routing rules, using cumulative scoring at the conclusion of each MST stage to inform routing decisions results in the most precise measures of test taker ability (Dallas, 2014).

Different MST routing module selection methods have been devised that can be used with either NC or dynamic routing rules. According to Kim, Chung, Park, and Dodd (2013), the two most frequently used MST routing module selection methods in research

are the defined population intervals (DPI) and the approximate maximum information (AMI) methods. Both methods are introduced and discussed in detail in Luecht, Brumfield, & Breithaupt, 2006. A brief overview of these two popular module selection methods is presented next.

Under the DPI module selection method, a pre-determined proportion of test takers are routed to different modules within the same test stage. One motivation for choosing to implement a DPI module selection method may be a desire by test developers to protect module items from overexposure (Dallas, 2014). The DPI routing module selection algorithm is implemented in two steps. First, interim measures of ability are calculated at the conclusion of the current MST test stage. This permits the DPI algorithm to rank order test takers according to their current estimated ability measure. Test takers may be rank ordered according to ability across all modules within the current test stage or just within a single module. Second, using the rank ordered listing a pre-determined proportion of test takers by ability grouping are routed to the different next test stage modules.

Unlike DPI, there are no pre-determinations made regarding the proportion of test takers to be administered a module with an AMI module selection method. Instead, these methods use the intersection of IRT information functions as routing rule decisions points when deciding which module in the next MST stage to route test takers to. Under IRT assumptions, the information functions used when implementing AMI routing module selection methods could be generated at the test or module level. It follows that at the completion of a test stage, AMI module selection procedures route test takers to the most informative next stage module.

Research studies (Kim, Chung, & Dodd, 2010; Kim, Chung, Park, & Dodd, 2013; Zenisky, 2004) have compared the performance of DPI and AMI module selection methods using different MST panel designs. Both methods tend to perform similarly with one exception: at the module level, the AMI method was shown to perform slightly better than other module selection methods (Kim, Chung, & Dodd, 2010).

Because an MST system has fewer adaptation points for recovery from routing errors than would be available in a CAT of similar test length, Zwick and Bridgeman (2014) note that investigation of influences that may potentially impact MST routing performance, and consequently final ability estimates, are key research areas. Yet, routing rules and module selection methods research has not investigated the performance of MST system operations when combined cumulative patterns of IPD and LID effects exist. Thus, this dissertation seeks to fill this gap in the MST and IPD literature by investigating the influence of IPD amplification and cancellation to the testlet-level on the performance of testlet-based MST routing efficiency, both when LID exists and when it does not. Decision classification accuracy measures are reported as an overall outcome of resulting ability measurement and test taker routing decisions. The detailed summary of the purpose of this study and the research questions are presented in the next section.

1.10 Purpose of this Study

MST test developers desire to meet test fairness goals by designing systems that generate final ability estimates at an acceptable degree of accuracy for all test takers across all testing occasions. How well an MST system produces precise ability estimates may be impacted in at least two ways. First, when measurement model usage

assumptions are violated, a lack of fit between the item response data structure and the measurement model could result. This mismatch between response data and measurement model may adversely impact final ability estimate accuracy. In practice, one testlet-based MST testing situation where a lack of model fit to response data may occur is when LID-affected testlet response data are modeled using standard IRT models.

Secondly, the presence of item parameter variance due to amplification and cancellation of IPD at the testlet level may potentially impact testlet-level statistics and scoring. When testlet-based MST system routing operations that depend on testlet-level scoring and statistics for making routing decisions are implemented, IPD amplification or cancellation to the testlet-level may have an indirect impact on final ability estimate accuracy due to increased routing error rates. Moreover, unacceptable routing error rates combined with imprecise ability estimation may potentially adversely impact overall test misclassification rates.

Numerous researchers have argued that when testlets are the testing unit, influences with the potential to impact test fairness should be studied at the testlet-level rather than at the individual item level (Bao**,** Dayton, & Hendrickson, 2009; Douglas, Roussos, & Stout, 1996; Nandakumar, 1993; Sireci, Wainer, & Thissen, 1991; Wainer, 1995). Even so, to date, the influence of testlet-level IPD and its impact on final ability estimates has not been studied within a testlet-based MST system where LID is assumed to exist. Moreover no studies have examined the impact of ignoring the LID effect on the operational performance of a testlet-based MST routing system in the presence of amplification and cancellation of IPD to the testlet-level. For test developers to administer testlet-based MST systems under conditions that may largely ignore the

potential for cumulative IPD and LID effects, investigations are needed to show that standard IRT final ability estimation and MST routing operations are robust under these conditions. This dissertation seeks to investigate cumulative patterns of IPD and LID combined effects, under various testing conditions, to provide empirical evidence of the extent of testlet-based MST ability estimation, routing, and decision classification sensitivities to these effects.

There are at least three reasons why combined cumulative IPD and LID influences on MST ability estimation and routing performance under both standard 2PL IRT and its extended 2PL TRT model assumptions are selected to be examined in this study. First, there exists previous MST and amplification and cancellation research that has been performed using both models. Second, standard IRT models are the most likely measurement models to be used in practice when constructing and scoring an MST even when testlets are used as the item type. Hence it is important for study conditions to be examined under both 2PL IRT and 2PL TRT assumptions. Third, since LID due to testlet effects can be parsimoniously modeled by varying the TRT testlet-effect variance, the 2PL TRT model well serves as the assumed true model throughout this study.

1.11 Research Questions

The following research questions are explored in this dissertation:

1) In a testlet-based MST system design, if the 2PL TRT model is the true measurement model, how is the degree of accuracy in final ability estimates calculated under 2PL IRT assumptions impacted by the magnitude, type, direction, and MST test stage where IPD amplification and cancellation at the testlet level exists?

2) Under a NC routing rule using an AMI module selection method, if the 2PL TRT model is the true measurement model, do testlet-based MST routing error and misclassification rates differ when IPD amplification and cancellation at the testlet level exists?
3) When the module average item difficulty difference conditions within an MST stage are adjusted, if the 2PL TRT model is the true measurement model, are testlet-based MST routing error and misclassification rates impacted when IPD amplification and cancellation to the testlet-level exists?
4) Would use of the 2PL TRT model, that can account for the impact of LID due to testlet effects, improve the overall measurement accuracy in final ability estimates, testlet-based MST routing error, and misclassification rates versus the use of the 2PL IRT model in the presence of IPD that exists as amplification or cancellation to the testlet-level?

To answer these research questions, a simulation study within a testlet-based MST system was carried out. During the simulation, various conditions were imposed to study the impact of combined cumulative testlet-level IPD and LID effects on final ability estimates and MST routing errors. Manipulated IPD related study factors include the type of IPD, the MST stage within which cumulative IPD exists, the magnitude of IPD, the percent of IPD items, and the direction of the IPD. The magnitude of the testlet-effect variance is the manipulated LID study factor. The module average item difficulty difference between adjacent modules within the same MST test stage also varies. All investigations are implemented using an NC routing rule under an AMI module selection method. Finally, to examine and report the extent of differences generated, the degree of

accuracy in ability estimates, routing error rates, and misclassification rates are compared between 2PL IRT and 2PL TRT model usage assumptions.

1.12 Organization of the Study

This study is presented in five chapters. Chapter 1 presented the background and purpose of the study as well as the research questions. Chapter 2 provides a literature review that concentrates on four aspects including (1) the basis of IRT and IRT based IPD investigations, (2) the problem of IPD amplification and cancellation to the testlet level, (3) the 2PL TRT model as an extension of the 2PL IRT model, and (4) the components and construction of an MST with focus on the MST routing procedure. Chapter 3 describes the research design, study conditions, manipulated factors, parameter generation, scoring, routing, and the evaluation approach used. Chapter 4 provides a detailed report of the study results for the analyses introduced in Chapter 3. Chapter 5 summarizes the findings, discusses the implications and limitations of the study, and provides some direction for future research.

Chapter 2: Literature Review

2.1 Introduction

This chapter includes four major sections. Section 1 presents an introduction to the chapter content. Section 2 reviews IRT-based IPD research including item level, compound, and cumulative IPD investigations. Amplification and cancellation at the testlet level are also discussed with emphasis on previous DIF related research and the potential problem of IPD amplification and cancellation at the testlet level. Section 3 provides information about the MST framework including component construction and panel assembly with focus on the MST routing system. This section also presents potential problems for ability estimation underlying the MST system when LID in testlets exists and a standard IRT measurement model is used. An overview is also presented of the 2PL TRT model, as an extension of the standard 2PL IRT model that is derived to account for the presence of LID in testlet response data. Finally, Section 4 summarizes this chapter and reiterates the purpose of the study.

Item parameter invariance is a key standard IRT model usage assumption. Yet, testing situations exist where violations of this usage assumption may unexpectedly occur. That is, final ability estimates and testing system operations may be impacted by item parameter variance. Mellenbergh (1989) presents a detailed discussion on measurement invariance that can be used to describe unexpected item parameter variance as a conditional independence of item functioning. When this occurs, unanticipated influences such as IPD may not accurately be accounted for by the selected measurement model when test taker abilities are estimated. These unaccounted for influences have the potential to threaten test fairness. It follows that if item parameters drift significantly

from the values assigned at calibration, a previously assumed correct standard IRT measurement model may no longer suitably represent the relation between test taker ability and the probability of a correct response (Wells, Subkoviak, & Serlin, 2002).

Under the 2PL IRT model, IPD could potentially be present as unexpected changes in the difficulty or discrimination item parameter. This is often referred to in the IPD literature as b-drift and a-drift respectively. Furthermore, it is also possible that a 2PL IRT item may exhibit ab-drift where IPD is simultaneously present in both the difficulty and the discrimination parameters.

When tests are constructed using statistically independent items, research has shown that the magnitude of IPD may compound in a single item (Bock, Muraki, & Pfeiffenberger, 1988; Sykes & Fitzpatrick, 1992) or accumulate in multiple items over time so as to be detectable at the test level (Hans & Wells, 2007; Han, Wells, & Sireci, 2012). For this reason, it is not improbable that IPD magnitudes could also accumulate to the testlet level; however, this phenomenon has not received similar attention in the literature. Instead, IPD studies have primarily focused on investigating non-cumulative item level effects under testing conditions where items are assumed to be statistically independent and modeled under IRT. An overview of item level IPD, compounded IPD over time, and cumulative IPD at the test level literature is presented in the next section.

2.2 IPD Research Overview

2.2.1 Item level IPD

Since items with significant magnitudes of IPD of 0.5 logits or higher are typically targeted to be flagged by detection procedures (Han & Guo, 2011), IPD studies have focused primarily on investigating the effects of insignificant magnitudes of IPD

present in item response data modeled under IRT. Even so, the IPD investigative literature tends not to be as comprehensive as the DIF literature. Wollack, Sung, and Kang (2005) note possibly that one reason IPD investigative literature may be fewer is that findings, given the study conditions imposed, have largely reported the impact of item level IPD on IRT final ability estimates to be negligible. A number of IPD studies have been conducted under linear test conditions where statistically independent items modeled under IRT are used (Rupp & Zumbo, 2006; Wainer and Thissen, 1987; Wells, Subkoviak & Serlin, 2002; Witt, Stahl, Bergstrom, & Muckle, 2003). These studies have shown that unless the percentage of affected IPD items was unusually large, the impact on final ability estimates tends to be insignificant.

Witt, Stahl, Bergstrom, and Muckle (2003) examined IPD effects under the Rasch model in a linear test constructed using statistically independent items. The study had six simulated insignificant unidirectional IPD shift conditions in the form of b-drift. They found that even under the assumptions of non-normal distributions of test takers, nearly 25% of the items needed to exhibit insignificant levels of IPD before a significant effect was detected. Similarly, Wells, Subkoviak, and Serlin (2002) found under the 2PL IRT model that at least 20% of linear test items exhibiting either b-drift or a-drift at magnitudes of 0.5 and 0.4 logits respectively were needed to meaningfully impact IRT final ability estimates. Wells, Subkoviak, and Serlin's (2002) findings were further supported by Rupp and Zumbo (2006) who, by mathematical and graphical means, demonstrated that the Wells, Subkoviak, and Serlin simulated changes to the statistically independent item parameters resulted in response probability differences that would have minimal impact on IRT final ability estimates.

Other findings showing insignificant magnitudes of IPD having negligible impact on IRT final ability estimates have been reported under MST and CAT study conditions when statistically independent items are administered. Wei (2013) investigated the impact of insignificant magnitudes of unidirectional IPD in the form of ab-drift on final ability estimates under a 3PL IRT-based MST system. IPD was simulated over conditions of 5%, 10%, and 20% of the total test length. Wei found that the simulated IPD had negligible impact on IRT final ability estimates unless at least 20% of items were IPD affected. Han and Guo (2011) investigated the impact of insignificant magnitudes of IPD in the form of uni-directional b-drift under a 3PL IRT model in a CAT system. To model IPD due to practice and curriculum changes, IPD was exhibited in partial group simulations of 0%, 10%, 20%, 30%, 40%, and 50%. Furthermore, IPD was simulated in 20% of items at a magnitude of -0.5 logits since according to Han and Guo (2011), practice and curriculum changes usually made items easier for test takers. Their findings were that the IPD effects did not meaningfully impact IRT final ability estimates. However, they did advise caution regarding the potential for impact on IRT final ability estimates due to compound IPD effects in items over time.

2.2.2 Compounded IPD

Compounded IPD results when a single item is repeatedly affected by IPD over time such that the overall detectable IPD magnitude increases. Numerous studies investigating compounded IPD effects using real data from linear tests have been performed; however, the findings have been mixed (Bock, Muraki, & Pfeiffenberger, 1988, Chan, Drasgow, & Sawin, 1999; De Mars, 2004; Han, Wells, & Sireci, 2012; Sykes & Fitzpatrick, 1992). For instance, Bock, Muraki, and Pfeiffenberger (1988) over a

10-year period using real data and under a three-parameter time dependent IRT model and Sykes and Fitzpatrick (1992) over a 5 year period under the Rasch model examined compounded IPD effects. Significant compound difficulty parameter IPD influences on the degree of accuracy in IRT final ability estimates were detected. Both studies suggested the possibility that content-related changes in curriculum emphasis may have resulted in the compounded IPD effect observed. Of note is that, in the Sykes and Fitzpatrick study, close to 91% of the items on the test were passage-based testlets with items modeled under IRT assumptions as statistically independent.

Chan, Drasgow, & Sawin (1999) using real data over a 16-year period and De Mars (2004) with simulated and real data conditions over a 4-year period also investigated compounded IPD effects under a 3PL IRT model. They found that the impact of the insignificant magnitudes of compound IPD on overall test scores was minimal due to canceling effects caused by the bidirectional changes in IPD magnitude. Moreover, De Mars (2004) was not able to conclude that the compounded IPD effects observed were associated with content differences as Bock, Muraki, and Pfeiffenberger (1988) and Sykes and Fitzpatrick (1992) had done. Finally, Deng and Melican (2009) performed a 4-year compound IPD investigation under the 3PL IRT model within a CAT large-scale placement test. They compared transformed item parameters with original item parameters. They reported that no detection of significant IPD was found except in two of the items.

2.2.3 Cumulative IPD

Unlike compounded IPD, cumulative IPD effects are due to the combined influence of multiple IPD-affected test items. Han, Wells, and Sireci (2012) investigated

the impact of cumulative IPD across multiple items at the test level on ability estimates and transformation coefficients. Transformation coefficients are often needed when multiple test forms are administered since it is typically the case that the test forms may differ slightly in terms of item difficulty. Thus, in order to ensure that test scores across multiple test forms can be compared, a test score equating procedure is performed to place the item parameters and the test scores from the multiple forms on to the same measurement scale (Cook & Eignor, 1991; Kolen & Brennan, 2004). Numerous equating designs have been derived that can be used when constructing exchangeable test forms (von Davier, 2010); however, the equating design used most often in large-scale testing is the anchor test design. With this design, a common set of items exist across the multiple test forms to be equated (Cook & Eignor, 1991). The study by Han, Wells, and Sireci (2012) investigated cumulative IPD effects detectable at the test level on equating with anchor item design and found that transformation coefficients were impacted.

Han, Wells, and Sireci (2012) noted that with respect to a learning effect, items could become easier or harder depending on their emphasis in the curriculum over time. This could result in multidirectional IPD existing across multiple items in the item bank. Hence Han, Wells, and Sireci (2012) were interested in examining the cumulative direct effect of bidirectional insignificant magnitudes of IPD in the form of b-drift in a linear test on item parameter transformation procedures, as well as its indirect effect on transformed ability estimates. Therefore, this study could be more generally classified as an IPD cancellation at the test level study. Under the 3PL IRT model, Han, Wells, and Sireci (2012) investigated the vulnerability of transformation coefficients derived using three different scaling techniques. Insignificant magnitudes of bidirectional IPD were

imposed on 40% of the item parameter estimates for the common items. They found that in the linear test, the extent of impact on final scores depended on the pattern of cumulative IPD imposed as well as the method used for computing the transformation coefficients.

Han and Wells (2007) investigated the impact of insignificant magnitudes of unidirectional IPD existing in common items in the form of b-drift on equating. This study could be more generally classified as IPD amplification at the test level study. They concluded that equating results could be significantly impacted when just 10% of items in the common item set are affected by insignificant magnitudes of IPD. Neither the Han and Wells (2007) nor the Han, Wells, and Sireci (2012) studies included investigations of cumulative patterns of IPD effects at the testlet-level. However, their work does show that cumulative insignificant magnitudes of IPD may become significant at the test level potentially affecting the accuracy of the statistical values calculated at the test level. That is, their study shows that cumulative IPD effects may have impact on calculated statistics even when IPD magnitudes are insignificant.

One limitation to consider when generalizing Han, Wells, and Sireci (2012) and Han and Wells (2007) is that they confined examination of cumulative IPD effects at the test level under the assumption of local item independence with item response data modeled under IRT. There are many large-scale assessments, particularly in the context of language and reading testing, in which the potential for LID due to the use of testlets is common (Min & He, 2014). Hence for these testing situations, questions remain regarding the extent, type, or magnitude do cumulative patterns of IPD at the testlet-level impact final ability estimates and/or testing operations in the presence of LID.

Some studies have shown insignificant magnitudes of item level DIF within a testlet potentially aggregating to produce unacceptable levels of DIF at the testlet level (Bao, Dayton & Hendrickson, 2009; Douglas, Roussos, & Stout, 1996; Nandakumar, 1993; Wainer, 1995). Wainer (1995) detected and investigated DIF amplification and cancellation effects present in testlet response data generated from both the reading comprehension and analytical reasoning sections of the Law School Admissions Test. For the investigation, the model fit to the data was the polytomous IRT nominal response model (Bock, 1972). After examining the differences in the TCC curves, Wainer (1995) detected statistically significant DIF cancellation at the testlet level. However, from a practical perspective and in terms of posing a threat to test fairness, the magnitudes of the differences detected were negligible. Bao, Dayton, and Hendrickson (2009) detected and investigated DIF amplification and cancellation effects present in a real response data set from a linear national reading test where all items were testlets. The models used in the study were multiple group 2PL IRT and 2PL TRT models. Inclusion of an additional grouping index on the model parameters permitted investigation of how testlets functioned differently due to test taker group membership. Bao, Dayton, and Hendrickson found testlet effects in the response data and item characteristics interacted with the testlet effects and resulted in DIF amplification and cancellation at the testlet level.

This dissertation proposes that like DIF, cumulative patterns of IPD can exist as amplification or cancellation at the testlet level. In a simulated linear test, Bryant and Jiao (2016) examined cumulative IPD effects at the testlet level as amplification and cancellation in the form of b-drift. The IPD study effects were investigated under both assuming and not assuming LID. The linear test consisted of 40 items assembled into 4

testlets of 10 items each. Just 15% (6 out of 40) of the total test items were simulated to exhibit IPD. Under study conditions, the magnitude of IPD was simulated at $\pm .3$ or $\pm .5$ logits. Bryant and Jiao found that the impact on final ability estimates was negligible. However, testlet-level statistics, such as average item difficulty and standard deviation, tended to increase as the magnitude of IPD increased, depending on the cumulative IPD pattern imposed. Unlike the linear test simulated in this study, there are testing systems with operations that may depend on testlet-level statistics remaining stable across testing occasions. An MST is one such system.

2.3 Multistage Tests

MST systems are becoming increasing popular as computerized adaptive testing solutions (Luecht, 2014; Jiao & Lissitz, 2012). Some examples of popular MST testing solutions include the Law School Admissions Test (Schnipke & Reese, 1999), the Uniform CPA Examination (Breithaupt, Ariel, & Veldkamp , 2005), the Medical Council of Canada test (Yan, von Davier, & Lewis, 2014), the Massachusetts Adult Proficiency Test (Sireci, Thissen, & Wainer, 2008), and the National Assessment of Educational Progress (Bock & Zimowski, 1998).

A testing system where adaptations made to test difficulty levels occur in stages, as is the case with an MST, is not a newly devised testing format. In fact, Cronbach and Gleser (1965) discussed a two-stage paper-based test (PBT) for classification testing. Under their design, all test takers receive the stage one test; however, the stage two test is only administered to test takers with scores close to the cut score used for making classification decisions. Results showed higher decision accuracy than that obtained when a traditional linear PBT is used. Lord (1971) investigated the performance of a two-

stage PBT design where test takers were administered a routing test in the first stage. Estimates of test taker ability were calculated based on scores they received after completing the routing test. During the second stage, test takers were administered a measurement test containing items selected at a difficulty level that was appropriate given their estimated ability level. Here, the two-stage PBT provided better ability measurement accuracy than a traditional linear PBT of equal length. Today, with advances in testing technology and the use of IRT, most multistage tests are computer administered (Zwick & Bridgeman, 2014).

When constructing modules for MST test forms, IRT item parameters can be used prior to or during testing to select items for inclusion that meet target test specifications (Han & Guo, 2014). Collectively, items selected for module inclusion have a desired average difficulty and level of information. Item information calculation can be accomplished under IRT with the use of an item information function (IIF; Birnbaum, 1968). An IIF quantifies how well an item measures test taker ability across the ability continuum. Under IRT, IIF values can be calculated prior to operational testing at each level of test taker ability.

The 2PL IRT formulation for the IIF is:

$$\mathrm{I}(\theta) = \mathrm{a}_{\mathrm{n}}^{2}\mathrm{P}_{\mathrm{n}}(\theta)\mathrm{Q}_{n}(\theta) \qquad (6)$$

where P_{n} is the probability of a correct response to item n and Q_{n} is the probability of an incorrect response to item n. Both probabilities are calculated at ability level θ where a test taker is administered the n^{th} item that was calibrated under the 2PL IRT model. As

can be seen from Equation 6, the magnitude of the discrimination parameter a_n is quite influential when calculating item information for the 2PL IRT model.

The 2PL IRT IIF curve is bell shaped and the vertical axis represents the amount of item information associated with an ability level. An example 2PL IRT IIF curve is shown in Figure 3.

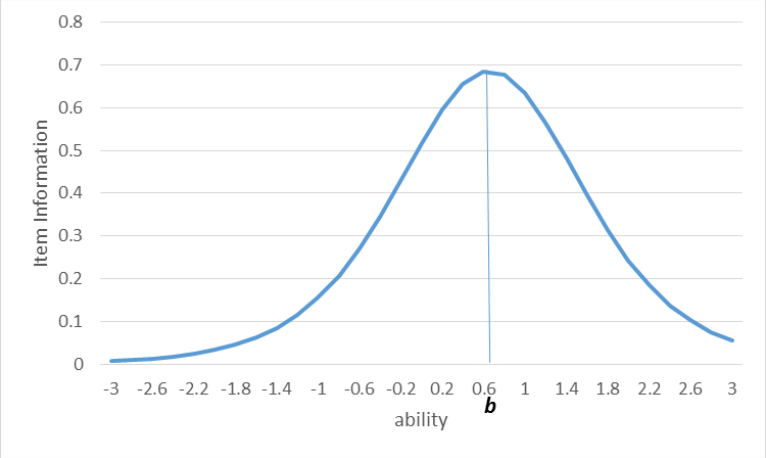

Figure 3. Graph of 2PL IRT IIF

Since the IIF maximum is at the item difficulty b location, items with difficulty levels close to the ability level of the test taker provide the most item information under the 2PL IRT model. It follows that when an MST is constructed under 2PL IRT, test developers may purposely select items that are highly discriminating within the neighborhood of the desired module difficulty.

Because the IIF varies along the ability continuum, an item may measure the ability at some levels along the continuum more precisely than at other levels. The formulation for the standard error (SE) of the IRT ability estimate $\hat{\theta}$, which is an

indication of how imprecisely the test measures at a particular ability level, is inversely related to item information and is generally formulated as:

$$\mathrm{SE}(\hat{\theta}) = \frac{1}{\sqrt{\mathrm{I}(\hat{\theta})}} \qquad (7)$$

More available information at a particular level of $\hat{\theta}$ implies that there is less error in the associated IRT ability estimate. In other words, the amount of error associated with an ability level depends on the amount of item information available at that level. Hence the IIF is also used in practice to pre-rank items for inclusion into MST modules not only according to item difficulty, but also according to how precisely the item measures within a desired range of ability.

As with a test item, a test that measures well within a particular neighborhood of ability has high test information within that neighborhood. The relationship between test taker ability levels and test information is shown in Figure 4.

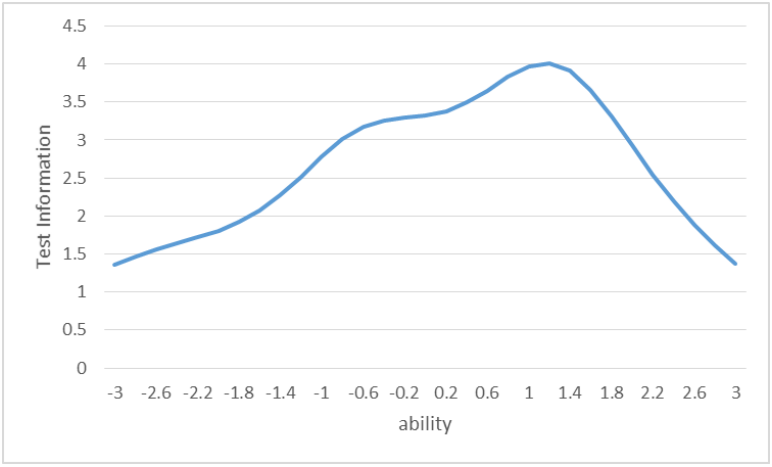

Figure 4. Graph of 2PL IRT TIF

The values for the test information function (TIF) are calculated by summing the IIFs for all test items. Its general formulation is as follows:

$$TIF(\hat{\theta}) = \sum I(\hat{\theta}) \qquad (8)$$

A module-level information function (MIF; Luecht, 2014) can be generated by summing the IIFs for all the items within a module. TIFs and MIFs provide an indication of the measurement accuracy along the ability continuum for the test and module respectively and not just for a single item as is the case with an IIF.

Prior knowledge of item parameter values is an important consideration in MST component pre-construction. Thus, limitations in the range of item difficulty or discrimination parameter values available in an MST item bank is an influential constraint to MST test form construction (Luecht, 2014; Zheng, Wang, Culbertson, & Chang, 2014). Van der Linden (2005), Luecht (2014), and Veldkamp (2014) recommend designing an MST item bank blueprint over simply replicating characteristics of an existing item bank. The purpose of the item blueprint is to pre-specify the characteristics of items that are needed to meet the MST design specifications.

Van der Linden and Glas (2000) introduced the idea of an item bank blueprint design approach where the bank is assembled based on test specification needs. Veldkamp (2014) extended this approach to MST item bank blueprint design. He suggested an approach that begins with determining the quantity of MST items needed for the entire item bank by classifying needed items by their content, statistical, and logical attributes. Veldkamp (2014) also suggests that since MST systems have several interrelated components, specifying how many distinct MST components are needed should also be considered in the MST item bank blueprint design. Assuming availability of an item bank with the range and depth of item attributes needed to support the MST design, the next section presents an overview of MST component construction.

2.3.1 Construction of MST Components

Luecht and Nungester (1998) and Zenisky and Hambleton (2014) present descriptions of MST components only summarized briefly here. MST components include modules, test stages, routing rules, pathways, and panels. Modules are a set of items that have been grouped based on pre-determined content and statistical specifications and they are scored and administered to test takers as a unit. When the set of module items is related to a common stimulus, the modules are referred to as testlets (Wainer & Kiely, 1987). More than one testlet can be included within a module. The items within a module can be either statistically independent or conditionally dependent. Under IRT, *m*odules can be constructed prior to operational testing or assembled algorithmically during administration of an MST (Han & Guo, 2014).

Once constructed, one or more modules can be assigned to each test stage of an MST. Most MSTs have been designed to have two to four test stages (Yan, von Davier, & Lewis, 2014). A minimum of three test stages has been preferred since it permits a recovery stage for test takers by allowing their ability estimates to be corrected if misrouting occurs (Zenisky & Hambleton, 2014). Research also shows that a maximum of four modules within a test stage and a three-stage test is sufficient for most MST applications (Armstrong, Jones, Koppel, & Pashley, 2004).

MST routing rules are followed to route test takers at each point of test adaptation to the appropriate next test stage module. The collection of modules traversed by a test taker through to test completion is called an MST pathway. Numerous pathways may exist through an MST panel based on the panel design imposed. A single MST panel contains its own pathways, routing rule, test stages, and modules. Since administration of

only a single MST panel may overexpose some items from the item bank and underexpose others, in practice multiple approximately parallel MST panels are constructed that can be assigned to different test takers on the same testing occasion (Yan, von Davier, & Lewis, 2014).

MST pathways, or test forms, can be assembled by using an automated test assembly (ATA) software or using a heuristic approach. Many ATA programs utilize a top-down or bottom-up approach (Luecht & Nungester, 1998) to generate test forms. Under IRT, both approaches assume that statistical specifications have been pre-determined at the level of measurement accuracy desired. With the top-down approach, test specifications based on target TIFs exist at the test form level. For each test form, items selected from the item bank for assignment to modules are such that the sum of all item IIFs comes sufficiently close to the target TIF.

With the bottom-up approach, target MIFs representing module-level information specifications are used. Items can be selected for inclusion into a module that collectively meet the MIF specifications. The top-down approach offers greater control for test developers to implement MST panel-level specifications (Zheng, Wang, Culbertson, & Chang, 2014), but the bottom-up approach is implemented more often in practice (Luecht, 2014; Zheng, Wang, Culbertson, & Chang, 2014). The bottom-up approach also eases implementation of other desired module-level specifications including content, average item difficulty, or constraints in the range of item difficulty for a specific module.

How best to forge a collective balance and integration of numerous statistical and non-statistical MST targets with large item banks and quality control goals (Luecht,

2014) are important test form construction considerations. Thus in practice, the recommendation for large-scale MST systems is to assemble MST test forms using an ATA system (Luecht, 2014). In an ATA, all MST test form assembly occurs simultaneously during a single optimization procedure (Zheng, Wang, Culbertson, & Chang, 2014).

This dissertation uses a heuristic approach to MST component assembly as an alternative to using an ATA. Zheng, Wang, Culbertson, and Chang (2014) describes the heuristic approach to MST assembly as breaking the test assembly process down to a sequential optimization problem by adding one item at a time to each module while also attempting to balance module quality across all test forms. Use of a heuristic approach to test form assembly and module construction for this dissertation permits the flexibility needed to freely impose and vary the conditions that are under study. Chapter 3 presents step-by-step the heuristic approach taken to MST test form assembly in this dissertation.

2.3.2 MST Panel Design

There are numerous ways to assemble MST components to form an MST panel design that aligns with the purpose of the test. For instance, Figure 5 presents an example of a 1-2-2 MST panel design. This design is often used in classification testing where test takers are to be sorted into one of two possible categories. The shorthand numeric notation describing the MST panel presents the count of how many modules have been assigned to each test stage. The 1-2-2 MST panel design notation is interpreted to mean that one module has been assigned to the first stage and two modules have been assigned to the second and third stages.

Figure 5 also shows that the 1-2-2 MST panel contains multiple overlapping test forms or pathways. For example, module B is a component administered to test takers traversing both the ABC and ABE pathways.

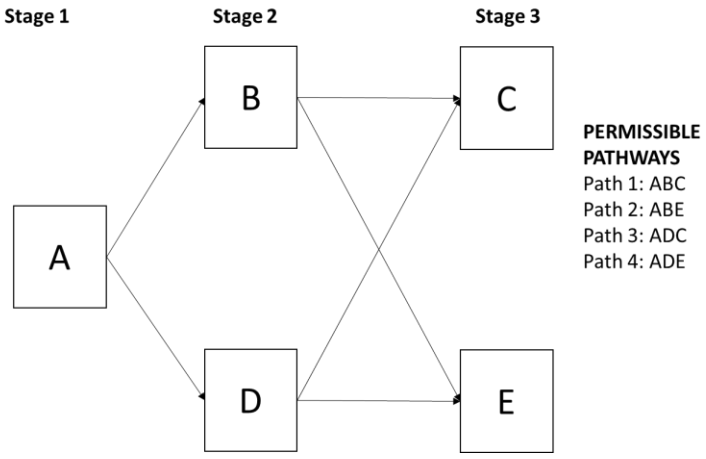

Figure 5. 1-2-2 MST Panel design

2.3.3 MST Routing

MST panel design and routing are interrelated since the accessible, or permissible, routes that test takers traverse are determined by the routing rule imposed. Routing rules can be set such that all or only selected pathways through an MST panel are permissible for use by some or all test takers. The arrowed lines in Figure 2.3 illustrate that there are four permissible pathways, or test forms, to route test takers through the MST panel.

The test difficulty adaptation, which customizes module selection within a test stage given the module-level ability estimate of a test taker, is controlled by the routing rule and associated module selection methods (Dallas, 2014). Research has shown that MST misrouting may impact classification decision accuracy and consistency as well as bias the standard errors of test taker final ability estimates (Zenisky & Hambleton, 2014).

Hence, the degree of efficiency at which MST routing operates is a test fairness consideration.

Since it is known that IPD exhibits in MST item banks (Wei, 2013), it would be of interest to test developers to understand to what extent or type and at what level or magnitude cumulative IPD may impact MST misrouting rates. Although no study has examined cumulative IPD effects on the performance of testlet-based MST routing operations, this section presents a brief overview of several study findings that have compared performance of the two most popular MST routing module selection methods, the AMI and DPI methods.

Kim, Chung, Park, and Dodd (2013) compared the performance of AMI and DPI MST routing module selection methods under four different MST panel designs for a classification test. The item response data was modeled using a polytomous IRT model. Results showed that the accuracy of classification decisions were similar under all routing methods where the test length was the same. This study did not report on final ability estimation differences.

Dallas (2014) noted that very little research has been performed which directly addresses the impact of a particular approach to MST routing on IRT final ability estimates. However, of the studies that have been performed under IRT, findings have been mixed. For instance, Dallas (2014) investigated the accuracy of IRT ability estimates in a testlet-based MST system after varying the panel design and the routing scoring and module selection methods under a 3PL IRT model. MST routing study conditions included both NC and dynamic routing rules under both DPI and AMI module-level selection methods. He found that more precise ability estimates were

obtained under the AMI routing module selection method that with DPI. This study also showed that ability estimate accuracy differences (differences between when the NC and dynamic routing rules were used) were negligible. Although testlets were used as item types in the Dallas (2014) study, the within-testlet items were assumed statistically independent and were modeled under IRT.

Under the 2PL IRT model, Kim and Moses (2014) investigate the performance of an MST system consisting of a 1-3 panel design. An NC routing rule was imposed with a DPI routing module selection method. To ensure representation of ability levels across the ability continuum, test takers with different ability levels were generated from the interval of -3 to 3 at increments of .15 logits. All simulated test takers were administered all MST modules along all possible pathways. This was done to examine how the final scores of different ability test takers may be impacted if they were administered modules not suitable for their ability level. Kim and Moses found that even for test takers with ability levels close to the routing decision cut points, final ability estimates were not significantly impacted if test takers were incorrectly routed to subsequent modules. Kim and Moses hinted that a more dramatic impact on ability estimates may occur if very high or low test takers were administered an incorrect module. They further suggested that perhaps in practice, the impact of routing error could be substantial if the quality of MST items used diminished.

Keng (2008) investigated item diminished quality concern in a testlet-based MST performance study. The presence of LID was assumed when comparing the performances of an item level CAT which adapted within the testlet, a testlet-based CAT which adapted at the testlet level, and a testlet-based MST with a 1-3-3 panel design. Keng imposed the

simulated LID influence condition as a testlet effect modeled under the 3PL TRT model. Although more precise ability estimation are realized under the item level adaptive testlet-based CAT, Keng found that the difference in accuracy observed over that obtained for the testlet-based CAT and the testlet-based MST was not of practical significance. A dynamic routing rule with AMI module-level selection was imposed for all study conditions. Thus, MST routing module selection and scoring rules were not variables manipulated in the Keng (2008) study. However, this study demonstrated the feasibility of using TRT in research to model LID when administering testlets in an MST system.

Kim and Moses (2014) and Kim, Moses, and Yoo (2015), under a 1-3 MST panel design with response data modeled under the 2PL IRT model, investigated the influence of changes in module-level statistics on routing efficiency and final ability estimation accuracy. They considered the impact of small and large differences in average item difficulty across MST modules. The performance of MST routing under an NC and a dynamic routing rule with a DPI module selection method was examined. After examining over 1000 MST panels used in practice, they imposed in their study the small, moderate, and large average item difficulty conditions presented in Table 1.

Table 1

Module Difficulty by Difficulty Level and Module Difference Condition

Module Difficulty level	Small	Moderate	Large
Easy	-.75	-1	-1.25
Moderate	0	0	0
Hard	.75	1	1.25

They found the impact of the differing module difficulty conditions on MST misrouting was minimal. Numerous other MST studies investigating routing performance

have imposed the moderate module difference condition presented in Table 1 as a fixed study condition (Dallas, 2014; Kim, Chung, Park, & Dodd, 2013; Jodoin, 2003; Hambleton & Xing, 2006). Findings were that performances were similar across routing rules and routing module selection methods imposed.

Evaluating the performance of MST routing procedures under the combined influences of cumulative IPD patterns and LID given differing module difficulty difference conditions were not considerations of the aforementioned MST routing studies. This dissertation seeks to add to the current MST routing research by investigating effects of IPD amplification and cancellation at the testlet-level under differing module difference conditions imposed within the same stage. Hence, given the findings of previous MST research, this dissertation investigates the performance of the AMI routing module selection method under an NC routing rule under conditions of varying patterns of IPD amplification or cancellation when LID is assumed and when it is not.

2.3.4 MST Final Ability Estimates

A different scoring technique than that used to calculate module-level estimates of ability can be used to calculate final ability estimatesfor the MST test form (Weissman, 2014). For example, an MST routing rule may be based on NC scoring, but final ability estimates could be calculated using IRT-based ability estimation methods. If standard IRT models are used to calculate final ability estimates, then the properties of item parameter invariance and local item independence are assumed to hold.

Final ability estimation under IRT for an MST test form can generally be conducted similarly to how scoring is performed in a linear test (Weissmann, 2014).

Thus, the Maximum Likelihood Estimation (MLE) and IRT Bayesian ability estimation techniques popularly used in linear test scoring can also be used for scoring MST test forms (Haberman & von Davier, 2014). Assuming that IRT item parameters are known, MLE based techniques used for calculating IRT ability estimates result in a likelihood function which shows the probability of an item response pattern given the ability level of the test taker (de Ayala, 2009). The maximum of the likelihood function is the estimate location of the test taker ability level assumed most likely to have produced the item response pattern observed. For the resulting joint probability to be a good representation of how likely the response pattern was generated at the ability level identified, the IRT model used should be a good fit to the item response data where items are assumed locally independent.

Bayesian IRT ability estimation methods make inferences on unknown parameters using prior beliefs and information contained in an observed data set. Therefore, when implementing Bayesian ability estimation methods, prior distributions for ability parameters are specified and then multiplied by a likelihood function. Since none of the standard IRT models make assumptions regarding the characteristics of the test taker ability distribution, there is flexibility in determining which prior distribution to select when implementing a Bayesian approach to ability estimation. However, the standard normal distribution is the most widely used Bayesian IRT ability estimation prior in educational testing (de Ayala, 2009).

Research comparing the performance of MLE and Bayesian estimation methods show that Bayesian methods tend to produce more precise ability estimates than MLE.

This is because using the prior contributes additional information beyond that which is available from administration of the items alone (Bock & Mislevy, 1982).

With Bayesian estimation, the prior distribution is multiplied by the likelihood function to obtain the posterior distribution resulting in an updated ability estimate distribution. When Bayesian estimation is used in an MST system to calculate module-level ability estimates, the updated posterior distribution is calculated prior to each test adaptation. Resulting posterior distributions become the priors used during calculations of the module-level ability estimate at the next adaptation point. Where the IRT ability estimate is located in the posterior distribution is what distinguishes the different Bayesian methods. The maximum a posteriori (Bock & Aitken, 1981) uses the mode of the posterior distribution as the ability estimate location while the expected a posteriori (Bock & Aitken, 1981) uses the mean of the posterior distribution.

Markov Chain Monte Carlo (MCMC) procedures have also been successfully used to estimate IRT ability estimates (Patz & Junker, 1999). Under MCMC, observations are sampled with respect to target parameter posterior distributions in order to determine distribution characteristics, such as the mean or variance. The target parameter posterior distribution is approximated by drawing samples from a distribution assumed to be similar in shape. Then using an iterative process, each previously generated posterior distribution is continually updated as new item response data becomes available. The iterative updating process continues until successive updated posterior distributions are sufficiently close in shape.

The steps used in the MCMC parameter estimation process can be generally summarized as (Wainer, Bradlow, & Wang, 2007):

- Step 1: Collect response data by administering items to a calibration sample.
- Step 2: Specify priors that are believed to be close in shape to the target item, ability, and testlet effect posterior distributions.
- Step 3: Select initial starting value estimates for all item, ability, and testlet effect parameters.
- Step 4: From the starting values, select a subset of the parameters such as not to include the parameter of interest to be estimated first.
- Step 5: Draw from the conditional distributions for the initial parameter to be estimated. The draw is conditional on the initial values of all other remaining parameters and the available response data.
- Step 6: Continue to draw in turn from the conditional distributions of each parameter to be estimated, where each draw continues to be conditional on the previous values of all other remaining parameters and the response data.
- Step 7: Continue the draws for a number of complete cycles (referred to as the burn-in).
- Step 8: After the burn-in period, the draws are considered to be from the target parameter distributions of interest.
- Step 9: The remaining draws can be used as a sample from which statistics, such as the mean or variance of the target parameter distributions, can be calculated and used as estimates.

When testlets are administered in an MST and LID exists, modeling and/or scoring the testlet response data under standard IRT model usage assumptions may not be the most feasible choice. Although testlets are not the only item type for which LID has been

observed, it is the item type most investigated for LID in the literature due to its frequency in use on tests (Min & He, 2014). When testlets are used as the testing unit and LID due to item interactions are assumed to exist, detection can be carried out using TRT models. Bradlow, Wainer, and Wang (1999) showed that when item dependencies exist in the testlet response data, the 2PL TRT was able to recover item parameters better than the 2PL IRT model.

Under the 2PL TRT model, the variance of the testlet effect is assumed to be constant. A testlet effect variance within the neighborhood of 0.25 is an indicator that, on average, low levels of LID is detected, while a testlet effect variance of 1.0 or greater indicates, on average, high levels of LID (Wainer, Bradlow, & Wang, 2007). Bradlow, Wainer, and Wang (1999) showed that when the testlet effect is ignored by modeling testlet response data with the 2PL IRT model, the size of item parameter estimation bias was a function of the magnitude of the testlet variance. That is, ignoring the testlet effect could result in item parameter estimation instability. Using 2PL TRT, as opposed to 2PL IRT, to model testlet response data may help diminish item parameter estimate bias by accounting for the testlet effect. However, the model change alone does not account for unexpected item parameter variance that may result due to un-modeled influences that may exist in the response data structure.

2.3.5 TRT Parameter Estimation

Since local item independence under 2PL TRT is not assumed among the items within the testlet, the popular MLE-based or IRT Bayesian approaches for item and ability estimation previously described for standard IRT models are not used for parameter estimation under TRT. TRT parameter estimates can however be obtained by

drawing samples from target marginal posterior distributions using a form of MCMC procedure (Wainer, Bradlow, & Wang, 2007). In fact, Wainer, Bradlow, and Wang (2007) showed through simulation studies that the 2PL TRT model implemented under the MCMC methods was able to account for item dependencies in testlet response data. Thus, MCMC is an approach that is regularly used in practice to estimate TRT item, testlet, and ability parameters.

The set of 2PL TRT priors on the ability θ, discrimination a, difficulty b, and testlet effect γ distributions include:

$$\theta_i \sim N(0,1.0)$$

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

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

$$\gamma_{id(n)} \sim N(0, \sigma_\gamma^2)$$

$N\ (\mu, \sigma^2)$ denotes a normal distribution with mean of μ and variance σ^2. The hyperpriors for the distributions of means and variances are non-informative normal distributions and inverse-gamma distributions respectively.

Although the 2PL TRT model is the only TRT model under investigation in this dissertation, there have been a number of other TRT models proposed. For instance, normal ogive versions of the 2PL TRT model have been devised (Li, Bolt, & Fu, 2006; Tao, Xu, Shi, & Jiao, 2013). The 2PL TRT model has also been extended to a 3PL TRT model (Wainer, Bradlow, & Du, 2000) which includes a guessing parameter and allows variation in the random effects to occur across testlets. A mixture model (Wang, Bradlow, & Wainer, 2002) and a covariates model (Wainer, Bradlow, & Wang, 2007)

have also been proposed. Wainer, Bradlow, and Wang (2007) provides a detailed discussion regarding these model extension formulations and their usage assumptions.

2.4 Summary

Standard IRT measurement models are widely used in the construction, operation, and scoring of testlet-based MST systems (Yan, von Davier, & Lewis, 2014). This is the case even as testlets are commonly used as the testing units in MST systems. If LID due to item interactions exist in testlet response data, standard IRT models do not include parameters to detect and account for this potential influence on ability estimates. Under these testing conditions, the accuracy in IRT ability estimates may be questionable.

One approach that has been devised to account for this type of LID when calculating final ability estimates calls for the use of TRT models in place of standard IRT models. However, the gain in the degree of accuracy in TRT ability estimates may potentially be impacted by amplification and cancellation of IPD to the testlet level.

Many currently available IPD detection techniques (Bock, Murkai, and Pfeiffenberger, 1988; DeMars, 2004; Donoghue & Isham, 1998; Veerkamp & Glas, 2000) are designed for item-level IPD detection and therefore may miss detection of cumulated insignificant magnitudes of IPD that has amplified to the testlet level. Therefore, it is important to investigate the extent of how vulnerable testlet-based MST routing procedures, final ability estimates, and, by extension, classification decisions are in the presence of these effects when LID is assumed and when it is not. More specifically, investigations should be performed not only under TRT model usage assumptions but also under standard IRT model usage assumptions since standard IRT

models are the most widely used measurement models in MST large-scale educational testing.

Thus far, research has primarily examined the impact of IPD on final ability estimates in IRT-based systems. These investigations have mainly been performed under linear testing situations that do not depend operationally on the efficient execution of routing procedures as does an MST. Thus, the generalizability of IPD linear test study findings to testlet-based MST systems may be limited. To date, no study has examined the influence of combined cumulative IPD and LID influences on testlet-based MST final ability estimation, routing system performance, and classification accuracy. This dissertation seeks to fill this gap by evaluating the impact of IPD amplification and cancellation at the testlet level on the performance of a testlet-based MST system.

Chapter 3: Methodology

3.1 Introduction

The conditions under study were simulated to represent the administration of a testlet-based MST over two testing occasions. It is not unusual in an IPD study to simulate and examine the impact of IPD effects at insignificant magnitudes across two test administrations (Rupp & Zumbo, 2006; Wainer & Thissen, 1987; Wells, Subkoviak & Serlin, 2002; Witt, Stahl, Bergstrom, & Muckle, 2003). All IPD magnitudes are simulated at insignificant levels which are generally shifts in item parameter values of 0.5 logits or less (Han & Guo, 2005). Wells, Subkoviak, and Serlin (2002) under the 2PL IRT model simulated IPD at similar magnitudes.

In this study, the first MST administration, assumed to occur immediately after item calibration, was delivered under non-IPD conditions when LID was assumed and when it was not. In this study, these were referred to as the baseline conditions. The second MST administration occurred some time later where insignificant magnitudes of IPD arose as patterns of amplification and cancellation but had been ignored by test developers. For the MST administrations with IPD, MST ability estimates, routing error, and misclassification rates were calculated and compared. The impact of IPD magnitude, directions (amplification or cancellation), LID magnitudes, and the location of IPD on these outcome measures were evaluated.

3.2 MST true item parameter generation

Two sets of estimated testlet parameters were used during the construction of MST system components. The MST testlet parameter generation procedure was:

- Step 1: Using R software (R Core Team, 2013), specify the 2PL TRT model to generate 6 sets of 7 testlet parameters by combining two levels of module difficulty difference with the three levels of the testlet effect factor.
- Step 2: Using R, randomly draw 3000 simulated test takers from a standard normal distribution to respond to items in all six sets of seven testlets generated in Step 1
- Step 3: Using the response data generated in Step 2, calibrate the 6 sets of 7 testlets under 2PL IRT and 2PL TRT using the MCMC item parameter estimation method in R. The number of iterations run for MCMC estimation was set at 10,000 with a burn-in of 5000. The iteration and burn-in values were determined after conducting pilot calibration runs that resulted in parameter estimates with a corresponding Gelman-Rubin MCMC convergence statistic within a $\pm.001$ neighborhood of 1.0. The degree of convergence of a random Markov Chain can be estimated using this statistic; values close to 1.0 are considered adequate and values much greater than 1 indicate inadequate convergence (Brooks & Gelman, 1998).
- Step 4: To create 10 panels for each MST, repeat Step 3 10 times. Thus, 60 sets of 7 testlets are generated based on 2PL IRT and 2PL TRT respectively.
- Step 5: Using the 120 (60 x 2) sets of 7 calibrated testlets generated in Step 4, construct 60 MST panels under 2PL IRT and 60 MST panels under 2PL TRT (see section 3.5 MST panel assembly for details on how MST panels are constructed).

3.3 Study Design

3.3.1 Fixed Factors

Several factors were fixed as follows. Four factors held constant in this study were test length, MST panel design, MST module selection method, and NC scoring.

3.3.1.1 Test Length

Research has shown that an MST with a test length of 30 items or more typically lead to ability estimates at accuracy levels comparable to a CAT (Luecht, Nungester, & Hadadi, 1996). Furthermore, the number of within-testlet items usually ranges from 4 to 12 items (Wainer, Bradlow, & Wang, 2007). To that end, the test length of 30 items with 10 items per testlet selected for use in this study fell within these acceptable ranges. Moreover, since Hambleton and Patsula (1999) found that varying the number of items per MST module has a negligible effect on ability estimate accuracy, the number of items per testlet in this study was fixed at 10.

3.3.1.2 MST Panel Design

There are numerous MST panel designs that have been used in MST studies. The 1-3-3 MST panel design is frequently used (**Zenisky** & Hambleton, 2014) and was the design selected for use in this study. The 1-3-3 MST panel design contained three module difficulty levels: easy, moderate, and hard as illustrated in Figure 6.

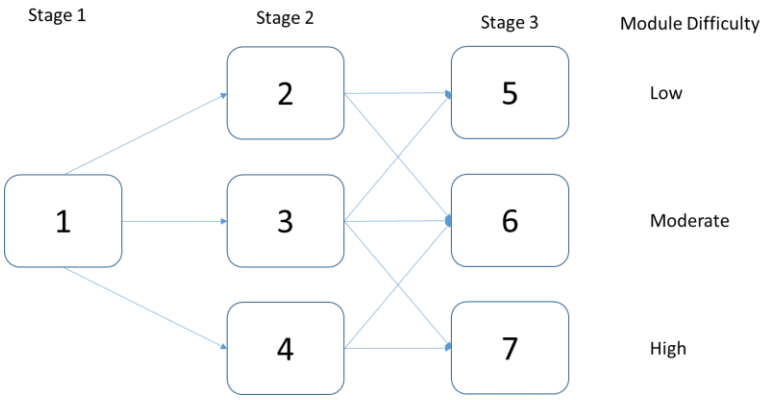

Figure 6. 1-3-3 MST panel design

Figure 6 shows the 1-3-3 MST panel as consisting of seven modules and three test stages. Test stage 1 contained one module of moderate difficulty and test stage 2 and 3 each contained one easy, moderate, and hard module. In this study, only one testlet was assigned to each MST module.

During the simulation, each test taker was first administered the test stage 1 module. Then according to the MST NC routing rule, the test taker was routed to and administered a stage 2 module followed by being routed to and administered a stage 3 module. The pointed arrows in Figure 6 show the MST permissible pathways through which test takers may have been routed in this study. For example, modules 1, 3, and 6 made up one permissible pathway. The seven permissible pathways through the MST were pathways 125, 126, 135, 136, 137,146, and 147. It is not unusual in MST research and practice to only permit routing to adjacent modules to occur (Keng, 2008; Kim, Chung, Park, & Dodd, 2013)

Table 2 shows the number of testlet items per panel that were generated to satisfy MST 1-3-3 panel design study specifications.

Table 2

MST Item Bank Blueprint for a Single 1-3-3 MST Panel

Module Difficulty Level	Number of testlets per module	Number of testlets per panel	Testlet length per module	Number of items per module with various difficulty levels needed to construct a panel
Easy	1	2	10	20
Moderate	1	3	10	30
Hard	1	2	10	20
Totals	3	7	30	70

3.3.1.3 MST Module Selection Method

The AMI routing module selection method was used in this study. This method was selected since in MST routing performance studies, the AMI method has been shown to perform slightly better than alternative routing methods (Dallas, 2014). Under the AMI routing module selection method, all routing decision points were determined prior to testing by first finding the MIF intersections of adjacent modules within each panel. This approach to AMI routing has been executed in previous MST studies (Dallas, 2014; Kim, Chung, Park, & Dodd, 2013) and has been shown to perform better than using test level TIFs to determine routing decision points (Kim, Chung, Park, & Dodd, 2013).

3.3.1.4 NC Scoring Method

Studies have shown NC scoring to result in acceptable MST final ability estimation accuracy and decision classification accuracy (Dallas, 2014; Kim, Chung, Park, & Dodd, 2013; Kim & Moses, 2014; Kim, Moses, & Yoo, 2015). NC scoring is also the preferred scoring procedure used in practice for MST panels (Weissman, 2014). For these reasons, an NC scoring rule was implemented in this study. The NC scoring procedure was executed in two stages with one stage occurring prior to testing and the

other during MST administration. The steps completed at each NC scoring stage are listed below.

3.3.1.5 The NC scoring procedure

- Stage 1: Prior to test administration
    - Step 1: Obtain two sets of model parameter estimates by calibrating under 2PL IRT and 2PL TRT.
    - Step 2: Construct two sets of MST panels using the two sets of calibrated testlets from Step 1 (see section 3.5 MST panel assembly for details on how MST panels are constructed).
    - Step 3: Under 2PL IRT and 2PL TRT respectively, identify θ routing decision points for test stage 1 and test stage 2 for each MST panel constructed in Step 2.
    - Step 4: For each MST panel, use module-level TCCs to convert all θ routing decision points identified in Step 3 to NC routing decision points.
    - Step 5: For each MST panel, construct an NC lookup table using panel-specific NC routing decision points from Step 4.
- Stage 2: During test administration
    - Step 1: After administration of all items assigned to a particular module, calculate the module NC score for each test taker.
    - Step 2: For each test taker, calculate the cumulative NC score by summing all previous and current module NC scores.

- Step 3: Use the current cumulative NC score and the panel NC lookup table to determine which next test stage module within the panel to route the test taker to.
- Step 4: Repeat Step 1 to Step 3 until the final test stage module has been administered.

3.3.2 Manipulated Factors

Table 3 lists the factors that are manipulated in this study.

Table 3

Study Condition Values by Manipulated Factor and Factor Levels

Manipulated Factor	Level 1	Level 2	Level 3	Total Levels
Testlet Effect Variance	0	.25	1	3
IPD item parameter	a	b	a and b	3
Magnitude of IPD	0	.3	.5	3
Direction of IPD	Amplification (Unidirectional)	Cancellation (Bidirectional)	-	2
Module difference	Small (0.5)	Large (1.0)	-	2
IPD test stage	Stage 1	Stage 2	Stage 3	3
Percent of IPD affected testlet items	13%	20%	-	2

3.3.2.1 Testlet Effect Variance

In this study, 2PL TRT was assumed to be the true model. The LID conditions simulated in the study were imposed by varying the magnitude of the testlet effect variance. Three levels of testlet effect variance were simulated at 0.0, 0.25, and 1.0. A testlet effect variance magnitude of 0.0 implies that no LID exists, whereas the magnitudes of 0.25 and 1.0 were considered small and large testlet effects respectively. The testlet effect variance magnitudes under consideration in this study have also been used in prior research (Eckes,2013; Lu, 2010; van der Linden, Glas, Wainer, & Bradlow, 2000; Wang, Bradlow, & Wainer, 2002; Zhang, 2010).

3.3.2.2 IPD affected item parameter

Studies have shown that item characteristics can be affected by IPD differently. For instance, according to Han and Guo (2005), long before an item becomes more or less discriminating, the item becomes easier or harder for the test taker. Moreover, Bock, Murkai, and Pfeiffenberger (1988) suggested that IPD more strongly affects the difficulty parameter, but also the discrimination parameter to a lesser extent.

Item parameter characteristics also may impact the accuracy of ability estimates and/or statistics differently. For instance, discrimination parameter values are quite influential when calculating item information used to generate MIFs. MIFs play a key role under the AMI routing module selection method when routing decision points are identified. When discrimination parameter drift is present, it is possible that the expected shape of a MIF may not be accurate for all test takers at all stages of the MST. Moreover, according to Kim and Plake (1993), some MST module item difficulty levels have been shown to have greater impact on final ability estimate accuracy than others. Given the potential differing impact of shifting item parameters on MST ability estimate accuracy and routing accuracy, this study simulated IPD in the difficulty and discrimination separately, as well as simultaneously in both parameters.

3.3.2.3 Magnitude of IPD

Once items selected to be affected by IPD were determined, either a 0.3 or 0.5 logit shift in item parameter values were imposed as b-drift, a-drift, or ab-drift. The IPD magnitudes imposed in this study were considered insignificant and were magnitudes used in previous IPD studies (Guo & Han, 2005; Wei, 2013; Wells, Subkoviak, & Serlin, 2002). The patterns of IPD simulated in this study are presented in the next section.

3.3.2.4 Direction of IPD

Cumulative patterns of IPD were simulated in item parameters as either amplification (unidirectional) or cancellation (bidirectional) to the testlet level. The amplification IPD was simulated as always positive or increasing. Under cancellation conditions, if both item parameters for the same item were to be affected, the direction of the IPD in both parameters was unidirectional. This is done so that the direction of IPD imposed on one parameter did not cancel out the effect of the IPD imposed on the remaining parameter (DeMars, 2004). Table 4 shows the direction and magnitude of the IPD patterns that were imposed in this study.

Table 4

IPD Patterns Imposed by Item Parameter, Magnitude, and Direction

	IPD magnitude imposed	
Direction of IPD	Difficulty parameter	Discrimination parameter
Baseline (No IPD)		
	0	0
Amplification (unidirectional)		
	0.3	0
	0.5	0
	0.3	0.3
	0.5	0.5
	0	0.3
	0	0.5
Cancellation (bidirectional)		
	±.3	0
	±.5	0
	±.3	±.3
	±.5	±.5
	0	±.3
	0	±.5

3.3.2.5 Module difficulty difference conditions

The two module difficulty difference conditions simulated in this study are presented in Table 5.

Table 5

Average Module Difficulty by Module Difficulty Level and Module Difference Factor Level

	Module Difference	
Module Difficulty level	Small	Large
Easy	-.5	-1
Moderate	0	0
Hard	.5	1

Study conditions were investigated under each of the two module difficulty difference conditions (small and large) presented in Table 5. The differences in module average item difficulty were simulated across modules within the same test stage. The module difficulty difference conditions used in this study have been used and compared in previous MST routing investigative studies (Kim & Moses, 2014; Kim, Moses, & Yoo, 2015).

3.3.2.6 MST stage with IPD

One MST test stage per pathway administration was selected to contain the affected IPD amplification or cancellation testlets. That is, under all non-baseline study conditions, each of the three test stages did at one time contain an IPD amplification- or cancellation-affected testlet in all modules assigned to that test stage. Once an MST test stage had been selected, either 13% or 20% of the total pathway items were randomly selected within the targeted testlet to be affected by IPD. If bidirectional IPD was to be applied, shifts in half of the affected within-testlet item parameters were simulated to increase, while the remaining half were simulated to decrease. During MST

administration, all IPD and non-IPD items within the affected testlet were administered to all test takers routed to the testlet.

3.3.2.7 Percent of IPD affected items

Wells, Subkoviak, and Serlin (2002) found that at least 20% of test items affected with insignificant magnitudes of IPD were needed to impact final ability estimates. For comparative purposed, tests with 13% (4 items per testlet) and 20% (6 items per testlet) of IPD-affected items were simulated.

3.4 Data Generation

3.4.1 Testlet effect parameters

Testlet effect parameters were generated from a normal distribution with a mean of zero and variance set equal to 0.0, 0.25, or 1.0. For example, large testlet effects were generated from $N(0, 1)$.

3.4.2 Difficulty parameters

The true difficulty parameters for each testlet were simulated from normal distributions where the means were set equal to the desired average item difficulty level for the module to which the testlet was to be assigned. For example, item difficulty parameters assigned to easy modules (constructed based on the small module difficulty difference specification) were drawn from a normal distribution with a mean of -0.5 and a standard deviation of 0.25 (variance of 0.0625). Pre-simulation investigations showed that setting the standard deviation at 0.25 generated item difficulty parameters that reasonably met the MST module design specifications. Following previous studies (Dallas, 2014; Hambleton & Xing, 2006, Jodoin, 2003; Kim, Chung, Park, & Dodd, 2013; Kim, Moses, & Yoo, 2015), Table 6 presents the module difficulty levels, module

differences, and module difficulty distributions from which true difficulty parameters were generated for this study.

Table 6

Item Difficulty Distribution, With Mean and Variance, by Module Difficulty and Module Difference

Module difference	Module Item Difficulty level	Average item difficulty for module	Distribution
Small			
	Easy	-0.5	N(-0.5, .0625)
	Moderate	0	N(0, . 0625)
	Hard	0.5	N(0.5, . 0625)
Moderate			
	Easy	-1	N(-1, . 0625)
	Moderate	0	N(0, . 0625)
	Hard	1	N(1, . 0625)

3.4.3 Discrimination parameter

The true discrimination parameters were randomly drawn from a uniform distribution within the interval range of 0.4 to 1.5 logits. This was done to achieve a uniform balance in the range of available discrimination parameters so as to have an overabundance of neither high nor low discriminating items. Although desired, it is not always that highly discriminating items are available for use when assembling MST modules in practice. The range selected for the discrimination parameters has been used in previous MST studies (Lu, 2010; Xing & Hambleton, 2004).

3.4.4 Ability parameters

A calibration sample of 3000 test takers were drawn from a standard normal distribution. This calibration sample size and ability distribution have been used in a previous MST study (Lu, 2010).

To ensure a representative distribution of ability levels across the ability continuum, 1000 simulated test takers were generated at each θ point over the interval

from -3 to 3 logits at increments of 0.15. This approach has been taken in previous MST studies (Keng, 2008; Kim & Moses, 2014; Wei, 2013; Wollack, Sung, & Kang, 2006). Generating simulated test takers using θ abilities within a set interval and at fixed increments has also been done in previous MST studies (Kim & Moses, 2014; Kim, Moses, & Yoo, 2015; Lu, 2010). This approach to replicating simulated test takers permits the replication of study results to occur within the same simulation run. A total of 41,000 test takers were simulated.

3.5 MST Panel Assembly

Even when ATA panel assembly programs are used, if the item bank has not been optimized with items having the necessary attributes to meet MST design needs, it may be challenging for the ATA program to meet all targeted specifications for all ability levels (Jodion, Zenisky, & Hambleton, 2006). However, for the purposes of this investigation, a heuristic approach to panel assembly was implemented where item parameter difficulties were generated and the testlets were assembled to meet the MST design specifications presented in Table 2, Table 4, and Table 5. 120 panels were constructed by fully crossing the levels of three factors: (1) calibrated sets of item parameters (10 sets at 2 model levels), (2) module difficulty difference (2 levels), and (3) testlet effect variance (3 levels). All amplification and cancellation conditions were imposed during administration of the MST.

For each panel, the MST panel assembly procedure was:

- Step 1: Generate testlets assembled to meet the desired combination of module difficulty and testlet effect variance level for the panel.

- Step 2: For testlets meeting moderate module difficulty level specifications, calculate the MIF value at the ability level of zero.
- Step 3: Rank order the testlets from Step 2 to identify the most informative testlet at the ability level of zero.
- Step 4: Using Figure 6 as a guide, assign from Step 3 the most informative testlet to the test stage 1 module.
- Step 5: Using Figure 6 as a guide, assign the next informative testlet with moderate difficulty from Step 3 to the test stage 2 module and finally assign the least informative testlet with moderate difficulty to the test stage 3 module.
- Step 6: Rank order the easy and hard testlets using the same procedure presented in step 2 and step 3. Using Figure 6 as a guide, assign the most informative, easy, and hard testlets to the test stage 2 module and then assign the least informative, easy, and hard testlets to the test stage 3 module.
- Step 7: Once all panel modules have been assigned and in keeping with the AMI routing procedure, identify the two θ decision points located at the intersections of each adjacent MIF for test stage 1 and then for test stage 2
- Step 8: Use TCC curves, derived under 2PL TRT or 2PL IRT, to determine the true score for each θ decision point identified in Step 7. Use the true scores to calculate cumulative NC decision points for use as replacements for the θ decision points calculated in Step 7.
- Step 9: Using the NC scores from Step 8, construct a NC lookup table for the panel that maps cumulative NC scores to next test stage modules.

Implementing content control is one of the many benefits of pre-constructed MST panels (Yan, von Davier, & Lewis, 2014). That is, in practice, content control is an important consideration that ensures that, collectively, all selected items are inclusive of the intended subject content. There was no content control implemented in this study since it was assumed that content control is addressed during the MST panel pre-construction procedure. Content control has also not been implemented in other MST studies (Hambleton & Xing, 2006; Lu, 2010; Wei, 2013).

Finally, MST panels were randomly selected, at a panel exposure rate of .25, to be administered to test takers in this study. This was done so as not to intentionally introduce a source of systematic error. The panel exposure rate of .25 has been used in a prior MST study (Lu, 2010). However, in practice many other considerations are made when implementing exposure control in an MST system including item attrition estimates and how often it is expected that a particular MST module may be viewed by test takers (Luecht, 2014).

3.6 Evaluation

The 2PL IRT and 2PL TRT final ability estimates were calculated using MCMC estimation procedures. All final ability estimates were compared to the true ability parameter values in order to evaluate estimation errors in terms of bias, root mean squared error, the standard error, and correlation.

3.6.1 Bias

Bias, which accounts for the systematic over or under estimation of the true ability parameters, was calculated for each ability level using the following formulation:

$$Bias = \frac{\sum_{i=1}^{K}(true_i - estimate_i)}{K}, \quad (9)$$

K is equal to 1000 because at each θ point there are 1000 simulated test takers.

Due to the use of simulated test taker θ points generated at set increments along the ability continuum, a weighting procedure similar to the one used in Lu (2010) was used in this study to calculate a weighted bias. The *dnorm* function in R was used to calculate the bias weights associated with each of the 41 test taker θ points. Using all 41,000 (1000 simulated test takers at each θ point) item responses to approximate a standard normal distribution, the overall bias for the standard normal distribution with each simulation condition was calculated as follows:

$$Overall\ Bias = \frac{\sum Bias * weight}{\sum weight} \quad (10)$$

3.6.2 Root Mean Square Error

The root mean square error (RMSE) represents the overall accuracy of ability estimates and can be used to compare the estimated ability parameters to the true ability parameters. The general formulation used for calculating the RMSE for each ability level is as follows:

$$RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(true_i - estimate_i)^2}\,. \quad (11)$$

A similar weighting procedure, as is used to calculate the weighted bias above, was used to calculate weighted RMSE as well as the overall RMSE for each simulation study condition:

$$Overall\ RMSE = \sqrt{\frac{\sum (RMSE) * weight}{\sum weight}}, \quad (12)$$

3.6.3 Standard Error

The standard error in this study was the standard error of measurement (SEM). SEM is used to estimate the range of random error within which it is expected that the true value of the ability estimate falls. There are numerous applications of the SEM index in educational testing, most notably when calculating the upper and lower bounds of test score confidence intervals. It can be seen in Equation 13 and 14 below that both the 2PL IRT and 2PL TRT SEM formulations are derived using model specific item information functions which depend on item parameter estimate values.

Given that model specific item parameter estimates are used in the SEM formulation, the accuracy of confidence interval bounds depends indirectly on the accuracy of the item parameter estimates. In this study, the 2PL IRT and 2PL TRT model parameters were calibrated over response data simulated based on the 2PL TRT model. Under these estimation conditions, prior research suggests that the resulting 2PL IRT parameters may be biased. Hence, the evaluation of the SEM values resulting from the use of model parameters generated under these study conditions may provide some insight into estimation error, and consequently confidence interval bounds, of item parameter estimates.

The standard error of θ under 2PL IRT was calculated using all N pathway items as follows:

$$SE(\theta) = \frac{1}{\sqrt{\sum_{n=1}^{N} a_n^2 P_n(\theta)(1-P_n(\theta))}}, \quad (13)$$

Under 2PL TRT, to account for the presence of the testlet effect parameter when calculating the standard error, a procedure applied in Lu (2010) was also used in this

dissertation. Since the distributions for the testlet effect were assumed normal with a mean of 0.0 and with variances of 0.0, 0.25, or 1.0, the continuous normal distribution could be approximated using quadrature points and weights. The method proposed in Lu (2010)—of generating equally spaced quadrature points over the interval of ability—was used here. That is, 13 quadrature points were calculated over the interval -3 to 3. Then, the *dnorm* function in R was used to calculate the weights associated with each of the 13 quadrature points.

For each item n under 2PL TRT, item information within a testlet was calculated as:

$$I_n(\theta) = \sum_{k=1}^{15}\{a_n^2\left(\frac{e^{(a_n(\theta-b_n-P_k(\gamma_{id(n)}))}}{1+e^{(a_n(\theta-b_n-P_k(\gamma_{id(n)}))}}\right)\left(1-\frac{e^{(a_n(\theta-b_n-P_k(\gamma_{id(n)}))}}{1+e^{(a_n(\theta-b_n-P_k(\gamma_{id(n)}))}}\right)W(P_k(\gamma_{id(n)}))\}, \quad (14)$$

where $P_k(\gamma_{id(n)})$ is the kth quadrature point and $W(P_k(\gamma_{id(n)}))$ is the corresponding weight. Hence, the standard error under 2PL TRT was calculated as:

$$SE(\theta) = \frac{1}{\sqrt{\sum_{n=1}^{N} I_n(\theta)}}. \quad (15)$$

3.6.4 Correlation

Pearson correlations between the sets of ability estimates by model and the true ability parameters were also calculated and analyzed. Equation 16 shows the formulation used in the calculation of correlation ($r_{\hat{\theta}\theta}$) between theta estimates ($\hat{\theta}$) and true abilities (θ) of n test takers.

$$r_{\hat{\theta}\theta} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i-\bar{\hat{\theta}})(\theta_i-\bar{\theta})}{\sqrt{\sum_{i=1}^{n}(\hat{\theta}_i-\bar{\hat{\theta}})^2\sum_{i=1}^{n}(\theta_i-\bar{\theta})^2}}. \quad (16)$$

3.6.6 MST routing error rates

Two sets of MST test stage routing error rates were calculated based on test stage 1 and test stage 2 misrouting results. An incorrect test stage routing decision occurred when a test taker was routed to a next stage module that was not suitable given their true ability. For example, when a test taker was routed to a test stage 2 moderate module but their true ability falls within an easy module, a test stage 1 routing error was made. Thus, for each panel, routing error rates by test stage were calculated using:

$$test\ stage\ \ routing\ error\ rate = \frac{number\ of\ incorrect\ routing\ decisions\ for\ the\ stage}{total\ number\ of\ test\ takers} \quad (17)$$

3.6.7 MST misclassification rates

All MST panels, regardless of the average item difficulty difference condition imposed, had the same two final decision classification cut scores at $\theta = 0.0$ and 1.0. These classification cut scores have been used in previous MST routing comparative performance studies (Kim, Chung, Park, & Dodd, 2013). A misclassification occurs when a test taker is incorrectly classified, either above or below the cut scores, based on calculated ability estimates. The misclassification rate was calculated for each ability level θ as follows:

$$\theta\ misclassification\ rate\ = \frac{number\ of\ incorrect\ classifications\ above\ or\ below\ a\ cut\ score}{K}, \quad (18)$$

where K is the number of test takers (1000) at the θ ability level. Using the same weighting procedure discussed previously, the misclassification rate for an MST pathway was calculated as:

$$Pathway\ misclassificaiton\ rate\ = \frac{\sum \theta\ misclassification\ rate * weight}{\sum weight}, \quad (19)$$

Chapter 4: Results

4.1 Introduction

Under the 2PL IRT and 2PL TRT models, this chapter summarizes the results of the simulation study outlined in Chapter 3. It is divided into five sections with the first four summarizing the effects of the studied factors and their impact on ability estimation accuracy, routing accuracy, and classification accuracy. Formulations for measures used in the evaluation of study results can be found in Chapter 3. The final section of this chapter provides a summary of results as they relate to the research questions.

To better utilize space when summarizing results, descriptive statistics and ANOVA tables generated to investigate the significance of relations between studied factors can be found in the Appendices. ANOVA test results are summarized in this chapter only for statistically significant effects with moderate or large effect sizes. A power analysis showed that with a total of 970 subjects, there was at least an 80% chance of correctly rejecting the ANOVA null hypotheses of no difference with an alpha of .05 for moderate effect size. Thus, the sample size of 41,000 simulated test takers was more than adequate for the hypothesis testing. Although results could have been analyzed using regression models, this approach was determined to be impractical given the large number of study factors.

The magnitude of the partial eta squared (η_p^2) index was used in the ANOVA analyses to identify which significant effects are of moderate or large effect sizes. The η_p^2 index was selected for use since it is widely reported in educational research as an effect size index in the factorial ANOVA designs (Richardson, 2011). For all tests, the alpha level was set at 0.05 for a two-tailed test to control for Type I error.

The η_p^2 index is a measure of what proportion of variance in the outcome variable can be attributable to the manipulated factor of interest. Richardson (2011) notes that the η_p^2 index can be benchmarked against Cohen's (1969) criteria of small (.01), moderate (.06), or large (.14) effect sizes. Accordingly, this same η_p^2 effect size criteria was used in this chapter to help interpret the ANOVA analyses results and also to identify and report the moderate and large statistically significant effects.

All omnibus ANOVA tests reported in this section were performed under the following eight study factors: (1) the testlet effect (none, small=.25, large=1.0), (2) module difference (small=.05, large=1.0), (3) model (2PL IRT, 2PL TRT), (4) IPD magnitude (.3, .5), (5) IPD affected parameter (a, ab, b), (6) IPD direction (amplification, cancellation), (7) IPD percent (13%, 20%), and (8) IPD affected test stage (test stage 1, test stage 2, test stage 3). All ANOVA tests were run using IBM SPSS version 22 (IBM Corp, 2013).

In terms of the factorial ANOVA assumptions, the measurements reported in this chapter were continuous and observations were mutually independent from each other. The factor variables used in the analyses were also independent from each other. A Kolmogorov-Smirnov test for normality was performed on each set of measurement data. Results are significant for all tests, indicating that the measurement data sets are not statistically normal. Figures 7 to 12 present the Normal Q-Q plots for each set of measurements.

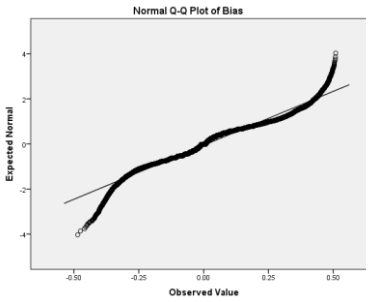

Figure 7. Normal Q-Q plot for bias

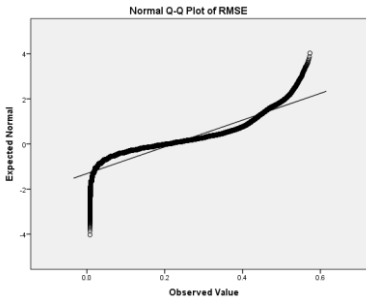

Figure 8. Normal Q-Q plot forRMSE

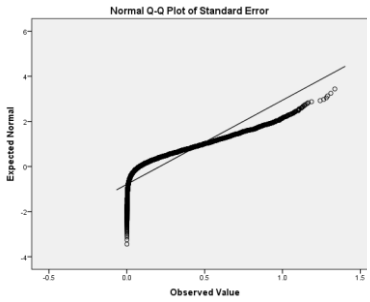

Figure 9. Normal Q-Q plot for SE

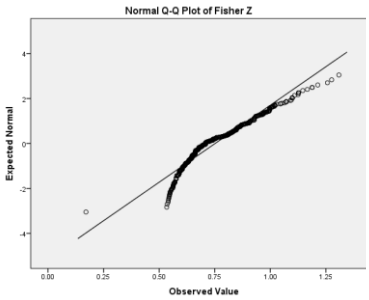

Figure 10. Normal Q-Q plot for Fisher Z

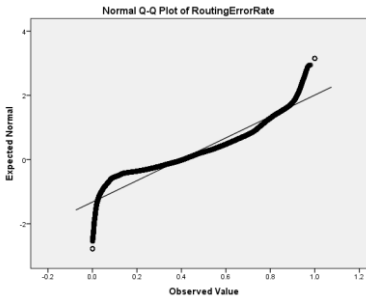

Figure 11. Normal Q-Q plot for routing error rate

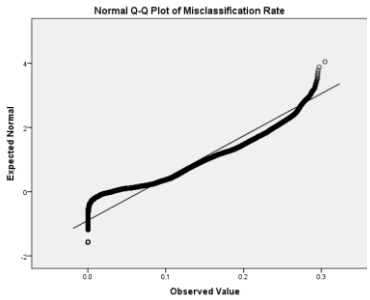

Figure 12. Normal Q-Q plot for misclassification rate

Figures 7 to 12 show that the data sets are either heavy tailed or skewed but do not have extreme non-normality. Factorial ANOVA designs are known to be robust when non-normality is not too extreme (Schmider, Ziegler, Daney, Beyer, & Buhner, 2010) and also if there are a large and equal number of observations in each cell as is the case with the data for this study. Moreover, given the large equal sample size of each factor group, the equal population variances ANOVA assumption is not needed for this study. Hence, the assumptions underlying the factorial ANOVA are satisfactorily met.

4.2 Ability Estimate Accuracy

The accuracy of ability estimates generated in this simulation study were evaluated by determining how well the testlet-based MST, under each study condition, recovers the true theta (θ) abilities of the simulated test takers. Study results for four

measures of ability estimate accuracy (bias, RMSE, standard error, and correlation) are summarized in this section.

4.2.1 Bias

Bias was the outcome measure in an eight-way ANOVA that included all eight study factors listed in the introduction to this chapter. No significant main effects nor interaction effects were detected. A follow-up one-way ANOVA was run to examine the differences between 13 IPD Type factor levels (Baseline, .3ab, .5ab, .3b, .5b, .3a, .5a, ±.3ab, ±.5ab, ±.3b, ±.5b, ±.3a, ±.5a) which includes the baseline no-IPD condition. No significant effect was detected.

4.2.2 RMSE

An eight-way ANOVA, which included all eight study factors, was run for RMSE. No significant main nor interaction effects were detected. RMSE values were also used as the dependent variable in a one-way ANOVA to examine RMSE differences in the IPD Type factor levels. No significant effect was detected.

4.2.3 Standard Error

SE was used as the outcome measures in a four-way ANOVA that included the module difference, testlet effect, pathway, and model factors. These four factors are the non-IPD condition factors used in this study. No significant main nor interaction effects were detected.

4.2.4 Correlation

Correlations were calculated to quantify the strength of relations between the true and model estimated test taker abilities. All correlations calculated and reported in this

section were tested and found to be significant. Prior to conducting analyses, all correlations were transformed to the Fisher Z scale and all analyses were performed using these transformed values. Analysis results were converted back to correlation values prior to including them in this report.

Fisher Z values were submitted to an eight-way ANOVA, using the same set of study factors listed in the introduction to this chapter, to identify significant effects for correlation. Results yielded three significant main effects. First, the analysis for the testlet effect, $F(2,863) = 972.46$, $p < .05$, $\eta_p^2 = .62$ showed a significantly large effect. Post-hoc Tukey multiple comparison results showed that all mean correlations were significantly different across all levels of the testlet effect factor. The mean correlation was the highest when the testlet effect was large (M=.71, SD=.14) than when it was small (M=.63, SD=.10) or when no testlet effect was present (M=.56, SD=.06). Second, the module difference effect, $F(1,863) = 45.65$, $p < .05$, $\eta_p^2 = .07$, was shown to be moderately significant with the mean correlation highest when the module difference was large (M=.65, SD=.15) than when the module difference was small (M=.62, SD = .14). Finally, the magnitude main effect, $F(1,863) = 36.67$, $p < .05$, $\eta_p^2 = .06$, was moderately significant. The mean correlation was the highest when the IPD magnitude was 0.3 logits (M=.65, SD=.15) than when the IPD magnitude was 0.5 logits (M=.62, SD = .13).

The ANOVA also showed the four-way interaction effect between module difference, testlet effect, the percent of IPD affected items, and the IPD affected test stage was moderately significant, $F(4, 863) = 66.92$, $p<.05$, $\eta_p^2 = .10$. Further, a five-way interaction effect between module difference, testlet effect, percent of IPD affected items,

IPD affected test stage, and model factors, $F(4, 863) = 91.20$, $p<.05$, $\eta_p^2 = .13$, was moderately significant.

A series of simple main effects analyses results showed that, for the 2PL TRT model, when the testlet effect was large or small, significant mean correlation differences *(p<.05)* exist across IPD affected test stages when the module difference condition was both large and small. For the 2PL TRT model, Figure 13 presents the mean correlation interaction plot among module difference, percent of IPD affected items, testlet effect, and test stage.

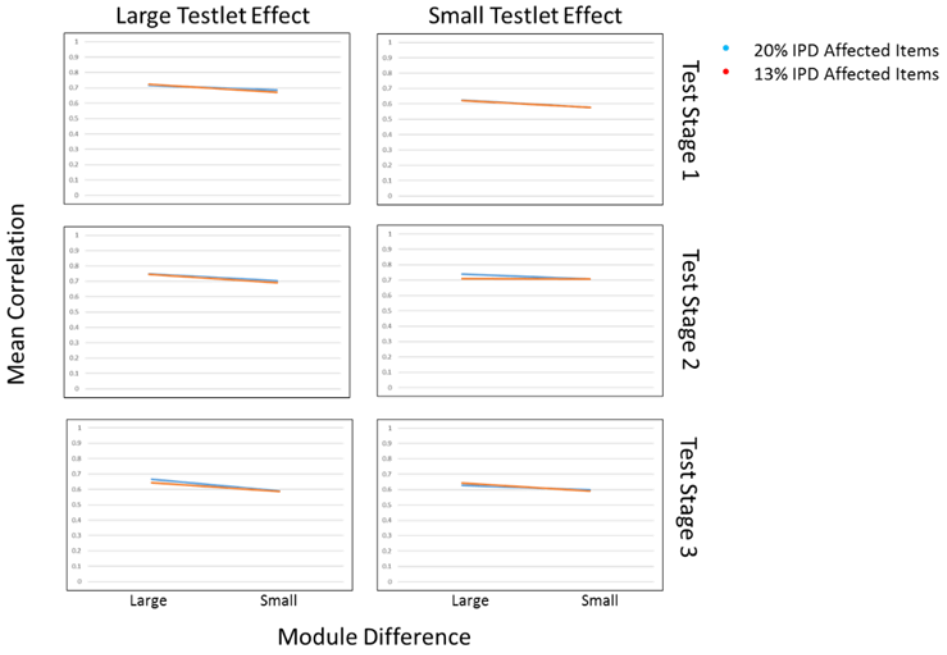

Figure 13. 2PL TRT mean correlation interaction plot for testlet effect, module difference, IPD affected test stage, and percent of IPD affected items

Figure 13 shows that when the 2PL TRT model is used, the interaction was ordinal where the line patterns displayed did not depend on the percentage of affected IPD items.

However, the large testlet effect consistently had the most influence towards increasing mean correlations.

A series of simple main effect analyses showed that when the 2PL IRT model was used, the percentage of affected IPD and module difference interactions was disordinal where the impact on the mean correlations depended on the level of the test stage and testlet effect factors. Using 2PL IRT, the mean correlations were the highest when 13% of the test items were IPD affected and the module difference was large, except when the testlet effect was large and IPD was present in test stage 1. When the testlet effect was large and IPD was in stage 1, the mean correlations were the highest when 20% of test items were IPD affected and the module difference was large.

Moreover, when the 2PL IRT model was used, the mean correlations were the highest when 13% of test items were IPD affected and the module difference was small, except when (1) the testlet effect was large and the test stage 3 testlet was affected by IPD or (2) when the testlet effect was small and the test stage 1 testlet was affected by IPD. When the testlet effect was large and IPD was present in stage 3 (1) or when the testlet effect was small and IPD was stage 1 (2), the mean correlations were the highest when 20% of test items were IPD affected and the module difference were small.

For the 2PL IRT model, Figure 14 presents the mean correlation interaction plot among module difference, percent of IPD affected items, testlet effect, and test stage.

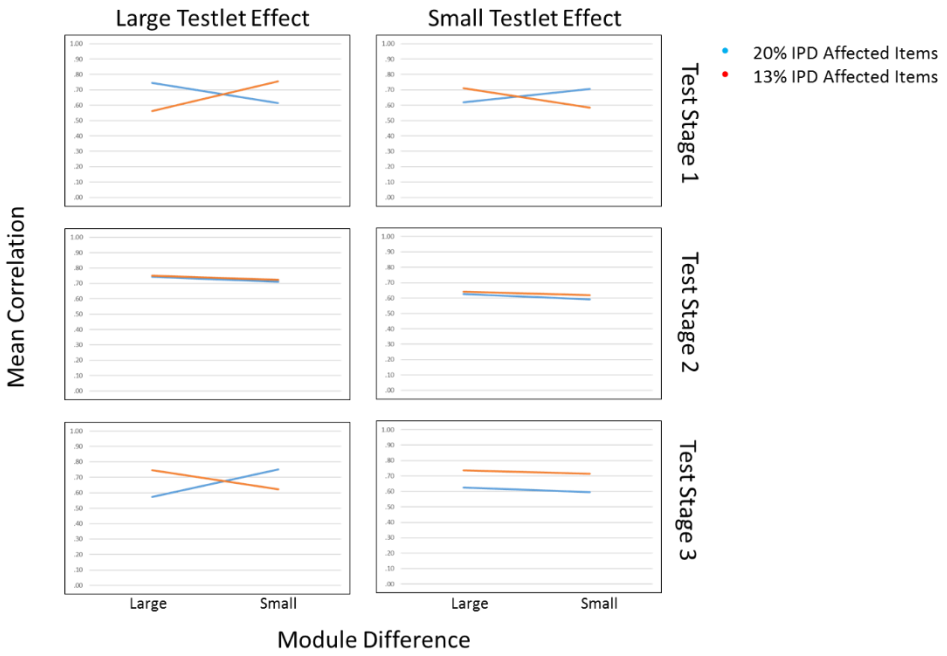

Figure 14. 2PL IRT mean correlation interaction plot for testlet effect, module difference, IPD affected test stage, and percent of IPD affected items

Unlike when the 2PL TRT model is used, Figure 14 does not show an overall definitive pattern indicating any one of the factors as being the most influential towards increasing mean correlations.

A one-way ANOVA was also run to examine the differences between IPD Type factor levels on the correlations. Analysis yielded a significantly moderate effect, $F(12,923) = 6.16$, $p < .05$, $\eta_p^2 = .08$. Associated Tukey multiple comparison results showed that the .3 b-drift group mean correlation (M=.69, SD=.18) was significantly larger than the baseline group mean correlation (M=.63, SD=.12).

4.3 MST Routing Accuracy

To enact test taker routing decisions, the testlet-based MST system design in this study implemented NC scoring with AMI module selection. Two sets of MST test stage level routing error rates were calculated by panel and by pathway. Test takers were

routed along either primary pathways (125, 136, or 147) or non-primary pathways (126, 135, 137, 146) through the MST. Primary pathways consisted of modules of the same difficulty level whereas non-primary pathways did not. Non-primary moderate difficulty pathways (126 and 146) consisted of one non-moderate difficulty module.

Test stage 1 routing error rates were calculated using counts of test takers misrouted from a test stage 1 module to a test stage 2 module. Test stage 2 routing error rates were calculated using counts of test takers misrouted from a test stage 2 module to a test stage 3 module. Moreover, descriptive statistics were also generated to examine mean routing error rates by the ability group level of misrouted test takers in either test stage 1 or test stage 2. The test taker ability group level ranges used were low ($\theta \leq 0$), moderate ($0 < \theta < 1$) , or high ($\theta \geq 1$). The ranges for the ability level groupings were selected to coincide with the two classification cut scores of $\theta = 0$ and $\theta = 1$ used with the testlet-based MST system under study.

A nine-factor ANOVA was conducted to evaluate the impact on panel routing error rates due to the combined effects of the same eight factors as above with one additional factor introduced, the routing error test stage factor. The routing error test stage factor consisted of two levels (Stage1, Stage2). The levels were associated with the proportion of test taker routing errors that occurred after the administration and scoring of testlets assigned to modules in either test stage 1 or test stage 2.

The nine-factor ANOVA yielded no significant interaction effects; yet the test did yield two moderate to large significant main effects. First, the analysis for the testlet effect factor, $F(2,11377) = 818.80$, $p < .05$, $\eta_p^2 = .12$, was moderately significant. The Tukey multiple comparison results showed that the mean panel routing error rate is

significantly different across all levels of the testlet effect factor. The mean panel routing error rate was higher when the testlet effect was large (M = .13, SD = .05) than when it was small (M=.10, .05) or when no testlet effect was present (M=.09, .04). Figure 15 shows the percent of misrouted test takers by ability group level and testlet effect.

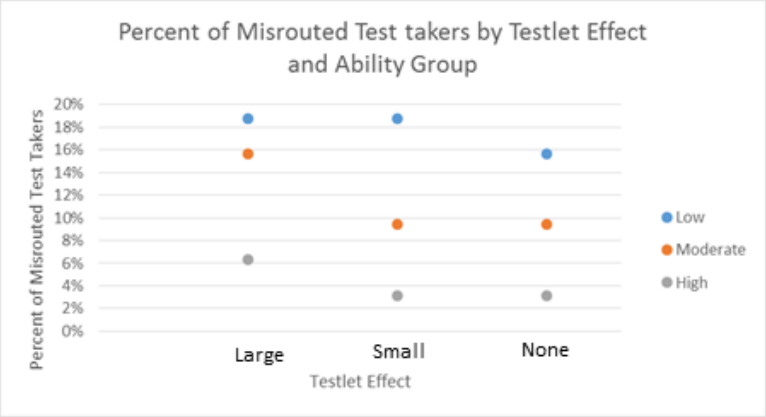

Figure 15. Percent of misrouted test takers by ability group and testlet effect level

In Figure 15, it appears that the testlet effect was more influential towards increasing the percentage of low ability test takers that were misrouted than increasing the percentage of moderate or high ability test takers that were misrouted. It also appears that the large testlet effect had a more intense impact on increasing the percentage of misrouted moderate ability test takers than the presence of the small or no testlet effect does.

Second, the routing error stage main effect was shown to be significantly large, $F(1,11377) = 2545.01$, $p < .05$, $\eta_p^2 = .17$, such that mean routing error rates were largest after the scoring of the test stage 1 module (M=.13, SD=.05) than after the scoring of the test stage 2 module (M=.09, SD=.04). Figure 16 presents the percent of misrouted test takers by ability group and routing error test stage.

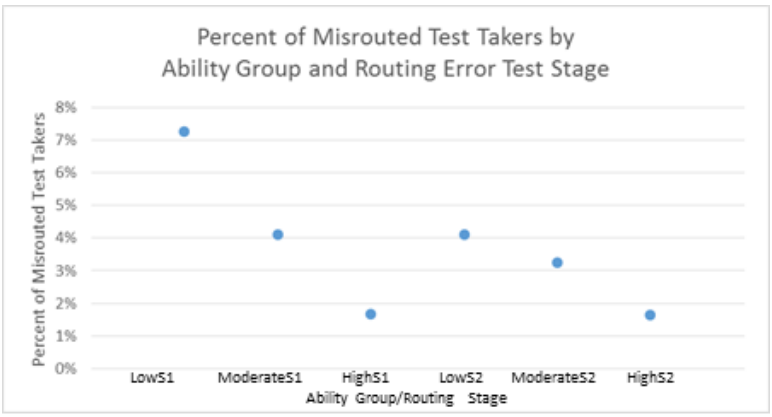

Figure 16. Percent of misrouted test takers by ability group and routing error test stage

Figure 16 illustrates that a larger percentage of low ability test takers were misrouted after administration of the test stage 1 module than moderate or high ability test takers.

A ten-factor ANOVA was also conducted to evaluate the impact on pathway routing error rates due to the combined effects of the same nine factors above with one additional factor introduced, the pathway factor. The pathway factor consisted of seven levels (125, 126, 135, 136, 137, 146, 147) that coincided with the routing error rates for test takers routed along each of the seven MST pathways.

The ten-factor ANOVA yielded two large significant main effects. First, the analysis for the significant pathway factor main effect, $F(6,82945) = 7417.2$, $p < .05$, $\eta_p^2 = .24$, was such that the mean routing error rate was highest for pathway 136 (M = .03, SD = .01) than for the remaining pathways. Pathway 136 was the primary moderate difficulty pathway through the MST. Second, the main effect of the stage factor on the routing error, $F(1, 82945) = 5385.89$, $p < .05$, $\eta_p^2 = .06$, was statistically significant with a moderate effect. The pathway mean routing error rate was higher for Stage1 (M = .02, SD = .02) than for Stage2 (M = .01, SD = .01).

The ANOVA also showed the two-way interaction effect between module difference and pathway was moderately significant, $F(6, 82945) = 1787.96$, $p<.05$, $\eta_p^2 = .11$. Because the interaction between module difference and pathway was significant, the simple main effects of module difference were examined. That is, the differences between pathway mean routing error rates at each level of the module difference factor were explored.

There were significant differences between mean pathway routing error rates across all levels of the pathway factor when the module difference was large $F(6,84658) = 4533.10$, $p<.05$, $\eta_p^2 = .24$, and when it was small $F(6,84658) = 3599.53$, $p<.05$, $\eta_p^2 = .20$. Figure 17 shows the interaction between the module difference and pathway factors.

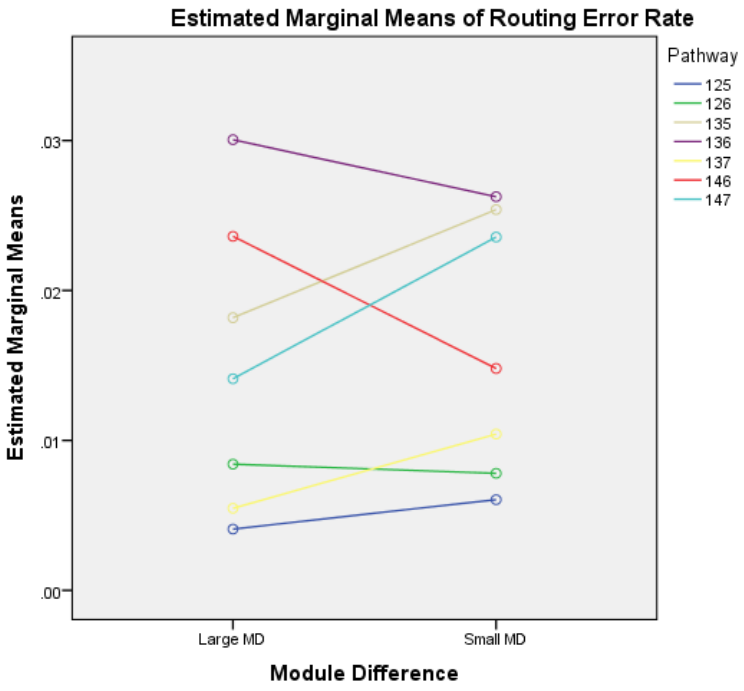

Figure 17. Pathway and module difference interaction plot

Figure 17 illustrates the interaction was disordinal, suggesting that the influence of each level of the module difference effect differed by pathway. First, it appears in Figure 17 that the large module difference effect was most influential towards increasing mean routing error rates for test takers routed along pathways 126, 136 and 146. On the other hand, the small module difference effect was most influential towards increasing mean routing error rates for test takers routed along pathways 135, 137, and 147. Figure 18 presents a visual representation of pathway mean routing error rates by module difference factor level.

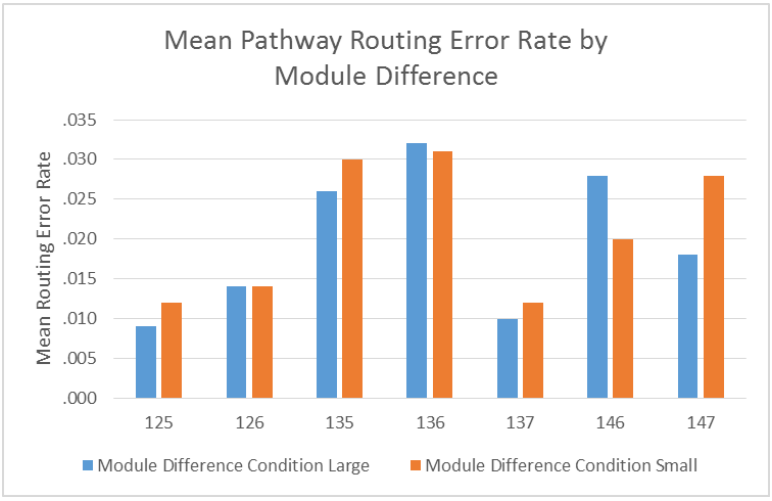

Figure 18. Pathway mean routing error rates by module difference

Figure 18 shows that the small module difference effect was more influential than the large module difference effect on increasing mean pathway routing error rates for the lowest and highest difficulty pathways, pathways 125 and 147 respectively.

ANOVA results also showed that a two-way interaction effect between the pathway and routing error stage was significantly large, $F(6, 82945) = 6854.95$, $p<.05$, $\eta_p^2 = .33$. Simple main effects for the pathway factor were also examined at levels of the routing error stage factor. Analysis showed that significant differences, $F(6,84658) =$

36104.86, p<.05, η_p^2 *=.30,* in Stage1 and Stage2 pathway routing error stage mean routing error rates existed only for pathway 135. Pathway 135 was the only pathway through the MST panel where test stage 2 recovery could take place for test stage 1 misrouted low ability test takers. Since a larger percentage of low ability level test takers were misrouted in test stage 1 (see Figure 16), it is logical that the Stage1 mean routing error rates for pathway 135 were significantly higher than Stage2 pathway 135 mean routing error rates.

Figure 19 presents the pathway and routing error stage interaction plot.

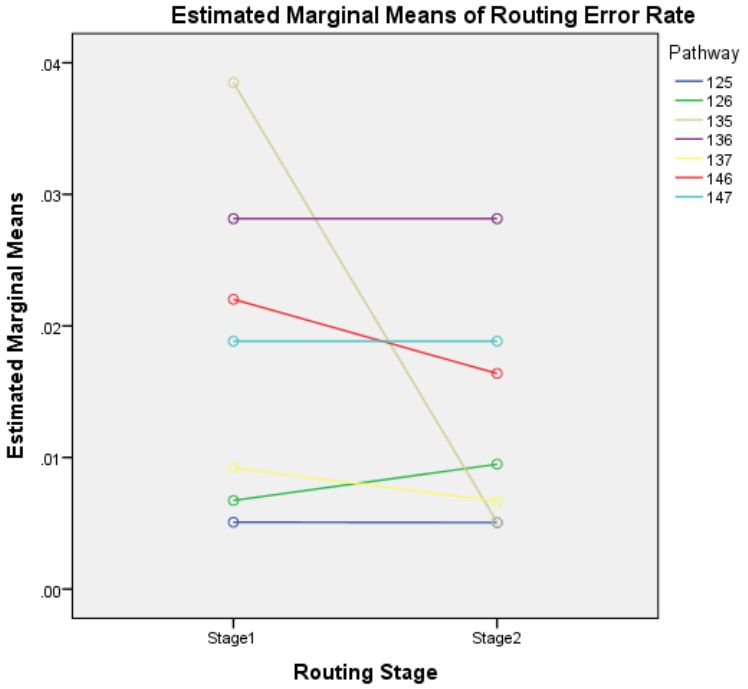

Figure 19. Pathway and routing error stage interaction plot

In Figure 19 shows that pathway 135 mean routing error rates were higher after administration of the test stage 1 module than after administration of the test stage 2 module.

Simple main effects for the routing error stage factor were also examined at levels of the pathway factor. There were significant differences across all pathway mean routing error rates after administration of the test stage 1 module, $F(6,84658) = 9816.04$, $p<.05$, $\eta_p^2 = .41$. Significant differences in pathway mean routing error rates also existed after administration of the test stage 2 module, $F(6,84658) = 4884.27$, $p<.05$, $\eta_p^2 = .26$, with the exception of pathways 125 and 135. Figure 20 presents a visual representation of pathway mean routing error rates by routing error stage factor level.

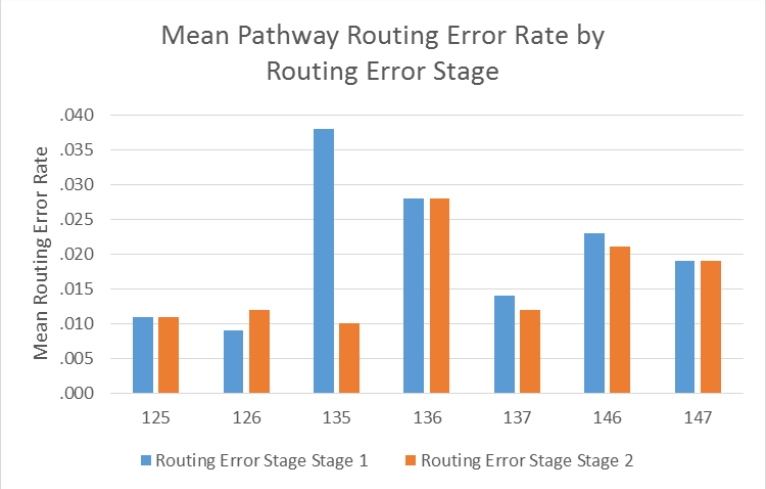

Figure 20. Mean pathway routing error rate by routing error stage

In Figure 20, both the Stage1 and Stage2 routing error stage effects appear to have similar levels of influence towards increasing mean routing error rates along the three primary MST pathways, pathways 125, 136, and 147. Furthermore, the Stage1 routing error rate effect appears to be most influential towards increasing mean routing error rates for non-primary pathways 135, 137, and 146 than for the remaining pathways.

ANOVA results also showed a two-way interaction effect between the pathway and testlet effect factors that was moderately significant, $F(12, 82945) = 531.54$, $p<.05$, $\eta_p^2 = .07$. The simple main effects of testlet effect across levels of the pathway factor

were examined. There were significant differences between all pathway mean routing error rates when either the testlet effect was large, $F(6,84651) = 2059.59$, $p<.05$, $\eta_p^2=.13$, small, $F(6,84651) = 2652.48$, $p<.05$, $\eta_p^2=.16$, or when no testlet effect was present, $F(6,84651) = 3069.81$, $p<.05$, $\eta_p^2=.18$. That is, associated pairwise comparison results confirm that mean routing error rates across all pathways were significantly different at each level of the testlet effect factor. Figure 21 presents the mean plot for the interaction between pathway and testlet effects.

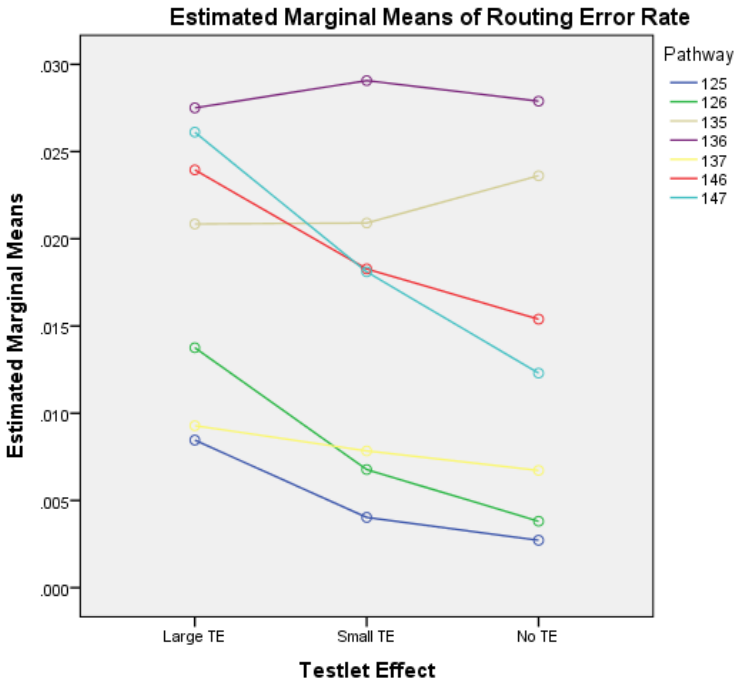

Figure 21. Pathway and testlet effect interaction plot

Figure 21 illustrates that the large testlet effect appeared to have more influence on increasing mean routing error rates across majority of the pathways, with the exception of pathways 135 and 136, than when the testlet effect was small or no testlet effect was present.

Figure 22 presents a visual representation of pathway mean routing error rates by testlet effect factor level.

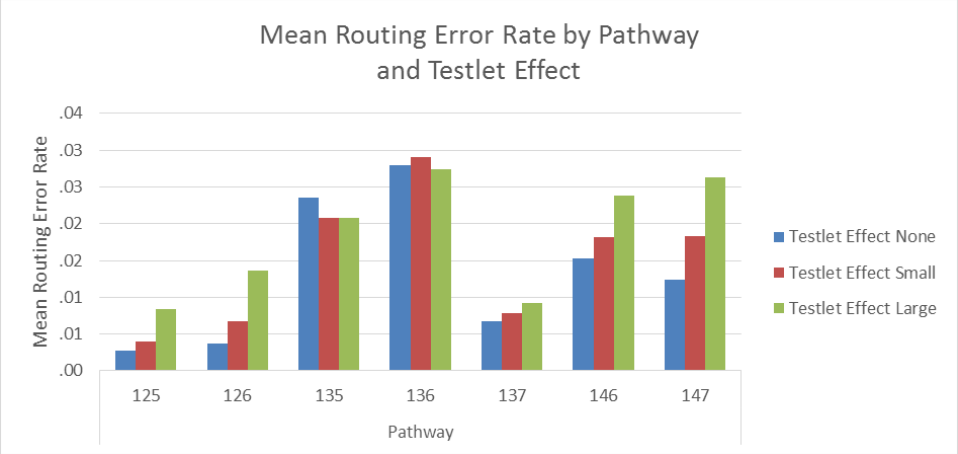

Figure 22. Pathway mean routing error rate by testlet effect level

From Figure 22, it appears that the large testlet effect was less influential towards increasing pathway mean routing error rates when the test stage 2 module was of moderate difficulty than when the test stage 2 module was of low or high difficulty.

4.4 Misclassification Rate

The misclassification rates for pass/fail decisions were calculated above and below the cut score of 0 and above and below the cut score of 1. All pathway misclassification rates were equal to 0.0 above the cut score of 1. That is, classification decisions for those above the cut score of 1 were correctly made and all incorrect pathway misclassification assignments occurred below the cut score of 1. However, when the ability level of test takers were considered, results showed that incorrect classifications occur within each ability level group.

Misclassification rates above or below the cut score of 0 were used as an outcome variable in a ten-way ANOVA using the same eight set of factors listed above with the

addition of the pathway and ability group factors. The pathway factor had seven levels (125, 126, 135, 136, 137, 145, 146) that coincided with the misclassification rates for test takers routed along each of the seven MST pathways. The ability group factor had three levels (low ($\theta \leq 0$), moderate ($0 < \theta < 1$) , and high ($\theta \geq 1$)) that coincided with test taker ability ranges partitioned by the classification decision cut points of $\theta = 0$ and $\theta = 1$.

The ANOVA test yielded two significantly large main effects for the pathway and ability factors. First, analysis for the pathway factor, *F(6, 20233) = 1267.21, p < .05,* η_p^2*=.18,* showed a significantly large effect. Associated Tukey multiple comparison results showed that all pathway misclassification rates were significantly different from each other. However, mean pathway misclassification rates were highest for pathways 137 and 146 (M=.14, SD=.07). Second, analysis for the ability group factor, *F(2, 20233) = 14554.32, p < .05,* η_p^2*.45,* also showed significantly large effect. Associated Tukey multiple comparison results showed that mean misclassification rates were significantly different for all levels of the ability group factor. The mean pathway misclassification rate was higher for moderate ability test takers (M=.18, SD=.05) than for low (M=.06, SD=.06) or high (M=.03, SD=.05) ability test takers.

ANOVA results showed the two-way interaction effect between the pathway and ability group factors was moderately significant, *F(12, 20233) = 270.42, p<.05,* η_p^2 *=.08*. The simple main effects of ability group across levels of the pathway variable factor were examined. Results showed significant differences in ability group mean misclassification rates *(p<.05)* across all levels of the pathway factor for low and high ability test takers. Figure 23 presents the ability group level and pathway interaction plot.

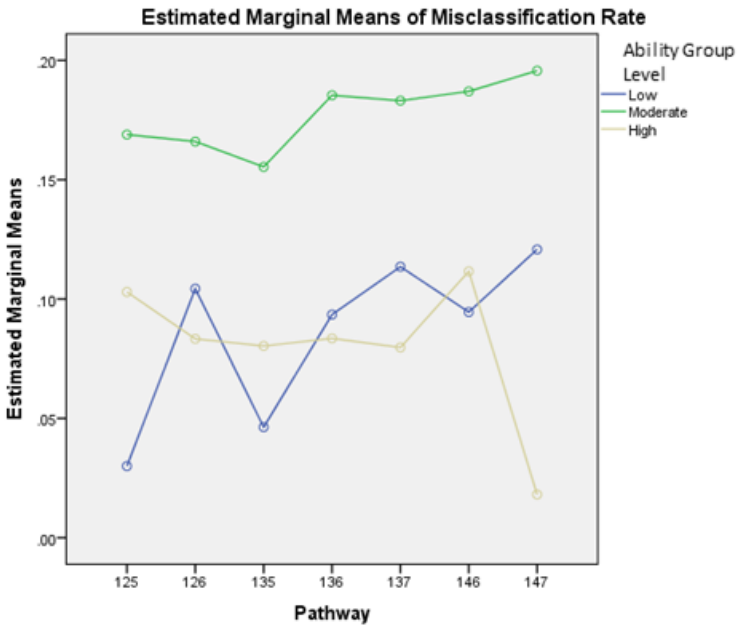

Figure 23. Ability group level and pathway interaction plot

Figure 23 shows the moderate ability test takers were more likely to be misclassified than low or high ability test takers.

Simple main effects for the pathway factor were also examined for levels of the ability group factor. Significant differences *($p < .05$)* between ability group mean misclassification rates across pathways, with the exception of pathways 126 and 135, were found. Moreover, the difference in ability group mean classification rates were moderately significant for all remaining pathways, with the exception of pathway 147. For pathway 147, mean misclassification rates were significantly different across all levels of the ability factor. Figure 24 presents a visual representation of the interaction of the pathway by ability group on the mean misclassification rates.

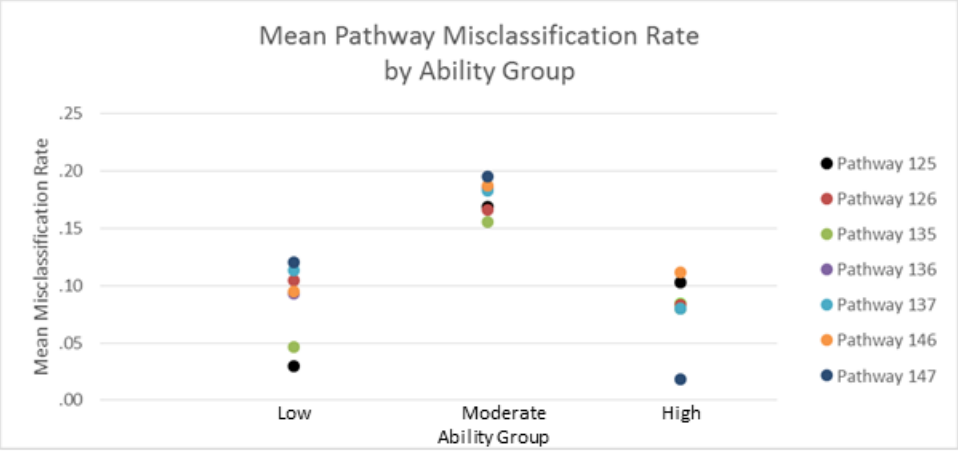

Figure 24. Mean pathway misclassification rate by ability group

In Figure 24, it appears that the moderate ability effect has more influence towards increasing mean pathway misclassification rates than low and high ability effects.

4.5 Summary

For the testlet-based MST system design being studied, research goals were to compare how three measures (ability estimates, routing error rates, and misclassification rates) were impacted by the combined presence of LID, IPD amplification , or IPD cancellation at the testlet level when either the 2PL IRT or 2PL TRT models were used. Studied factors included testlet effect, module difference, percent of IPD affected test items, IPD affected test stage, model, IPD parameter, IPD magnitude, IPD direction, routing error stage, ability group, and pathway. Results reported in this chapter for each of the three impacted measures are summarized in this section.

ANOVA results for bias, RMSE, and standard error showed no significant effects indicating that none of the study factors impact these ability estimation accuracy measures under either model in any meaningful way. Even so, ANOVA results did show

that the correlations were impacted by the presence of the testlet effect where the highest correlations, on average, occur when the testlet effect is large.

Results showed that for both models, mean correlations were significantly impacted by testlet effect, module difference, IPD affected test stage, and percent of IPD affected test items. Finally, the mean correlations were significantly higher than the baseline mean correlation when testlet items contained IPD on item difficulty with a magnitude of 0.3.

The mean routing error rates were not impacted due to amplification or cancellation study conditions, but were impacted by the presence of LID and MST structural features under non-IPD study conditions in several ways. First, the results showed that the mean routing error rates across the MST panels were the highest when the testlet effect was large. Moreover, test takers were more likely to be misrouted after the administration of the test stage 1 module than the test stage 2 module. Finally, the panel level routing error rates revealed that a larger percentage of low ability test takers were misrouted than moderate or high ability test takers.

When examining the mean routing error rates by MST pathway, significant differences in the mean routing error rates across all pathways were detected when the module difference condition was large and when it was small. Upon further examination, the mean routing error rates for the primary and non-primary pathways appeared to be influenced differently by the levels of the module difference effect. For instance, pathways 126,136, and 146 appeared to have increased mean routing error rates influenced more by the presence of the large module difference effect than the small. On the other hand, increases in the mean routing error rates for pathways 135, 137, and 147

appeared to be influenced more due to the presence of the small module difference effect than the large module difference effect.

Results also showed that overall, Stage1 mean pathway routing error rates tended to be larger than Stage2 rates. Yet when considered by individual pathway, pathway 135 was the only pathway that had Stage1 mean pathway routing error rates that differ significantly from Stage2 rates.

In addition to the module difference and routing error stage effects, results also showed that each pathway mean routing error rate differs significantly from the others when the testlet effect factor is fixed. Furthermore, the large testlet effect appeared to be more influential towards increasing pathway mean routing error rates than the small or no testlet effects were.

Finally, the study results showed that the misclassification rates were impacted differently by the combined influences of the pathway and test taker ability level effects. Moderate ability test takers were more likely to be misclassified than low or high low ability test takers while traversing any pathway within the MST system.

Chapter 5: Discussion

5.1 Introduction

This chapter discusses the study results in seven main sections. The first section serves as an introduction while the second section presents a restatement of the four research questions and a brief summary of the methodology used in the study. Using results from the study, the next three sections present and discuss the impact, due to study conditions effects, on measurement accuracy, routing accuracy, and decision classification accuracy for the testlet-based MST system. Next, conclusions for and implications of the study findings are presented. Finally, the limitations of the study and future research are discussed.

5.2 Restatement of Research Questions

Under various testing conditions, this study sought to investigate the impact of cumulative patterns of IPD, which manifest as amplification or cancellation at the testlet level, combined with LID effects on test taker ability estimation, routing accuracy, and decision classification accuracy. This was a simulation study conducted within a 1-3-3 testlet-based MST system where modules consisted of either a low, moderate, or high average item difficulty testlet.

To that end, this study contributes to the IPD literature and MST developers by providing information in response to the following four research questions:

1. In a testlet-based MST system design, if the 2PL TRT model is the true measurement model, how is the degree of accuracy in final ability estimates calculated under 2PL IRT assumptions impacted by the magnitude, type,

direction, and MST test stage where IPD amplification and cancellation at the testlet level exists?

2. Under an NC routing rule using an AMI module selection method, if the 2PL TRT model is the true measurement model, do testlet based MST routing error and misclassification rates differ when IPD amplification and cancellation at the testlet level exists?
3. When the module average item difficulty difference conditions within an MST stage are adjusted, if the 2PL TRT model is the true measurement model, are testlet based MST routing error and misclassification rates impacted when IPD amplification and cancellation at the testlet level exists?
4. Would use of the 2PL TRT model, which can account for the impact of LID due to testlet effects, improve the overall measurement accuracy in final ability estimates, testlet based MST routing error, and misclassification rates versus the use of the 2PL IRT model in the presence of IPD that exists as amplification or cancellation at the testlet level?

In this study, two 1-3-3 testlet-based MST systems constructed using R software, were compared in terms of ability estimation accuracy, routing accuracy, and classification decision accuracy under simulated study conditions. There were a total of 864 study conditions. Both MST systems, under all study conditions, were administered to all simulated test takers. There were 41,000 simulated test takers, defined as 1000 replicates at each θ point ranging from -3 to 3 in an increment of 0.15. This resulted in a uniform distribution of simulated test takers. However, in practice, for most education testing situations the test taker population is assumed to be normally distributed. Thus, in order

to approximate a standard normal distribution of test takers for this study, weighting was applied to all theta points generated. Hence, reported results should have been similar to those observed in practice when the normal distribution for the ability parameter is assumed.

Also under all study conditions, ability accuracy, routing error rates, and misclassification rate measurements were calculated and evaluated using ANOVA tests, descriptive statistics, and graphs. All ANOVA tests and associated indices were performed and calculated in SPSS. Graphs were created in SPSS or Microsoft Excel.

Both MST systems were constructed using panels across two module difference conditions. The average module difficulties within the second and third test stage modules were spaced at .5 logits for the small module difference condition and 1.0 logits for the large module difference condition. The amplification and cancellation effects were applied to one or both item parameters within a single testlet. The treatments were applied to either 13% or 20% of total test items residing within a single testlet. Affected testlets were assumed to be one of the three stages, (stage 1, stage 2, or stage 3). Moreover, an NC scoring rule and an AMI module selection method were implemented in this study as is currently done in practice. The NC scoring rule was used to determine module-level scores needed to inform the routing system. The AMI module selection method was used at each test stage to identify the location of the routing decision cut scores. The two-decision classification cut scores were set at $\theta = 0$ and $\theta = 1$.

The two MST systems differed in MST construction and upon which measurement model test taker routing and scoring was based. One was based on the 2PL IRT model, while the other on 2PL TRT. While the 2PL TRT model is the true measurement model,

both models were used for calibrations. Three sets of response data based on the 2PL TRT model were generated where a 0.0, 0.25, or 1.0 testlet effect variance magnitude is imposed. Under each model, 10 MST panels were constructed for each of the 6 module difference condition and testlet effect study combinations.

During the simulation, IPD amplification and cancellation conditions were imposed when item responses were generated. A panel exposure rate during MST administration was set at .25.

5.3 Ability Measurement Accuracy

ANOVA results for bias, RMSE, and standard error showed no significant effects indicating that none of the study factors impact these ability estimation accuracy measures under either model in any meaningful way. Under amplification and cancellation study conditions when no LID is present, this result can be supported by prior research: prior linear test, CAT, and MST standard IRT-based IPD research found that ability estimation is not significantly impacted when insignificant magnitudes of 20% or less of test items are IPD affected.

Yet in the presence of large LID effects with no IPD, the present study results showed that ability estimate accuracy is not significantly impacted under either model. It could be argued that since the 2PL TRT model is the true model for the study, the expectation is that this model should perform well in terms of ability estimation accuracy in the presence of LID and the absence of IPD amplification or cancellation. However, for the 2PL IRT model, given results of other standard IRT-based MST research investigations (e.g., Lu, 2010), the expectation is that ability estimation accuracy should be impacted in the presence of large LID effects.

There may be at least two possible reasons why in this present study ability estimation accuracy was not significantly impacted when the 2PL IRT model was used in the presence of LID. First, prior research has shown that using standard IRT models when LID is present may lead to the overestimation of the accuracy of ability estimates (Wainer, Bradlow, & Wang, 2007). Therefore, the degree of ability estimation accuracy, under use of the 2PL IRT model in the presence of LID, that is reported in the present study may be questionable.

Another reason for the lack of significant impact on ability estimates in the presence of large LID effects may be related to the MST panel structural design used here compared to panel structures used in other MST LID studies. For instance, Lu (2010) investigated the influence of LID effects within a 1-2-2 panel MST system: when the magnitude of LID was large, the accuracy of ability estimates was impacted under a standard IRT model. This present study was simulated within a 1-3-3 MST panel design structure. Unlike the 1-2-2 panel design which provides just two module difficulties levels, the 1-3-3 panel provides three different module difficulty levels and therefore provides more test adaptation options. Given that the ability estimation accuracy improves as the difficulty of a test becomes more aligned with a test taker's ability level, the 1-3-3 panel design is an explanation for the lack of significant impact on ability estimates in the presence of large LID effects

Results also showed that under combined LID and amplification and/or cancellation effects, ability estimation accuracy was not significantly impacted under either model. This may be because the combined influence of LID and insignificant magnitudes of IPD may actually lower the potentially adverse influence of LID on ability

estimation for some test takers. In the case where LID and IPD effects are both directionally favorable or unfavorable to the test taker, an unexpected increase in the likelihood of answering the items correctly or incorrectly results. However, in instances where the LID and IPD effects are in opposite directions, the combined influence may diminish due to the cancellation of the two effects. This could potentially result in a diminished influence of the LID effect when IPD is present compared with when it is not present, which may possibly lead to improved ability estimation accuracy.

Study results showed that for both the 2PL IRT and 2PL TRT models, mean correlations are impacted by the module difference, testlet effect, and IPD-affected test stage effects. In fact, the highest mean correlations tended to occur when the testlet effect was large rather than small or non-existent. This result is likely due to the performance of the 2PL TRT model in estimating test taker abilities when the testlet effect is large. Associated 2PL TRT mean correlation interaction plots showed that the large testlet effect appears to have the most influence on increasing mean correlations.

Unlike the 2PL TRT model, when the 2PL IRT model is used, the mean correlation interaction plots showed a disordinal pattern. That is, the interaction plots did not show an overall definitive pattern of impact on mean correlations by the factors. This is likely due to model misfit given that the 2PL TRT model is the true model for this study.

When compared to other studies with IPD magnitude shifts in item parameters, the results also showed that the highest mean correlation occurred when testlet items had a 0.3 IPD magnitude in item difficulty. In terms of the 2PL IRT model, the results were consistent with findings from previous research which have shown that a 0.3 shift to the

b-parameter of 20% or less of test items did not significantly impact ability estimation accuracy. In terms of the 2PL TRT model, it is possible that a 0.3 IPD in difficulty parameter may be detected by this model because of its status as the true model and thus may be another variance source. Prior research has considered that item characteristics and the testlet effect could be considered as non-construct related sources of variance (Bao ,2007; Bao, Dayton, & Hendrickon, 2007). Therefore, one reason for this study result under the 2PL TRT model may be that the unexpected 0.3 IPD magnitude shift in the b-parameter was not large enough to influence the performance of the affected testlet. That is, this magnitude shift may not be significant in adversely impacting the 2PL TRT consistent testlet effect variance assumption across testlets, thus resulting in estimates that correlate highly with true ability estimates.

On the other hand, this study found that the mean correlations are higher when the testlet items are affected by IPD with a magnitude of 0.3 than by IPD with a magnitude of 0.5. Therefore, an unexpected 0.5 IPD magnitude shift in the b-parameter may be large enough to impact 2PL TRT performance.

5.4 Routing Accuracy

Kim and Moses (2014) suggested that perhaps in practice, the impact of routing error could be substantial if the quality of MST items used is not high. The presence of LID is a testlet item quality concern. This study indicated that the presence of the testlet effect increases panel and pathway mean routing error rates under both models particularly when the testlet effect is large.

There may be several reasons for this. One reason may be related to the use of the 2PL IRT model fitted to response data in the presence of LID. Lu (2010) found that use

of a standard IRT model in the presence of a large testlet effect influenced MST test stage 1 interim ability estimates used to inform routing procedures. Another reason may be related to biased item parameter estimates due to LID under the 2PL IRT model. Bradlow, Wainer, and Wang (1999) showed that when the testlet effect is ignored by the 2PL IRT model, the size of item parameter estimation bias was related to the magnitude of the testlet variance. Hence, the biased item parameter estimates combined with the presence of a large testlet effect may significantly impact the interim item scoring performance of the 2PL IRT. Inaccuracy in testlet item scoring could lead to inaccurate NC scores for the module and subsequently to possible increases in routing error rates.

In general, stage 1 panel mean routing error rates were found to be significantly larger than those in stage 2. The reason might be that more item response data is available to inform the NC scoring rule after administration of the test stage 2 module than after the test stage 1 module.

A larger percentage of low ability test takers were misrouted across MST panels than moderate or high ability test takers. Given that study results also showed that test takers were more likely to be misrouted after administration of the test stage 1 module, this is a logical result given that test takers of all abilities are administered a moderate difficulty stage 1 module. Hence it is more likely that low ability test takers will respond incorrectly to these test stage 1 items than either moderate or high ability test takers. This could consequently lead to increases in the percentage of low ability subjects misrouted.

The mean routing error rates in different pathways differed significantly when the levels of the module difference were fixed. In fact, test takers were more likely to be misrouted along moderate difficulty primary and non-primary pathways when the module

difference is large rather than small. One reason for this may be that the larger the difference in the average module difficulty between adjacent modules in the same test stage, the more distinct the modules are. This could result in MIFs with less overlap. Under the AMI module selection method used in this study, less overlap in adjacent module MIFs resulted in the generation of θ routing decision cut points, located at the intersections of adjacent MIFs, that are farther apart than if adjacent modules were less distinct in average difficulty. It follows that the NC cut scores generated from these θ routing decision cut points would also be farther apart.

Under the NC scoring rule used in this study, when the module difference was large, the difference was wider between the two NC cut scores used in routing decision making within the same test stage than when the module difference was small. The result is a wider range of acceptable NC scores to route test takers to moderate difficulty next stage modules. This also resulted in a narrower range of acceptable NC scores to route test takers to either low or high difficulty next stage modules. Therefore, when the module difference is large, it is more likely that a larger percentage of low and high ability test takers may be misrouted onto a moderate difficulty pathway. This results in an increase in routing errors occurring along moderate difficulty pathways.

On the other hand, test takers routed along pathways that are not moderate difficulty (pathways 135, 137, and 147) appear to be more likely to experience increases in routing errors when the module difference condition is small than when it is large. When the module difference condition was small, there was a narrower range of acceptable NC scores that can be obtained in order for the routing system to route test takers to moderate difficulty next stage modules. This also resulted in a larger acceptable

NC score range availability for routing test takers to a low or high difficulty next stage module. Hence when the module difference is small, it makes sense that moderate ability test takers routed along pathways 135, 137, or 147 are likely to experience an increase in routing errors (to low and high difficulty next stage modules) than when the module difference condition is small. Of note is that the test stage 3 modules for pathways 135, 137, and 147 are either low or high difficulty modules.

This study found that testlet effects impacted the mean routing error rates for all pathways. However, the large testlet effect appears to have more influence towards increasing pathway mean routing error rates than small or no testlet effects do. Similar findings have been reported in other MST studies (Lu, 200). Diminished interim measurement accuracy at the module level could lead to less accurate NC scores and consequently an increase in routing error rates.

In this study, one reason that the testlet effect may have more impact on interim measurement accuracy than on overall final ability estimation accuracy may be the difference in the number of item responses at the module and test level. At the module level, item response data from just 10 testlet items were available to inform the NC scoring procedure, as opposed to 30 items at the test level informing the final ability estimation procedure. That is, less available item response information combined with the presence of a large testlet effect at the module level likely diminishes interim measurement accuracy than would occur when more item response data is available.

Similar logic could be applied in explaining why the routing error stage 1 effect has more influence towards increasing pathway mean routing error rates than the routing error stage 2 effect. Yet, study results show that pathway 135 is the only pathway where

the difference in pathway mean routing error rates for test stage 1 and pathway mean routing error rates for test stage 2 is significantly different.

5.5 Classification Accuracy

Results show that moderate ability test takers were more likely than low or high low ability test takers to be misclassified. Prior research has shown that MST misrouting may impact classification decision accuracy (Zenisky & Hambleton, 2014). This study showed that the highest pathway mean routing error rates occur along the primary moderate difficulty pathway, pathway 136. Pathway 136 is the pathway for moderate ability test takers under favorable routing conditions. The pathway mean routing error rates along pathway 136 are likely to be impacted by the widening and narrowing of the acceptable range of NC score, influenced by the large and small module difference respectively. Hence it makes sense that the mean pathway misclassification rates may increase for moderate ability test takers due to the high routing error rates reported for these test takers.

The same reasoning can be used to explain why differences in mean misclassification rates across pathways are significantly different for low and high ability test takers. Depending on the pathway and the level of the module difference imposed, the mean routing error rates for low and high ability test takers are likely different and, consequently, the mean misclassification rates for low and high ability test takers differ across pathways as well.

5.6 Conclusions and Implications

This study found that the accuracy in ability estimation and classification are not significantly impacted by the studied factors. However, the panel and pathway mean

routing error rates were shown to be impacted by the presence of LID particularly when the testlet effect is large. Pathway mean routing error rates were also shown to be significantly impacted by the module difference effect. Furthermore, the direct impact of the module difference effect on pathway mean routing error rates may likely have had an indirect impact on the pathway classification rates for moderate ability test takers.

Results also showed that the effects of testlet effect, module difference, and cumulative IPD study conditions impact patterns in mean correlations. The mean correlation patterns observed under the under the 2PL TRT model appear to be more consistent than patterns observed under the 2PL IRT model. For instance, regardless of which MST test stage is IPD affected, under the 2PL TRT model mean correlations tend to increase as the module difference increases. To the contrary, under the 2PL IRT model the relation between mean correlation and module difference varies across IPD affected test stage. Thus, for this MST system, the 2PL TRT model may be the more feasible model choice over the 2PL IRT model when controlling for adverse mean correlation effects is an important consideration.

Therefore, study results align with previous MST research in two ways. First, when less than 20% of test items in the same testlet are amplified or cancelled at the testlet level with insignificant magnitudes of IPD, neither ability estimation, routing, nor classification rates are shown to be significantly impacted. Second, in the presence of LID, interim ability estimation could be impacted, resulting in an increase in mean routing error rates particularly when the testlet effect is large.

This study contributes to the literature by providing empirical evidence that the module difference MST structural feature may have a significant impact on routing error

rates, particularly along moderate difficulty pathways. This in turn could indirectly impact misclassification rates for moderate ability test takers. Related findings in previous studies performed under the 2PL IRT model showed that changes in module difference did not significantly impact routing error rates (Kim and Moses, 2014; Kim, Moses, and Yoo, 2015). However, these previous studies were conducted under the assumption of no LID. Furthermore, the previous studies were also conducted using a two-stage MST panel design as opposed to the three-stage MST panel design used in this study. It is possible that these different panel designs, and the subsequent difference in the number of available MST adaptation points for each study, may have contributed to the different routing error rate detection outcomes.

The findings of this study suggest that the impact on routing accuracy associated with LID and the module difference effect should be test fairness concerns as there are a number of testing applications, such as diagnostic or placement tests, where module routing and classification decision accuracy may play a more significant role in decision making than do final ability estimates. When a testlet-based MST, modeled under 2PL IRT or 2PL TRT, is selected for administration of such tests, this study offers empirical evidence that ignoring influences of LID or module difference levels may bring into question the validity of the inferences made. Nevertheless, this study does suggest several actions that may be taken by test developers to help diminish the impact of the significant study effects on the routing and pathway classification performance of the 1-3-3 testlet-based MST under study.

First, when constructing MST panels under NC routing rules, test developers should optimize the difference in the average module difficulty between adjacent

modules so as to support accurate routing and pathway classification for as wide a range of test taker abilities as possible. This study showed that maintaining a .5 logit difference in average module difficulty throughout the panel tends to directly increase mean routing error rates along the low and high difficulty non-primary and primary pathways. Furthermore, study results also suggest that the mean routing error rates increase along the moderate difficulty non-primary and primary pathways when the module difficulty difference between adjacent modules is 1.0 logit.

It is possible that a gradual increase in module difference, as the test stage increases, may work to more evenly balance mean routing error rates across pathways. That is, if the test stage 2 module difference is not too wide, this would permit more overlap in testlet item difficulty across modules. Study results suggest that a .5 logit module difference may be too small, and a 1.0 logit difference may be too large for some test taker ability levels. Hence preliminary investigations prior to MST implementation can be performed to determine a module difference value between 0.5 and 1.0 logits that may be acceptable for a test stage 2 average module difficulty difference.

This study also found test takers were more likely to be misrouted after the administration of the test stage 1 module. If the module difference in test stage 2 is acceptably close, test takers misrouted from test stage 1 will be more likely to be administered test stage 2 testlet items that are more aligned to their ability levels than if the module difference in test stage 2 is unacceptably wide.

Moreover, if the module difference in test stage 3 is larger than the module difference in test stage 2, the MST modules will become more distinct to support better customized routing of test takers to a test stage 3 module more aligned with their ability

level. Furthermore, NC scoring used to inform routing to the test stage 3 modules (with a wider module difficulty) would be based on testlet item response information from two previous test stage modules. This should work to increase interim test stage 2 measurement accuracy. Given the more accurately available routing information, a test stage 3 module difficulty within the neighborhood of 1.0 may be acceptable.

The study results also suggest that diminishing the influence of LID particularly in test stage 1 may help to diminish mean routing error rates. One approach to address this may be to consider primarily administering statistically independent items in test stage 1. With such an item set, LID effects would be minimal.

Another possible approach that may help improve routing test mean routing error rates may be to use more than one stage 1 module. The benefits of such an approach are discussed in Yan, von Davier, and Lewis (2014). In short, including more than one routing module increases the likelihood that the ability estimate will be more precise than when only a single routing test is used. Moreover, more item responses are available to inform the test stage 1 routing decision than if only one module is used.

5.7 Limitations and Future Considerations

This study is only an initial step in exploring the ability estimation accuracy, routing system accuracy, and decision classification accuracy of a testlet-based MST with combinations of amplification, cancellation, and LID-affected testlets present. To keep this study manageable, factors and variable levels were limited in scope. First, only one testlet was assigned to every module and there were no statistically independent items assigned to any of the MST modules. In practice, there are MST applications where combinations of testlets and individual items are used to construct modules. Under such

an MST module design structure, the impact of amplification and cancellation effects can be examined at the module level and not just the testlet level. Hence future studies could investigate MST performance in the presence of LID when both testlets and individual items cumulatively contribute to either amplification or cancellation to the module level.

Second, the assignment of testlets to modules in this study was fixed per panel throughout the administration of MST. That is, there was no opportunity for a non-panel testlet better matched to the ability of the test taker to be dynamically selected from a testlet pool and administered to that test taker. MST designs of this type are discussed in Yan, von Davier, and Lewis (2014). If test takers could be adaptively administered testlets that are better matched to their ability level, improvements in routing error rates may be realized under either of the models investigated in this study. Therefore, future studies could examine—under both models—the performance of an MST system routing system with dynamic testlet selection features when combinations of amplification, cancellation, and LID are assumed present.

Third, only the NC static scoring method was used to inform the MST routing decision procedure in this study. Research has found static and dynamic routing rules to be feasible alternatives to each other (Armstrong, 2002; Dallas, 2014; Zenisky, 2004). Still, Weismann (2014) noted that dynamic routing rules are more efficient at incorporating currently available performance information into routing decisions than NC static routing rules. Dallas (2016) found no significant differences in the MST routing accuracy when either NC or IRT-based scoring was used. However, Dallas (2016) investigated this effect under the assumption of no LID. When LID is present, a more precise interim measure of ability may be needed to ensure acceptable performance of the

testlet-based MST routing system under study. Future studies could compare MST routing system performance under static and dynamic routing methods to determine which performs best in the presence of LID effects.

Fourth, this study only examined fixed module difficulty differences between adjacent modules. Hence, future studies could vary the module difference at each test stage to determine the impact of the effect on the routing error rates. Moreover, future studies might also include investigations varying module difference and MST panel design effects to determine the extent that these combined effects may impact routing error rates.

Fifth, this study only examined the impact of study effects under one module selection method, the AMI method. Under conditions imposed in this study, the percentage of misrouted moderate ability test takers was higher than for low or high ability test takers. Unlike AMI, there are module selection methods that take the ability group membership of test takers into consideration when making routing decisions. For instance, under the DPI module selection method, a pre-determined proportion of test takers are routed to the different modules within the same test stage. As a result, test developers have control over the percent of routed test takers to next stage modules with ability grouping as a consideration. Hence, future studies could investigate the impact on routing error rates due to varying the module selection method under similar study conditions.

Appendices

Appendix I

Descriptive Statistics and ANOVA Tables

Bias

There are no significant interaction effects detected for bias

Table 7

Descriptive Statistics for Bias

Manipulated Factors	Factor Levels	Mean	Standard Deviation	Minimum	Maximum
Testlet Effect (TE)					
	No TE	.01	.19	-.39	.49
	Small	.01	.21	-.45	.46
	Large	.01	.23	-.49	.51
Module Difference					
	Small	.01	.21	-.46	.51
	Large	.01	.21	-.49	.51
Percent of IPD Affected Items					
	13%	.01	.21	-.44	.51
	20%	.01	.21	-.49	.51
IPD Affected Test Stage					
	Test Stage 1	.01	.21	-.49	.51
	Test Stage 2	.01	.21	-.42	.50
	Test Stage 3	.01	.21	-.42	.51
Model					
	2PLIRT	.01	.21	-.45	.51
	2PLTRT	.01	.21	-.49	.51
Magnitude					
	No IPD	.01	.21	-.41	.50
	.3	.01	.21	-.45	.50
	.5	.01	.21	-.49	.51
Direction					
	No IPD	.01	.21	-.41	.50
	Amplification	.01	.21	-.49	.51
	Cancellation	.01	.21	-.42	.51
IPD Parameter					
	ab	.01	.21	-.49	.51
	b	.01	.21	-.42	.50
	a	.01	.21	-.42	.51

Table 8

ANOVA Results for Bias

Source	df	F	p	η_p^2
Testlet Effect	2	.52	.60	.00
Module Difference	1	1.70	.19	.00
Percent of IPD Affected Items	1	.01	.94	.00
IPD Affected Test Stage	2	.57	.56	.00
Model	1	.58	.45	.00
Magnitude	1	.00	.98	.00
Direction	1	1.13	.29	.00
IPD Parameter	2	.16	.85	.00

Table 9

One-Way ANOVA Results for Bias

Source	df	F	p	η_p^2
IPD Type	12	.20	.99	.00

RMSE

There are no significant interaction effects detected for RMSE

Table 10

Descriptive Statistics for RMSE

Manipulated Factors	Factor Levels	Mean	Standard Deviation	Minimum	Maximum
Testlet Effect (TE)					
	No TE	.23	.18	.01	.57
	Small	.22	.17	.01	.54
	Large	.22	.16	.01	.55
Module Difference					
	Small	.22	.17	.01	.56
	Large	.22	.17	.01	.57
Percent of IPD Affected Items					
	13%	.22	.17	.01	.57
	20%	.22	.17	.01	.57
IPD Affected Test Stage					
	Test Stage 1	.22	.17	.01	.57
	Test Stage 2	.22	.17	.01	.57
	Test Stage 3	.22	.17	.01	.57
Model					
	2PLIRT	.23	.17	.01	.57
	2PLTRT	.21	.16	.01	.57
Magnitude					
	No IPD	.23	.17	.01	.56
	.3	.22	.17	.01	.57
	.5	.22	.17	.01	.57
Direction					
	No IPD	.23	.17	.01	.56
	Amplification	.22	.17	.01	.57
	Cancellation	.22	.17	.01	.56
IPD Parameter					
	ab	.22	.17	.01	.57
	b	.23	.17	.01	.56
	a	.22	.17	.01	.57

Table 11

ANOVA Results for RMSE

Source	df	F	p	η_p^2
Testlet Effect	2	14.40	.00	.00
Module Difference	1	1.42	.23	.00
Percent of IPD Affected Items	1	.02	.89	.00
IPD Affected Test Stage	2	.17	.84	.00
Model	1	56.37	.00	.00
Magnitude	1	1.23	.27	.00
Direction	1	.59	.44	.00
IPD Parameter	2	4.68	.01	.00

Table 12

One-Way ANOVA Results for RMSE

Source	df	F	p	η_p^2
IPD Type	12	4.93	1.00	.00

Standard Error

There are no significant interaction effects detected for standard error

Table 13

Descriptive Statistics for Standard Error

Manipulated Factors	Factor Levels	Mean	Standard Deviation	Minimum	Maximum
Testlet Effect (TE)					
	No TE	.19	.23	.00	1.16
	Small	.20	.26	.00	1.28
	Large	.25	.31	.00	1.34
Module Difference					
	Small	.21	.27	.00	1.34
	Large	.22	.27	.00	1.13
Model					
	2PLIRT	.24	.29	.00	1.31
	2PLTRT	.18	.24	.00	1.34
Pathway					
	125	.21	.28	.00	1.29
	126	.21	.26	.00	1.08
	135	.21	.26	.00	1.09
	136	.22	.27	.00	1.03
	137	.21	.27	.00	1.15
	146	.21	.26	.00	1.13
	147	.21	.28	.00	1.34

Table 14

ANOVA Results for Standard Error

Source	df	F	p	η_p^2
Model	1	35.06	.00	.01
Module Difference	1	.89	.35	.00
Testlet Effect	2	15.34	.00	.01
Pathway	6	.22	.97	.00

Correlation

There are no significant interaction effects detected for correlation

Table 15

Descriptive Statistics for Correlation

Manipulated Factors	Factor Levels	Mean	Standard Deviation	Minimum	Maximum
Testlet Effect (TE)					
	No TE	.56	.06	.49	.69
	Small	.63	.10	.53	.84
	Large	.71	.14	.17	.86
Module Difference					
	Small	.62	.14	.49	.86
	Large	.65	.15	.17	.86
Percent of IPD Affected Items					
	13%	.64	.15	.17	.86
	20%	.63	.14	.49	.85
IPD Affected Test Stage					
	Test Stage 1	.63	.14	.49	.86
	Test Stage 2	.64	.15	.50	.86
	Test Stage 3	.64	.15	.17	.83
Model					
	2PLIRT	.64	.15	.49	.86
	2PLTRT	.63	.14	.17	.86
Direction					
	No IPD	.62	.12	.50	.81
	Amplification	.64	.15	.17	.86
	Cancellation	.64	.14	.50	.86
Magnitude					
	No IPD	.62	.12	.50	.81
	.3	.65	.15	.17	.86
	.5	.62	.13	.49	.84
IPD Parameter					
	ab	.63	.14	.17	.83
	b	.65	.16	.50	.86
	a	.63	.14	.49	.84

Table 16

ANOVA Results for Correlation

Source	df	F	p	η_p^2
Testlet Effect	2	972.46	.00	.62
Module Difference	1	45.65	.00	.07
Percent of IPD Affected Items	1	4.10	.04	.01
IPD Affected Test Stage	2	6.79	.01	.01
Model	1	3.13	.08	.01
Direction	1	.15	.70	.00
Magnitude	1	36.67	.00	.06
IPD Parameter	2	.01	.93	.00
Testlet Effect*Module Difference*IPD Percent*IPD Test Stage	4	66.92	.00	.10
Testlet Effect*Module Difference*IPD Percent*IPD Test Stage*Model	4	91.20	.00	.13

Table 17

One-Way ANOVA Results for Correlation

Source	df	F	p	η_p^2
IPD Type	12	6.16	.00	.07

Table 18

Simple Effect Significant Results for Correlation

Factor	Interaction	df	F	p	η_p^2
IPD Affected Test Stage					
	2PLIRT*Large TE*Large MD*20% IPD	2	64.85	.00	.14
	2PLIRT*Large TE*Large MD*13% IPD	2	77.53	.00	.16
	2PLIRT*Large TE*Small MD*20% IPD	2	35.78	.00	.08
	2PLIRT*Large TE*Small MD*13% IPD	2	35.86	.00	.08
	2PLIRT*Small TE*Small MD*20% IPD	2	26.56	.00	.06
	2PLIRT*Small TE*Small MD*13% IPD	2	28.30	.00	.07
Percent of IPD Affected Items					
	2PLIRT*Large TE*Large MD*Test Stage 1	1	111.83	.00	.12
	2PLIRT*Large TE*Large MD*Test Stage 3	1	100.30	.00	.11
	2PLIRT*Large TE*Small MD*Test Stage 1	1	75.37	.00	.09
	2PLIRT*Large TE*Small MD*Test Stage 3	1	61.05	.00	.07
Module Difference					
	2PLIRT*Large TE*20% IPD*Test Stage 1	1	63.84	.00	.07
	2PLIRT*Large TE*20% IPD*Test Stage 2	1	4.88	.03	.01
	2PLIRT*Large TE*20% IPD*Test Stage 3	1	105.23	.00	.12
	2PLIRT*Large TE*13% IPD*Test Stage 1	1	126.94	.00	.14
	2PLIRT*Large TE*13% IPD*Test Stage 3	1	57.31	.00	.07
	2PLIRT*Small TE*13% IPD*Test Stage 1	1	48.90	.00	.06
Testlet Effect					
	2PLIRT*Large MD*20% IPD*Test Stage 1	2	73.08	.00	.16
	2PLIRT*Large MD*20% IPD*Test Stage 2	2	51.89	.00	.12
	2PLIRT*Large MD*13% IPD*Test Stage 1	2	36.20	.00	.08
	2PLIRT*Large MD*13% IPD*Test Stage 2	2	53.78	.00	.12
	2PLIRT*Large MD*13% IPD*Test Stage 3	2	81.03	.00	.17
	2PLIRT*Small MD*20% IPD*Test Stage 1	2	45.38	.00	.10
	2PLIRT*Small MD*20% IPD*Test Stage 2	2	48.28	.00	.11
	2PLIRT*Small MD*20% IPD*Test Stage 3	2	79.62	.00	.17
	2PLIRT*Small MD*13% IPD*Test Stage 1	2	79.57	.00	.17
	2PLIRT*Small MD*13% IPD*Test Stage 2	2	49.14	.00	.11
	2PLIRT*Small MD*13% IPD*Test Stage 3	2	31.39	.00	.07
	2PLTRT*Large MD*20% IPD*Test Stage 1	2	37.53	.00	.09
	2PLTRT*Large MD*20% IPD*Test Stage 2	2	46.16	.00	.10
	2PLTRT*Large MD*20% IPD*Test Stage 3	2	47.95	.00	.11
	2PLTRT*Large MD*13% IPD*Test Stage 1	2	42.65	.00	.10
	2PLTRT*Large MD*13% IPD*Test Stage 2	2	50.50	.00	.11
	2PLTRT*Large MD*13% IPD*Test Stage 3	2	28.03	.00	.07
	2PLTRT*Small MD*20% IPD*Test Stage 1	2	30.22	.00	.07
	2PLTRT*Small MD*20% IPD*Test Stage 2	2	37.06	.00	.09
	2PLTRT*Small MD*20% IPD*Test Stage 3	2	38.61	.00	.09
	2PLTRT*Small MD*13% IPD*Test Stage 1	2	21.99	.00	.05
	2PLTRT*Small MD*13% IPD*Test Stage 2	2	29.40	.00	.07
	2PLTRT*Small MD*13% IPD*Test Stage 3	2	37.73	.00	.09

Routing Error Rates

No no significant interaction, beyond two-way, effects detected for routing error rates

Table 19

Descriptive Statistics for Panel Routing Error Rate

Manipulated Factor Level	Mean	Standard Deviation	Minimum	Maximum
No Testlet Effect	.13	.05	.04	.42
Small Testlet Effect	.10	.05	.02	.34
Large Testlet Effect	.09	.04	.02	.24
Large Module Difference	.10	.05	.02	.42
Small Module Difference	.11	.05	.02	.34
13% IPD Affected Items	.11	.05	.02	.42
20% IPD Affected Items	.11	.05	.02	.28
IPD Affected Test Stage 1	.11	.05	.02	.42
IPD Affected Test Stage 2	.11	.05	.02	.29
IPD Affected Test Stage 3	.11	.05	.02	.34
2PLIRT	.11	.05	.02	.42
2PLTRT	.11	.05	.02	.34
Amplification	.11	.05	.02	.34
Cancellation	.11	.05	.02	.42
Magnitude: .3	.11	.05	.02	.28
Magnitude: .5	.11	.05	.02	.42
IPD Parameter: ab	.11	.05	.02	.27
IPD Parameter: b	.11	.05	.02	.42
IPD Parameter: a	.11	.05	.02	.26
Routing Error Stage: Stage1	.13	.05	.03	.42
Routing Error Stage: Stage2	.09	.04	.02	.27
No TE*Large MD	.09	.04	.02	.22
No TE*Small MD	.10	.05	.02	.24
No TE*0.13	.09	.04	.02	.22
No TE*0.2	.09	.04	.02	.24
No TE*Test Stage 1	.09	.04	.02	.24
No TE*Test Stage 2	.09	.05	.02	.22
No TE*Test Stage 3	.09	.04	.02	.23
No TE*2PLIRT	.09	.04	.02	.22
No TE*2PLTRT	.10	.04	.02	.24
No TE*Amplification	.09	.04	.02	.23
No TE*Cancellation	.09	.05	.02	.24
No TE*.3	.09	.04	.02	.21
No TE*.5	.10	.05	.02	.24

Table 19 continued

No TE*ab	.09	.04	.02	.22
No TE*b	.09	.04	.02	.24
No TE*a	.09	.04	.02	.21
No TE*Stage1	.11	.05	.03	.24
No TE*Stage2	.07	.03	.02	.14
Small TE*Large MD	.10	.04	.02	.28
Small TE*Small MD	.11	.05	.02	.34
Small TE*0.13	.11	.05	.03	.34
Small TE*0.2	.10	.05	.02	.26
Small TE*Test Stage 1	.10	.04	.03	.26
Small TE*Test Stage 2	.11	.05	.02	.28
Small TE*Test Stage 3	.11	.05	.02	.34
Small TE*2PLIRT	.10	.04	.02	.24
Small TE*2PLTRT	.11	.05	.03	.34
Small TE*Amplification	.10	.05	.02	.34
Small TE*Cancellation	.11	.04	.03	.26
Small TE*.3	.10	.04	.03	.28
Small TE*.5	.11	.05	.02	.34
Small TE*ab	.10	.04	.02	.24
TSmall TE*b	.11	.05	.03	.34
Small TE*a	.10	.04	.03	.24
Small TE*Stage1	.12	.05	.03	.34
Small TE*Stage2	.09	.03	.02	.25
Large TE*Large MD	.13	.05	.04	.42
Large TE*Large MD	.13	.05	.04	.29
Large TE*0.13	.13	.05	.04	.42
Large TE*0.2	.13	.05	.04	.28
Large TE*Test Stage 1	.13	.05	.04	.42
Large TE*Test Stage 2	.13	.05	.04	.29
Large TE*Test Stage 3	.13	.05	.04	.28
Large TE*2PLIRT	.13	.05	.04	.42
Large TE*2PLTRT	.13	.05	.04	.29
Large TE*Amplification	.13	.05	.04	.29
Large TE*Cancellation	.13	.05	.04	.42
Large TE*.3	.13	.05	.04	.26
Large TE*.5	.13	.05	.04	.42
Large TE*ab	.13	.05	.04	.27
Large TE*b	.14	.05	.04	.42
Large TE*a	.13	.05	.04	.26
Large TE*Stage1	.15	.05	.05	.42

Table 19 continued

Large TE*Stage2	.11	.04	.04	.27
Large MD*13% IPD	.10	.05	.02	.42
Large MD*20% IPD	.10	.05	.02	.27
Large MD*Test Stage 1	.10	.05	.03	.42
Large MD*Test Stage 2	.11	.05	.02	.28
Large MD*Test Stage 3	.10	.05	.02	.27
Large MD*2PLIRT	.10	.04	.02	.42
Large MD*2PLTRT	.11	.05	.03	.28
Large MD*Amplification	.10	.05	.02	.28
Large MD*Cancellation	.11	.05	.03	.42
Large MD*.3	.10	.04	.03	.28
Large MD*.5	.11	.05	.02	.42
Large MD*ab	.10	.04	.02	.23
Large MD*b	.11	.05	.03	.42
Large MD*a	.10	.04	.03	.23
Large MD*Stage1	.12	.05	.03	.42
Large MD*Stage2	.09	.04	.02	.27
Small MD*13% IPD	.11	.05	.02	.34
Small MD*20% IPD	.11	.05	.02	.28
Small MD*Test Stage 1	.11	.05	.02	.28
Small MD*Test Stage 2	.11	.05	.02	.29
Small MD*Test Stage 3	.12	.05	.02	.34
Small MD*2PLIRT	.11	.05	.02	.27
Small MD*2PLTRT	.12	.05	.02	.34
Small MD*Amplification	.11	.05	.02	.34
Small MD*Cancellation	.12	.05	.02	.28
Small MD*.3	.11	.05	.02	.26
Small MD*.5	.12	.05	.02	.34
Small MD*ab	.11	.05	.02	.27
Small MD*b	.12	.05	.02	.34
Small MD*a	.11	.05	.02	.26
Small MD*Stage1	.14	.05	.03	.34
Small MD*Stage2	.09	.04	.02	.25
20% IPD*Test Stage 1	.11	.05	.02	.42
20% IPD*Test Stage 2	.11	.05	.02	.29
20% IPD*Test Stage 3	.11	.05	.02	.34
20% IPD*2PLIRT	.10	.05	.02	.42
20% IPD*2PLTRT	.11	.05	.02	.34
20% IPD*Amplification	.11	.05	.02	.34
20% IPD*Cancellation	.11	.05	.02	.42

Table 19 continued

20% IPD*.3	.11	.05	.02	.28
20% IPD*.5	.11	.05	.02	.42
20% IPD*ab	.10	.05	.02	.25
20% IPD*b	.11	.05	.02	.42
20% IPD*a	.11	.05	.02	.26
20% IPD*Stage1	.13	.05	.03	.42
20% IPD*Stage2	.09	.04	.02	.27
13% IPD*Test Stage 1	.11	.05	.02	.27
13% IPD*Test Stage 2	.11	.05	.02	.27
13% IPD*Test Stage 3	.11	.05	.02	.28
13% IPD*2PLIRT	.11	.05	.02	.27
13% IPD*2PLTRT	.11	.05	.02	.28
13% IPD*Amplification	.11	.05	.02	.28
13% IPD*Cancellation	.11	.05	.02	.27
13% IPD*.3	.11	.05	.02	.26
13% IPD*.5	.11	.05	.02	.28
13% IPD*ab	.11	.05	.02	.27
13% IPD*b	.11	.05	.02	.28
13% IPD*a	.11	.05	.02	.26
13% IPD*Stage1	.13	.05	.03	.28
13% IPD*Stage2	.09	.04	.02	.22
Test Stage 1*2PLIRT	.10	.05	.02	.42
Test Stage 1*2PLTRT	.11	.05	.02	.28
Test Stage 1*Amplification	.10	.05	.02	.26
Test Stage 1*Cancellation	.11	.05	.02	.42
Test Stage 1*.3	.10	.05	.02	.26
Test Stage 1*.5	.11	.05	.02	.42
Test Stage 1*ab	.10	.05	.03	.25
Test Stage 1*b	.11	.05	.02	.42
Test Stage 1*a	.11	.05	.02	.26
Test Stage 1*Stage1	.13	.05	.03	.42
Test Stage 1*Stage2	.09	.04	.02	.27
Test Stage 2*2PLIRT	.11	.05	.02	.27
Test Stage 2*2PLTRT	.11	.05	.02	.29
Test Stage 2*Amplification	.11	.05	.02	.29
Test Stage 2*Cancellation	.11	.05	.02	.27
Test Stage 2*.3	.11	.05	.02	.28
Test Stage 2*.5	.11	.05	.02	.29
Test Stage 2*ab	.11	.05	.02	.27
Test Stage 2*b	.11	.05	.02	.29

Table 19 continued

Test Stage 2*a	.11	.05	.02	.25
Test Stage 2*Stage1	.13	.05	.03	.29
Test Stage 2*Stage2	.09	.04	.02	.22
Test Stage 3*2PLIRT	.11	.05	.02	.27
Test Stage 3*2PLTRT	.11	.05	.02	.34
Test Stage 3*Amplification	.11	.05	.02	.34
Test Stage 3*Cancellation	.11	.05	.02	.27
Test Stage 3*.3	.11	.05	.02	.26
Test Stage 3*.5	.11	.05	.02	.34
Test Stage 3*ab	.11	.05	.02	.27
Test Stage 3*b	.11	.05	.02	.34
Test Stage 3*a	.11	.05	.02	.25
Test Stage 3*Stage1	.13	.05	.03	.34
Test Stage 3*Stage2	.09	.04	.02	.25
2PLIRT*Amplification	.10	.05	.02	.25
2PLIRT*Cancellation	.11	.05	.02	.42
2PLIRT*.3	.10	.05	.02	.25
2PLIRT*.5	.11	.05	.02	.42
2PLIRT*ab	.10	.05	.02	.27
2PLIRT*b	.11	.05	.02	.42
2PLIRT*a	.11	.05	.02	.26
2PLIRT*Stage1	.12	.05	.03	.42
2PLIRT*Stage2	.09	.04	.02	.27
2PLTRT*Amplification	.11	.05	.02	.34
2PLTRT*Cancellation	.11	.05	.02	.28
2PLTRT*.3	.11	.05	.02	.28
2PLTRT*.5	.12	.05	.02	.34
2PLTRT*ab	.11	.05	.02	.26
2PLTRT*b	.12	.05	.02	.34
2PLTRT*a	.11	.05	.02	.26
2PLTRT*Stage1	.13	.05	.03	.34
2PLTRT*Stage2	.09	.04	.02	.25
Amplification*.3	.10	.05	.02	.28
Amplification*.5	.11	.05	.02	.34
Amplification*ab	.10	.05	.02	.25
Amplification*b	.11	.05	.02	.34
Amplification*a	.11	.05	.02	.25
Amplification*Stage1	.13	.05	.03	.34
Amplification*Stage2	.09	.04	.02	.25
Cancellation*.3	.11	.05	.02	.26

Table 19 continued

Cancellation*.5	.11	.05	.02	.42
Cancellation*ab	.11	.05	.02	.27
Cancellation*b	.11	.05	.03	.42
Cancellation*a	.11	.05	.02	.26
Cancellation*Stage1	.13	.05	.04	.42
Cancellation*Stage2	.09	.04	.02	.27
.3*ab	.10	.05	.02	.26
.3*b	.11	.05	.02	.28
.3*a	.11	.05	.02	.26
.3*Stage1	.12	.05	.04	.28
.3*Stage2	.09	.04	.02	.20
.5*ab	.11	.05	.02	.27
.5*b	.12	.06	.02	.42
.5*a	.11	.05	.02	.26
.5*Stage1	.13	.05	.03	.42
.5*Stage2	.09	.04	.02	.27
ab*Stage1	.13	.05	.03	.27
ab*Stage2	.09	.04	.02	.21
b*Stage1	.13	.05	.04	.42
b*Stage2	.09	.04	.02	.27
a*Stage1	.13	.05	.03	.26
a*Stage2	.09	.04	.02	.19

Table 20

ANOVA results for Panel Routing Error Rate

Source	df	F	p	η_p^2
Testlet Effect	2	818.80	.00	.12
Module Difference	1	180.71	.00	.01
Percent of IPD Affected Items	1	.37	.54	.00
Model	1	87.03	.00	.01
IPD Affected Test Stage	2	6.82	.00	.00
Routing Error Stage	1	2545.01	.00	.17
IPD Parameter	2	36.92	.00	.01
Magnitude	1	70.21	.00	.01
Direction	1	31.23	.00	.00
IPD Affected Test Stage * Direction	2	34.03	.00	.01
IPD Parameter * Direction	2	16.59	.00	.00
Magnitude * Direction	1	3.26	.07	.00
Model * Direction	1	2.83	.09	.00
Module Difference * Direction	1	2.95	.09	.00

Table 20 continued

Percent of IPD Affected Items * Direction	1	.01	.90	.00
Routing Error Stage * Direction	1	2.22	.14	.00
Testlet Effect * Direction	2	.64	.53	.00
IPD Affected Test Stage * IPD Parameter	4	1.20	.31	.00
IPD Affected Test Stage * Magnitude	2	.27	.76	.00
Model * IPD Affected Test Stage	2	.02	.98	.00
Module Difference * IPD Affected Test Stage	2	1.48	.23	.00
Percent of IPD Affected Items * IPD Affected Test Stage	2	.04	.96	.00
IPD Affected Test Stage * Routing Error Stage	2	1.38	.25	.00
Testlet Effect * IPD Affected Test Stage	4	.96	.43	.00
IPD Parameter * Magnitude	2	26.31	.00	.00
Model * IPD Parameter	2	34.58	.00	.01
Module Difference * IPD Parameter	2	.29	.75	.00
Percent of IPD Affected Items * IPD Parameter	2	3.61	.03	.00
Routing Error Stage * IPD Parameter	2	1.02	.36	.00
Testlet Effect * IPD Parameter	4	3.17	.01	.00
Model * Magnitude	1	30.85	.00	.00
Module Difference * Magnitude	1	1.42	.23	.00
Percent of IPD Affected Items * Magnitude	1	.13	.72	.00
Routing Error Stage * Magnitude	1	2.70	.10	.00
Testlet Effect * Magnitude	2	1.65	.19	.00
Module Difference * Model	1	9.42	.00	.00
Percent of IPD Affected Items * Model	1	4.77	.03	.00
Model * Routing Error Stage	1	2.89	.09	.00
Testlet Effect * Model	2	2.51	.08	.00
Module Difference * Percent of IPD Affected Items	1	.71	.40	.00
Module Difference * Routing Error Stage	1	118.83	.00	.01
Testlet Effect * Module Difference	2	1.05	.35	.00
Percent of IPD Affected Items * Routing Error Stage	1	.01	.93	.00
Testlet Effect * Percent of IPD Affected Items	2	.45	.64	.00
Testlet Effect * Routing Error Stage	2	19.50	.00	.00

Table 21

Simple Effects Results for Module Difference and Pathway Routing Error Rate Interaction

Manipulated Factor	Factor Level	df	F	p.	η_p^2
Pathway					
	125	1	92.81	.00	.00
	126	1	8.99	.00	.00
	135	1	1251.73	.00	.01
	136	1	348.32	.00	.00
	137	1	590.28	.00	.01
	146	1	1869.37	.00	.02
	147	1	2149.27	.00	.02
Module Difference					
	Small MD	6	4533.10	.00	.24
	Large MD	6	3599.53	.00	.20

Table 22

Simple Effects Results for Routing Error Stage and Pathway Routing Error Rate Interaction

Manipulated Factor	Factor Level	df	F	p.	η_p^2
Pathway					
	125	1	.02	.90	.00
	126	1	246.66	.00	.00
	135	1	36104.86	.00	.30
	136	1	.00	1.00	.00
	137	1	206.88	.00	.00
	146	1	1026.77	.00	.01
	147	1	.00	1.00	.00
Routing Error Stage					
	Stage1	6	9816.04	.00	.41
	Stage2	6	4884.27	.00	.26

Table 23

Simple Effects Results for Testlet Effect and Pathway Routing Error Rate Interaction

Manipulated Factor	Factor Level	df	F	p.	η_p^2
Pathway					
	125	2	292.54	.00	.01
	126	2	842.70	.00	.02
	135	2	80.78	.00	.00
	136	2	21.50	.00	.00
	137	2	53.56	.00	.00
	146	2	612.29	.00	.01
	147	2	1554.37	.00	.04
Testlet Effect					
	Large TE	6	2059.59	.00	.13
	Small TE	6	2652.48	.00	.16
	No TE	6	3069.81	.00	.18

Table 24

One-Way ANOVA Results for Panel Routing Error Rate

Source	df	F	p	η_p^2
IPD Type	12	19.05	.00	.02

Table 25

Descriptive Statistics for Pathway Routing Error Rate

Manipulated Factor Level	Mean	Standard Deviation	Minimum	Maximum
No Testlet Effect	.01	.01	.00	.11
Small Testlet Effect	.01	.01	.00	.15
Large Testlet Effect	.02	.01	.00	.19
Large Module Difference	.01	.01	.00	.19
Small Module Difference	.02	.01	.00	.15
13% IPD Affected Items	.02	.01	.00	.19
20% IPD Affected Items	.02	.01	.00	.11
IPD Affected Test Stage 1	.02	.01	.00	.19
IPD Affected Test Stage 2	.02	.01	.00	.10
IPD Affected Test Stage 3	.02	.01	.00	.15
2PLIRT	.02	.01	.00	.19
2PLTRT	.02	.01	.00	.15
Amplification	.02	.01	.00	.15
Cancellation	.02	.01	.00	.19
Magnitude: .3	.02	.01	.00	.10
Magnitude: .5	.02	.01	.00	.19
IPD Parameter: ab	.02	.01	.00	.11
IPD Parameter: b	.02	.01	.00	.19
IPD Parameter: a	.02	.01	.00	.10
Routing Error Stage: Stage1	.02	.02	.00	.19
Routing Error Stage: Stage2	.01	.01	.00	.10
Pathway 125	.01	.00	.00	.06
Pathway 126	.01	.01	.00	.05
Pathway 135	.02	.02	.00	.19
Pathway 136	.03	.01	.00	.08
Pathway 137	.01	.01	.00	.06
Pathway 146	.02	.01	.00	.10
Pathway 147	.02	.01	.00	.10
No TE*Large MD	.01	.01	.00	.10
No TE*Small MD	.01	.01	.00	.11
No TE*0.13	.01	.01	.00	.11
No TE*0.2	.01	.01	.00	.11
No TE*Test Stage 1	.01	.01	.00	.11
No TE*Test Stage 2	.01	.02	.00	.10
No TE*Test Stage 3	.01	.01	.00	.10
No TE*2PLIRT	.01	.01	.00	.10
No TE*2PLTRT	.01	.01	.00	.11
No TE*Amplification	.01	.01	.00	.10
No TE*Cancellation	.01	.02	.00	.11

Table 25 continued

No TE*.3	.01	.01	.00	.09
No TE*.5	.01	.02	.00	.11
No TE*ab	.01	.01	.00	.11
No TE*b	.01	.01	.00	.11
No TE*a	.01	.01	.00	.10
No TE*Stage1	.02	.02	.00	.11
No TE*Stage2	.01	.01	.00	.06
No TE*125	.00	.00	.00	.06
No TE*126	.00	.00	.00	.02
No TE*135	.02	.03	.00	.11
No TE*136	.03	.01	.00	.06
No TE*137	.01	.00	.00	.03
No TE*146	.02	.01	.00	.05
No TE*147	.01	.01	.00	.06
Small TE*Large MD	.01	.01	.00	.10
Small TE*Small MD	.02	.01	.00	.15
Small TE*0.13	.01	.01	.00	.15
Small TE*0.2	.02	.01	.00	.09
Small TE*Test Stage 1	.01	.01	.00	.09
Small TE*Test Stage 2	.02	.01	.00	.10
Small TE*Test Stage 3	.02	.01	.00	.15
Small TE*2PLIRT	.01	.01	.00	.09
Small TE*2PLTRT	.02	.01	.00	.15
Small TE*Amplification	.01	.01	.00	.15
Small TE*Cancellation	.02	.01	.00	.09
Small TE*.3	.01	.01	.00	.10
Small TE*.5	.02	.01	.00	.15
Small TE*ab	.01	.01	.00	.09
Small TE*b	.02	.01	.00	.15
Small TE*a	.01	.01	.00	.09
Small TE*Stage1	.02	.02	.00	.15
Small TE*Stage2	.01	.01	.00	.10
Small TE*125	.00	.00	.00	.02
Small TE*126	.01	.00	.00	.02
Small TE*135	.02	.02	.00	.15
Small TE*136	.03	.01	.00	.08
Small TE*137	.01	.01	.00	.06
Small TE*146	.02	.01	.00	.06
Small TE*147	.02	.01	.00	.10
Large TE*Large MD	.02	.01	.00	.19
Large TE*Large MD	.02	.01	.00	.08
Large TE*0.13	.02	.01	.00	.07

Table 25 continued

Large TE*0.2	.02	.01	.00	.19
Large TE*Test Stage 1	.02	.01	.00	.19
Large TE*Test Stage 2	.02	.01	.00	.08
Large TE*Test Stage 3	.02	.01	.00	.07
Large TE*2PLIRT	.02	.01	.00	.19
Large TE*2PLTRT	.02	.01	.00	.08
Large TE*Amplification	.02	.01	.00	.08
Large TE*Cancellation	.02	.01	.00	.19
Large TE*.3	.02	.01	.00	.07
Large TE*.5	.02	.01	.00	.19
Large TE*ab	.02	.01	.00	.08
Large TE*b	.02	.01	.00	.19
Large TE*a	.02	.01	.00	.07
Large TE*Stage1	.02	.01	.00	.19
Large TE*Stage2	.02	.01	.00	.09
Large TE*125	.01	.00	.00	.03
Large TE*126	.01	.01	.00	.05
Large TE*135	.02	.02	.00	.19
Large TE*136	.03	.01	.01	.07
Large TE*137	.01	.01	.00	.04
Large TE*146	.02	.01	.00	.10
Large TE*147	.03	.01	.00	.06
Large MD*13% IPD	.01	.01	.00	.09
Large MD*20% IPD	.01	.01	.00	.19
Large MD*Test Stage 1	.01	.01	.00	.19
Large MD*Test Stage 2	.02	.01	.00	.10
Large MD*Test Stage 3	.01	.01	.00	.08
Large MD*2PLIRT	.01	.01	.00	.19
Large MD*2PLTRT	.02	.01	.00	.10
Large MD*Amplification	.01	.01	.00	.10
Large MD*Cancellation	.02	.01	.00	.19
Large MD*.3	.01	.01	.00	.10
Large MD*.5	.02	.01	.00	.19
Large MD*ab	.01	.01	.00	.10
Large MD*b	.02	.01	.00	.19
Large MD*a	.01	.01	.00	.08
Large MD*Stage1	.02	.01	.00	.19
Large MD*Stage2	.01	.01	.00	.09
Large MD*125	.00	.00	.00	.06
Large MD*126	.01	.01	.00	.05
Large MD*135	.02	.02	.00	.19
Large MD*136	.03	.01	.00	.08

Table 25 continued

Large MD*137	.01	.00	.00	.03
Large MD*146	.02	.01	.00	.10
Large MD*147	.01	.01	.00	.06
Small MD*13% IPD	.02	.01	.00	.11
Small MD*20% IPD	.02	.01	.00	.15
Small MD*Test Stage 1	.02	.01	.00	.11
Small MD*Test Stage 2	.02	.01	.00	.10
Small MD*Test Stage 3	.02	.01	.00	.15
Small MD*2PLIRT	.02	.01	.00	.10
Small MD*2PLTRT	.02	.01	.00	.15
Small MD*Amplification	.02	.01	.00	.15
Small MD*Cancellation	.02	.01	.00	.11
Small MD*.3	.02	.01	.00	.09
Small MD*.5	.02	.01	.00	.15
Small MD*ab	.02	.01	.00	.11
Small MD*b	.02	.01	.00	.15
Small MD*a	.02	.01	.00	.10
Small MD*Stage1	.02	.02	.00	.15
Small MD*Stage2	.01	.01	.00	.10
Small MD*125	.01	.00	.00	.03
Small MD*126	.01	.01	.00	.04
Small MD*135	.03	.02	.00	.15
Small MD*136	.03	.01	.00	.06
Small MD*137	.01	.01	.00	.06
Small MD*146	.01	.01	.00	.05
Small MD*147	.02	.01	.00	.10
20% IPD*Test Stage 1	.02	.01	.00	.19
20% IPD*Test Stage 2	.02	.01	.00	.10
20% IPD*Test Stage 3	.02	.01	.00	.15
20% IPD*2PLIRT	.01	.01	.00	.19
20% IPD*2PLTRT	.02	.01	.00	.15
20% IPD*Amplification	.02	.01	.00	.15
20% IPD*Cancellation	.02	.01	.00	.19
20% IPD*.3	.02	.01	.00	.10
20% IPD*.5	.02	.01	.00	.19
20% IPD*ab	.01	.01	.00	.11
20% IPD*b	.02	.01	.00	.19
20% IPD*a	.02	.01	.00	.09
20% IPD*Stage1	.02	.02	.00	.19
20% IPD*Stage2	.01	.01	.00	.10
20% IPD*125	.01	.00	.00	.03
20% IPD*126	.01	.01	.00	.05

Table 25 continued

20% IPD*135	.02	.02	.00	.19
20% IPD*136	.03	.01	.00	.08
20% IPD*137	.01	.01	.00	.06
20% IPD*146	.02	.01	.00	.10
20% IPD*147	.02	.01	.00	.10
13% IPD*Test Stage 1	.02	.01	.00	.11
13% IPD*Test Stage 2	.02	.01	.00	.09
13% IPD*Test Stage 3	.02	.01	.00	.10
13% IPD*2PLIRT	.02	.01	.00	.10
13% IPD*2PLTRT	.02	.01	.00	.11
13% IPD*Amplification	.02	.01	.00	.10
13% IPD*Cancellation	.02	.01	.00	.11
13% IPD*.3	.02	.01	.00	.09
13% IPD*.5	.02	.01	.00	.11
13% IPD*ab	.02	.01	.00	.10
13% IPD*b	.02	.01	.00	.11
13% IPD*a	.02	.01	.00	.10
13% IPD*Stage1	.02	.02	.00	.11
13% IPD*Stage2	.01	.01	.00	.07
13% IPD*125	.00	.00	.00	.06
13% IPD*126	.01	.01	.00	.04
13% IPD*135	.02	.02	.00	.11
13% IPD*136	.03	.01	.00	.06
13% IPD*137	.01	.01	.00	.04
13% IPD*146	.02	.01	.00	.07
13% IPD*147	.02	.01	.00	.06
Test Stage 1*2PLIRT	.01	.01	.00	.19
Test Stage 1*2PLTRT	.02	.01	.00	.11
Test Stage 1*Amplification	.01	.01	.00	.08
Test Stage 1*Cancellation	.02	.02	.00	.19
Test Stage 1*.3	.01	.01	.00	.09
Test Stage 1*.5	.02	.01	.00	.19
Test Stage 1*ab	.01	.01	.00	.11
Test Stage 1*b	.02	.01	.00	.19
Test Stage 1*a	.02	.01	.00	.10
Test Stage 1*Stage1	.02	.02	.00	.19
Test Stage 1*Stage2	.01	.01	.00	.09
Test Stage 1*125	.01	.00	.00	.03
Test Stage 1*126	.01	.01	.00	.05
Test Stage 1*135	.02	.02	.00	.19
Test Stage 1*136	.03	.01	.00	.07
Test Stage 1*137	.01	.01	.00	.04

Table 25 continued

Test Stage 1*146	.02	.01	.00	.10
Test Stage 1*147	.02	.01	.00	.06
Test Stage 2*2PLIRT	.02	.01	.00	.10
Test Stage 2*2PLTRT	.02	.01	.00	.10
Test Stage 2*Amplification	.02	.01	.00	.10
Test Stage 2*Cancellation	.02	.01	.00	.08
Test Stage 2*.3	.02	.01	.00	.10
Test Stage 2*.5	.02	.01	.00	.10
Test Stage 2*ab	.02	.01	.00	.10
Test Stage 2*b	.02	.01	.00	.10
Test Stage 2*a	.02	.01	.00	.09
Test Stage 2*Stage1	.02	.02	.00	.10
Test Stage 2*Stage2	.01	.01	.00	.08
Test Stage 2*125	.00	.00	.00	.02
Test Stage 2*126	.01	.01	.00	.03
Test Stage 2*135	.02	.02	.00	.10
Test Stage 2*136	.03	.01	.00	.08
Test Stage 2*137	.01	.00	.00	.04
Test Stage 2*146	.02	.01	.00	.07
Test Stage 2*147	.02	.01	.00	.06
Test Stage 3*2PLIRT	.02	.01	.00	.09
Test Stage 3*2PLTRT	.02	.01	.00	.15
Test Stage 3*Amplification	.02	.01	.00	.15
Test Stage 3*Cancellation	.02	.01	.00	.09
Test Stage 3*.3	.02	.01	.00	.08
Test Stage 3*.5	.02	.01	.00	.15
Test Stage 3*ab	.02	.01	.00	.09
Test Stage 3*b	.02	.01	.00	.15
Test Stage 3*a	.02	.01	.00	.08
Test Stage 3*Stage1	.02	.02	.00	.15
Test Stage 3*Stage2	.01	.01	.00	.10
Test Stage 3*125	.00	.00	.00	.06
Test Stage 3*126	.01	.01	.00	.04
Test Stage 3*135	.02	.02	.00	.15
Test Stage 3*136	.03	.01	.01	.06
Test Stage 3*137	.01	.01	.00	.06
Test Stage 3*146	.02	.01	.00	.07
Test Stage 3*147	.02	.01	.00	.10
2PLIRT*Amplification	.01	.01	.00	.10
2PLIRT*Cancellation	.02	.01	.00	.19
2PLIRT*.3	.01	.01	.00	.09
2PLIRT*.5	.02	.01	.00	.19

Table 25 continued

2PLIRT*ab	.01	.01	.00	.10
2PLIRT*b	.02	.01	.00	.19
2PLIRT*a	.02	.01	.00	.10
2PLIRT*Stage1	.02	.02	.00	.19
2PLIRT*Stage2	.01	.01	.00	.09
2PLIRT*125	.01	.00	.00	.03
2PLIRT*126	.01	.01	.00	.05
2PLIRT*135	.02	.02	.00	.19
2PLIRT*136	.03	.01	.00	.06
2PLIRT*137	.01	.01	.00	.04
2PLIRT*146	.02	.01	.00	.10
2PLIRT*147	.02	.01	.00	.06
2PLTRT*Amplification	.02	.01	.00	.15
2PLTRT*Cancellation	.02	.01	.00	.11
2PLTRT*.3	.02	.01	.00	.10
2PLTRT*.5	.02	.01	.00	.15
2PLTRT*ab	.02	.01	.00	.11
2PLTRT*b	.02	.02	.00	.15
2PLTRT*a	.02	.01	.00	.09
2PLTRT*Stage1	.02	.02	.00	.15
2PLTRT*Stage2	.01	.01	.00	.10
2PLTRT*125	.00	.00	.00	.06
2PLTRT*126	.01	.01	.00	.04
2PLTRT*135	.02	.02	.00	.15
2PLTRT*136	.03	.01	.01	.08
2PLTRT*137	.01	.01	.00	.06
2PLTRT*146	.02	.01	.00	.07
2PLTRT*147	.02	.01	.00	.10
Amplification*.3	.01	.01	.00	.10
Amplification*.5	.02	.01	.00	.15
Amplification*ab	.01	.01	.00	.10
Amplification*b	.02	.01	.00	.15
Amplification*a	.02	.01	.00	.09
Amplification*Stage1	.02	.02	.00	.15
Amplification*Stage2	.01	.01	.00	.10
Amplification*125	.01	.00	.00	.06
Amplification*126	.01	.01	.00	.04
Amplification*135	.02	.02	.00	.15
Amplification*136	.03	.01	.00	.08
Amplification*137	.01	.01	.00	.06
Amplification*146	.02	.01	.00	.07
Amplification*147	.02	.01	.00	.10

Table 25 continued

Cancellation*.3	.02	.01	.00	.09
Cancellation*.5	.02	.01	.00	.19
Cancellation*ab	.02	.01	.00	.11
Cancellation*b	.02	.01	.00	.19
Cancellation*a	.02	.01	.00	.10
Cancellation*Stage1	.02	.02	.00	.19
Cancellation*Stage2	.01	.01	.00	.09
Cancellation*125	.00	.00	.00	.02
Cancellation*126	.01	.01	.00	.05
Cancellation*135	.02	.02	.00	.19
Cancellation*136	.03	.01	.01	.06
Cancellation*137	.01	.01	.00	.04
Cancellation*146	.02	.01	.00	.10
Cancellation*147	.02	.01	.00	.06
.3*ab	.01	.01	.00	.09
.3*b	.02	.01	.00	.10
.3*a	.02	.01	.00	.09
.3*Stage1	.02	.02	.00	.10
.3*Stage2	.01	.01	.00	.08
.3*125	.00	.00	.00	.02
.3*126	.01	.01	.00	.03
.3*135	.02	.02	.00	.10
.3*136	.03	.01	.00	.08
.3*137	.01	.00	.00	.03
.3*146	.02	.01	.00	.06
.3*147	.02	.01	.00	.06
.5*ab	.02	.01	.00	.11
.5*b	.02	.02	.00	.19
.5*a	.02	.01	.00	.10
.5*Stage1	.02	.02	.00	.19
.5*Stage2	.01	.01	.00	.10
.5*125	.01	.00	.00	.06
.5*126	.01	.01	.00	.05
.5*135	.02	.02	.00	.19
.5*136	.03	.01	.00	.06
.5*137	.01	.01	.00	.06
.5*146	.02	.01	.00	.10
.5*147	.02	.01	.00	.10
ab*Stage1	.02	.02	.00	.11
ab*Stage2	.01	.01	.00	.07
ab*125	.01	.00	.00	.06
ab*126	.01	.01	.00	.04

Table 25 continued

ab*135	.02	.02	.00	.11
ab*136	.03	.01	.00	.06
ab*137	.01	.01	.00	.04
ab*146	.02	.01	.00	.07
ab*147	.02	.01	.00	.06
b*Stage1	.02	.02	.00	.19
b*Stage2	.01	.01	.00	.10
b*125	.01	.00	.00	.02
b*126	.01	.01	.00	.05
b*135	.02	.02	.00	.19
b*136	.03	.01	.01	.08
b*137	.01	.01	.00	.06
b*146	.02	.01	.00	.10
b*147	.02	.01	.00	.10
a*Stage1	.02	.02	.00	.10
a*Stage2	.01	.01	.00	.06
a*125	.00	.00	.00	.02
a*126	.01	.01	.00	.03
a*135	.02	.02	.00	.10
a*136	.03	.01	.00	.06
a*137	.01	.01	.00	.03
a*146	.02	.01	.00	.06
a*147	.02	.01	.00	.06

Table 26

ANOVA Results for Pathway Routing Error Rates

Source	df	F	p	η_p^2
Testlet Effect	2	1732.79	.00	.04
Module Difference	1	382.42	.00	.00
Percent of IPD Affected Items	1	.79	.37	.00
Model	1	184.18	.00	.00
IPD Affected Test Stage	2	14.44	.00	.00
Routing Error Stage	1	5385.89	.00	.06
IPD Parameter	2	78.14	.00	.00
Magnitude	1	148.58	.00	.00
Direction	1	66.09	.00	.00
Pathway	6	7417.23	.00	.34
IPD Affected Test Stage * Direction	2	120.88	.00	.00
IPD Parameter * Direction	2	58.93	.00	.00
Magnitude * Direction	1	11.58	.00	.00
Model * Direction	1	10.03	.00	.00
Module Difference * Direction	1	10.49	.00	.00
Direction * Pathway	6	100.26	.00	.01
Percent of IPD Affected Items * Direction	1	.05	.82	.00
Routing Error Stage * Direction	1	7.87	.01	.00
Testlet Effect * Direction	2	2.26	.10	.00
IPD Affected Test Stage * IPD Parameter	4	4.25	.00	.00
IPD Affected Test Stage * Magnitude	2	.95	.39	.00
Model * IPD Affected Test Stage	2	.06	.94	.00
Module Difference * IPD Affected Test Stage	2	5.26	.01	.00
IPD Affected Test Stage * Pathway	12	56.55	.00	.01
Percent of IPD Affected Items * IPD Affected Test Stage	2	.14	.87	.00
IPD Affected Test Stage * Routing Error Stage	2	4.91	.01	.00
Testlet Effect * IPD Affected Test Stage	4	3.42	.01	.00
IPD Parameter * Magnitude	2	93.43	.00	.00
Model * IPD Parameter	2	122.82	.00	.00
Module Difference * IPD Parameter	2	1.02	.36	.00
IPD Parameter * Pathway	12	19.80	.00	.00
Percent of IPD Affected Items * IPD Parameter	2	12.81	.00	.00
Routing Error Stage * IPD Parameter	2	3.63	.03	.00
Testlet Effect * IPD Parameter	4	11.26	.00	.00
Model * Magnitude	1	109.58	.00	.00
Module Difference * Magnitude	1	5.05	.02	.00
Magnitude * Pathway	6	7.76	.00	.00
Percent of IPD Affected Items * Magnitude	1	.46	.50	.00
Routing Error Stage * Magnitude	1	9.59	.00	.00

Table 26 continued

Testlet Effect * Magnitude	2	5.87	.00	.00
Module Difference * Model	1	33.44	.00	.00
Model * Pathway	6	177.35	.00	.01
Percent of IPD Affected Items * Model	1	16.96	.00	.00
Model * Routing Error Stage	1	10.28	.00	.00
Testlet Effect * Model	2	8.90	.00	.00
Module Difference * Pathway	6	1787.96	.00	.11
Module Difference * Percent of IPD Affected Items	1	2.51	.11	.00
Module Difference * Routing Error Stage	1	422.06	.00	.00
Testlet Effect * Module Difference	2	3.73	.02	.00
Percent of IPD Affected Items * Pathway	6	2.17	.04	.00
Routing Error Stage * Pathway	6	6854.95	.00	.33
Testlet Effect * Pathway	12	531.54	.00	.07
Percent of IPD Affected Items * Routing Error Stage	1	.03	.87	.00
Testlet Effect * Percent of IPD Affected Items	2	1.61	.20	.00
Testlet Effect * Routing Error Stage	2	69.27	.00	.00

Table 27

Two-Way ANOVA Results for Pathway Routing Error Rate

Source	df	F	p	η_p^2
IPDType	12	50.07	.00	.01
Pathway	6	7388.90	.00	.33
IPDType * Pathway	72	14.86	.00	.01

Misclassification Rates

There are no significant interaction, beyond two-way interaction, effects detected for Misclassification Rates

Table 28

Descriptive Statistics for Pathway Misclassification Rate

Manipulated Factor Level	Mean	Standard Deviation	Minimum	Maximum
No Testlet Effect	.06	.07	.00	.23
Small Testlet Effect	.07	.08	.00	.28
Large Testlet Effect	.07	.08	.00	.30
Large Module Difference	.07	.08	.00	.30
Small Module Difference	.07	.07	.00	.26
13% IPD Affected Items	.07	.08	.00	.30
20% IPD Affected Items	.07	.08	.00	.30
IPD Affected Test Stage 1	.07	.08	.00	.30
IPD Affected Test Stage 2	.07	.08	.00	.29
IPD Affected Test Stage 3	.07	.08	.00	.30
2PLIRT	.07	.08	.00	.30
2PLTRT	.07	.07	.00	.30
Amplification	.07	.08	.00	.30
Cancellation	.07	.08	.00	.30
Magnitude: .3	.07	.08	.00	.30
Magnitude: .5	.07	.07	.00	.29
IPD Parameter: ab	.07	.08	.00	.29
IPD Parameter: b	.07	.08	.00	.30
IPD Parameter: a	.07	.08	.00	.30
Pathway 125	.04	.06	.00	.27
Pathway 126	.12	.05	.00	.26
Pathway 135	.05	.06	.00	.28
Pathway 136	.12	.07	.00	.30
Pathway 137	.14	.07	.00	.30
Pathway 146	.14	.07	.00	.30
Pathway 147	.04	.07	.00	.30
Ability Group: Low	.06	.06	.00	.23
Ability Group: Moderate	.18	.05	.03	.30
Ability Group: High	.03	.05	.00	.21
No TE*Large MD	.06	.07	.00	.23
No TE*Small MD	.06	.07	.00	.22
No TE*0.13	.06	.07	.00	.22
No TE*0.2	.06	.07	.00	.23

Table 28 continued

No TE*Test Stage 1	.06	.07	.00	.23
No TE*Test Stage 2	.06	.07	.00	.22
No TE*Test Stage 3	.06	.07	.00	.23
No TE*2PLIRT	.07	.07	.00	.23
No TE*2PLTRT	.06	.06	.00	.22
No TE*Amplification	.06	.07	.00	.23
No TE*Cancellation	.06	.07	.00	.23
No TE*.3	.06	.07	.00	.22
No TE*.5	.06	.07	.00	.23
No TE*ab	.06	.07	.00	.23
No TE*b	.06	.07	.00	.23
No TE*a	.06	.07	.00	.23
No TE*125	.04	.06	.00	.23
No TE*126	.12	.06	.00	.23
No TE*135	.05	.06	.00	.22
No TE*136	.10	.06	.00	.23
No TE*137	.10	.06	.00	.22
No TE*146	.10	.06	.00	.22
No TE*147	.03	.06	.00	.22
No TE*Low	.05	.05	.00	.17
No TE*Moderate	.17	.03	.07	.23
No TE*High	.02	.04	.00	.18
Small TE*Large MD	.07	.08	.00	.28
Small TE*Small MD	.07	.07	.00	.26
Small TE*0.13	.07	.08	.00	.28
Small TE*0.2	.07	.08	.00	.28
Small TE*Test Stage 1	.07	.08	.00	.28
Small TE*Test Stage 2	.07	.08	.00	.28
Small TE*Test Stage 3	.07	.08	.00	.28
Small TE*2PLIRT	.07	.08	.00	.28
Small TE*2PLTRT	.07	.08	.00	.28
Small TE*Amplification	.07	.08	.00	.28
Small TE*Cancellation	.07	.08	.00	.28
Small TE*.3	.07	.08	.00	.28
Small TE*.5	.07	.08	.00	.28
Small TE*ab	.07	.08	.00	.28
Small TE*b	.07	.08	.00	.28
Small TE*a	.07	.08	.00	.28
Small TE*125	.03	.05	.00	.27
Small TE*126	.12	.04	.00	.26
Small TE*135	.05	.06	.00	.25
Small TE*136	.14	.07	.00	.28

Table 28 continued

Small TE*137	.15	.06	.00	.28
Small TE*146	.15	.06	.00	.28
Small TE*147	.05	.07	.00	.28
Small TE*Low	.06	.06	.00	.20
Small TE*Moderate	.19	.05	.06	.28
Small TE*High	.03	.05	.00	.18
Large TE*Large MD	.07	.09	.00	.30
Large TE*Large MD	.07	.07	.00	.26
Large TE*0.13	.07	.08	.00	.30
Large TE*0.2	.07	.08	.00	.30
Large TE*Test Stage 1	.07	.08	.00	.30
Large TE*Test Stage 2	.07	.08	.00	.29
Large TE*Test Stage 3	.07	.08	.00	.30
Large TE*2PLIRT	.07	.08	.00	.30
Large TE*2PLTRT	.07	.08	.00	.30
Large TE*Amplification	.07	.08	.00	.30
Large TE*Cancellation	.07	.08	.00	.30
Large TE*.3	.07	.08	.00	.30
Large TE*.5	.07	.08	.00	.29
Large TE*ab	.07	.08	.00	.29
Large TE*b	.07	.08	.00	.30
Large TE*a	.07	.08	.00	.30
Large TE*125	.03	.05	.00	.26
Large TE*126	.13	.03	.04	.24
Large TE*135	.05	.06	.00	.28
Large TE*136	.13	.07	.00	.30
Large TE*137	.17	.06	.03	.30
Large TE*146	.17	.06	.03	.30
Large TE*147	.05	.08	.00	.30
Large TE*Low	.06	.06	.00	.23
Large TE*Moderate	.19	.06	.03	.30
Large TE*High	.03	.06	.00	.21
Large MD*13% IPD	.07	.08	.00	.30
Large MD*20% IPD	.07	.08	.00	.30
Large MD*Test Stage 1	.07	.08	.00	.30
Large MD*Test Stage 2	.07	.08	.00	.29
Large MD*Test Stage 3	.07	.08	.00	.30
Large MD*2PLIRT	.07	.08	.00	.30
Large MD*2PLTRT	.07	.08	.00	.30
Large MD*Amplification	.07	.08	.00	.30
Large MD*Cancellation	.07	.08	.00	.30
Large MD*.3	.07	.08	.00	.30

Table 28 continued

Large MD*.5	.07	.08	.00	.29
Large MD*ab	.07	.08	.00	.29
Large MD*b	.07	.08	.00	.30
Large MD*a	.07	.08	.00	.30
Large MD*125	.04	.06	.00	.27
Large MD*126	.13	.05	.00	.26
Large MD*135	.06	.06	.00	.28
Large MD*136	.13	.07	.00	.30
Large MD*137	.15	.08	.00	.30
Large MD*146	.13	.07	.00	.30
Large MD*147	.05	.08	.00	.30
Large MD*Low	.06	.06	.00	.23
Large MD*Moderate	.19	.05	.06	.30
Large MD*High	.02	.04	.00	.20
Small MD*13% IPD	.07	.07	.00	.26
Small MD*20% IPD	.07	.07	.00	.26
Small MD*Test Stage 1	.07	.07	.00	.26
Small MD*Test Stage 2	.07	.07	.00	.26
Small MD*Test Stage 3	.07	.07	.00	.26
Small MD*2PLIRT	.07	.07	.00	.26
Small MD*2PLTRT	.06	.07	.00	.26
Small MD*Amplification	.07	.07	.00	.26
Small MD*Cancellation	.07	.07	.00	.26
Small MD*.3	.07	.07	.00	.26
Small MD*.5	.07	.07	.00	.26
Small MD*ab	.07	.07	.00	.26
Small MD*b	.07	.07	.00	.26
Small MD*a	.07	.07	.00	.26
Small MD*125	.04	.05	.00	.21
Small MD*126	.11	.05	.00	.22
Small MD*135	.05	.05	.00	.22
Small MD*136	.12	.06	.00	.26
Small MD*137	.13	.06	.00	.26
Small MD*146	.14	.05	.00	.26
Small MD*147	.04	.07	.00	.26
Small MD*Low	.06	.05	.00	.19
Small MD*Moderate	.17	.04	.03	.26
Small MD*High	.03	.05	.00	.21
20% IPD*Test Stage 1	.07	.08	.00	.30
20% IPD*Test Stage 2	.07	.08	.00	.29
20% IPD*Test Stage 3	.07	.08	.00	.29
20% IPD*2PLIRT	.07	.08	.00	.30

Table 28 continued

20% IPD*2PLTRT	.07	.07	.00	.30
20% IPD*Amplification	.07	.08	.00	.30
20% IPD*Cancellation	.07	.08	.00	.29
20% IPD*.3	.07	.08	.00	.30
20% IPD*.5	.07	.08	.00	.29
20% IPD*ab	.07	.08	.00	.29
20% IPD*b	.07	.08	.00	.29
20% IPD*a	.07	.08	.00	.30
20% IPD*125	.04	.06	.00	.26
20% IPD*126	.12	.05	.00	.24
20% IPD*135	.05	.06	.00	.28
20% IPD*136	.12	.07	.00	.30
20% IPD*137	.14	.07	.00	.30
20% IPD*146	.14	.07	.00	.30
20% IPD*147	.04	.07	.00	.29
20% IPD*Low	.06	.06	.00	.23
20% IPD*Moderate	.18	.05	.03	.30
20% IPD*High	.03	.05	.00	.21
13% IPD*Test Stage 1	.07	.08	.00	.30
13% IPD*Test Stage 2	.07	.08	.00	.29
13% IPD*Test Stage 3	.07	.08	.00	.30
13% IPD*2PLIRT	.07	.08	.00	.29
13% IPD*2PLTRT	.07	.07	.00	.30
13% IPD*Amplification	.07	.08	.00	.29
13% IPD*Cancellation	.07	.08	.00	.30
13% IPD*.3	.07	.08	.00	.30
13% IPD*.5	.07	.07	.00	.29
13% IPD*ab	.07	.08	.00	.29
13% IPD*b	.07	.08	.00	.30
13% IPD*a	.07	.08	.00	.29
13% IPD*125	.04	.06	.00	.27
13% IPD*126	.12	.05	.00	.26
13% IPD*135	.05	.06	.00	.23
13% IPD*136	.12	.07	.00	.29
13% IPD*137	.14	.07	.00	.29
13% IPD*146	.14	.07	.00	.29
13% IPD*147	.04	.07	.00	.30
13% IPD*Low	.06	.06	.00	.21
13% IPD*Moderate	.18	.05	.03	.30
13% IPD*High	.03	.05	.00	.21
Test Stage 1*2PLIRT	.07	.08	.00	.30
Test Stage 1*2PLTRT	.07	.07	.00	.30

Table 28 continued

Test Stage 1*Amplification	.07	.08	.00	.30
Test Stage 1*Cancellation	.07	.08	.00	.30
Test Stage 1*.3	.07	.08	.00	.30
Test Stage 1*.5	.07	.08	.00	.29
Test Stage 1*ab	.07	.07	.00	.29
Test Stage 1*b	.07	.08	.00	.30
Test Stage 1*a	.07	.08	.00	.30
Test Stage 1*125	.04	.06	.00	.25
Test Stage 1*126	.12	.05	.00	.25
Test Stage 1*135	.05	.06	.00	.28
Test Stage 1*136	.12	.07	.00	.30
Test Stage 1*137	.14	.07	.00	.30
Test Stage 1*146	.14	.07	.00	.30
Test Stage 1*147	.04	.07	.00	.30
Test Stage 1*Low	.06	.06	.00	.23
Test Stage 1*Moderate	.18	.05	.03	.30
Test Stage 1*High	.03	.05	.00	.21
Test Stage 2*2PLIRT	.07	.08	.00	.29
Test Stage 2*2PLTRT	.06	.07	.00	.29
Test Stage 2*Amplification	.07	.07	.00	.29
Test Stage 2*Cancellation	.07	.08	.00	.29
Test Stage 2*.3	.07	.08	.00	.29
Test Stage 2*.5	.07	.07	.00	.29
Test Stage 2*ab	.07	.07	.00	.29
Test Stage 2*b	.07	.08	.00	.29
Test Stage 2*a	.07	.08	.00	.29
Test Stage 2*125	.04	.06	.00	.26
Test Stage 2*126	.12	.05	.00	.26
Test Stage 2*135	.05	.06	.00	.25
Test Stage 2*136	.12	.07	.00	.29
Test Stage 2*137	.14	.07	.00	.29
Test Stage 2*146	.13	.06	.00	.29
Test Stage 2*147	.04	.07	.00	.29
Test Stage 2*Low	.06	.06	.00	.18
Test Stage 2*Moderate	.18	.05	.04	.29
Test Stage 2*High	.03	.05	.00	.21
Test Stage 3*2PLIRT	.07	.08	.00	.29
Test Stage 3*2PLTRT	.07	.07	.00	.30
Test Stage 3*Amplification	.07	.08	.00	.29
Test Stage 3*Cancellation	.07	.08	.00	.30
Test Stage 3*.3	.07	.08	.00	.30
Test Stage 3*.5	.07	.08	.00	.29

Table 28 continued

Test Stage 3*ab	.07	.08	.00	.29
Test Stage 3*b	.07	.08	.00	.30
Test Stage 3*a	.07	.07	.00	.29
Test Stage 3*125	.04	.06	.00	.27
Test Stage 3*126	.12	.05	.00	.24
Test Stage 3*135	.05	.06	.00	.25
Test Stage 3*136	.13	.07	.00	.29
Test Stage 3*137	.14	.07	.00	.28
Test Stage 3*146	.14	.07	.00	.29
Test Stage 3*147	.05	.07	.00	.30
Test Stage 3*Low	.06	.06	.00	.18
Test Stage 3*Moderate	.19	.05	.04	.30
Test Stage 3*High	.02	.05	.00	.21
2PLIRT*Amplification	.07	.08	.00	.30
2PLIRT*Cancellation	.07	.08	.00	.29
2PLIRT*.3	.07	.08	.00	.30
2PLIRT*.5	.07	.08	.00	.29
2PLIRT*ab	.07	.08	.00	.29
2PLIRT*b	.07	.08	.00	.29
2PLIRT*a	.07	.08	.00	.30
2PLIRT*125	.04	.06	.00	.27
2PLIRT*126	.13	.05	.00	.26
2PLIRT*135	.06	.06	.00	.22
2PLIRT*136	.13	.07	.00	.30
2PLIRT*137	.14	.07	.00	.29
2PLIRT*146	.14	.07	.00	.29
2PLIRT*147	.05	.07	.00	.29
2PLIRT*Low	.06	.06	.00	.23
2PLIRT*Moderate	.19	.05	.06	.30
2PLIRT*High	.03	.05	.00	.21
2PLTRT*Amplification	.06	.07	.00	.30
2PLTRT*Cancellation	.07	.07	.00	.30
2PLTRT*.3	.07	.07	.00	.30
2PLTRT*.5	.06	.07	.00	.29
2PLTRT*ab	.06	.07	.00	.29
2PLTRT*b	.07	.07	.00	.30
2PLTRT*a	.06	.07	.00	.30
2PLTRT*125	.04	.05	.00	.25
2PLTRT*126	.11	.05	.00	.25
2PLTRT*135	.05	.06	.00	.28
2PLTRT*136	.12	.06	.00	.29
2PLTRT*137	.14	.07	.00	.30

Table 28 continued

2PLTRT*146	.13	.07	.00	.30
2PLTRT*147	.04	.07	.00	.30
2PLTRT*Low	.05	.06	.00	.21
2PLTRT*Moderate	.18	.05	.03	.30
2PLTRT*High	.03	.05	.00	.21
Amplification*.3	.07	.08	.00	.30
Amplification*.5	.07	.07	.00	.29
Amplification*ab	.07	.08	.00	.29
Amplification*b	.07	.07	.00	.29
Amplification*a	.07	.07	.00	.30
Amplification*125	.04	.06	.00	.25
Amplification*126	.12	.05	.00	.26
Amplification*135	.06	.06	.00	.28
Amplification*136	.13	.06	.00	.30
Amplification*137	.14	.07	.00	.30
Amplification*146	.14	.07	.00	.30
Amplification*147	.04	.07	.00	.29
amplification*Low	.06	.06	.00	.23
amplification*Moderate	.18	.05	.04	.30
amplification*High	.03	.05	.00	.21
Cancellation*.3	.07	.08	.00	.30
Cancellation*.5	.07	.08	.00	.29
Cancellation*ab	.07	.07	.00	.29
Cancellation*b	.07	.08	.00	.30
Cancellation*a	.07	.08	.00	.29
Cancellation*125	.04	.06	.00	.27
Cancellation*126	.12	.05	.00	.25
Cancellation*135	.05	.06	.00	.25
Cancellation*136	.12	.07	.00	.29
Cancellation*137	.14	.07	.00	.29
Cancellation*146	.14	.07	.00	.29
Cancellation*147	.04	.07	.00	.30
cancellation*Low	.06	.06	.00	.18
cancellation*Moderate	.19	.05	.03	.30
cancellation*High	.02	.05	.00	.20
.3*ab	.07	.08	.00	.29
.3*b	.07	.08	.00	.30
.3*a	.07	.08	.00	.30
.3*125	.04	.06	.00	.27
.3*126	.12	.05	.00	.26
.3*135	.05	.06	.00	.25
.3*136	.13	.07	.00	.30

Table 28 continued

.3*137	.14	.07	.00	.30
.3*146	.14	.07	.00	.30
.3*147	.04	.07	.00	.30
.3*Low	.06	.06	.00	.21
.3*Moderate	.19	.05	.03	.30
.3*High	.03	.05	.00	.21
.5*ab	.07	.07	.00	.29
.5*b	.07	.07	.00	.29
.5*a	.07	.08	.00	.29
.5*125	.03	.05	.00	.25
.5*126	.12	.05	.00	.25
.5*135	.05	.06	.00	.28
.5*136	.12	.06	.00	.29
.5*137	.14	.07	.00	.29
.5*146	.13	.06	.00	.29
.5*147	.04	.07	.00	.29
.5*Low	.06	.06	.00	.23
.5*Moderate	.18	.05	.03	.29
.5*High	.03	.05	.00	.21
ab*125	.04	.06	.00	.25
ab*126	.12	.05	.00	.25
ab*135	.05	.06	.00	.23
ab*136	.13	.07	.00	.29
ab*137	.14	.07	.00	.28
ab*146	.14	.07	.00	.29
ab*147	.04	.07	.00	.29
ab*Low	.06	.06	.00	.23
ab*Moderate	.18	.05	.03	.29
ab*High	.03	.05	.00	.21
b*125	.04	.06	.00	.27
b*126	.12	.05	.00	.24
b*135	.05	.06	.00	.25
b*136	.12	.06	.00	.29
b*137	.14	.07	.00	.29
b*146	.14	.07	.00	.29
b*147	.04	.07	.00	.30
b*Low	.06	.06	.00	.21
b*Moderate	.18	.05	.05	.30
b*High	.03	.05	.00	.21
a*125	.04	.06	.00	.25
a*126	.12	.05	.00	.26
a*135	.05	.06	.00	.28

Table 28 continued

a*136	.12	.07	.00	.30
a*137	.14	.07	.00	.30
a*146	.14	.07	.00	.30
a*147	.04	.07	.00	.29
a*Low	.06	.06	.00	.19
a*Moderate	.19	.05	.04	.30
a*High	.02	.05	.00	.21

Table 29

ANOVA Results for Pathway Misclassification Rates

Source	df	F	p	η_p^2
Testlet Effect	2	193.73	.00	.01
Module Difference	1	109.69	.00	.00
Percent of IPD Items	1	.01	.92	.00
Model	1	99.71	.00	.00
IPD Affected Test Stage	2	1.22	.30	.00
Pathway	6	1267.47	.00	.18
Direction	1	1.06	.30	.00
IPD Parameter	2	1.42	.24	.00
Magnitude	1	8.46	.00	.00
Ability Group	2	14553.66	.00	.45
Direction * Ability Group	2	30.57	.00	.00
IPD Affected Test Stage * Ability Group	4	8.35	.00	.00
IPD Parameter * Ability Group	4	2.69	.03	.00
Magnitude * Ability Group	2	3.75	.02	.00
Model * Ability Group	2	50.86	.00	.00
Module Difference * Ability Group	2	148.73	.00	.01
Pathway * Ability Group	12	269.58	.00	.08
Percent of IPD Items * Ability Group	2	.09	.91	.00
Testlet Effect * Ability Group	4	18.15	.00	.00
IPD Affected Test Stage * Direction	2	5.81	.00	.00
Direction * IPD Parameter	2	26.92	.00	.00
Direction * Magnitude	1	.13	.72	.00
Model * Direction	1	.09	.76	.00
Module Difference * Direction	1	1.74	.19	.00
Pathway * Direction	6	2.43	.02	.00
Percent of IPD Items * Direction	1	.00	.98	.00
Testlet Effect * Direction	2	.81	.44	.00
IPD Affected Test Stage * IPD Parameter	4	2.20	.07	.00
IPD Affected Test Stage * Magnitude	2	.46	.63	.00
Model * IPD Affected Test Stage	2	.08	.92	.00
Module Difference * IPD Affected Test Stage	2	2.95	.05	.00
IPD Affected Test Stage * Pathway	12	1.49	.12	.00
Percent of IPD Items * IPD Affected Test Stage	2	.16	.85	.00
Testlet Effect * IPD Affected Test Stage	4	.49	.74	.00

Table 29 continued

IPD Parameter * Magnitude	2	6.85	.00	.00
Model * IPD Parameter	2	3.68	.03	.00
Module Difference * IPD Parameter	2	.40	.67	.00
Pathway * IPD Parameter	12	.62	.83	.00
Percent of IPD Items * IPD Parameter	2	.78	.46	.00
Testlet Effect * IPD Parameter	4	.10	.98	.00
Model * Magnitude	1	4.38	.04	.00
Module Difference * Magnitude	1	.40	.53	.00
Pathway * Magnitude	6	.28	.94	.00
Percent of IPD Items * Magnitude	1	.11	.74	.00
Testlet Effect * Magnitude	2	.56	.57	.00
Module Difference * Model	1	.35	.55	.00
Model * Pathway	6	12.62	.00	.00
Percent of IPD Items * Model	1	.34	.56	.00
Testlet Effect * Model	2	.42	.66	.00
Module Difference * Pathway	6	21.28	.00	.00
Module Difference * Percent of IPD Items	1	.00	.96	.00
Testlet Effect * Module Difference	2	4.30	.01	.00
Percent of IPD Items * Pathway	6	.57	.75	.00
Testlet Effect * Pathway	12	40.93	.00	.01
Testlet Effect * Percent of IPD Items	2	1.13	.32	.00

Table 30

Two-Way ANOVA Results for Pathway Misclassification Rates

Source	df	F	p	η_p^2
IPDType	12	6.48	.00	.00
Pathway	6	2309.49	.00	.27
IPDType * Pathway	72	1.40	.01	.00

Table 31

Simple Effects Results for Ability Group and Pathway Misclassification Rate Interaction

Manipulated Factor	Factor Level	df	F	p	
Pathway					
	125	2	1538.20	.00	.08
	126	2	377.23	.00	.02
	135	2	778.61	.00	.04
	136	2	1502.15	.00	.08
	137	2	1111.12	.00	.06
	146	2	1223.29	.00	.06
	147	2	10703.69	.00	.38
Abiiity Group					
	Low	6	1283.11	.00	.18
	Moderate	6	56.20	.00	.01
	High	6	446.88	.00	.07

Appendix II

Sample MST R Code

```
### 1-3-3 MST Under 2PLTRT and 2PLIRT With LID, Amplification, Cancellation #
#############################################################################
##Define Variables
#############################################################################
library(psych)
library(sirt)
nTestlets                                  #Number of testlets in a pathway
nModules                                   # Number of modules per panel
NitemsPerModule                           #Number of items per testlet
nStages                                    #Number of MST test stages
nPanels                                   # Number of MST panels
nDifferenceConditions                      # Number of module difference levels
uThetas                                   # Number of unique thetas
TestLength                                 # Number of items per test (per pathway)
teffect                                   # small testlet effect magnitude
DriftStage                                 # set the test stage to impose IPD
nPersons                                 # number of calibration sample
nVarianceConditions                        # number of testlet effect levels
replications

#############################################################################
##Define Models
#############################################################################

#### 2PLTRT  #####
TRTModel2 <- function(b,a,theta,t)
{R_TRT <- 1/(1+exp(-a*(theta+t-b)))
return(R_TRT)
}

#####2PLIRT  #####
IRTModel2 <- function(b,a,theta)
{R_IRT <- 1/(1+exp(-a*(theta-b)))
return(R_IRT)
}

#############################################################################
##Administer the MST
#############################################################################

##Simulate Test Takers ###
```

```
w<-seq(-3,3,by=.15)
w<-rep(w,5)
theta.list = as.matrix(w)
T<-nrow(theta.list)
theta_outfile <- paste("Theta_List.dat")
write.table(theta.list,theta_outfile,quote=FALSE,sep=" ",row.names=FALSE,
col.names=FALSE)

###Set matrices

for (r in 1:replication)
{

rm(TestAdministered,ResponsesForCalibration,ModulebyStage,NCbyStage)
gc()

TestAdministered <- replicate(nStages,matrix(99,nrow=T,ncol=NitemsPerModule+1))
TestAdministered <- replicate(13, TestAdministered ) ### probabilities associated with
each drift condition
TestAdministered <- replicate(nVarianceConditions, TestAdministered )
TestAdministered <- replicate(nDifferenceConditions, TestAdministered )
TestAdministered <- replicate(2, TestAdministered ) ##13% and 20% drift conditions
TestAdministered <- replicate(2, TestAdministered )
TestAdministered <- replicate(3, TestAdministered )
TestAdministered <- replicate(3, TestAdministered ) ## to capture the 3 parameter
estimates (b,a,testlet effect)

ModulebyStage <- replicate(13,matrix(0,nrow=T,ncol=nStages))## for the 3 possible
drift affected test stages
ModulebyStage <- replicate(nVarianceConditions, ModulebyStage)
ModulebyStage <- replicate(nDifferenceConditions, ModulebyStage)
ModulebyStage <- replicate(2, ModulebyStage)
ModulebyStage <- replicate(2, ModulebyStage)
ModulebyStage <- replicate(3, ModulebyStage) # 3 possible test stages where drift is
administered
ModulebyStage[,1,,,,,,]<-1

NCbyStage <- replicate(13, matrix(0,nrow=T,ncol=nStages))
NCbyStage <- replicate(nVarianceConditions, NCbyStage)
NCbyStage <- replicate(nDifferenceConditions, NCbyStage)
NCbyStage <- replicate(2, NCbyStage)
NCbyStage <- replicate(2, NCbyStage)
NCbyStage <- replicate(3, NCbyStage)

ResponseProb <- replicate(nStages,matrix(99,nrow=T,ncol=NitemsPerModule))
```

```
ResponseProb <- replicate(13, ResponseProb) ### probabilities associated with each drift condition
ResponseProb<- replicate(nVarianceConditions, ResponseProb)
ResponseProb<- replicate(nDifferenceConditions, ResponseProb)
ResponseProb<- replicate(2, ResponseProb)## 13% and 20% drift conditions
ResponseProb<- replicate(2, ResponseProb)
ResponseProb<- replicate(3, ResponseProb)

Responses <- replicate(nStages,matrix(99,nrow=T,ncol=NitemsPerModule))
Responses <- replicate(13, Responses) ### probabilities associated with each drift condition
Responses <- replicate(nVarianceConditions, Responses)
Responses <- replicate(nDifferenceConditions, Responses)
Responses <- replicate(2, Responses) ##13% and 20% of test items drift conditions
Responses <- replicate(2, Responses)
Responses <- replicate(3, Responses)

tscnr<-matrix(0,nrow=nrow(theta.list),ncol=NitemsPerModule)
tscnr<-replicate(nModules, tscnr)
tscnr<-replicate(nVarianceConditions, tscnr)
tscnr<-replicate(nPanels, tscnr)
tscnr<-replicate(nDifferenceConditions, tscnr)
tscnr<-replicate(2, tscnr)
tscnr<-replicate(2, tscnr)

for (b in 1:2)
{
for (t in 1:2)
{
for (d in 1:nDifferenceConditions)
{
for (p in 1:nPanels)
{
for (j in 1:T)
{
tn1<- as.matrix((rnorm( nModules, 0, 1 )))
tn2<- as.matrix((rnorm( nModules, 0, .5 )))


tn1<-as.matrix(rep(tn1, each=NitemsPerModule), ncol=nModules*NitemsPerModule)
tn1<-t(tn1)
tn2<-as.matrix(rep(tn2, each=NitemsPerModule), ncol=nModules*NitemsPerModule)
tn2<-t(tn2)


n<-1
```

```
y<-NitemsPerModule
for ( m in 1:nModules)
{
k<-1
for ( i in n:y)
{
tscnr[j,k,m,1,p,d,t,b]<-round(tn1[1,i], digits=3)
tscnr[j,k,m,2,p,d,t,b]<-round(tn2[1,i], digits=3)
k<-k+1
}
n<-n+NitemsPerModule
y<-y+NitemsPerModule
}
}
}
}
}
}

for (b in 1:2)
{
for (d in 1:nDifferenceConditions)
{
for (c in 1:nVarianceConditions)
{
for (j in 1:T)
{
theta<-theta.list[j]

##select a random panel to administer ###
n<-as.matrix(sample(1:nPanels))
k<-1
h<-n[k,1]
while (Panel_Gen[h,1,c,d,b,r]>50)
{k<-k+1
h<-n[k,1]}
Panel_Gen[h,1,c,d,b,r]<-Panel_Gen[h,1,c,d,b,r]+1
p<-h


i<-1
f<-1
s<-1

for (Z in 1:3)
{
```

```
DriftStage<-Z


for (s in 1:nStages)
{

for (f in 1:13)
{

for (i in 1:NitemsPerModule)
{

for (t in 1:2)
{

if (s == DriftStage)
{
d1<-drift[i,1,f,t]
d2<-drift[i,2,f,t]
}

if (s!= DriftStage)
{
d1<-0
d2<-0
}
bb1<-Test_Bank[i,1,ModulebyStage[j,s,f,c,d,t,b,DriftStage],p,c,d,b] + d1
aa1<-Test_Bank[i,2,ModulebyStage[j,s,f,c,d,t,b,DriftStage],p,c,d,b] + d2
tsnr1<-tscnr[j,i,ModulebyStage[j,s,f,c,d,t,b,DriftStage],c,p,d,t,b]

ResponseProb[j,i,s,f,c,d,t,b,DriftStage] =TRTModel2(bb1,aa1,theta,tsnr1)
rini<-runif(1)
if(rini>ResponseProb[j,i,s,f,c,d,t,b,DriftStage])
{
Responses[j,i,s,f,c,d,t,b,DriftStage]<-0
TestAdministered[j,11,s,f,c,d,t,b,DriftStage,1]<-p
TestAdministered[j,i,s,f,c,d,t,b,DriftStage,1]<-bb1-d1
TestAdministered[j,i,s,f,c,d,t,b,DriftStage,2]<-aa1-d2
TestAdministered[j,i,s,f,c,d,t,b,DriftStage,3]<-tsnr1}
if(rini<=ResponseProb[j,i,s,f,c,d,t,b,DriftStage])
{
Responses[j,i,s,f,c,d,t,b,DriftStage]<-1
TestAdministered[j,11,s,f,c,d,t,b,DriftStage,1]<-p
TestAdministered[j,i,s,f,c,d,t,b,DriftStage,1]<-bb1-d1
TestAdministered[j,i,s,f,c,d,t,b,DriftStage,2]<-aa1-d2
TestAdministered[j,i,s,f,c,d,t,b,DriftStage,3]<-tsnr1}
```

```
}

}
}

##Calculate number correct score

f<-1
for (f in 1:13)
{

xa1<-Responses[j,,s,f,c,d,1,b,DriftStage]
xa2<-Responses[j,,s,f,c,d,2,b,DriftStage]

ModuleNC131<-sum(xa1)
ModuleNC132<-sum(xa2)

NCbyStage[j,s,f,c,d,1,b,DriftStage]<-ModuleNC131
NCbyStage[j,s,f,c,d,2,b,DriftStage]<-ModuleNC132

if (s != 3)
{
y21<-NCbyStage[j,,f,c,d,1,b,DriftStage]
y22<-NCbyStage[j,,f,c,d,2,b,DriftStage]

CumulativeNC21<-sum(y21)
CumulativeNC22<-sum(y22)

CumulativeNC21<-CumulativeNC21+1
CumulativeNC22<-CumulativeNC22+1


ModulebyStage[j,s+1,f,c,d,1,b,DriftStage]<-
Route[CumulativeNC21,2,ModulebyStage[j,s,f,c,d,1,b,DriftStage],p,c,d,b] ### module to
route to for next stage
ModulebyStage[j,s+1,f,c,d,2,b,DriftStage]<-
Route[CumulativeNC22,2,ModulebyStage[j,s,f,c,d,2,b,DriftStage],p,c,d,b] ### module to
route to for next stage

}
}
}
}
}
}
}
```

```
}


#############################################################################
##Generate Theta Estimates
#############################################################################

# define testlets
testlets <- rep(1:nTestlets , each=NitemsPerModule  )
burnin <- 5000
iter <- 10000

####Calculate IRT estimates ###

if (b == 1)
{

mod2PL_IRT_Theta <- mcmc.3pno.testlet( dat=Respondbyf , est.slope=TRUE ,
est.guess=FALSE , burnin=burnin, iter=iter , save.theta = TRUE )

summary(mod2PL_IRT_Theta)

summmod2PL_IRT_Theta <- mod2PL_IRT_Theta$summary.mcmcobj

TMI_List<-summmod2PL_IRT_Theta[ "Mean" ]
TMIRhat_List<-summmod2PL_IRT_Theta[ "Rhat" ]

Parameter_List<-as.matrix(cbind(TMI_List[2:61,1],TMIRhat_List[2:61,1]))
Theta_List<-
as.matrix(cbind(TMI_List[62:nrow(TMI_List),1],TMIRhat_List[62:nrow(TMI_List),1]))


IRT_outfile_Theta <- paste(r,f,c,d,t,b,Z,"_MCMCTheta_irt_estimates.txt")
write.table(Theta_List,IRT_outfile_Theta,quote=FALSE,sep=" ",row.names=FALSE,
col.names=FALSE)


IRT_outfile_Parameters <- paste(r,f,c,d,t,b,Z,"_MCMCParameters_irt_estimates.txt")
write.table(Parameter_List,IRT_outfile_Parameters ,quote=FALSE,sep="
",row.names=FALSE, col.names=FALSE)
rm(Respondbyf,mod2PL_IRT_Theta,summmod2PL_IRT_Theta,TMI_List,Parameter_Li
st,Theta_List,TMIRhat_List)
gc()
}
```

```
### Calculate TRT estimates ###

if (b == 2)
{

mod2PL_TRT_Theta <- mcmc.3pno.testlet( dat=Respondbyf , est.slope=TRUE ,
est.guess=FALSE , burnin=burnin, iter=iter , testlets= testlets, save.theta = TRUE )

summary(mod2PL_TRT_Theta)

summmod2PL_TRT_Theta <- mod2PL_TRT_Theta$summary.mcmcobj

TM_List<-summmod2PL_TRT_Theta[ "Mean" ]
TM_List<-as.matrix(TM_List)

TMRhat_List<-summmod2PL_TRT_Theta[ "Rhat" ]

ParameterT_List<-as.matrix(cbind(TM_List[2:61,1],TMRhat_List[2:61,1]))
ThetaT_List<-
as.matrix(cbind(TM_List[62:nrow(TM_List),1],TMRhat_List[62:nrow(TM_List),1]))


TRT_outfile_Theta <- paste(r,f,c,d,t,b,Z,"_MCMCTheta_trt_estimates.txt")
write.table(ThetaT_List,TRT_outfile_Theta,quote=FALSE,sep=" ",row.names=FALSE,
col.names=FALSE)


TRT_outfile_Parameters <- paste(r,f,c,d,t,b,Z,"_MCMCParameters_trt_estimates.txt")
write.table(ParameterT_List,TRT_outfile_Parameters ,quote=FALSE,sep="
",row.names=FALSE, col.names=FALSE)


rm(Respondbyf,mod2PL_TRT_Theta,summmod2PL_TRT_Theta,TM_List,ParameterT_
List,ThetaT_List,TMRhat_List)

gc()

}

}
```

Bibliography

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.

Angoff, W. (2012). Perspectives on differential item functioning and methodology. In Holland & Wainer, *Differential Item Functioning* (Chapter 1). Routledge.

Armstrong, R. D. (2002). *Routing rules for multiple-form structures* (Computerized Testing Report No. 02-08.) Newtown, PA: Law School Admission Council.

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.

Baker, F. B. & Kim, S.H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Ballou, D., & Springer, M. (2015). Teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44(2),* 77-86.

Ban, J., Hanson, B., Wang, T., Yi, Q., & Harris, D. (2001). A comparative study of on-line pretest item calibration-scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38,* 191-212.

Bao, H. (2007). *Investigating differential item function amplification and cancellation in application of item response testlet models* (Order No. 3283409). Available from ProQuest Dissertations & Theses Global. (304850930). Retrieved from http://search.proquest.com/docview/304850930?accountid=14696.

Bao, H., Dayton, C. M., & Hendrickson, A. B. (2009). Differential item functioning amplification and cancellation in a reading test. *Practical Assessment, Research & Evaluation*, *14*(19), 2.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223.

Birnbaum, 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* (pp. 397-479). Reading, MA: Addison-Wesley.

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

Bock, R., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

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

Bock, R. D. (1997). A brief history of item theory response. *Educational measurement: issues and practice, 16*(4), 21-33.

Bock, R., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25(4),* 275-285.

Bock, R. D., & Zimowski, M. F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP*. American Institutes for Research in the Behavioral Sciences.

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* (Order No. 3110732). Available from ProQuest Dissertations & Theses Global. (305295445). Retrieved from http://search.proquest.com/docview/305295445?accountid=14696.

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

Breithaupt , K., Ariel, A., & Veldkamp, B. (2005). Automated Simultaneous Assembly for Multistage Testing, *International Journal of Testing, 5(3),* 319-330, DOI: 10.1207/s15327574ijt0503_8.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics , 7* (4), 434-455.

Brossman, B. G., & Guille, R. A. (2014). A Comparison of Multi-Stage and Linear Test Designs for Medium-Size Licensure and Certification Examinations. *Journal of Computerized Adaptive Testing, 2*(3), 18-36.

Bryant, R., & Jiao, H. (2016, April). *Investigating IPD Amplification and Cancellation at the Testlet-level on Model Parameter Estimation*. Presented at the annual meeting of the National Council on Measurement in Education, Washington D.C.

Buyske, S. (1998). Optimal design for item calibration in computerized adaptive testing: The 2PL case. *New Developments and Applications in Experimental Design, 34,* 115-125.

Cappaert, K. (2014). *Dissecting the Impact of DIF/DBF on Ability Estimation and Person Fit*. (Order No. 3684871). Available from ProQuest Dissertations & Theses Global. (1660972693). Retrieved from http://search.proquest.com/docview/1660972693?accountid=14696.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*(4), 610.

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

Chen, J. (2014). *Model selection for IRT equating of testlet-based tests in the random groups design* (Order No. 3680050). Available from ProQuest Dissertations & Theses Global. (1652862787). Retrieved from http://search.proquest.com/docview/1652862787?accountid=14696.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Cohen, J. (1969). *Statistical power analysis for behavioral sciences*. New York: Academic Press.

Cook, L., Eignor, D. & Taft, H. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25(1),* 31-45.

Cook, L., & Eignor, D. (1991). IRT equating methods. *Educational measurement: Issues and practice, 10*(3), 37-45.

Cronbach , L., & Gleser, G. (1965). *Psychological tests and personnel decisions*. Oxford, England. University of Illinois Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Curtis, S. McKay. 2010. "BUGS Code for Item Response Theory." Journal of Statistical Software 36(1).

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The Consequences of Student Testing for Teaching and Teacher Quality. *Yearbook of the National Society for the Study of Education (Wiley-Blackwell),* 104(2), 289-319. doi:10.1111/j.1744-7984.2005.00034.x .

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* (Order No. 3624184). Available from ProQuest Dissertations & Theses Global. (1553002150). Retrieved from http://search.proquest.com/docview/1553002150?accountid=14696.

Davey, T., & Lee, Y. H. (2011). Potential impact of context effects on the scoring and equating of the multistage GRE® revised General Test. *ETS Research Report Series*, *2011*(2), i-44.

de Ayala, R. (2009). *The Theory and Practice of Item Response Theory*. Guilford Publications, NY .

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*(3), 265-300.

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. Retrieved from http://www.jstor.org/stable/20461818.

Deng, H., & Melican, G. (2009). *An investigation of scale drift in computer adaptive test*. Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Donoghue, J. R., & Isham, S. P. (1996). Comparing the Effectiveness of Procedures to Detect Item Parameter Drift. *ETS Research Report Series*,*1996*(2), i-46.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-Bundle DIF Hypothesis Testing: Identifying Suspect Bundles and Assessing Their Differential Functioning. *Journal of Educational Measurement*, *33*(4), 465-484.

Eckes, T. (2013). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. Language Testing, OnlineFirst, published on July 11, 2013 as doi: 10.1177/0265532213492969.

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

Finch, W. (2011). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement 71(4)*, p663–683.

Gibbons , R., & Hedeker, D. (1992) Full-information item bi-factor analysis, *Psychometrika, 57(3)*, 423-436.

Gierl, M. J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, *19*(2-3), 188-203.

Glas, G. A. (2000). Item calibration and parameter drift. In *Computerized adaptive testing: Theory and practice* (pp. 183-199). Springer Netherlands.

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

Goldstein, H. (1983). Measuring Change in Educational Attainment over Time: Problems and Possibilities. *Journal of Educational Measurement*, 20, 369-377.

Haberman, S., & von Davier, A. (2014). Considerations on item parameter estimation, scoring, and linking in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 15). New York, NY: CRC Press.

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

Hambleton, R. K., Jones, R.W, & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30(2),* 143-155. Retrieved from http://www.jstor.org/stable/1435461.

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., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices, *Journal of Applied Testing Technology, 1(1),* 1-13.

Han, K. & Kosinski, M. (2014). Software tools for multistage testing simulations. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 26). New York, NY: CRC Press.

Han , K., Wells, C., & Sireci, S. (2012) The Impact of Multidirectional Item Parameter Drift on IRT Scaling Coefficients and Proficiency Estimates. *Applied Measurement in Education, 25:2,* 97-117, DOI: 10.1080/08957347.2012.660000.

Han, K. T., & Wells, C. S. (2007). *Impact of differential item functioning (DIF) on test equating and proficiency estimates*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Han, K. T., & Guo, F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing. *GMAC Research Reports, RR-11, 2*.

Han, K. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 8). New York, NY: CRC Press.

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (Eds.). (2012). *Differential item functioning*. Routledge. New York, New York.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Jiao, H., Kamata, A., Wang, S., & Jin, Y.(2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49, 82-100. doi: 10.1111/j.1745-3984.2011.00161.x.

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., & Chen, Y. (2014). Differential item and testlet functioning analysis. In Kunnan, A., *The Companion to Language Assessment, First Edition* (Chapter 76). John Wiley & Sons, Inc.

Jin, K. & Wang, W. (2105). Item response theory models for carry-over effect across different scales. *Applied Psychological Measurement, 39(5),* p 406–425.

Jodoin, M. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40(1),* 1-15.

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

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38, 79-93.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50(1),* 1-73. doi:10.1111/jedm.12000.

Keller, L. (2000). *Ability estimation procedures in computerized adaptive testing*. Technical report from the American Institute of Certified Public Accountants Research Consortium. Retrieved from http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/TechnicalReports/DownloadableDocuments/Keller_AbilityEstimation.pdf.

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

Kim, S., & Moses, T. (2014). An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study. *ETS Research Report Series*, *2014*(1), 1-13.

Kim, J., Chung, H., Park, R., & Dodd, B. G. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior research methods*, *45*(4), 1087-1098.

Kim, J., Chung, H., & Dodd, B. G. (2010). *Comparing routing methods in the multistage test based on the partial credit model*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Kim, H. & Plake, B. (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, 2, 359-375. Doi:10.1207/s15324818ame0204_6.

Kim, J. & Dodd, B. (2014). Mixed format multistage tests: Issues and methods. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 4). New York, NY: CRC Press.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212-228.

Kim, S., & Moses, T. (2014). An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study. *ETS Research Report Series*, *2014*(1), 1-13.

Kim, S., Moses, T., & Yoo, H. H. (2015). Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing. *ETS Research Report Series*, *2015*(1), 1-19.

Kingsbury, G. (2009). *Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test*. In the Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Kolen, M. & Brennan, R. (2004). *Test equating, scaling, and linking: methods and practices*, 2nd edition, Springer New York.

Lee, Y., Lewis, C., & von Davier, A. (2014). Test security and quality control for multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 5). New York, NY: CRC Press.

Li, D. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design* (Order No. 3359398). Available from ProQuest Dissertations & Theses Global. (304918545). Retrieved from http://search.proquest.com/docview/304918545?accountid=14696.

Li, Y., Bolt, D. M., & Fu, J. (2006). A Comparison of Alternative Models for Testlets. *Applied Psychological Measurement, 30,* 1, 3-21.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions. *Psicológica: Revista de metodología y psicología experimental*, *30*(2), 343-370.

Lissitz, R. W., & Jiao, H. (2012). *Computers and their impact on state assessment: Recent history and predictions for the future*. Charlotte: Information Age Publishing Inc.

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

Lord, F. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23(2),* 157-162. Retrieved from http://www.jstor.org/stable/1434513.

Lu, R. (2010). *Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions* (Order No. 3443478). Available from ProQuest Dissertations & Theses Global. (854982968). Retrieved from http://search.proquest.com/docview/854982968?accountid=14696.

Luecht, R. (2014). Design and implementation of large-scale multistage testing systems. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 5). New York, NY: CRC Press.

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 on Measurement in Education (NCME), New York.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229–249.

Luecht, R. M. , Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19, 189-202. doi:10.1207/s15324818ame193_2.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique and future directions (with discussion), *Statistics in Medicine* 28, 3049--3082.

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

Melican, G., & Deng, H. (2009). An Investigation of Scale Drift in Computer Adaptive Test Paper presented at the annual meeting of the National Council on Measurement in Education, .

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13, 127–143.

Min, S. & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing 2014, 31(4),* 453–477.

Mislevy, R. & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement, 49(2),* p148-166.

Murphy, D., Dodd, B., Vaughn, B. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement, 34(6),* 424-437.

Muthe´n, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, 28, 1-22.

Nandakumar, R. (1993). Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

Oshima, T. & Miller, M. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. *Journal of Educational Measurement, 27(3),* 273-283.

Paap, Muirne C.S. and Glas, Cees A.W. and Veldkamp, Bernard P. (2013) *An Overview of Research on the Testlet Effect: Associated Features, Implications for Test Assembly, and the Impact of Model Choice on Ability Estimates.* Law School Admission Council Research Report 13-03. *Retrieved from http://www.lsac.org/docs/default-source/research-(lsac-resources)/rr-13-03.pdf.*

Patsula, L. N. (2000). *A comparison of computerized adaptive testing and multistage testing*. Dissertation Abstracts International: Section B: the Sciences & Engineering, 60, 5829.

Patz, R., & Junker, B. (1999). Straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

Raju, N. S., van der Linden, W.; Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, Vol 19(4),* 353-368.

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

Ravand, H. (2015). Assessing Testlet Effect, Impact, Differential Testlet, and Item Functioning Using Cross-Classified Multilevel Measurement Modeling. *SAGE Open*, *5*(2), 2158244015585607.

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

Roznowski, M. (1988). Review of test validity. *Journal of Educational Measurement, 25,* 357-361.

Rupp, A. & Zumba, B. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational Psychological Measurement, 66,* 63-84.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34(4, Pt. 2),* 100.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Schnipke, D. L., & Reese, L. M. (1997). *Comparison of testlet-based test designs for computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Schmider, E., Ziegler, M., Daney, E., Beyer, L., Buhner, M. (2010). Is it really robust?. *Methodology*, 6(4), 147-151.

Shewhart, W. (1931). *Economic control of quality of manufactured product*, ASQ Quality Press.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *ETS Research Report Series*, *1991*(1), i-15.

Stocking, M. (1988). Scale drift in on-line calibration. ETS report, RR-88-28-ONR. Retrieved from www.dtic.mil/dtic/tr/fulltext/u2/a196439.pdf.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4),* 361–370. Retrieved from http://www.jstor.org/stable/1434855.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement*, 201-211.

Tao, J., Xu, B., Shi, N. Z., & Jiao, H. (2013). Refining the two-parameter testlet response model by introducing testlet discrimination parameters. *Japanese Psychological Research*, *55*(3), 284-291.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological methods*, *6*(2), 181.

Veerkamp, W., & Glas, C. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25(4)*, 373-389.

Veldkamp, B.P. & van der Linden, W. (2000). Designing item pools for computerized adaptive testing. In W.J. van der Linden & C.A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 149-162). Boston, MA: Kluwer Academic Publishers.

Veldkamp, B.P. (2014). Item pool design and maintenance for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 3). New York, NY: CRC Press.

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

van der Linden, W., Glas, C., Wainer, H. & Bradlow, L. (2000). MML and EAP estimates for the testlet response model. In van der Linden W.J. , Glas C.A.W. (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp.271-287). Boston, MA: Kluwer-Nijhoff Publishing.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

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

von Davier, A. (2010). *Statistical models for test equating, scaling, and linking*. Springer Science & Business Media.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Routledge.

Wainer, H. (1995). Accuracy 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. (1987). Item Clusters and Computerized Adaptive Testing: A Case for Testlets. *Journal of Educational Measurement,24(3),* 185-201.

Wainer, H., Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics,12(4),* 339-368.

Wainer, H. & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57(5),* 741-758.

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

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 27, 1-14.

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

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

Wainer, H., Sireci, S., Thissen, D. (1991). Differential Testlet Functioning: Definitions and Detection. *Journal of Educational Measurement, 28(3),* 197-219.

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.

Wang, X., Bradlow, E., Wainer, H. (2002). A general Bayesian model for testlets theory and applications. GRE Board Report No 98-01P, *ETS Research Report, 02-02*.

Wang, W. C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*(4), 549-576.

Wei, X. E. (2013). *Impacts of Item Parameter Drift on Person Ability Estimation in Multistage Testing*. CPA Examination Technical Reports Retrieved from http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/TechnicalReports/Pages/default.aspx .

Wise, S. & Kingsbury, G. (2000). Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program. *Psicológica 21*,135-155. Retrieved from http://www.uv.es/revispsi/articulos1y2.00/wise.pdf.

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

Weissman, A. (2014). IRT-based multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 10). New York, NY: CRC Press.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77-87.

Witt, E., Stahl, J., Bergstrom, B., & Muckle, T. (2003). *Impact of item drift with non-normal distributions*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Wollack, J., Sung, H., & Kang, T. (2005). *Longitudinal effects of item parameter drift*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Wollack, J. A., Sung, H. J., & Kang, T. (2006). *The impact of compounding item parameter drift on ability estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Woodruff, D., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement*. ACT Research Report Series 2013(7).

Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Xing, D. (2001). *Impact of several computer -based testing variables on the psychometric properties of credentialing examinations* (Order No. 3012196). Available from

ProQuest Dissertations & Theses Global. (304702493). Retrieved from http://search.proquest.com/docview/304702493?accountid=14696.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*, 5–21.

Yan, D., von Davier, A., Lewis, C. (2014). Computerized Multistage Testing: Theory and Applications. *Computerized multistage testing: Theory and applications* (Chapter 6). New York, NY: CRC Press.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 8(3)*, 187-213.

Yin, M., & Sims, J. (2006). *Diagnostic Language Testing for Taiwanese University Students: The Online English Assessment System (OEAS) Project*. Paper presented in the 2nd International Conference on English Instruction and Assessment. National Chung Cheng University, Taiwan.

Zenisky, A. & Hambleton, R. (2014). Multistage research designs: moving research results into practice. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 2). New York, NY: CRC Press.

Zenisky, A. L. (2004). *Evaluating the effects of several multi -stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Order No. 3136800). Available from ProQuest Dissertations & Theses Global. (305176302). Retrieved from http://search.proquest.com/docview/305176302?accountid=14696.

Zenisky, A., Sireci, S., Martone, A., Baldwin, P., & Lam, W. (2009). *Massachusetts Adult Proficiency Tests Technical Manual Supplement: 2008-2009*. Center for Educational Assessment, University of Massachusetts Amherst.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27(1),* 119–140.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 6). New York, NY: CRC Press.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage Adaptive Testing for a Large-Scale Classification Test: Design, Heuristic Assembly, and Comparison with Other Testing Modes*. ACT Research Report Series, 2012 (6). *ACT, Inc.*

Zwick, R., Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 18). New York, NY: CRC Press.