The Dissertation Committee for Ian Fredrick Hembry Certifies that this is the approved version of the following dissertation:

Operational Characteristics of Mixed-format Multistage Tests Using the 3PL Testlet Response Theory Model

Committee:

Barbara Dodd, Supervisor

S. Natasha Beretvas

Jodi Casabianca

Leslie Keng

Tiffany Whittaker

Operational Characteristics of Mixed-format Multistage Tests Using the 3PL Testlet Response Theory Model

by

Ian Fredrick Hembry, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2014

Dedication

I want to dedicate this dissertation to my departed father Fredrick Lyle Hembry and recently born son Nile Fredrick Hembry. It has been said that “new life makes losing life easier to understand.” It has been a blessing to have both father and now son in my life.

Acknowledgements

There are several people that have positively influenced my life that I wish to thank. I have been blessed by a loving family, raised in a wonderful childhood town, and been taught at a number of excellent institutions throughout my life.

I would first like to express my deepest admiration and gratitude towards my dissertation advisor, Barbara Dodd. Dr. Dodd provided me both intellectual and personal support throughout my doctoral studies at The University of Texas at Austin. My thorough understanding of psychometrics is directly related to the coursework and research I was able to accomplish under her guidance. In addition, Dr. Dodd continually motivated me to endure the ebb and flow of the dissertation process, which ultimately allowed me to keep making progress through such a demanding, yet rewarding journey.

I would also like to thank Tasha Beretvas for bestowing upon me her wisdom and support during my doctoral studies. As both her research assistant and teaching assistant, Dr. Beretvas provided me with responsibilities that fostered my development in becoming an independent learner and researcher. In particular, our interactions not only imparted a wealth of knowledge on me, but helped instill confidence in my own abilities as a young researcher in a complex field.

Additionally, I would like to thank the remainder of my dissertation committee, Dr. Jodi Casabianca, Dr. Leslie Keng, and Dr. Tiffany Whittaker. Each person has been most accommodating with their time and feedback. Each committee member's insightful suggestions has aided the overall quality of the study and broadened my own perspective and interests in our field.

A special thanks to Pearson for appointing me as a psychometric intern. Specifically, I would like to thank Dr. Laurie Davis and Dr. Daniel Murphy. Both Dr. Davis and Dr. Murphy allowed me the autonomy to explore exciting and innovative research, while helping to develop my skillset necessary for success in an operational setting. Only through experiences like mine at Pearson could I have been exposed to the intricacies of applied work. My experience with Pearson helped validate my decision to study in a field I barely knew existed when I first began my graduate studies.

Next, I would like to thank the faculty and staff in the Educational Measurement and Statistics program at The University of Iowa. My academic exposure and assistantship introduced me to a new field within education that I barely realized existed. The faculty at Iowa provided my foundation of psychometric knowledge and helped me prepare for a career long aspiration of positively influencing and improving the psychometrics and the broader education field.

I would also like to thank the Clinton Community School District for my experience as a high school mathematics teacher. My experience in the k-12 setting has afforded me a unique perspective to the research I conduct. I have not participated in a study where I do not somehow relate the methodological work to the applicability in a real-life educational setting.

My life has been blessed with supportive parents and siblings. My parents Fredrick and Cheryl Hembry always supported me throughout my childhood and adult life. I could not have asked for a better set of parents who were patient, supportive and loving. I have learned the meaning of hard work, dedication, and compassion directly from their demeanor towards life. In addition, my brother Che Hembry and sister Carly Hembry have grown to be not only family, but two of my best friends. I would not be

where I am today without their encouragement, friendship, and love that can only be given by a brother and sister.

To my son, Nile, I can no longer imagine life without you. Even in my most sleep deprived state, you show me the true purpose in life. Someday I hope I can inspire and teach you as much as you have already shown me in such a short time in my life.

To my lovely wife Tracey, I could not ask for a better partner. You support me when I need support, you motivate me when I need a push, and you love me always. I hope the first five years of our life together are a predictive measure of how commitment, devotion, and happiness will continue to grow within our family. You have always been understanding about the stresses of graduate school. You were supportive when I needed to sit and work at night, even when that interfered with our ability to spend time together. I could not ask for a better person to navigate through life. I love you and thank you!

Operational Characteristics of Mixed-format Multistage Tests Using the 3PL Testlet Response Theory Model

Ian Fredrick Hembry, Ph. D

The University of Texas at Austin, 2014

Supervisor: Barbara G. Dodd

Multistage tests (MSTs) have received renewed interest in recent years as an effective compromise between fixed-length linear tests and computerized adaptive test. Most MSTs studies scored the assessments based on item response theory (IRT) methods. Many assessments are currently being developed as mixed-format assessments that administer both standalone items and clusters of items associated with a common stimulus called testlets. By the nature of a testlet, a natural dependency occurs between the items within the testlet that violates the local independence of items. Local independence is a fundamental assumption of the IRT models. Using dichotomous IRT methods on a mixed-format testlet-based assessment knowingly violates local independence. By combining the score points within a testlet, researchers have successfully applied polytomous IRT models. However, the use of such models loses information by not using the unique response patterns provided by each item within a testlet. The three-parameter logistic testlet response theory (3PL-TRT) model is a measurement model developed to retain the uniqueness in response patterns of each item, while accounting for the local dependency exhibited by a testlet, or testlet effect.

Because few studies have examined mixed-format MSTs administration under the 3PL-TRT model, the dissertation performed a simulation to investigate the administration of a mixed-format testlet based MSTs under the 3PL-TRT model. Simulee responses were generated based on the 3PL-TRT calibrated item parameters from a real large-scale passage based standardized assessment. The manipulated testing conditions considered four panel designs, two test lengths, three routing procedures, and three conditions of local item dependence.

The study found functionally no bias across testing conditions. All conditions showed adequate measurement properties, but a few differences did occur between some of the testing conditions. The measurement precision was impacted by panel design, test length and the magnitude of local item dependence. The three-stage MSTs consistently illustrated slightly lower measurement precision than the two-stage MSTs. As expected, the longer test length conditions had better measurement precision than the shorter test length conditions. Conditions with the largest magnitude of local item dependency showed the worst measurement precision. The routing procedure had little impact on the measurement effectiveness.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The use of computers and the development of item response theory (IRT) exemplify two of the most paramount changes in test delivery platforms in recent history (Brennan, 2006). Although less sophisticated forms of adaptive testing have been developed using classical test theory approaches (e.g. Stanford-Binet), the interaction between computing power and IRT has resulted in tailored testing with equal if not greater reliability than fixed item length testing forms. A testing delivery platform has implications on a number of test specifications such as test scoring, scaling, equating, platform comparability, item pool development, among others (Schmeiser & Welch, 2006). The current study focuses on a particular delivery platform, namely multistage tests (MSTs), in order to fully understand the implication of using this particular test delivery platform and its capability of capturing an examinee's current ability being measured. Investigations, such as this dissertation, must be conducted to provide information about the best delivery method to test developers for the needs of their program. The current introduction briefly reviews different testing platforms and the advancement in computers and IRT that have led to advancement in adaptive testing. Then, a description of the impact item types have on both delivery platforms and IRT models is given.

MODERN TEST DELIVERY PLATFORMS

Today numerous test delivery platforms are available to a testing program. The advent of computers and their increased computing power has enhanced the way in which a test can be delivered to an examinee from that of traditional paper-pencil testing.

Paper-pencil Based Testing

Large-scale assessments are sometimes delivered in the form that may be regarded as a "traditional" testing format. For the current discussion, a standardized paper-pencil test is being referenced as a traditional testing format. For example, an examinee receives a hard copy of a test form consisting of a set of fixed-length items in conjunction with an answering document. The form length is fixed, in that, every examinee receiving the form is administered the same number of items. The examinee then records their answers to an answer document, such as a Scantron bubble sheet, where items are answered by filling in a bubble that corresponds to the perceived correct item's response option from the test form. Upon completion, examinees' answer documents are sent to a scoring center, where the answer documents are scored.

Computer-Based Testing

Computer-based testing entails any platform that uses a desktop, laptops, or other versions of a microcomputer to administer a test. Drasgow, Luecht, and Bennett (2006) discuss varying computer-based delivery platforms such as computerized linear fixed-length tests (LFTs), item-level computerized adaptive tests (CATs), clusterized adaptive tests, and structured computer-adaptive multistage tests (MSTs). Two major differences in administration from traditional based testing, besides the user-interface, are the need for a large and accessible item bank and potential on-site scoring. Computer-based tests' item banks are generally stored on a network server or the administering computer for accessibility purposes during test administration. Another difference found is in the test scoring procedures. Scoring in most cases of computer-based tests can be done on-site and no longer needs to be shipped to a scoring center. The scoring procedures does impact the platform in which a test may be delivered (Wainer, 2000b).

A computerized LFT is analogous to the paper-pencil based test described above. The difference is that administration takes place on a computer. Because computerized LFT are administered as intact forms, scoring can either be done during the testing experience or after. Computerized adaptive testing (CAT) is the first style of testing presented that distinguishes itself from testing for a broad range of abilities by tailoring the test items to the ability of the examinee (Wainer, 2000b). The tailoring is executed by scoring responses in real time and using this information to select the next item for administration (Drasgow et al., 2006).

The remaining platforms discussed are alternative forms of CAT. *Clusterized adaptive tests* were introduced by Wainer and Kiely (1987). They provide a framework for the adaptive unit at the cluster level of items developed around a central content category. For clusterized adaptive tests, an examinee is administered a content item cluster. Upon completion of the cluster, the examinee's ability is updated based on all responses and a new cluster is selected. *Adaptive MSTs* are similar to the clusterized adaptive tests in that sets of items are used as the building block for a test (Zenisky, Hambleton, & Luecht, 2010). In MSTs, an examinee is administered a set of items in stages, where the set of items is not primarily centered on content categories but rather will piecemeal the content categories within a stage. A typical administration might proceed as follows. The first stage, i.e. first set of items, is administered to the examinees. Then based on the item responses in the first stage and the MST routing procedures, an examinee navigates to the second stage, where another set of items is selected from a series of subtests. The routing procedure is a predetermined subtest selection procedure that matches an ability level to a subset of items, while adhering to a program's policies. The examinee will then respond to the new set of items. Then the pattern continues for

each subsequent stage until the examinee reaches the termination point in the final stage of the exam.

Implementing various delivery platforms impacts the psychometric properties of a test. Lord (1971) noted that administering a test appropriate to an ability level yields more accurate measurements of an examinee. One major distinction between LFTs and adaptive tests is that adaptive tests attempt to tailor the administration to the examinee's ability. In doing so, CATs have been shown to provide similar measurement precision to LFTs while significantly reducing the administered test length (Wainer, 1993). However, CATs have their own complications. Two distinct difficulties in implementing CATs are item exposure rate issues and the inability for an examinee to review items. An item exposure rate is the rate at which an item is administered to the examinee population. Security issues can be exasperated by fully adaptive tests when many of the same items get administered to most examinees, which leads to overexposure of some items and the underutilization of others (Thissen & Mislevy, 2000). MSTs offer test designers a compromise between allowing an examinee the ability to review subsets of items while building in exposure control strategies and tailoring to an examinee's ability (Zenisky et al., 2010). The current study focuses on the use and implementation of adaptive MSTs. Because MSTs are adaptive tests, the statistical theory underlying adaptive testing will be briefly reviewed.

MODERN TESTING THEORY

Current forms of adaptive testing were strongly influenced by modern testing theory. Modern testing theory started with Lord's (1952) normal ogive models. Then major breakthroughs for applicability were developed by Rasch's (1960) and Birnbaum's (1968) extensions using logistic functions. Their work has now provided the framework

for the current family of item response theory (IRT) models that are capable of measuring an examinee's proficiency through test items rather than the test forms. Now, adaptive testing uses an item's statistical characteristics estimated from the IRT family to tailor a test to an examinee's estimated ability level.

The appropriateness of IRT models can be dictated by the examinee response patterns and the item types being administered on a test. The most common item type is a multiple-choice item. A multiple-choice item is an item with a prompt or question eliciting a response from a set of possible response options. Included in the response options is exactly one correct item and a set of distractors or incorrect options. The item is then either scored right or wrong and often coded 0 or 1, respectively. Due to the binary scoring of multiple-choice items, they are often referred to as dichotomous items. *Polytomous items* are items that consist of multiple categories of classifications (i.e. two or more) for a given response. For example, essay items score using a scoring rubric with two or more possible score points, survey style Likert items, or constructed response items that receive partial credit are among the varying style of polytomous item types. A polytomous item will receive a score *k* from the ordered category set of scores 0, 1, 2, …, m_j, where m_j is the highest score category for a given item. *Testlets* are another set of items that have elicited appropriate scoring methods in modern testing theory. A testlet is a cluster of items with prompts and response options developed around a common stimulus (Wainer, Bradlow, & Wang, 2007). A common stimulus might consist of a reading passage, graph, or data table. Like dichotomous items, testlet-based items can be scored as right or wrong within the testlet.

A test can be constructed of standalone dichotomous, standalone polytomous, testlet-based items, or any combination of the item types. When combinations of item types are used for administration, it will be referred to as a *mixed-format test*. The

current study uses a mixed-format testlet-based item pool that consists of standalone multiple choice items and testlet-based multiple choice items.

Since the inception of IRT, adaptive testing has taken advantage of its ability to measure a person's ability at the item level. However, specific models are only appropriate for specific types of items. Since some items considered for the current study are testlet-based, the appropriateness of the IRT family must be scrutinized. IRT models are statistically based models. As with any statistical model, assumptions are made when defining a model. For IRT based models, one of the assumptions that can be violated for a testlet-based exam is the assumption of local independence (Lord, 1980). However, in the presence of a testlet, an item's local independence within-testlets can be violated and lead to bias in an items statistical properties (Sireci, Thissen, & Wainer, 1991; Wainer et al., 2007). The violation in local independence has lead researchers to develop a number of testlet response theory (TRT) models. Because of the nature of the item pool, the current study employs an MST under the three parameter logistic TRT (3PL-TRT) model.

Currently, only a handful of studies have used the 3PL-TRT model for an MST administration (Galindo, Park, & Dodd, 2013; Keng, 2008; Lu, 2010). Both Keng (2008) and Galindo, et al. (2013) used an item pool that only consisted of testlet-based items. Lu (2010) studies various levels of mixed-format administration. The study found that higher levels of a testlet effect, or a manifest variable that violates local independence within a testlet, created a greater need to use the 3PL-TRT model. However, Lu (2010) only investigated one of two scenarios suggested for appropriate implementation of a 3PL-TRT model which aligns more with a clusterized adaptive test than an MST (Wainer et al., 2007). Specifically, Lu (2010) used a subset of items from the larger total set of items associated with a given testlet. In practice, it is common to see smaller testlet sizes

that are administered as a whole set, say a testlet sizes between five and ten items, such as the GRE and SAT as was investigated by Wainer, Bradlow, and Du (2000).

Therefore, the purpose of the current research is to use a simulation study to investigate the operational characteristics of an MST under the 3PL-TRT model for a mixed-format testlet-based item pool that administers standalone items and whole testlets. The simulation uses item parameters from an existing testing program to generate item responses in order to represent a realistic testing scenario. To assess the operational characteristics of an MST in this context, the precision of ability estimates and the routing of simulees during administration is examined. One aspect of the item responses for the item pool that was manipulated is the testlet effect. MST components that varied are the panel design, test length, and routing procedures. Gaining insight into the impact of these various testing conditions for an MST administration helps inform testing programs interested in employing MSTs about the associated advantages and disadvantages for testlet-based item pools.

Chapter 2: Literature Review

The following literature review describes the relevant research and background information for the current study. The first section provides an introduction to item response theory (IRT). The second section then outlines the use of testlet response theory (TRT). Then the basic components of multistage testing are described. Next, an overview of research pertaining to MSTs is explained. The final section states the problem of main interest for this dissertation.

ITEM RESPONSE THEORY

Item response theory (IRT) is a scaling method that consists of a family of models that relates one's item response to a latent variable, or trait ability, generally referred to as *theta*, or θ (de Ayala, 2009). In doing so, IRT treats latent traits and item characteristics as predictors of observed item responses.

Generally speaking, most IRT models consist of three assumptions that characterize modeling the probability of getting an item correct conditional on modeling an ability level. IRT models have a dimensionality assumption, where the response data is an indicator of one or more latent construct(s) or factor(s) (Reckase, 2009). By and large, this refers to the model being representative of the construct(s) being assessed in order to inform the probability of a correct response. The models for the current study will only consider that of a unidimensional structure, or a person's ability being represented by one latent dimension or factor. The second assumption is that of conditional independence, sometimes referred to as local independence. The weak form of local independence assumes that the items are uncorrelated conditional on ability. The strong form of local independence assumes that the responses are statistically independent given an ability level. The final assumption common to IRT models is that

of functional form, or that a mathematical statement can be made relating the person's response and ability to an item (see Lord, 1980). All three assumptions hold throughout the discussion of dichotomous and polytomous IRT models.

Dichotomous Models

Dichotomously-scored items consider only two response options. Typical responses can be right/wrong but could include agree/disagree or true/false. Item types such as these often occur as multiple choice items, where an item stem, or the intended question to be answered, is presented to an examinee eliciting a response from which the options include one correct option and the remaining distractor options are incorrect. Therefore, the mathematical models to be introduced represent the relationship between the item's parameters and a person's ability parameter, where only binary options are considered for the response patterns and are generally coded 0 or 1.

Three of the most widely used dichotomous models in the IRT literature are: 1) Rasch/one-parameter logistic (1PL) model (Rasch, 1960); 2) two-parameter logistic (2PL; Birnbaum, 1968); and 3) three-parameter logistic (3PL; Birnbaum, 1968). Collectively, these dichotomous IRT models differ based on three item parameters which distinguish the probabilistic functioning of each model.

Rasch/1PL

The Rasch/1PL formulation of a response function models the probability of answering an item correctly given a person's ability as follows:

$$p\left(u_j = 1 \mid \theta, b_j\right) = \frac{\exp\left(\theta - b_j\right)}{1 + \exp\left(\theta - b_j\right)}, \tag{1}$$

where the function in Equation 1 models the probability of a correct response, *u* equals 1, on item *j* for a given ability θ with relative item difficulty b_j. The item difficulty

parameter is the θ value that corresponds to the inflection point of the probability trace line. For the 1PL model, the point of inflection is always at the .50 probability of answering an item correct. Additional assumptions for the model are equal discriminations across items and no guessing occurring when answering an item. See item 1 in Figure 1 for an example of a 1PL item characteristics curve (ICC).

2PL

The 2PL model is an extension of the 1PL model developed by Birnbaum (1968). The extension to the 2PL model occurs by including a unique item discrimination parameter for every item:

$$p\left(u_j = 1 | \theta, a_j, b_j\right) = \frac{\exp(a_j(\theta - b_j))}{1 + \exp(a_j(\theta - b_j))}, \qquad (2)$$

where b_j and θ retain their original interpretation as the item difficulty and person ability, respectively. The model also includes the item discrimination parameter a, modeling a unique discrimination index for each item *j*. The item discrimination parameter is a function of the slope at the point of inflection of the ICC. Descriptively items with higher *a* parameters have steeper slopes than items with lower *a* parameters. An additional assumption for the 2PL model is that no guessing occurs when answering an item. For an example of the effects of the discrimination parameter, Figure 1 compares two items, Items 1 and 2, where Item 1 has item discrimination equal to 1 and the item discrimination for item 2 is 1.4. It can then be seen that around the point of inflection, $\theta = 0.5$, Item 2 accelerates to its upper asymptote more quickly than does Item 1.

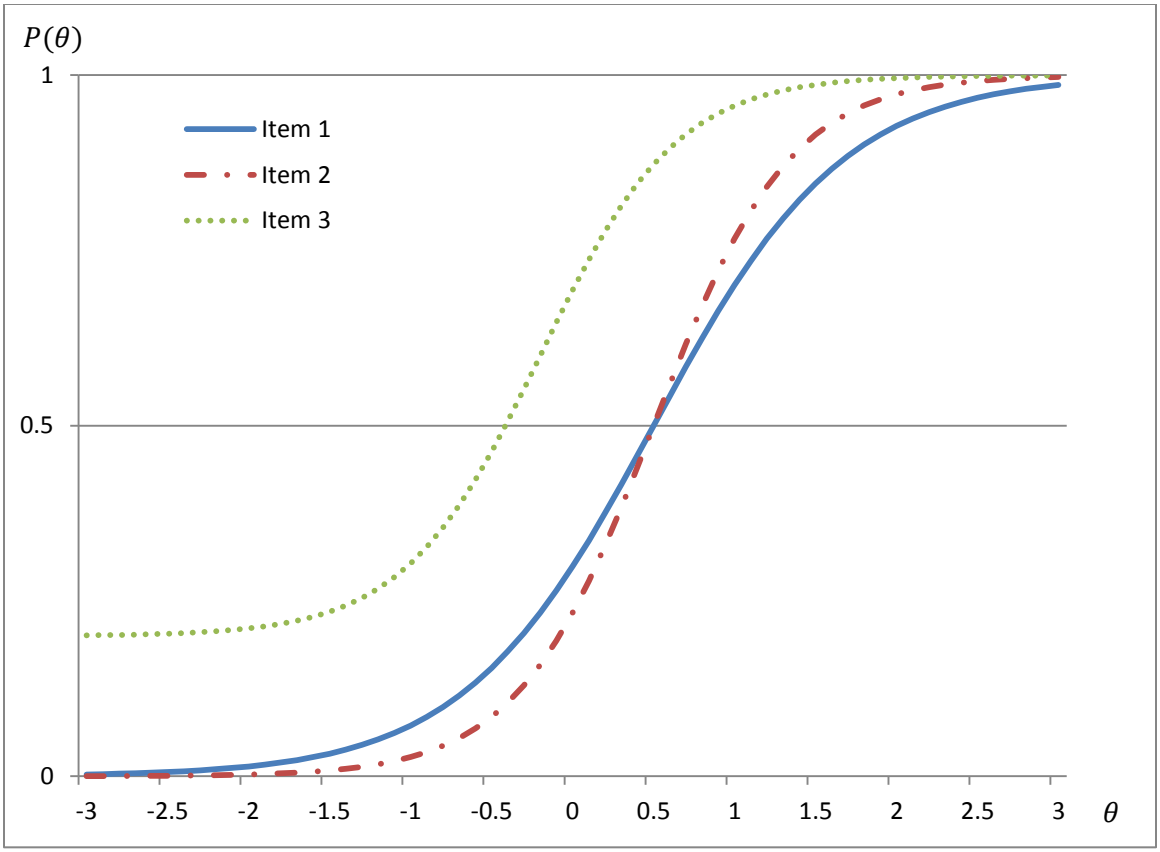

Figure 1. Item Characteristic Curve for 1PL, 2PL, and 3PL. Item 1 *a*=1, *b*=0.5, and c=0; Item 2 *a*=1.4, *b*=0.5, and *c*=0; Item 3 *a*=1.4, *b*=-0.2, and *c*=0.2.

3PL

Birnbaum (1968) further extended the 2PL model in to the 3PL model by introducing a pseudo-guessing parameter. The probability of a correct response conditional on ability for the 3PL model is modeled as follows:

$$p\left(u_j = 1 \mid \theta, a_j, b_j, c_j\right) = c_j + \left(1 - c_j\right)\frac{\exp(a_j(\theta - b_j))}{1 + \exp(a_j(\theta - b_j))}, \qquad (3)$$

where a_j is the item discrimination parameter, b_j is the item difficulty parameter, and c_j is the pseudo-guessing parameter for item *j*. The pseudo-guessing parameter, c_j, is representative of the probability that a person with ability approaching negative infinity will answer the item correctly. Mathematically, it acts as the lower asymptote for the function in Equation 3. The item difficulty parameter is still the ability level that corresponds to the point of inflection. When guessing is present, the point of inflection is the halfway point between the pseudo-guessing parameter, c_j, and the upper asymptote of 1, i.e. $(1 + c_j)/2$. Figure 1's Item 3 illustrates an item with a *c* value equal to 0.2, an *a* value equal to 1.4, and a *b* value of -0.2. In this instance, the probability of answering the item correctly for an ability at the point of inflection is $\frac{1.2}{2}$ or .6. It can be seen that the lower asymptote approaches 0.2 and that the point of inflection has shifted down on the theta scale, or Item 3 is easier when comparing it to Items 1 and 2. It must be noted that if the pseudo-guessing parameter equals zero, the models in Equations 2 and 3 are equivalent and the 3PL model reduces down to the 2PL model.

Item and Test Information

Any statistical model has a level of uncertainty associated with the parameters of interest and is generally referred to as the standard error (SE). Psychometric work is designed to provide evidence for a person's ability on a given construct within a certain degree of (un)certainty. The uncertainty of an ability estimate is known as the standard

error (SE) and is symbolized as $\sigma_e(\hat{\theta})$. IRT uses the contribution of each item to help reduce the uncertainty associated with a person's estimated theta by use of the *item information function* (Lord, 1980):

$$I_j(\theta) = \frac{p'^2_j}{p_j q_j} = a_j^2 \left[\frac{(p_j - c_j)^2}{(1-c_j)^2} \right] \left[\frac{1-p_j}{p_j} \right], \quad (4)$$

where p_j is the probability of answering item *j* correctly, q_j is the probability of answering item *j* incorrectly (and can be rewritten as $1-p_j$), a_j is the discrimination parameter for item *j*, c_j is the pseudo guessing parameter, and p'_j is the first derivative of p_j with respect to θ. Notice that information will vary across the ability continuum. Additionally, information with all other item parameters being held constant varies directly from to the square of the item discrimination parameter.

A test's total information, or *test information function*, is the sum of the administered item information functions:

$$I(\theta) = \sum_{i=1}^{L} I_i(\theta), \quad (5)$$

where it has been shown to be the upper bound of information attained from item responses and is equivalent to the squared inverse of the SE (Birnbaum, 1968). Therefore the SE for person's ability can be defined by the inverse of the square root of the test information function as follows:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}. \quad (6)$$

Polytomous IRT Models

Polytomously-scored items consist of multiple category classifications (i.e. two or more). For example, essay items graded on a scoring rubric of two or more scores, Likert attitude scale items, or constructed response items that receive partial credit are examples

of polytomous item types. For a polytomous item *j* the score x(*k*) = {0,1,2,…m_j}, is the set of possible ordinal response categories. The polytomous models included in the discussion entail popular models appropriate for ordinal score categories. Due to the nature of the item responses, polytomous models represent the probability of a category score as a function of θ, and not the probability of answering an item correct as was the interpretation for dichotomous models. Polytomous IRT models are broadly categorized by two methods 1) difference models and 2) divide-by-total models (Thissen & Steinberg, 1986). One difference model and two divide-by-total models are discussed below, including the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), and the generalized partial credit model (Muraki, 1992), respectively.

Graded Response Model

The graded response model (Samejima, 1969) is a difference model which scores items with two or more ordered response categories. The graded response model first defines the probability of scoring in a particular category or higher as follows:

$$P^*_{j(k)}(\theta) = \frac{\exp\left[a_j\left(\theta - b_{j(k)}\right)\right]}{1+\exp\left[a_j\left(\theta - b_{j(k)}\right)\right]}, \tag{7}$$

where $P^*_{j(k)}$ is the probability for a person with ability θ scoring in category *k* or higher on item *j*, a_j is the item discrimination for item *j* and $b_{j(k)}$ is the *category boundary location* for category *k* in item *j* and are always in increasing order. The category boundary location, $b_{j(k)}$, is defined as the theta value that corresponds to .5 on the given $P^*_{j(k)}$ function. It can be interpreted as the category bound between categories *k* and *k-1*, and represents the point of inflection for the function in Equation 7. Lower and upper limits for the function are defined as $P^*_0 = 1$ and $P^*_{m_j+1} = 0$, respectively. Because $P^*_{j(k)}$ is the probability of scoring in category *k* or higher, to calculate the probability of

scoring in category *k* for a given ability, one must take the difference between the probability of category *k* and *k+1* as follows:

$$p_k(\theta_i) = P^*_k - P^*_{k+1}. \quad (8)$$

Hence, the graded response model is referred to as a difference model. Note, the model in Equation 7 reduces to the 2PL model in the case of only two response categories.

Partial Credit Model

The partial credit model is the first of the two divide-by-total models to be described. Masters (1982) extended the Rasch model to incorporate two or more score categories. The partial credit model's mathematical form is defined as:

$$p_{j(k)}\left(x_j = k \mid \theta, b_{j(k)}\right) = \frac{\exp\left[\sum_{k=0}^{x_j}(\theta - b_{j(k)})\right]}{\sum_{h=0}^{m_j}\exp\left[\sum_{k=0}^{h}(\theta - b_{j(k)})\right]}, \quad (9)$$

where $b_{j(k)}$ is the *step difficulty parameter* for item *j* with category score *k*. The step difficulty parameter acts as the transition point at which category *k*'s probability equals the probability of category *(k-1)*. The divide-by-total namesake originates from the notion that the probability of scoring in category *k* is calculated by dividing the numerator, or the unique proportion of area represented by category *k*, divided by the total area of the function or the sum of all legitimate category response functions, or the denominator. Note that the partial credit model reduces down to the dichotomous Rasch model in the case of only two response categories.

Generalized Partial Credit Model

The generalized partial credit model is the second divide-by-total model to be described and is an extension to the partial credit model (Muraki, 1992). Muraki's extension includes a unique item discrimination parameter and is formulated as follows:

$$p_{j(k)}\left(x_j = k \mid \theta, a_i, b_{ik}\right) = \frac{\exp\left[\sum_{k=0}^{x_j} a_j(\theta - b_{j(k)})\right]}{\sum_{h=0}^{m_j} \exp\left[\sum_{k=0}^{h} a_j(\theta - b_{j(k)})\right]}, \quad (10)$$

where a_j is item *j*'s item discrimination parameter, and $b_{j(k)}$ has the same interpretation as in the partial credit model. In the case of an item having only two response categories the model reduces down to the 2PL model.

Item and Test Information

Item information for polytomous models is more complex in that, the models are composed of multiple category parameters. Samejima's (1969) noted that to construct the item information for polytomous models one must first construct the category information function, then the item information function is the sum of the weighted category information functions:

$$I_j(\theta) = \sum_{k=0}^{m_j}\left[\frac{\left(p'_{j(k)}\right)^2}{p_{j(k)}}\right]. \quad (11)$$

where $p'_{j(k)}$ is the first derivative of the probability of scoring in category *k* with respect to θ, and $p_{j(k)}$ is the probability of scoring in category *k*. The test information is the same as in the dichotomous case seen in Equation 5, in that the test information function is the sum of the item information functions.

TESTLET RESPONSE THEORY

The two types of models discussed thus far have primarily dealt with the idea of scoring either dichotomous or polytomous items or a mixture of the two. However, testlets are not quite dichotomous and not quite polytomous. When testlets are developed, they are typically created to cover an array of content specification by using a number of multiple choice items stemming from the common stimulus (Wainer & Kiely,

1987). Due to the nature of the items' development, one could consider each item individually (i.e., dichotomously) and thus claiming their independence from each other. If a researcher considers the cluster of items as locally independent, one is knowingly overlooking the possibility of a context effect. A context effect is the influence an item contracts simply by its relationship with another item. This is an *item interaction*, where the item responses create a manifest variable which violates the local independence assumption (Tuerlinckx & De Boeck, 2001). Turlinckx and De Boek (2001) supported the implication of prior research (Sireci et al., 1991; Wainer & Thissen, 1996; Yen, 1993) by illustrating the positive bias present in the discrimination parameters for a manifest variable that is not modeled in some manner.

Rosenbaum's (1988) treatment of item bundles proved that local independence holds across testlets, even in the presence of testlet dependency. Rosenbaum's (1988) findings encouraged researchers to change the unit of measurement from the item level to the testlet by using summed testlet scores and applying unidimensional polytomous IRT models (Wainer & Lewis, 1990). Using polytomous IRT models has been found to work better than treating them as stand-alone dichotomous items in the presence of a testlet effect (Sireci et al., 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991; Wainer, 1995). Though using a polytomous model to score testlets as a unit has been shown to work well (Wainer, 1995), Wainer et al. (2007) discussed two circumstances in which the use of a testlet-based model is more appropriate than using a polytomous model.

The first circumstance discussed is the use of a within-testlet adaptive test (Wainer et al., 2007). Testing programs may consider the construction of large testlets, or a common stimulus with say 10-20 associated items. The intention of such a development is not that one examinee is expected to see all items within the testlet but to

adaptively select a smaller subset of items. In this scenario, the selection algorithm would have a predetermined within-testlet stopping rule in which the examinee might see, for example, seven testlet items; then a new testlet would be selected and the associated items from within the testlet would be adaptively administered. This process would continue until the desired psychometric properties were reached at which point the test would be terminated. The total score cannot be modeled in this particular scenario with polytomous models because an examinee is not receiving all of the items within a testlet. This occurs because the testlets would need to be calibrated based on the total number of categories even though a person would not see all within-testlet items. Therefore, the polytomous model would not carry the same meaning from examinee to examinee and, thus, lending itself uninterpretable.

Wainer et al. (2007) describe the second circumstance as one in which more information from the testlet is desired. When testlets are scored as individual items and within-testlet variability exists, the item discrimination parameters tend to be positively biased which results in overestimated testlet information (see Wainer, Bradlow, & Du, 2000; Wainer et al., 2007). When a testlet is represented as a summed score and modeled as a polytomous item, the dependency present in the testlet will naturally and more appropriately reduce the total information of the item cluster when compared with the sum of the dichotomously modeled items. Yet, modeling the item as polytomous loses some of the information gained by modeling individual response patterns. For example, if a five-item testlet is taken, there are 10 different ways in which an examinee could receive a total testlet score of three, but using a polytomous model is non-differentiable between any response patterns. In addition, no guessing can be parameterized for polytomous models, which appears likely for a testlet consisting of all multiple-choice items.

In each scenario, better approximation of response patterns is necessary. The current study focused on a data structure analogous to the latter scenario where additional information was desired based on the response patterns.

Model Specification

Testlet response theory (TRT) models were developed as a generalization of existing IRT models. They were developed in such a way that if no dependence within a testlet is present, the model can reduce back to the dichotomous IRT model (Wainer et al., 2007). Wainer and colleagues have developed a series of models for multiple choice item-based testlet for both the 2PL and 3PL context (see Bradlow, Wainer, & Wang, 1999; Wainer et al., 2000). Conceptually, the extensions are the same for each model and the focus here is on the 3PL dichotomous model.

The development of the TRT family of models was done so in a fully Bayesian framework (Wainer et al. 2007). Bayesian methods allow for the probability model to be constructed as joint probability distributions for both the likelihood of the data given the model and prior distributions for the parameters of interest. This can be more formally written as:

$$P(\theta \mid x) \propto L(x \mid \theta)\, p(\theta), \tag{12}$$

where the posterior distribution, $P(\theta \mid x)$, is used to provide estimates for the parameters of interest by expressing a prior knowledge about the parameters of interest, $p(\theta)$, on the likelihood function of the observed data, $L(x \mid \theta)$. By modeling parameter estimates in this fashion, Bayesian inference treats the parameters of interest as random variables rather than unknown fixed variables. A naturally occurring consequence of Bayesian estimation is the estimation of not only the point estimates for the parameters of interest, but also the inclusion of posterior distributions for each parameter in the model. In order

to estimate the posterior distributions, random sampling methods have been developed called Markov chain Monte Carlo (MCMC) methods, that numerically approximate probability distributions for large spaces (Kruschke, 2011).

The 3PL-TRT model still entails three item parameters that retain their original interpretations as referenced in Equation 3, namely the discrimination, *a*, the difficulty, *b*, and the pseudo-guessing, *c* parameters. To progress to a TRT model, the 3PL function may be rewritten more generally as:

$$P(y_{ij}=1)=c_j+(1-c_j)\frac{\exp(t_{ij})}{1+\exp(t_{ij})}, \tag{13}$$

where t_{ij} is the latent linear score predictor for person i on item j. For the 3PL model, the linear score predictor is defined as:

$$t_{ij}{}_{=}\ a_j\left(\theta_i-b_j\right). \tag{14}$$

In the case of testlets, Equation 14 does not capture the entirety of the data structure. In order to do so, the testlet effect must be captured. Modeling the items' within-testlet interaction, which accommodates the violation of local independence, is done by introducing an additional parameter to the linear score predictor as follows:

$$t_{ij}=a_j\left(\theta_i-b_j-\gamma_{id(j)}\right), \tag{15}$$

where $\gamma_{id(j)}$ is the *testlet effect parameter*. The testlet effect parameter is the interaction of person *i* and item *j* that is nested within testlet $d(j)$. Then $\gamma_{id(j)}$ represents a person *i*'s random effect for the testlet *d* taking item *j*. The testlet effect parameter then accounts for the communality of items associated with a common stimulus by including a person's random effect for the items within a testlet. The variance for $\gamma_{id(j)}$ is then estimated for each testlet and can be used as an indicator for within-testlet local item dependence (LID). Modeling the within-testlet's LID allows for the assumption of conditional independence between testlets.

For the formulation of the probability of scoring an item correctly Equation 15 is substituted into Equation 13 which results in the following:

$$P(x_{ij}=1)=c_j+(1-c_j)\frac{\exp\left(a_j\left(\theta_i-b_j-\gamma_{id(j)}\right)\right)}{1+\exp\left(a_j\left(\theta_i-b_j-\gamma_{id(j)}\right)\right)}. \quad (16)$$

The next step of the Bayesian analysis is to specify the prior distributions for the parameters of interest. In the case of the current model a prior must be specified for a_j, b_j, c_j, θ_i, and $\gamma_{id(j)}$. The prior distributions include:

$$\begin{aligned}
&\theta_i \sim N(0,1)\\
&\log(a_j) \sim N(\mu_a,\sigma_a^2)\\
&b_j \sim N(\mu_b,\sigma_b^2) \quad ,\\
&\text{logit}(c_j) \sim N(\mu_c,\sigma_c^2)\\
&\gamma_{id(j)} \sim N(0,\sigma^2_{d(j)})
\end{aligned}$$

Notice in the prior distributions defined above, more parameters are introduced. For example, b_j is normal with mean, μ_b, and variance, σ_b^2. Each new parameter specified in the prior distributions must also have a distribution assigned to them known as *hyperpriors*. Wainer et al. (2000) recommended hyperpriors to have normal-inverse gamma distributions acting as slightly informative conjugate priors for the mean and variance respectively for each parameter of interest. In the case of a_j and c_j, or the item discrimination and pseudo-guessing parameter, respectively, a transformation on the parameters is performed in order to use normal priors on the two distributions. This occurs because a_j acts similarly to an exponential function and c_j is a parameter bound from 0 to 1. As one inspects the prior for the testlet effect, $\gamma_{id(j)}$, one may observe that the variance of the parameter is testlet specific, or each testlet has its own unique variance estimate, $\hat{\sigma}^2_{d(j)}$, where in the case of a stand-alone dichotomous item, the variance equals zero and no testlet effect is present. The variance estimate $\hat{\sigma}^2_{d(j)}$ for each testlet represents the magnitude of LID within testlet $d(j)$.

To illustrate the efficacy of the 3PL-TRT model, Wainer et al. (2000) conducted a simulation study that compared four response modeling methods. The models included a 3PL IRT model estimated with marginalized maximum likelihood (MML), 3PL IRT model estimated with MCMC with Gibbs sampling (Geman & Geman, 1984), 3PL-TRT with a common estimated testlet variance using MCMC methods, and 3PL-TRT with unique estimated testlet variances using MCMC methods. A 70 item test with 30 independent items and four testlets of size 10 were crossed with three levels of testlet dependency, one with no dependencies, one with equal dependencies, and one with unequal dependencies. 1000 simulees were generated with binary responses. The results showed that all models recovered the parameter estimates similarly well in the condition with no testlet effects. For conditions with testlet effects, the two 3PL-TRT models outperformed both IRT based models. When unique testlet effects were generated, the 3PL-TRT estimating unique variances outperformed the simpler TRT model that assumed common testlet effect variability, along with all other models considered in the study.

Wainer et al. (2000) also conducted an applied study to data from the Scholastic Assessment Test (SAT) and Graduate Record Examination (GRE) with the same four models as mentioned in their simulation study. Each verbal test consisted of numerous independent items and four testlets of varying size. The SAT verbal test resulted in very small estimated testlet variance indicating the testlet items were written well with regards to conditional independence properties. The GRE verbal test, when modeled for unique testlet effect variability, showed much higher testlet effect variance estimates. Further analysis indicated significantly higher item discrimination parameters for the items associated with testlets under the 3PL IRT model.

The result of increased item discrimination encourages a further look at item information found in Equation 4 and the information function for TRT. Wainer et al. (2000) derived TRT item information as:

$$I_j(\theta) = a_j^2 \left(\frac{\exp\left(a_j\left(\theta_i - b_j - \gamma_{id(j)}\right)\right)}{1+\exp\left(a_j\left(\theta_i - b_j - \gamma_{id(j)}\right)\right)} \right)^2 \left(\frac{1-c_j}{c_j + \exp\left(a_j\left(\theta_i - b_j - \gamma_{id(j)}\right)\right)} \right). \quad (17)$$

Notice that Equation 4 and 17 are virtually equivalent except for the modeling of $\gamma_{id(j)}$ in the linear score predictor. Recall the direct relationship with the item discrimination parameters and item information. Specifically, items with higher a parameters provide more information than items with lower a parameters. This relationship is still present in the TRT information function. Also recall that information has been found to be overestimated because of the inflated item discrimination parameters when conditional independence of testlets is ignored (see (Sireci et al., 1991). Because (Wainer et al., 2000) found increased item discrimination parameters when local independence was ignored, Wainer et al. (2007) reasoned that models ignoring the testlet effect, create a context effect that overestimates a_j's. If the a_j's are overestimated, then the test information function is overestimated, and thus SE's of ability esimates are underestimated (Murphy, Dodd, & Vaughn, 2010; Sireci et al., 1991; Wainer et al., 2007; Wainer & Thissen, 1996; Yen, 1993). Therefore, the accuracy in measurement precision for the GRE verbal was more appropriately accounted for by modeling the testlet effect with the 3PL-TRT model.

In certain situations, past research has indicated that ignoring local dependency created by the use of testlets results in overestimated item discrimination parameters, which then results in inaccurate measurement precision. Therefore, the measurement model shows signs of utility when assessments administer testlet-based items.

MULTISTAGE TEST

Multistage tests (MSTs) have been described as a compromise between computerized adaptive tests (CATs) and linear fixed-length tests (LFTs) (Jodoin, Zenisky, & Hambleton, 2006; Zenisky et al., 2010). LFTs are tests where all examinees receiving a particular form receive the same set of items. On the other end of the spectrum, CATs adapt to an examinee's ability during administration at the item level, where there are potentially countless forms that could be administered to examinees. MSTs represent the middle of the continuum, where the test is pre-constructed like an LFT but adaptation takes place between *modules*, or a collection of preassembled items (Wainer, 2000a). MSTs have multiple naming conventions including computerized master testing (Lewis and Sheehan, 1990) computer adaptive sequential testing (Luecht & Nungester, 1998), multiple-form structures (Armstrong, Jones, Koppel, & Pashley, 2004), and bundled multistage adaptive testing (Luecht, 2003). Each naming convention provides a commonality by administering sub-units or clusters of items and adapting to an examinee's ability through a network of paths that select the modules based on the current estimate of an examinee's ability.

The concept of an MST has been around for some time. Cronbach and Gleser (1965) and Lord (1971a, 1971b, 1980) discussed the use of two-stage testing. The original idea behind administering a two-stage test was to administer an average difficulty first stage routing subtest to estimate a preliminary ability level. The second stage subtest test is then chosen among a set of subtests of varying difficulty, e.g. an easy, medium, or hard test. Wainer and Kiely (1987) then suggested developing sub-units of items around a common content category and adaptively administering on the sub-unit of interest. Bock and Mislevy (1988) renewed interest in two-stage tests as a multi-purpose means of scoring interpretation for both individual ability and larger unit achievement

scoring. Then research regarding MSTs went dormant as fully adaptive tests overshadowed MSTs as the preferred method of adaptive administration (Mead, 2006). More recently, resurgence in MSTs has occurred as the benefit for using MSTs offers quantitative advantages over traditional LFTs and qualitative advantages over fully adaptive delivery platforms.

Relative advantages of MSTs are apparent when comparing them to LFTs. MSTs provide increased measurement precision across the ability scale because of their adaptive nature. In many testing situations, the abilities of examinees widely range. LFTs are typically constructed in a fashion that concentrates the test construction around a specified level of difficulty, which reduces to effectively measure an examinee's capabilities across the distribution of examinees' (Kim & Plake, 1993; Lord, 1980). Additionally, MSTs lead to reduced testing time and potential on-site scoring (Jodoin et al., 2006; Zenisky et al., 2010). Modern MSTs also provide more flexible testing windows because administration takes place on computer (Hendrickson, 2007).

Multiple advantages also exist for MSTs over CAT. Some of the advantages are qualitative in nature and are related to quality assurance. MST forms can be created prior to administration allowing developers more control over test blueprint, item ordering, and item review (Hendrickson, 2007). MSTs also allow examinees to review items within a stage, which helps alleviate testing anxiety and allows examinees to maximize scores without compromising the adaptive selection criteria (Vispoel, 1998). In addition, item exposure control can be handled during MST test form construction, thus eliminating the need for special algorithms that combined with item selection procedures as is necessary in CATs (Georgiadou, Triantafillou, & Economides, 2007). Thissen, Steinberg, & Mooney (1989) argue that for a test consisting of testlets, unidimensionality and local independence is better assured for MSTs than CATs. Zenisky et al. (2010) also noted

that when incorporating mixed-format tests into the test design, the overall system design tends to be much easier to develop and implement.

MSTs do not come without any disadvantages when compared to CATs. First the fully-adaptive test have better measurement precision over a typical MST design when all model assumptions are met, because tailoring takes place at the item level in CAT and therefore adapts at more points. This leads to a need for more items to be given using MSTs in order to ensure equal measurement properties compared to CAT administrations. When replacement is needed for items within a testlet it can be difficult to exchange items because of dependency of testlet-based items (Wainer & Kiely, 1987). In addition, the potential for routing errors can be burdensome especially in the case of two-stage testing (Weiss, 1974).

Using MSTs are appealing because of its compromise between LFTs and CATs. MSTs include more adaptive points than traditional testing but less than item level CATs. Because MSTs are a compromise between both the LFT and CAT testing platforms, MSTs are now being implemented in practice. MSTs that have been researched or are currently in operational use include the Law School Admissions Test (LSAT), the Graduate Record Examination (GRE), the Uniform Certified Public Accountant Examination (UCPAE), the Test of English as a Foreign Language (TOEFL), the National Council of Architectural Registry Board (NCARB), the National Assessment of Educational Progress (NAEP), and the U.S. Medical Licensure Examination (USMLE) (Bock, Zimowski, & Panel, 1998; Davey & Lee, 2011; R. Luecht, Brumfield, & Breithaupt, 2006; R. M. Luecht & Nungester, 1998; Schnipke & Reese, 1997; Wainer, Lewis, Kaplan, & Braswell, 1990; Wainer & Lukhele, 1997).

Basic MST Components

Basic features for all adaptive tests are still relevant for MSTs. Fundamental decisions must be made that outline the structure of the test, such as, the total number of items in the test, the number of stages in the test, the number of items within each stage, and the overall test blueprint that all guide the development process. Considerations related to the adaptive nature of MSTs entail within and total test scoring, stage construction, between stage navigational procedures and algorithms, and the number of forms or *panels* available during an administration window.

The structure of an MST has been categorized into four main components: *panels*, *stages*, *modules*, and *pathways*. Together, the relationship between components dictates the administration of the test to an examinee. A panel is the overall set of items an examinee could be administered. The items are catalogued into sub-units of items called modules. Modules are generally constructed based on both qualitative and quantitative features, where both content specifications and the item's statistical properties are used to construct modules of varying difficulty. A stage can be thought of as the levels of a panel. A stage consists of one or more modules that typically vary in difficulty. The first stage in an MST is often, but not limited to, one module referred to as the routing test. Based on the performance on the routing test adaptation occurs between the first and second stage where an examinee is routed to one of the modules at the next stage. At the end of each stage are the adaptive points, where performance on the current and/or previously administered modules are used to identify the navigational pathway to the next module in the subsequent stage with exception of the final stage.

One commonly researched panel design cited for MSTs is a 1-3-3. (e.g. Hambleton & Xing, 2006; Jodoin et al., 2006; Keng, 2008; Kim, 2010; R. Luecht et al., 2006; R. M. Luecht & Nungester, 1998; Zenisky et al., 2010). This design has three

stages. One module is available at Stage 1, three modules are available at Stage 2, and three modules are available at Stage 3. Each respective stage beginning with Stage 1, has one module, three modules, and three modules that an examinee could potentially be administered. Figure 2 is an example of a 1-3-3 MST design with three panels. Here it can be seen that an examinee has three navigational pathways as they finish the routing test in Stage 1. The pathway chosen will lead the examinee to a module in Stage 2. For a person in Stage 2, easy and hard modules then have two pathways available to move from Stages 2 to 3 and a person in the medium module will have three possible pathways. Generally it has been suggested that only adjacent steps can be made by an examinee (Luecht & Nungester, 1998), and thus the reason for only two pathways for the easy and hard modules. The connective lines indicate potential pathways an examinee can navigate based on performance during administration and the policy decision related to routing procedures. The 1-3-3 panel design is not the only possible design but is probably the most prominent design structure across research studies. Variations in panel designs can potentially entail an increase or decrease in both modules and stages.

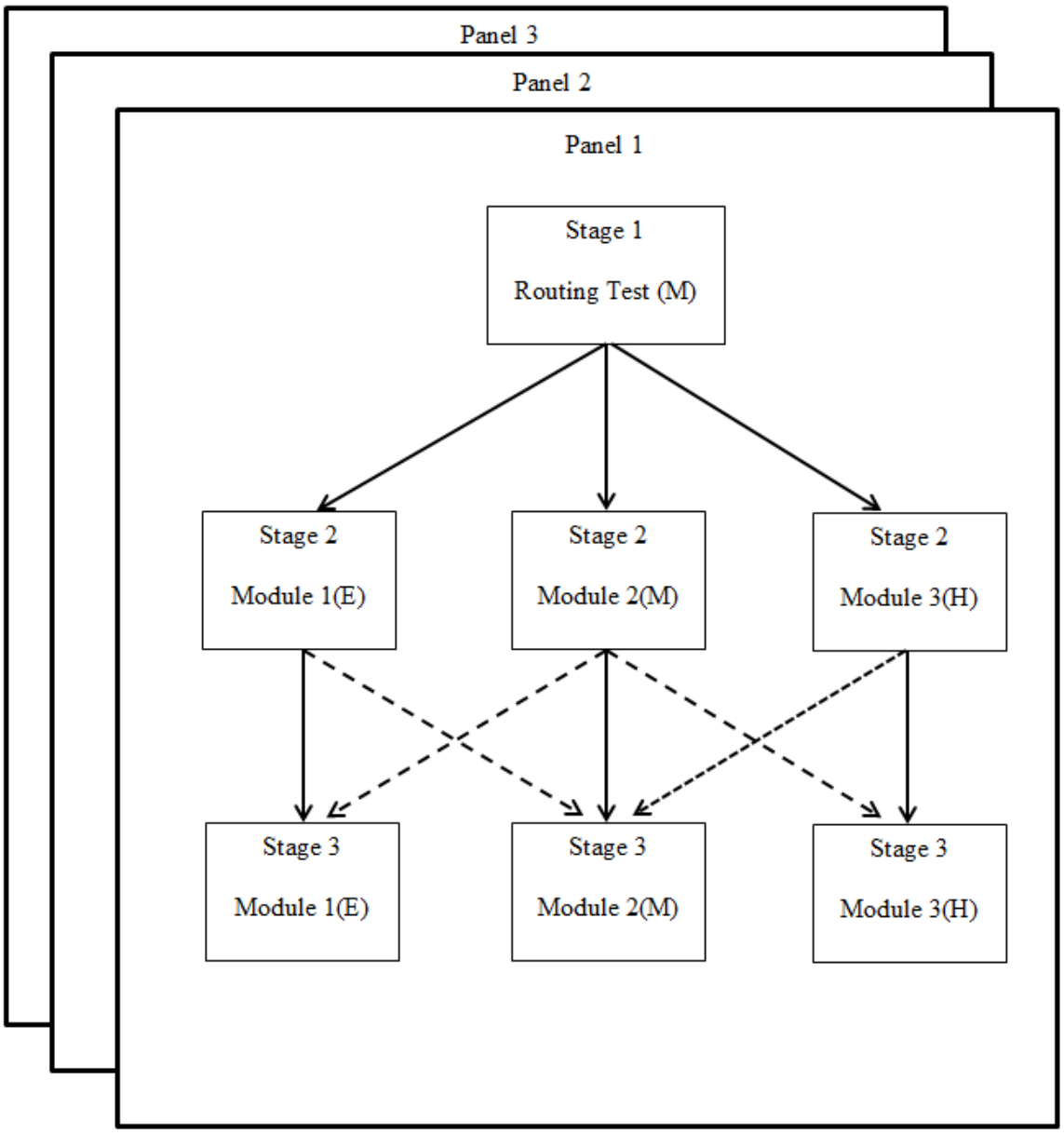

Figure 2. Design for a 1-3-3 adaptive multistage test with multiple panels. The dark bold lines are the primary routes and the dashed lines are the supplementary routes for an examinee. E = easy difficulty module; M = medium difficulty module; H = hard difficulty module.

The skills and content being assessed dictate the size and types of items that need to be developed for an item pool (Schmeiser & Welch, 2006). Policy decisions such as those regarding test blueprint have implications on the proportion of the item pool that relate to specific content. In addition, an item pool should be representative of the types of tasks being developed for the assessment. For instance, when a test consists of both stand-alone multiple-choice items and testlet-based multiple choice items, the depth or number of each type of item must be inspected in order to ensure the ability to create parallel modules and potentially parallel panels. Additionally, if a test is intended for a broad ability range, the item pool should have a range of difficulty and discrimination parameters (Xing & Hambleton, 2004).

Target Test Information Function

The use of IRT can construct modules of varying difficulty. Defining the desired statistical targets for a range of abilities can be achieved by using the information from individual item parameters (Breithaupt & Hare, 2007). To achieve a certain level of measurement precision, one can target specific θ's on the ability continuum as the *target test information functions* (TTIFs). Recall that the test information function is the sum of the item information as seen from Equation 5. A TTIF represents the amount of estimated error variance for a range of θ's that is willing to be tolerated for a panel and/or a sub-unit of items (Luecht & Nungester, 1998). Targeted information can then be used to inform the assembly mechanism about the overall desired test information for the panel and the "test" information of a sub-unit of items when building modules. TTIFs are one of the main components that guide the simultaneous assembly of modules and panels.

Automated Test Assembly

In an attempt to populate the modules with the desired psychometric properties and maintain the integrity of the testing blueprint, automated testing assembly (ATA) is an effective way to assemble modules from an existing item pool (Melican, Breithaupt, & Zhang, 2010). Four main features must be deliberated in order to specify the necessary constraints for a successful ATA: defining the objectives in assembly, the TTIFs, the number of modules, and the stringency of each test constraint. (Breithaupt, Ariel, & Veldkamp, 2005). Defining objectives essentially refers to defining the decision variables needed for assembly. The stringency of test constraints is the weighting of various constraints that exist in conjunction with the item pool. For instance, if one constructs 3 panels, will the panels be allowed to repeat items across panels? Is that of less concern than having duplicate testlets across panels? Researchers have been successful with such dilemmas when constructing MSTs by using ATA. Linear programming approaches have been notably successful at solving the many constraints that must considered when building the panels (Theunissen, 1985; Van Der Linden, Veldkamp, & Reese, 2000). Linear programming essentially creates optimal solutions to a complex set of equations. Mixed Integer Programming (MIP) has been one of the more prominent forms of successful linear programming. MIP approaches test assembly by defining constraints such as exposure control or content representation as decision variables (Cor, Alves, & Gierl, 2009). Ultimately, decision variables are defined in such a way that a binary decision is made by the program solver, or a variable given the binary options of a (1) is included and (0) is excluded in the final form. Mathematically this can be expressed as follows:

$$X_{it} = \begin{cases} 1 \text{ if item } i \text{ is included in test form } t \\ 0 \text{ if item } i \text{ is excluded in test form } t \end{cases}. \quad (18)$$

The next step is to express an objective function. An objective function is an equation that changes as decision variables change. For instance, it might be desired to target or maximize the information of panel for a particular value. The objective function can then be expressed as follows:

$$Maximize: \sum_t \sum_k \sum_i I_i(\theta_{kt}) x_{jt} , \tag{19}$$

where Equation (19) states that the information function *I*, across all items *j*, across the specified abilities, θ_k, across each test, *t*, times the decision variable x_{jt}, should be maximized.

The next types of equations are the item and test level constraints set forth by the test developers (Cor et al., 2009). Types of constraints for a 3PL model potentially include limits on the item discrimination, pseudo guessing parameters, or certain levels of item difficulty that will be allowed. For example, say a developer only wanted to allow items with item difficulty, b_j, between -1.5 and 1.5. This can be mathematically expressed as:

$$-1.5 \le x_{jt}(b_{jt}) \le 1.5, \text{for all items } j \text{ on all tests} t. \tag{20}$$

An instance of test level constraints may consist of content related constraints. An example may be that it is desired to have 9 items from the first content category, C_1, for each test form. The constraint could then be represented as:

$$\sum_{j \in V_{C_i}} x_{jt} = 9 , \tag{21}$$

or the sum of items that are a subset of content category C_1 is equal to nine.

Once all the decision rules are in place an ATA algorithm can be used to assemble a test. A number of successful MIP based ATAs software has been developed such as CASTISEL (R. M. Luecht, 1998), JPLEX (Park, Kim, Dodd, & Chung, 2011), or lp_Solve (Diao & van der Linden, 2011a).

Total Test Scoring and Ability Estimation

Scoring MSTs has two main considerations. The first concerns estimating an overall ability upon test completion. The second consideration for scoring modules during the actual test administration to inform the MST routing procedures. Final ability estimates are generally calculated using one of the following point estimate techniques, maximum likelihood estimates (MLE) or a Bayesian estimators like maximum a posteriori (MAP) and expected a posteriori (EAP).

The MLE procedure uses the maximum of the likelihood function for examinee *j* with a given ability θ_j as follows:

$$L(x_j \mid \theta_j) = \prod_{i=1}^{n} P_{ij}(\theta_j)^{x_{ij}} (1 - P_{ij}(\theta_j))^{1-x_{ij}}, \quad (22)$$

where *n* is the number of items, x_j is the vector of item responses for examinee *j*, with a given ability, θ_j , and $P_{ij}(\theta_j)$ is the probability of answering item *i* correctly for a given ability (Lord, 1980). MLE estimators are advantageous because they are asymptotically efficient unbiased estimators. Disadvantages occur for examinees who either answer all items correctly or incorrectly, which creates a situation where the likelihood does not have a mode or maximal point (Parshall, Spray, Kalohn, & Davey, 2002).

Bayesian procedures produce point estimator of θ by using the posterior distribution. Recall from Equation 12 that $P(\theta \mid x)$ is the posterior distribution for θ, the parameter of interest. MAP is an iterative method that uses the mode of the posterior distribution (see Lord, 1986; Mislevy, 1986). Alternatively, EAP uses the mean of the posterior distribution and is a non-iterative Bayesian approach (R. D. Bock & Mislevy, 1982). Unlike MLE techniques, Bayesian methods will provide a point estimate for response vectors that are all correct or incorrect. However, Bayesian methods tend to

underestimate ability estimates at high abilities and overestimate lower abilities (Parshall et al., 2002).

Module Scoring and Routing Procedures

Because of the adaptive nature of an MST between modules, scoring during test administration must also be considered to help inform the routing procedures put in place. Routing procedures dictate which pathway is used to navigate from one module to the next. The decision points of a routing procedure indicate which pathways will be navigated and therefore must be determined prior to administration. Generally, one of the advantages of an MST is the adaptive administration coupled with the capability of an examinee to review items within a module. Consequently, scoring takes place during administration once a module has been submitted. Then the current level of performance is used to select a module at a subsequent stage. One option, similar to a CAT, is to score and route examinees based on the selecting the maximum module information (i.e. test information for a module) for a given ability estimate based on one of the aforementioned point estimate methods like MLE (Chuah, Drasgow, & Luecht, 2010; Davis & Dodd, 2003; Jodoin et al., 2006; Kim & Plake, 1993), MAP (Schnipke & Reese, 1997) or EAP (Hambleton & Xing, 2006; Luecht et al., 2006). Other scoring procedures are dictated by the routing methods implemented through the use of either *true score routing* or *theta routing*.

Both the true score routing and theta routing are achieved by defining decision points on the ability continuum that will route an examinee from one module to the next. For theta routing, after a module has been submitted an ability estimate is performed through one of the aforementioned point estimate procedures. Then based on the decision points made prior to administration, the examinee is routed to the corresponding

module at the next stage. True score routing is achieved by using the decision points on the ability continuum to estimate an examinee at that point's true score (Luecht et al., 2006). Then based on the number correct (NC) for an examinee on all previous items, the examinee is routed to the corresponding module. For example, the decision point for two pathways from Stage 1 to Stage 2 is a true score of 5.6. If an examinee answers five items correct then of the two possible modules regarding the decision point, the examinee is routed to the easier of the two modules at Stage 2, but if an examinee answers 6 items correctly (s)he will be navigated to the harder of the two modules. It should be noted that both types of module scoring are intended to yield the same routing patterns, where the true score was previously used as a means to get around the computation intensity of in-test ability estimation.

The needs of the testing program impact the decision for which particular scoring method is implemented. For instance, it may be desired to control the overexposure of items and therefore send approximately the same proportion of people to all modules; it may be desired to select based on the optimal statistical properties of a module; or it may be desired to select a module based on the optimal statistical properties with respect to a particular cut point for decision based purposes. Multiple routing procedures have been developed to answer the needs of testing programs. The two most prominent methods in the literature are the approximate maximum information (AMI) and the defined population interval (DPI; Luecht et al., 2006). Originally these methods were developed using true score techniques but have since been adapted to use the theta scoring routing (e.g. Hambleton & Xing, 2006; Kim, Chung, Park, & Dodd, 2013).

The AMI procedure compares the TIF for adjacent modules and solves for the intersection point on the ability continuum. Once the intersection is located on the ability scale this becomes the routing decision point. Then an estimated true score or

theta estimate is compared to the decision point and the module at the subsequent stage is selected and administered.

The DPI is another routing procedure that aims to control the proportion of examinees that see each module. The DPI is particularly useful for programs interested in more directly dictating pool utilization and exposure rates of items (Luecht et al., 2006). Figure 2 illustrates the pathways available in a typical 1-3-3 panel design. Then if a testing program is interested in maintaining proportional module exposure, the program will define the routing procedure in such a way that sends one-third of examinees to the easy, medium, and hard modules after the routing test is submitted. This is accomplished by assuming a normal ability distribution and then finding the theta's associated with the 33rd and 67th percentile. Then using either true score routing or theta routing approximately proportional modules are administered. Similarly, to maintain proportionality the decision points are used at each stage.

Additional methods have been proposed such as the proximity method by Kim & Plake (1993). The proximity method compares average difficulty of adjacent modules to the current ability estimate. Whichever module has the minimum absolute deviation is selected for administration. However methods like the proximity method are problematic because they do not take into account all of the item parameters when selecting modules and therefore the current study will focus on the two most prominent methods in the AMI and DPI.

MST Findings

MST research can be classified into two types of studies. One type compares the MST to various test delivery platforms, usually consisting of a CAT and sometimes including an LFT. The second type of study only considers the MST platform and

compares various aspects of an MST's components to further understand its operational characteristics. The first section to be discussed concerns the testing platform, or how MST compares to both CATs and LFTs. Included in the testing platform are findings regarding measurement precision, decision consistency, and exposure control findings. Then general operational characteristics are in the second section. Specifically, the following MST components will be discussed: panel design, test length and total test information, stage level information, and routing procedures. Finally, MST research based on the 3PL-TRT model is discussed.

MST versus Other Test Delivery Platforms

The majority of MST studies have compared an MST to other test delivery platforms. The two primary comparisons are made between LFTs and CATs. As one might expect, the overall performance of an MST falls somewhere in the middle of LFTs and CATs. When comparing CATs to MSTs for dichotomous item pools CATs consistently outperform both MSTs and LFTs (Edwards, Flora, & Thissen, 2012; Hambleton & Xing, 2006; Kim & Plake, 1993; Patsula, 1999; Xing & Hambleton, 2004; Zheng, Nozawa, Gao, & Chang, 2012). These studies consistently offer two main findings. Item level CATs show higher levels of measurement precision and decision consistency than both MSTs and LFTs. In addition, MSTs generally outperform LFTs of the same test length. This supports the utility of MSTs because increased precision can be attained over an LFT, while allowing both the testing program and examinees the ability to review items.

When a test is delivered from a dichotomous item pool with no constraints the results for test delivery platforms are fairly straightforward. However, when a CAT is administered with constraints like exposure control or when item selection procedures

administer entire testlets, mixed findings have been recorded. Schnipke & Reese (1997) conducted a study that included LFTs, MSTs, and two CATs under the 3PL model. One CAT was fully adaptive at the item level, and the other was a testlet-based CAT administration where items were administered by testlets. Not surprisingly the item level CAT showed the highest level of measurement precision. When comparing the testlet-based CAT to the MST, results indicated similar performance between the two testing platforms. Keng (2008) compared an item level CAT, testlet-based CAT, and a 1-3-3 MST panel design for an item pool consisting of solely passage-based testlet items. All testing platforms in the study included content balancing based on the originating test. For longer test lengths, the MST outperformed the testlet-based CAT with respect to measurement precision, and for all other conditions the two platforms performed similarly. In addition, the MST consistently attained higher levels of pool usage, better exposure control rates, and consistent item overlap for similar examinees. However, when the study condition generated negatively-skewed ability distributions the module exposure rates rose dramatically, indicating a non-normal ability distribution may pose additional security concerns for a testing program when using AMI for module selection.

When polytomous item pools have been studied, mixed results were found when comparing the CAT administrations to the MSTs (Davis & Dodd, 2003). Note, understanding the impact of exposure control was the primary objective for the study. With respect to exposure controls, the MSTs outperformed the CATs with various exposure control procedures. In addition, smaller amounts of bias were observed for the MST system. The CATs did result in better root mean square error (RMSE) and the average absolute deviation (AAD), two measures of the measurement precision for ability estimates.

Kim (2010) compared a number of adaptive formats using a mixed-format item pool consisting of dichotomous and polytomous items. The primary focus was to compare the decision consistency and decision accuracy of the various testing platforms. The item-level CAT and sequential probability ratio test, an adaptive test that incorporates decision rates into the item selection process, both outperformed the MST. However, it should be noted that all platforms performed adequately with regards to pass/fail decisions. In support of MSTs, the study also found that the MST maintained adequate decision consistency with lower exposure rates and increased pool usage.

From the research that compares MSTs to other testing programs, two main findings supporting the use of MSTs can be stated: (1) The more constraints put on a CAT-based delivery such as content balancing, exposure control, and testlet-based administration, the more similar the psychometric performance of CATs and MSTs become. (2) MSTs consistently outperform LFT administrations.

MST Operational Characteristics

The previous section summarized the current understanding across test delivery platforms. The current section synthesizes findings to provide insight into the current understanding MST's operational characteristics.

Panel design and test length are two of the initial decisions developers must make when constructing an MST. Panel designs effect virtually all other aspects of the administration, such as exposure control, item pool size, or routing of examinees. The panel design is probably the most researched aspect of an MST, with the 1-3-3 being the most prominent throughout studies (e.g. Armstrong et al., 2004; Chen, 2010; Edwards et al., 2012; Galindo et al., 2013; Jodoin et al., 2006; Kim & Plake, 1993; Kim, Chung, Dodd, & Park, 2012; Patsula, 1999; Zenisky, 2004; Zheng et al., 2012). Two main

findings have been found regarding the panel design. The more stages and the more modules at each stage increase the measurement precision. This occurs because increased adaptation points create an administration more similar to a fully adaptive test. Researchers have then suggested that for the sake of implementing an MST, three stages and up to five modules will likely suffice for appropriate decision consistency and ability estimation (Brossman, 2014, Chen, 2010; Kim et al., 2012; Patsula, 1999; Xing & Hambleton, 2004; Zenisky, 2004; Zheng et al., 2012).

Intuitively, it has been known for decades, even before the development of modern test theory, that a longer test is more precise than a shorter test. Consistent with intuition, studies that directly compare test length, consistently show that a longer test, or a test with more information, is more precise (Chen, 2010; Galindo, Park, & Dodd, 2013; Hambleton & Xing, 2006; Jodoin et al., 2006; Kim & Plake, 1993; Kim et al., 2013; Kim, 2010; Zenisky, 2004; Zheng et al., 2012).

Mixed results have been observed when comparing the interaction between overall test information and stage-level information for dichotomous item pools. The most comprehensive studies in regards to stage-level information conditions were conducted by Patsula (1999) and Zheng et al. (2012). Both studies used increased information at the beginning, middle, and final stage levels respectively and compared them to equal stage information. Patsula found that equal stage information lead to the most accurate ability estimates. Zheng et al. (2012) found little differences between the various stage level differences, with the exception of the condition high levels of information in the middle stage, which yielded the least precise ability estimates.

Additional studies have been conducted that investigated varying stage-level information conditions. Kim and Plake (1993) studied a number of two-stage panels with varying number of two-stage modules (6, 7, or 8 modules) and various routing test target

TIFs for a dichotomous item pool. The study found little difference between the number of modules or test length used. The study did find increased measurement precision for routing tests with a uniform distribution of information across the targeted ability range. Zenisky (2004) compared both test level information and stage level information using a dichotomous item pool by comparing two distributions o stage-level information: (1) equal distribution across stages and increase information at Stage 1, or (2) routing stage information. The study found that tests with the highest levels of overall TIFs resulted in the highest levels of precision when the stages TIFs were equal. However, when lower overall TIF's were present, the panel design was more precise for higher levels of Stage 1 information. Like Zenisky (2004), Jodoin et al. (2006) also considered different stage-level information conditions. In one condition, stage information was equal, while in the other condition routing stage information was decreased while increasing Stages' 2 and 3 information. This study found little impact on the decision accuracy rates when comparing the stage information conditions. Edwards et al. (2012) used an MST with increased information at Stage 3 to compare to a longer LFT and found that the MST yielded higher reliability for the MST. The generalization of these studies is tenuous as the comparisons of stage level information were limited and findings were mixed.

Research has also considered the impact of stage-level information on polytomous item pools. Macken-Ruiz (2008) studied MSTs using a polytomous item pool under the generalized partial credit model. The study used three MST designs with an equal proportion of information design, an increasing subsequent stage information design, and a decreasing subsequent stage information design. The condition with information increasing with each stage resulted in the best measurement precision. Chen (2010) varied the routing test information under the generalized partial credit model for a polytomous item pool. Overall, the MST test structures produced similar results and little

difference was found between the two routing test conditions. Kim et al., (2012) varied the level of information in the routing stage for a mixed-item pool under the generalized partial credit model. The study found that more information in the routing stage resulted in better decision consistency. A more recent study, conducted by Galindo et al. (2013) investigated the effects of overall TIF with routing module information. The study indicated the strongest factor in measurement accuracy and precision was the overall length of the test and little differences were found between the routing modules with the same overall test length.

Overall, the more items on a test, the better the measurement precision. Varying stage level information has produced mixed results. Increased routing module information indicated better decision accuracy, while other studies have indicated increased precision when the final stage modules have increased information.

The overall performance of an MST is also influenced by routing procedures. Most research on routing procedures has taken place in the context of classification testing. Hambleton and Xing, (2006) used two forms of the DPI, one with consistent decision points as described by Luecht et al. (2006), the other derived decision points from the approximate standard error of the target TIF's center. No difference in the two methods was detected and both performed adequately with respect to decision consistency measures and exposure rates were controlled as expected with the DPI. Weissman, Belov, and Armstrong, (2007) compared the AMI and DPI. AMI outperformed DPI with respect to decision consistency measures. However, the DPI utilized the modules at a much higher rate than did the other methods. Zheng et al. (2012) compared theta routing with number correct routing for the AMI procedure. They found little difference in either routing procedure. Kim et al. (2013) conducted the most recent study which compared the AMI procedure with two DPI procedures for a mixed-

format dichotomous/polytomous item pool. The first DPI procedure proportionally routed examinees based on their stage-level provisional ability, and the second procedure routed examines based on their within-module rank order. The authors found little difference between the routing procedures with respect to decision consistency, but no bias or accuracy measures were reported for the examinees ability estimation.

Because these studies focused on classification testing, little consideration has been given to ability estimation for the broader population. Zenisky (2004) did compare the AMI, DPI, and the proximity method. The AMI resulted in the highest measurement precision, but all methods were considered adequate. With respect to exposure control rates, the DPI routed examinees to each module with equal proportions, while the AMI administered certain modules at higher rates.

Overall, the AMI has been shown to consistently outperform other routing procedures because it selects modules based on maximum information. However, the DPI controls the rate of module administration much better than the AMI, while maintaining adequate levels of decision consistency. More research needs to be conducted with regards to the recovery of abilities and routing procedures in the MST context.

MST Under the 3PL-TRT Model

To date few studies have compared MSTs under the 3PL-TRT model (Galindo et al., 2013; Keng, 2008; Lu, 2010). Moreover, only Lu (2010) investigated a mixed-format testlet-based item pool. Lu (2010) found that the higher the presence of LID the more a test needed to be scored under the 3PL-TRT model. Lu only administered a module standalone items, or a module of one testlet. A testing program might want to administer a combination of standalone items and testlets in a given stage.

In addition, a number of studies have used testlet-based item pools that combine the testlet responses into a polytomously-scored item and used a polytomous scoring model (Chen, 2010; Davis & Dodd, 2003; Davis, 2004; Kim, 2010) But as noted in the TRT section, using a polytomous model on testlet-based items rather than a TRT model results in a loss of information concerning the examinee's ability. Findings do not necessarily generalize across models and item pool types. Therefore, the use of the 3PL-TRT model with mixed-format data needs to be conducted for MSTs.

STATEMENT OF PROBLEM

The use of MSTs has been recommended as a compromise between traditional LFTs and fully adaptive CATs (Hendrickson, 2007). Most MST research has compared MSTs to other testing platforms. In addition, some studies have considered the operational characteristics of MSTs, but much of this work has been done with respect to classification testing. Although some classification studies examined both measurement precision and decision indices (e.g. Zenisky, 2004), the majority of these studies focused on the consistency and accuracy of decision making when using an MST. One of the benefits of MSTs is that a testing program can for control content specifications, item pool usage, and item exposure control rates with the panel design prior to administration (see Georgiadou, Triantafillou, & Economides, 2007). Although it has been consistently shown that item-level CATs generally result in higher measurement precision, some studies have noted that using exposure controls, content constraints, or a testlet-based CAT has produced comparable results between a CAT and an MST administration (Davis & Dodd, 2003; Keng, 2008; Schnipke & Reese, 1997). Because of the promising findings when comparing the comparability of MSTs and CATs, further studies need to be conducted that explicitly try to optimize the psychometric and operational

characteristics of MSTs as an alternative to a fully adaptive CAT. More specifically, more research is needed to consider panel designs, test length, routing procedures, and the impact of these MST components on measurement precision for MSTs.

Researchers have investigated a number of panel designs structures. Because many studies focused on classification testing, construction of panels and modules sometime target the various cut scores being used (see, e.g. Hambleton & Xing, 2006; Kim et al., 2013; Lu, 2010; Zenisky, 2004). While this approach functions for classification testing, the approach can lead to less measurement precision for the upper and lower extremes of the distribution when the test purpose focuses on ability estimation.

Generally, longer tests lead to higher reliability and decreased measurement error. This notion was mathematically proven in classical test theory (Brown, 1910; Spearman, 1910), and has been a supported benefit of adaptive MSTs (Chen, 2010; Edwards et al., 2012; Galindo et al., 2013; Keng, 2008; Kim, 2010; Lu, 2010; Zheng et al., 2012). But at what point does a test become too short? Stark and Chernyshenko (2006) suggested that test length in most studies has not been short enough to really find the point where MST components and measurement precision interact and breakdown. The current study explores different test lengths in order to find that point.

Routing procedures are critical components of MSTs that help testing programs maintain desired exposure rates while tailoring an examinee's test. To date only a handful of studies have directly compared routing procedures (Edwards et al., 2012; Hambleton & Xing, 2006; Kim et al., 2013; Zenisky, 2004; Zheng et al., 2012). Of those studies, only one did so with a mixed-format item pool containing dichotomous and polytomous items (Kim et al., 2013). However, Kim et al. (2013) only used two DPI procedures that proportionally routed examinees to non-adjacent modules beyond Stage

2. This is not generally found in applied settings because it is too far a jump for an examinee (Breithaupt & Hare, 2007; Luecht et al., 2006) and has been empirically demonstrated to occur very rarely (Hambleton and Xing, 2006). In addition, none of these studies examined routing procedures when using a 3PL-TRT model.

Many tests are mixed-format meaning they contain multiple item types. For example the Advanced Placement (AP) (College Board, 2004), National Assessment of Educational Progress, or state assessments like those administered in Wisconsin and North Carolina, (see Rosa, Swygert, Nelson, and Thissen, 2001) use mixed-format assessments. In some cases, mixed-format testing comes about because skills being assessed create a need for constructed responses or rubric-scored items coupled with multiple choice items. In other instances, mixed-format tests testing is created by developing stand-alone multiple-choice items alongside testlet-based items (Wainer & Kiely, 1987).

When testlets are scored as stand-alone items, local independence is violated and results in overestimated discrimination parameters, a_j, which in turn leads to overestimated TIF and underestimated SE's (Murphy et al., 2010; Sireci et al., 1991; Wainer et al., 2000). When items are scored polytomously, a loss of information can occur because the score pattern is not taken into account (Wainer et al., 2007). Yet, only three studies have investigated the use of a 3PL TRT model in the context of MSTs (Galindo et al., 2013; Keng, 2008; Lu, 2010). Moreover, only Lu (2010) studied the effects of the model with mixed-form testlet item pool. However, the simulation study conducted by Lu (2010) only considered the first scenario presented by Wainer et al. (2007), where a large number of testlet-items are developed and a subset of items from the testlet are adaptively administered. The current study investigated the latter scenario

presented by Wainer et al. (2007), where whole testlets mixed with stand-alone items are administered to examinees.

As more testing companies are implementing MSTs for purposes of estimating broader ability levels, understanding the operational characteristics of MSTs become increasingly important. The present research increases the body of literature on the operational characteristics of MSTs with a mixed-format testlet-based item pool. Additionally, numerous testing companies have utilized mixed-format testlet-based assessments. The IRT family is typically used to score such assessments. It behooves the measurement to investigate the fidelity of scoring a mixed-format testlet-based assessment under 3PL-TRT model in realistic testing scenarios. Studies that investigate the scoring under the 3PL-TRT model will help inform testing programs about the recovery of ability estimates when administering mixed-format assessments. Specifically, this study investigated measurement accuracy and precision of ability estimates for an MST with multiple panel designs under a 3PL TRT model administering a mixed-format testlet administration. In addition, routing procedures under a mixed-testlet structure have never been investigated under the 3PL TRT model. The magnitude of LID has been shown to impact the administration of MSTs and the recovery of ability estimation. Because testlets can violate the assumption of local independence, the current study used results from an operational form to estimate the magnitude of the testlet effect, as well as, varying the magnitude of local item dependence (LID). Covering a wider range of LID helps generalize the findings as they relate to item banks with various levels of LID when administering mixed-format testlet-based assessments. Investigating the various test designs, routing procedures, and LID under the 3PL TRT model will provide testing programs with guiding principles for the implementation of a mixed-format testlet-based MSTs.

Research Goal

The primary goal of this study is to examine the operational characteristics of MSTs for a mixed-format testlet-based item pool under the 3PL-TRT model. Standardized tests with mixed-format testlet-based item pools are being administered in practice such as the GRE and SAT (Wainer et al., 2000), where the GRE illustrated the presence of a potential testlet effect. Therefore, it behooves the psychometric community to conduct studies, such as this dissertation, to investigate the applicability of an MST under the 3PL-TRT model.

Research Questions

Four research questions are then needed to evaluate the operational characteristics of an MST under the 3PL-TRT model.

(1) *How does panel design impact the measurement precision of an MST with a mixed-format testlet-based item pool?*

(2) *How does test length impact the measurement precision of an MST with a mixed-format testlet-based item pool*?

(3) *How does the magnitude of LID effect the administration and ability estimation of an MST with a mixed-format testlet-based item pool*?

(4) *How do various routing procedures impact the ability estimation of an MST with a mixed-format testlet-based item pool*?

(5) *How do panel design, test length, LID, and routing procedures interact with respect to the accuracy and precision of ability estimation*?

Chapter 3: Method

Design Overview

Four multistage test (MST) designs were compared across several manipulated test conditions. The MST conditions include two, three-stage MSTs and two, two-stage MSTs. Additional test conditions include two total test lengths, three routing procedures, and three testlet effect conditions. When being routed from Stage 1 to Stage 2, the two defined population interval (DPI) procedures in the study are mathematically equivalent and result in the same administrative procedure. The resulting design for the three-stage tests is (2 panel design x 2 test lengths x 3 routing procedures x 3 testlet effects) 36 manipulated conditions. The results for two of the three routing procedures used in a two-stage test are equivalent. Therefore, the design for the two-stage tests yields (2 panel designs x 2 test lengths x 2 routing procedures x 3 testlet effects) 24 manipulated conditions. A total of 60 conditions were investigated for the current study. Table 1 contains the breakdown of the various conditions for the three-stage and two-stage panel design conditions.

The item pool and parameter estimation procedures are described first. Then the manipulated conditions are fully explained, followed by the procedures to generate simulees' responses. The penultimate section fully expound upon the MST simulation and administration, including the automated test assembly (ATA) and simulation's administrative procedures. Finally, the data analysis is described.

Table 1. Multistage Test Study Conditions.

Panel Design	LID	Test Length	Routing Procedure
1-5-5	$\sigma^2_{d(j)} = \begin{pmatrix} 0.0 \\ 0.8 \\ \hat{\sigma}^2_{d(j)} \end{pmatrix}$	Long	AMI
			ML-DPI
			SL-DPI
		Short	AMI
			ML-DPI
			SL-DPI
1-3-3	$\sigma^2_{d(j)} = \begin{pmatrix} 0.0 \\ 0.8 \\ \hat{\sigma}^2_{d(j)} \end{pmatrix}$	Long	AMI
			ML-DPI
			SL-DPI
		Short	AMI
			ML-DPI
			SL-DPI
1-5	$\sigma^2_{d(j)} = \begin{pmatrix} 0.0 \\ 0.8 \\ \hat{\sigma}^2_{d(j)} \end{pmatrix}$	Long	AMI
			SL-DPI=ML-DPI
		Short	AMI
			SL-DPI=ML-DPI
1-3	$\sigma^2_{d(j)} = \begin{pmatrix} 0.0 \\ 0.8 \\ \hat{\sigma}^2_{d(j)} \end{pmatrix}$	Long	AMI
			SL-DPI=ML-DPI
		Short	AMI
			SL-DPI=ML-DPI

Note. LID=Local Item Dependence; AMI=approximate maximum information; SL-DPI=Stage-level defined population interval; ML-DPI=module level defined population interval.

ITEM POOL DEVELOPMENT

The item pool development took place over four basic steps. The first was obtaining parameter estimates from real-data responses. Following the item parameter estimation, the item pool was concatenated four times to create an adequately sized pool for use in the current study. Item responses were generated once the pool was expanded. Finally, the generated item responses for the expanded pool were recalibrated.

Real Dataset

Three test forms from a nationally administered test were used for the study. Item responses from the large-scale passage-based exam were used to estimate the item parameters for the simulation study. The item responses consist of a random sample of approximately 100,000 examinees for each form. Actual responses were used to estimate item and testlet parameters. Each of the three forms consisted of 67 unique multiple-choice items, for a total number of 201 items. Each form was a mixed-format pool of standalone items and passage-based testlet items. Each form consisted of either 19 or 20 stand-alone items and 6 or 7 testlet based items. The items associated with a testlet range from 2 to 14 items.

At this point it is pertinent to the discussion to introduce the concept of test units. Each type of item is considered a test unit. If the item is a standalone that represents one type of test unit with two possible score points. If the test unit is testlet-based with two questions associated, it is a test unit with three possible score points (i.e. 0-2). If the testlet has 13 questions associated with a reading passage, it is a test unit with 14 possible score points (i.e 0-13). Nine different test units were identified across the three forms. Each of the original forms consisted of 26 total test units of various sizes. In total 58 test units were stand-alone items, five test units were two-item testlets, three test units were four-item testlets, two test units were six-item testlets, two test units were seven-item

testlets, two test units were nine-item testlets, two test units were twelve-item testlets, three test units were thirteen-item testlets, and one test unit was a fourteen-item testlet.

Parameter Estimation

Parameter estimation for items and testlets were estimated under the three-parameter logistic testlet response theory (3PL-TRT) model (Wainer et al., 2000) by using the SCORIGHT software (Wang, Bradlow, & Wainer, 2005). Under the 3PL-TRT model, item and testlet parameters were estimated for each item *j*, the discrimination a_j, the difficulty b_j, the pseudo-guessing c_j. In addition, the variance of the testlet effect, $\sigma^2_{d(j)}$, for each testlet $d(j)$ was estimated. Allowing for variability in each testlet $d(j)$ permits varying degrees of local dependency among the testlets associated items.

SCORIGHT is a general computer program that can model dichotomous and polytomous items or any combination of the two in a fully Bayesian framework. SCORIGHT estimates parameters using a Markov chain Monte Carlo (MCMC) techniques (Geman & Geman, 1984). All three forms were calibrated using the priors as previously specified in the TRT section and as suggested by Wainer et al. (2007). Two chains were used for each form with a total of 20,000 iterations. During each chain, the first 12,000 iterations were discarded, referred to as the *burn-in period*. A burn-in period is used to stabilize the posterior distribution. Final item parameter estimates were based on every eighth draw of the posterior distribution from the remaining 8,000 iterations, a process referred to as thinning. Similar procedures were used by Boyd (2003) and Keng (2008) when estimating the 3PL-TRT model.

Simulated Dataset

During the panel construction one goal was to maintain test unit proportionality. This objective was expressly meant to maintain the ratio of standalone items and testlet-

based items from the original. The most complex panel design was the 1-5-5. To construct a three 1-5-5 panels attempt constructing three panels with no repeated test units, the design required 245 unique test units. Therefore, the original mixed-format testlet-based item pool was increased in size by concatenating the original pool four times. This resulted in an item pool size of 1005 with 390 test units available for administration. Table 2 shows the distribution of test units for the entire mixed format testlet-based item pool after expansion.

Table 2. Final Mixed-format Testlet-based Item Pool

Items per Testlet	1	2	4	6	7	9	12	13	14	Totals
Total Test Units	290	25	15	10	10	10	10	15	5	390
Total Items	290	50	60	60	70	90	120	195	70	1005

Item Response Generation for Item Pool Recalibration

It was desired to not simply administer repeat items during the MST administration, so random error was introduced to the item parameters through a recalibration process. The recalibration process consisted of a sample size of 20,000 simulees generated from a standard normal distribution $\theta \sim N(0,1)$. Additionally, each simulee had a testlet effect parameter value $\gamma_{id(j)}$ for each testlet $d(j)$. Simulees' $\gamma_{d(j)}$ were randomly drawn from a normal distribution with mean equal to zero and the testlet specific variance or $\gamma_{d(j)} \sim N(0, \sigma^2_{\gamma_{d(j)}})$. Then response patterns were generated for all 1005 items. The probability of person *i* responding correctly to item *j* was calculated from the generated simulee's parameters θ and $\gamma_{d(j)}$, and from the item parameter estimates a_j, b_j, c_j obtained from the 3PL TRT calibration of the real dataset. Then a random number was generated for each person and item from a Uniform (0, 1)

distribution to introduce random error into the item responses. A simulee for item *j* was then given a correct response (i.e., 1) if the generated random probability from the uniform distribution was less than or equal to the calculated probability of a simulee answering the item correctly. Otherwise the simulee is given an incorrect response (i.e., 0). This procedure was used to generate responses to all 1005 items for all 20,000 simulees.

Recalibration

Once responses were generated, the mixed-format testlet-based item pool consisting of 1005 items was recalibrated based on the 20,000 simulee responses. Item parameters and the testlet effect variance, $\sigma^2_{d(j)}$, were estimated using SCORIGHT as described in the parameter estimation section. Because the SCORIGHT program centers the calibration on the examinees, the estimated theta distribution was approximately standard normal. Table 3 provides the descriptive statistics for the item parameter estimates for the final mixed-format testlet-based item pool. In addition, Figure 3 provides the pool's information when the gammas for each testlet equal zero.

Table 3. Mixed Format Testlet-Based Item Pool Descriptive Statistics.

Item Parameter	Mean	S.D.	Minimum	Maximum
A	1.549	0.542	0.431	5.932
B	-0.155	1.102	-2.937	2.104
C	0.101	0.080	0.008	0.489
$\sigma^2_{d(j)}$	0.245	0.209	0.061	1.053

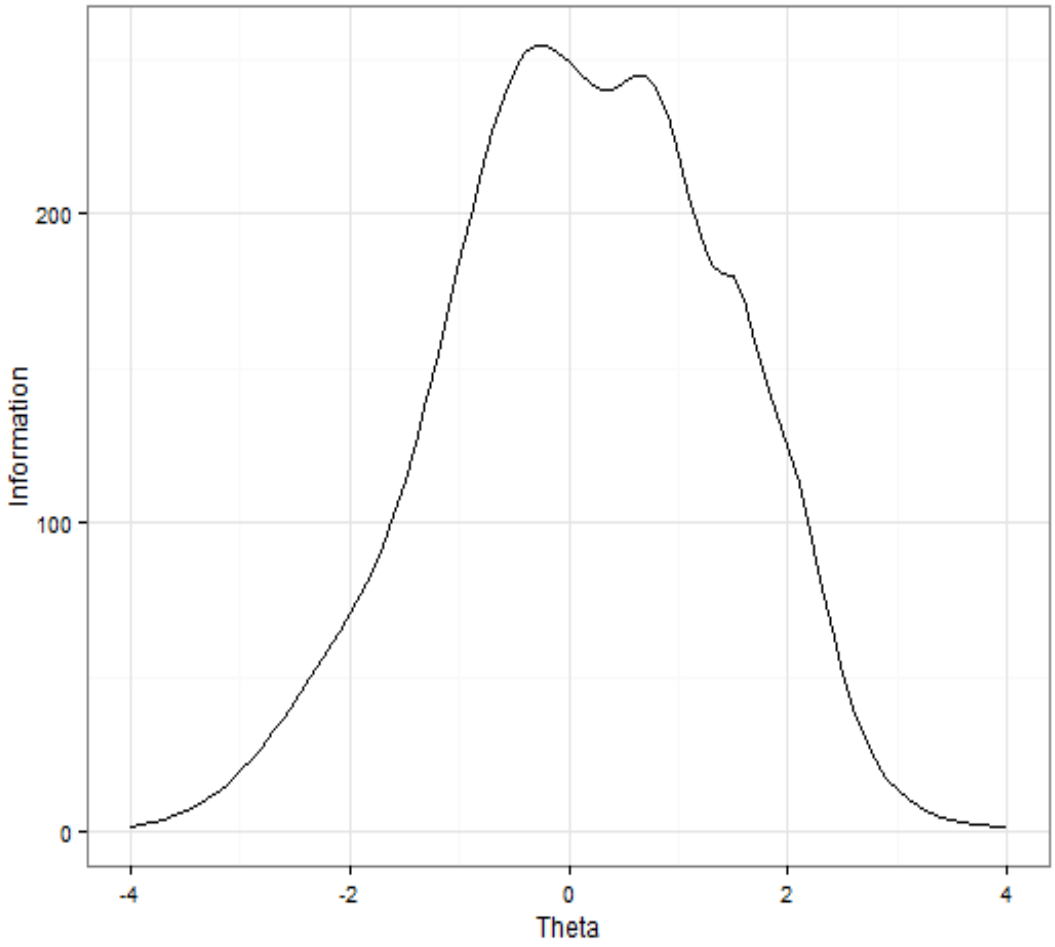

Figure 3. Pool information under the testlet response theory model when $\gamma_{(d)} = 0$.

MANIPULATED CONDITIONS

Panel Design

Four panel designs were constructed for the current study. Probably the most commonly cited panel design is the 1-3-3 (e.g. (Chen, 2010; Hambleton & Xing, 2006; Keng, 2008; Kim, Chung, Dodd, & Park, 2012; Kim, Chung, Park, & Dodd, 2013; Macken-Ruiz, 2008), where the routing stage or Stage 1 consists of one module available for all examinees. Stage 2 and Stage 3 then consist of three modules at each stage. The modules consisted of variable difficulty levels, which allow for the adaptive routing of an examinee. The three modules typically consist of an easy, medium, and hard difficulty level. Patsula (1999) found that adding both stages and modules increased the accuracy of ability estimates. Because the current study is primarily focused on estimating abilities for a broad range of abilities, a second three-stage panel design was considered, namely the 1-5-5. For this panel design, one module is again present in Stage 1 and five modules are available for administration at both Stage 2 and Stage 3. Again, each module within a stage ranged in difficulty from very easy to very hard difficulty modules. Three-stage panel designs are not the only designs that are found in the literature. Often it is of interest for researchers to consider a two-stage test. Some of the original research conducted on MSTs incorporated two-stage tests (see Kim & Plake, 1993; Lord, 1980; Patsula, 1999). In addition, current testing programs such as the GRE (ETS, 2011) are implementing two-stage MSTs. The study therefore used two two-stage panel designs, specifically a 1-3 panel design and a 1-5 panel design. Each panel design is depicted, including pathways, in Figure 4. The 1-3, 1-5, 1-3-3, and the 1-5-5 are located in the figure panes A, B, C, and D, respectively.

Test Length

The current study investigated two test lengths in part to see how short an MST can become in a mixed-format setting before losing the capacity to estimate abilities well. The data responses originated from three 67 item mixed-format test forms of standalone and testlet-based items. The goal of the current study was to create conditions of test length that maintained proportionality similar to the original test blueprints. Maintaining proportionality entailed a number of specifications including the number of standalone items in proportion to the number of testlets administered within modules and across stages.

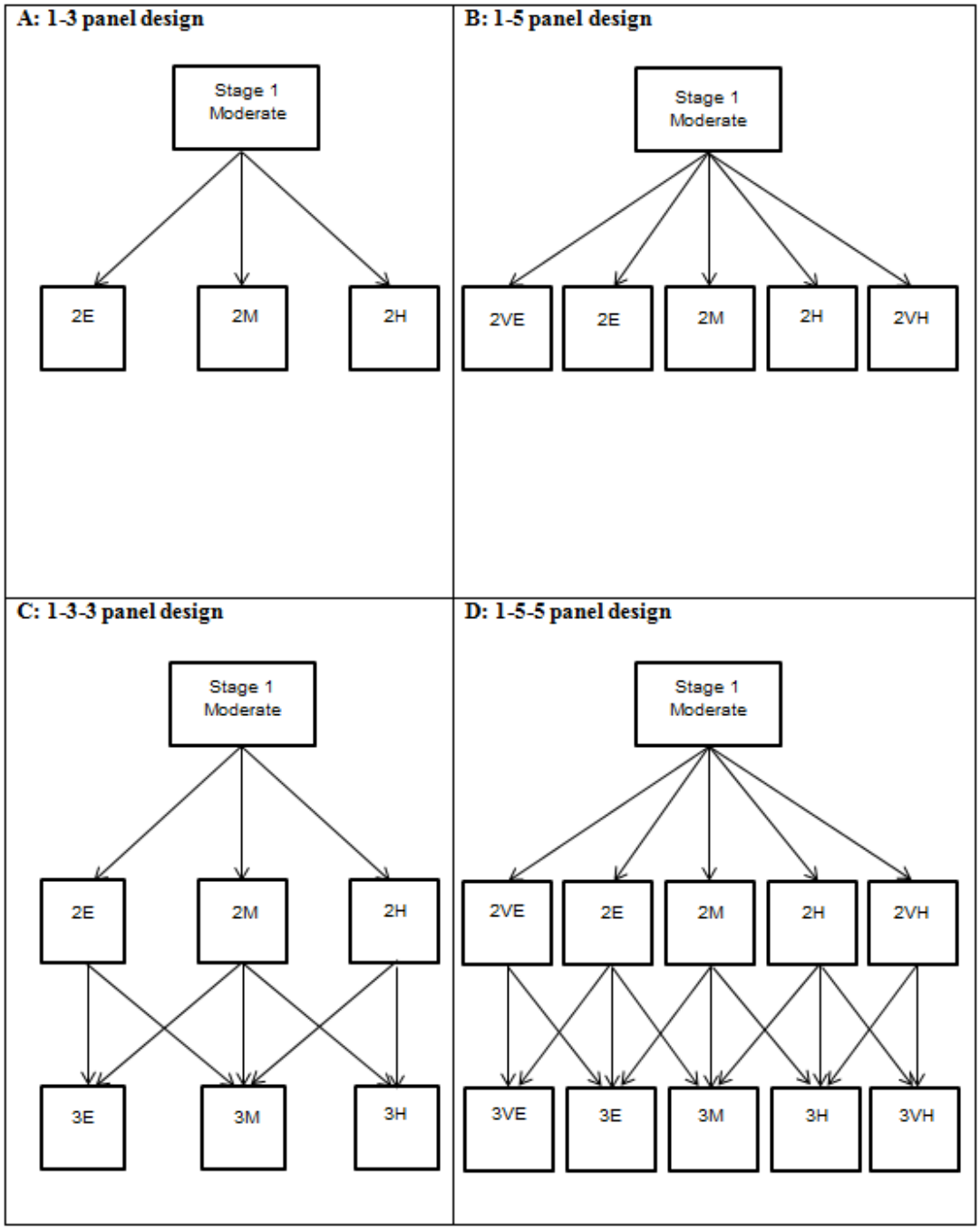

Figure 4. Panel designs with pathways for the study. VE=Very Easy difficulty E=Easy difficulty; M=Medium difficulty; H=Hard difficulty; VH=Very Hard difficulty

A test form had 19 or 20 standalone items and 47 or 48 passage-based items, respectively. To adhere to this constraint, the item pool was organized into test units. Recall each form consisted of 26 test units on the original forms. In terms of test units, each form consisted of 19 to 20 dichotomously test units and six or seven testlet-based test units. The study used two shorter test lengths approximately proportional to an original full length LFT. The longer of the two test lengths was 55 items which reasonably approximates to 20 or 21 test units. The shorter test length condition was a panel with each examinee receiving 44 items which reasonably approximates to and 17 or 18 test units. To inform the automated assembly program identifying which test units becomes of great import. Table 4 provides the proportions of test units in the pool, approximate number of test units for a given length, and panel combinations used in the ATA.

Notice that the test unit combinations do fluctuate as the ultimate goal was to approximate the test units while administering a specific number of items, namely 55 and 44 items. Flexibility in the number of test units proportional to the original test was allowed to float, so the composition of panels would have exact number of items. The test unit composition still maintains a sense of proportionality with respect to the originating test unit administration. In addition, identifying test units in this fashion increases the pool usage because varying testlet lengths are identified during the assembly process. The following combinations were considered to be parallel for the purpose of this study because the number of items is reaming constant during panel assembly.

Table 4. Distribution of Test Unit Combinations for Long and Short Test Length Panels.

No. of Items per Test Unit	1	2	4	6	7	9	12	13	14	TU(IT)
No. of Pool Test Units	290	25	15	10	10	10	10	15	5	390 (1005)
Long Combinations	15	1	1	0	0	1	1	1	0	20(55)
	16	1	1	1	0	0	0	1	1	21(55)
	15	2	1	0	1	0	1	1	0	21(55)
Short Combinations	13	1	1	0	0	0	1	1	0	17(44)
	14	1	0	1	0	1	0	1	0	18(44)

Note. No.=Number; IT=items; TU=test units.

Routing Procedures

Routing procedures are a variable of interests for the current study because they impact the security of an item pool. A decision must be made on the control of item exposure rates administered to the population of examinees. The study used three routing procedures. One routing procedure selects modules based on approximate maximum information (AMI) and the other two procedures are variants of the DPI. Kim et al. (2013) did not investigate the efficacy of ability estimation under any of the routing procedures for a mixed-format item pool. In addition, no studies have compared routing procedures in a mixed-format testlet pool under the 3PL-TRT model. Therefore, the current study explored the usefulness of routing procedures commonly seen in practice using the 3PL-TRT model for a mixed-format item pool.

Approximate Maximum Information

The AMI is the most commonly used routing procedure across studies (Zenisky et al., 2010). Similar to CAT's item selection on maximum information, the AMI selects modules that provide the highest amount of information for a person's current ability

estimate. The method generally results in the most accurate ability estimates. However, the method will over administer certain modules based on the nature of the ability distribution (Luecht et al., 2006). Once panels were constructed, decision point for each pathway were identified based on the intersection of adjacent module information functions. For the 1-3 and 1-3-3 panel design, two decision points at Stage 2 and Stage 3 were needed and are denoted as θ_L and θ_U. Then, if the provisional ability estimate was less than or equal to the decision point, $\hat{\theta} \leq \theta_L$, the simulee was routed to the easy module. If the provisional estimate was greater than or equal to the upper decision point, $\hat{\theta} \geq \theta_U$, the simulee was routed to the hard difficulty module. Otherwise, the simulee received the medium difficulty. Similar decisions were made for routing from Stage 2 to Stage 3. Recall from Figure 4, additional pathways can reroute simulees to adjacent modules. Again, the process of identifying the intersection of adjacent information functions was repeated for the decision points at Stage 3. From Stage 2 to Stage 3, a simulee was rerouted from the easy module difficulty if the current estimated ability was greater than the decision point, $\hat{\theta} > \theta_L$. A simulee in the medium difficulty module was rerouted if they fall below or above the decision points, $\hat{\theta} \leq \theta_L$ or $\hat{\theta} \geq \theta_U$, and was rerouted to the easy or hard module, respectively. A simulee in the hard module at Stage 2 was rerouted to the medium module at Stage 3, if their provisional ability estimate fell below the decision point, $\hat{\theta} < \theta_U$. Figure 5 illustrates the routing decisions for the AMI procedure on a 1-3-3 panel design. Similarly, four decision points must be identified for the 1-5 and 1-5-5 panel designs. The decision point between the very easy and easy module are labeled θ_{LL}; the decision point between the easy and medium module is θ_{LM}; the decision point between the medium and hard module is θ_{MU}; and the decision point between the hard and very hard module is θ_{UU}. A detailed table of the AMI decision points is presented in the panel assembly in the Results chapter.

Defined Population Interval

The DPI was developed as a means to specifically dictate the exposure rates of modules during the administration process. Most often this method is used to maintain equally proportioned module administration. If the policy of a testing program is to administer equal proportions of each module available after Stage 1, then a form of the DPI will need to be implemented (Luecht et al. 2006). Two methods of the DPI have been investigated in the classification setting (Kim et al., 2013). One such method is to consider stage-level routing, which is referred to as the stage level DPI (SL-DPI). SL-DPI essentially rank orders examinees across all modules and then routes them to the subsequent stage and module based on the overall rank order. The second method is to rank order examinees within modules and proportionally route each examinee within a module to the next stage's module. This method is referred to as the module level DPI (ML-DPI).

The goal for both DPI procedures was to administer equal proportions to all modules at all stages of administration. All examinees at Stage 1 see the routing module. Then one-third of examinees see the easy, medium, and hard modules at Stage 2 and Stage 3 in the 1-3 or 1-3-3 panel design; and one-fifth of the examinees see each module, including very easy, easy, medium, hard, and very hard difficulty modules in the 1-5 or 1-5-5 panel design at Stage 2 and Stage 3.

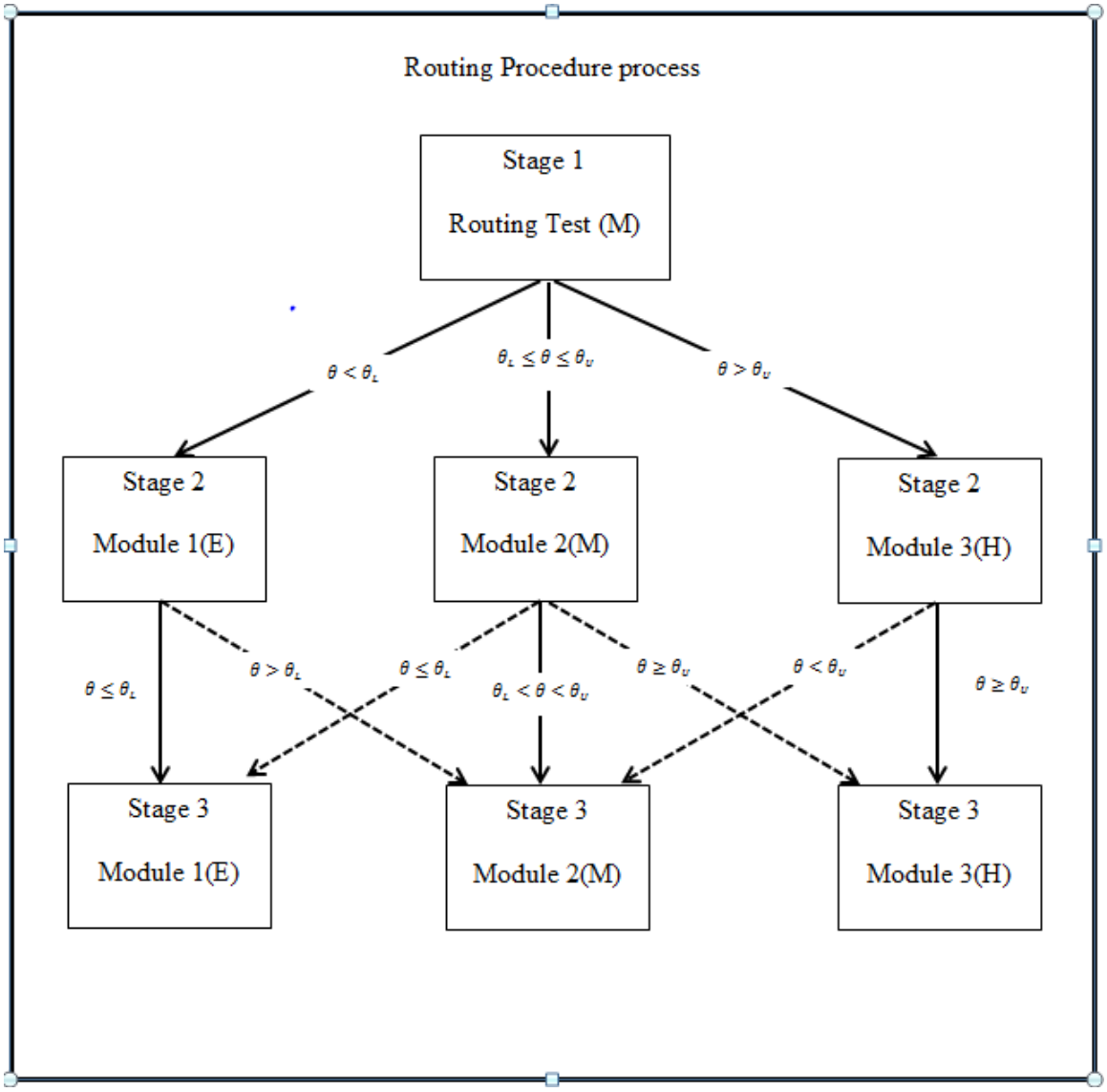

Figure 5. Example of AMI and SL-DPI procedure. AMI=approximate maximum information; SL-DPI=Stage level defined population interval. E=Easy difficulty; M=Medium difficulty; H=Hard difficulty.

To use the DPI procedures, it was assumed that the ability distribution was normally distributed. The details that distinguish the two procedures will now be further explained. The SL-DPI procedure similar to that described in Luecht et al. (2006) has the

same decision points at each stage. Assuming the ability distribution to be normal, theta decision points of -0.41 and 0.41 correspond to the 33rd and 67th percentile on a standard normal distribution and were used as θ_L and θ_U at each stage for the 1-3 and 1-3-3 panel designs, respectively. For the 1-5 and 1-5-5 panel designs, the 20th, 40th, 60th, and 80th percentile on the standard normal curve need to be identified. The corresponding theta decision points between each stage were then $\theta_{LL} = -0.85$, $\theta_{LM} = -0.26$, $\theta_{MU} = 0.26$, and $\theta_{UU} = 0.85$, respectively.

The ML-DPI used in Kim et al. (2013) considered pathways between all possible modules at each stage and defined an equal proportion of examinees to be (re)routed to and from each module at subsequent stages. However, in practice, it has been empirically demonstrated that very few examinees are rerouted to non-adjacent modules at subsequent stages (Hambleton & Xing, 2006). Therefore, the current study only routed examinees to adjacent modules from Stage 2 to Stage 3 as is illustrated in Figures 3. Notice for the 1-3 and 1-5 panel design, the SL-DPI and ML-DPI are equivalent, because no rerouting takes place within the MST administration. Therefore, the two DPI methods can only be compared in a three-stage MST setting.

The decision points from Stage 1 to Stage 2 are the same for the SL-DPI and ML-DPI procedure, and decision points for the ML-DPI between Stage 1 and Stage 2 were used as described for the SL-DPI. The proportion of reroutes was desired to be equal to the exposure rates for the overall test administration. This translates to one-third of within-module examinees to be rerouted to each adjacent module at the next stage for the 1-3-3 panel design, and one-fifth of within-module examinees to be rerouted for the 1-5-5 panel design. To achieve this proportional rerouting, more decision points were identified when navigating from Stage 2 and Stage 3 in both the 1-3-3 and 1-5-5 panel designs. Figure 6 illustrates the proportion of examinees within each module that are

routed to each module at subsequent stages. For a 1-3-3 panel design this means one-third of the examinees were routed to the easy, medium, and difficult modules at stage two. Then from Stage 2 to Stage 3, one-third of the examinees within a module were rerouted to each adjacent module in Stage 3.

To achieve a reroute of one-third of examinees within a module a change in cumulative probability needs to be considered. For the case of the 1-3-3 design each module contains approximately 33.3% of examinees. To reroute one-third of examinees to each adjacent module, a .111 change in the cumulative probability on a normal distribution was taken into account. The corresponding theta score decision points between Stage 2 and Stage 3 were identified as $\theta_L \leq -0.765$ to remain in the easy module; $-0.140 < \theta_M < 0.140$ to remain in the medium module, and $\theta_H \geq 0.765$ to remain in the hard module. For the 1-5-5 panel design, a reroute of one-fifth of examinees within a module corresponds to a cumulative probability change of .04 for each module. Accounting for this change between examinees induced eight decision points to select an examinees pathway from Stage 2 to Stage 3. The decision points to remain in the very easy module was $\theta_{LL} \leq -0.994$; the decision points to remain in the easy module were $-0.706 < \theta_{LM} < -0.358$; the decision points to remain in the medium module were $-0.151 < \theta_M < 0.151$; the decision points to remain in the hard module were $0.358 < \theta_{MH} < 0.706$; and finally, the decision point to remain in the hard module was $\theta_{HH} \geq 0.994$.

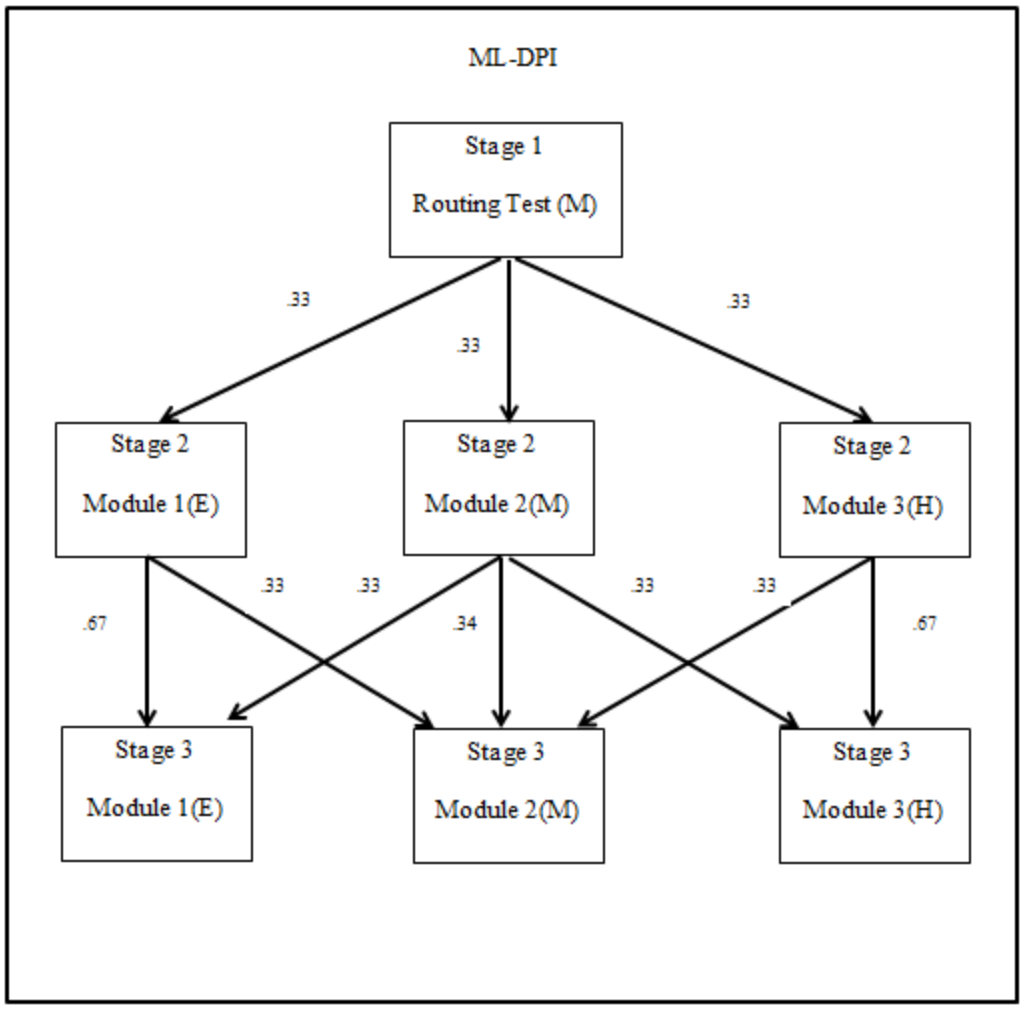

Figure 6. Illustration of the within-module proportions to be routed at subsequent stages for the module level defined population interval (ML-DPI).

Local Item Dependence

The study also included three levels of local item dependence (LID) within testlets. The presence of LID within testlets has been empirically shown to degrade model fit (Glas, Wainer, & Bradlow, 2000; Murphy et al., 2010; Wainer et al., 2000) and that decision accuracy and the measurement accuracy of ability estimation decreases as the magnitude of LID increases (Lu, 2010). The first magnitude considered for the study was the magnitude estimated during the calibration of the full size item pool. This

provides insight into the effects LID may have on tests currently being administered with similar psychometric properties to the originating test. The other two magnitudes will be 0.0 and 0.8, two levels previously used by Wainer et al. (2000). The three levels were used during simulation by defining $\sigma^2_{d(j)} = 0.0, 0.8,$ and $\hat{\sigma}^2_{d(j)}$. The LID was then used as corresponding testlet effect variance in the generation of response patterns for each item.

DATA GENERATION FOR MST SIMULATIONS

Two steps in the data generation process were necessary. The first step generates the simulated examinees' ability and testlet effects. Then the response patterns need to be generated. The simulated response patterns used a combination of the simulees' ability and the item parameters from the recalibrated testlet pool. A total of 100 replications were created. For each replication 1,000 simulee ability values, or known $\theta's$ were generated. The θ values were drawn from a standard normal distribution, $\theta \sim N(0,1)$. Additionally, each simulee needed to have the testlet effect parameter value $\gamma_{id(j)}$ for each testlet $d(j)$ generated. Simulees' $\gamma_{d(j)}$ were generated from a normal distribution with mean equal to zero and the testlet specific variance or $\gamma_{d(j)} \sim N(0, \sigma^2_{\gamma_{d(j)}})$ for the given LID condition. Then response patterns for all 1005 items were generated using the same process as described previously in the item pool development section.

Recall that for each condition there are three LID variance patterns. The same set of simulee θ distribution was used with the three testlet effect conditions. All $\gamma_{d(j)}$ will be generated from a normal distribution with mean equal to zero and the respective LID condition testlet variances. Therefore, three sets of responses were generated with each set consisting of 100 replications. The first set of responses was generated with no LID, i.e. $\gamma = 0$, or $\sigma^2_{d(j)} = 0$ for all simulees; the second set of generated responses had a large

constant LID for $\gamma_{d(j)}$ with mean equal to zero and $\sigma^2_{d(j)} = 0.8$; and the third set of generated responses had the estimated LID for $\gamma_{d(j)}$ with mean equal to zero and $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$, or the estimated variance for testlet $d(j)$.

MST ASSEMBLY

Recall that one of the manipulated variables for the current study regards panel design. Four panel designs were considered. The current study constructed four panel designs are 1-5-5, 1-3-3, 1-5, 1-3. Each panel design condition constructed parallel panels. The automated testing assembly (ATA) constructed panels to have two test length conditions, a 55 item and a 44 item panel. Content constraints were initially considered but were deemed to sparse to be realistic for the given testlet-based item pool. The current study used the bottom-up method of assembly, where the assembly strategy treats the process as a simultaneous building of all modules for the given panel design (Luecht et al., 2006). The proceeding discussion provides further constraints set in place for the ATA algorithm.

One of the main objectives that must be declared for ATA is defining the targets for the targeted test information function (TIF) for each module being assembled. The target TIFs were defined to be equal across modules within a stage. This method allows the assembly process to optimize the TIF within each module and is referred to as a relative targets (Diao & van der Linden, 2011). To obtain equal targeted TIFs within a stage, an additional targeting parameter must be defined for ATA denoted as the targeted θ_k values. The θ_k values for all conditions commonly range from -1.0 to 1.0 on the θ-scale and were used as the range in the current study (van der Linden, 2005). For the conditions with three modules, this will result in $\theta_k = \{-1.0, 0.0, 1.0\}$, and for the conditions with five modules, this will result in $\theta_k = \{-1.0, -0.5, 0.0, 0.5, 1.0\}$. The

target θ_k values tell the linear programming solver to maximize a module TIF at each θ_k, or in the case of the routing module, the TIF must obtain a uniform TIF criteria at across the defined θ_k values, for modules and panels.

MST constructions were conducted in R with the lpSolveAPI package (Konis, 2013). First the routing module specifications will be discussed, followed by the subsequent stages. The model for the simultaneous assembly of the routing module was assembled as an approximate uniform distribution for the relative targets (Diao & van der Linden, 2011b; Kim & Plake, 1993). For a multistage test, the objective function was based on a maximin principle (van der Linden, 2005). The maximin principle is defined so the minimum value of the module information function across the targeted θ_k values is maximized. The items in the pool are denoted as $j = 1,..., J$. The number of items in a testlet is denoted as $n_{d(j)}$ and the number of items for the module will be denoted as n_{ML}. The model that will be used is

$$\max y \tag{23}$$

which will be subject to the following constraints for decision variable x_j

$$\sum_{j=1}^{J} I_j(\theta_k) x_j \geq y, \text{for all } k, \tag{24}$$

$$\sum_{j=1}^{J} I_j(\theta_k) x_j \geq y + \delta, \text{for all } k, \tag{25}$$

$$\sum_{j \in d(j)} x_j = n_{d(j)}, \tag{26}$$

$$\sum_{j=1}^{J} x_j = n_{ML}, \tag{27}$$

$$x_j \in \{0,1\}, \tag{28}$$

$$y \geq 0. \tag{29}$$

The combination of Equation 23 and the constraints in Equation 24 implement the maximin principle. This will achieve an approximately uniform distribution of the

targeted θ_k values for the routing module. Equation 25 uses δ as a tolerance constraint to help ensure the obtainment of the uniform distribution and to set an upper bound on the information for the targeted θ_k. The testlet, and module length requirement are declared in Equations 26 and 27.

To simultaneously assemble the designated number of modules *f* for a given condition, The model at each possible stage must also be declared. At Stage 2, the ATA program will target peaked information functions for each module at θ_k for the specified panel design. The model will be the same as in Equation 23. The difference will be in the constraints that are specified for the modules

$$\sum_{j=1}^{J} I_j(\theta_k) x_{jf} \geq y, \text{ for all } k \text{ and all forms } f, \tag{30}$$

$$\sum_{j=1}^{J} I_j(\theta_k) x_{jf} \geq y + \delta, \text{ for all } k \text{ and all forms } f, \tag{31}$$

$$\sum_{f=1}^{m} x_{jf} \leq 1, \text{ for all } j, \tag{32}$$

$$\sum_{j=1}^{J} x_{jf} = n_{ML}, \text{ for all forms } f, \tag{33}$$

$$\sum_{j \in d(j)} x_{jf} \geq n_{d(j)}, \text{ for all forms } f, \tag{34}$$

$$x_{jf} = 0, \text{ for all } j \in S_1 \text{ or } S_3, \tag{35}$$

$$x_{if} \in \{0,1\}, \tag{36}$$

$$y \geq 0. \tag{37}$$

Notice in the following set of Equations 30-37 the additional subscript for the module, *f*. Similarly to the routing module Equations 30 and 31 define the target information and tolerance for each module. Equation 32 is used to ensure no item overlap for the forms at Stage 2. Again, Equations 33 and 34 are the testlet and module length constraints. Equation 35 is used to guarantee no item overlap across stages within a panel. Note that S_3 in Equation 35 will not be included for the two stage panel designs.

Like the model for the second stage, the constraints and model are defined exactly the same for the third stage with the exception of Equation 37. This constraint was then replaced by the following constraint

$$x_{jf} = 0, \text{for all } j \in S_1 \text{ or } S_2, \text{and all forms } f, \tag{40}$$

where Equation 40 instead includes the set of items used in Stage 2, S_2.

Minimizing the number of constraints can alleviate the optimization process in the branch-and-bound search method when identifying a solution. By specifying the test unit distribution, the number of branches is drastically reduced and becomes more efficient at identifying a workable solution (Galindo et al., 2013). This approach was used to achieve the constraints for Equations 26 and 34. Tables 5 and 6 provide the specific test unit identifications which are extensions to the integer combinations presented in Table 4. The extensions provide the test unit specifications at the stage level for each panel, when distinct. These combinations represent the combinations in which the ATA algorithm were able to find a solution.

Table 5. Distribution of Test Unit Combinations for Long Test Length Conditions for Distinct Stages and Panels.

No. of Items per Test Unit	Stage	1	2	4	6	7	9	12	13	14	TU(IT)
Pool		290	25	15	10	10	10	10	15	5	390 (1005)
1-5-5 Combinations	Stage 1	5	0	1	0	0	1	0	0	0	7(18)
	2	5	1	0	0	0	0	1	0	0	7(19)
	3	5	0	0	0	0	0	0	1	0	6(18)
1-3-3 Combinations Panel 1	Stage 1	5	1	0	0	0	0	1	0	0	7(19)
	2	5	0	1	0	0	1	0	0	0	7(18)
	3	5	0	0	0	0	0	0	1	0	6(18)
Panel 2	1	6	1	1	1	0	0	0	0	0	9(18)
	2	5	0	0	0	0	0	0	0	1	6(19)
	3	5	0	0	0	0	0	0	1	0	6(18)
Panel 3	1	5	1	1	0	1	0	0	0	0	8(18)
	2	5	1	0	0	0	0	1	0	0	7(19)
	3	5	0	0	0	0	0	0	1	0	6(18)
1-5 Combinations	Stage 1	8	1	0	1	0	0	1	0	0	11(28)
	2	8	1	1	0	0	0	0	1	0	11(27)
1-3 Combinations	Stage 1	8	1	0	1	0	0	1	0	0	11(28)
	2	8	1	1	0	0	0	0	1	0	11(27)

Note. IT=items; No.=Number; TU=test units.

Table 6. Distribution of Test Unit Combinations for the Short Test Length Conditions for Distinct Stages and Panels.

Test Unit Score Points	Stage	1	2	4	6	7	9	12	13	14	TU(IT)
1-5-5 Combinations	Stage 1	2	0	0	0	0	0	0	1	0	3(15)
	2	4	1	0	0	0	1	0	0	0	6(15)
	3	8	0	0	1	0	0	0	0	0	9(14)
1-3-3 Combinations	Stage 1	2	0	0	0	0	0	0	1	0	3(15)
	2	4	1	0	0	0	1	0	0	0	6(15)
	3	8	0	0	1	0	0	0	0	0	9(14)
1-5 Combinations	Stage 1	6	0	1	0	0	0	1	0	0	8(22)
	2	7	1	0	0	0	0	0	1	0	9(22)
1-3 Combinations	Stage 1	6	0	1	0	0	0	1	0	0	8(22)
	2	7	1	0	0	0	0	0	1	0	9(22)

Note. No.=Number; IT=items; TU=test units.

MST ADMINISTRATION

Two additional decisions beyond panel assembly were made that effect the administration of the MST to the simulee. They are the within test scoring, and routing procedures of the MST. EAP was used for all scoring for the MST administration, including both within and total MST scoring. The routing procedures were based on theta scoring. Three conditions were considered for the routing procedures but do not affect the steps in administration of modules to the simulee. The following procedure provides a step-by-step process of the MST administration for the simulation study:

1. The random number generator assigned one of three panels to the simulee.
2. The simulee were administered the routing module at Stage 1 from the chosen panel.
3. After the module from Stage 1 is completed, a provisional ability, $\hat{\theta}$, and their estimated testlet effects, $\hat{\gamma}$, were estimated using EAP.
4. The simulee was then routed to a module at Stage 2 based on the provisional ability $\hat{\theta}$, and the current routing procedure in place, namely AMI, DPI-1, or DPI-2.
5. For a two-stage test, administration terminates and a final ability estimate, $\hat{\theta}$, and testlet effect parameters, $\hat{\gamma}$, were calculated using EAP for all responses collected. For a three-stage test, an addition provisional ability, $\hat{\theta}$, and testlet effect parameters, $\hat{\gamma}$, were estimated, then based on the current routing procedure the simulee was routed to the appropriate module at Stage 3.
6. For the three-stage panel designs, after Stage 3's administration the final ability, $\hat{\theta}$, and testlet effect parameters, $\hat{\gamma}$, were estimated for all administered responses using EAP.

DATA ANALYSIS

The primary purpose of the dissertation was to investigate the operational characteristics of an MST administration for a mixed format testlet-based item pool. Specifically the study investigated multiple MST conditions and their effect on ability estimation for a broad population of examinees. In addition, the administrative characteristics of the MSTs were inspected for differences in panel and module administration.

Evaluating Research Questions

The following evaluation indices are related to answering the questions regarding ability estimation.

(1) *How does panel design impact the measurement precision of an MST with a mixed-format testlet-based item pool?*

(2) *How does test length impact the measurement precision of an MST with a mixed-format testlet-based item pool*?

(3) *How does the magnitude of LID effect the administration and ability estimation of an MST with a mixed-format testlet-based item pool*?

(4) *How do various routing procedures impact the ability estimation of an MST with a mixed-format testlet-based item pool*?

(5) *How do test length, LID, and routing procedures interact with respect to the accuracy and precision of ability estimation*?

The simulation study's evaluation criteria assessed the measurement accuracy and of recovering a simulee's known theta, θ. Simulee descriptive statistics, including estimated theta, $\hat{\theta}$, and standard error of ability estimates $\sigma_e(\hat{\theta})$ were calculated. The Pearson product-moment correlation between the known theta, θ and estimated theta, $\hat{\theta}$, was computed. The following indices were used to evaluate the measurement properties

of the MST designs: bias, root mean squared error (RMSE), and average absolute deviation (AAD) on ability estimation. Bias is calculated by the following formula:

$$Bias = \frac{\sum_{i=1}^{n} \hat{\theta}_i - \theta_i}{n}. \quad (41)$$

The formula for RMSE is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2}{n}}. \quad (42)$$

The formula for AAD is

$$AAD = \frac{\sum_{i=1}^{n} \left| \hat{\theta}_i - \theta_i \right|}{n}. \quad (43)$$

For each of the formulas in Equations 41-43, θ_i is the known theta estimate for the simulee *i*, and $\hat{\theta}_i$ is the estimated theta for simulee *i*. Where *n* is the total number of simulees in each condition. All indices described were averaged across the 100 replications.

To illustrate the accuracy and precision of the evaluative indices, conditional plots were constructed. In addition, the conditional plots for the $\sigma_e(\hat{\theta})$ were depicted. Conditional plots help visually illustrate the performance of accuracy and precision estimates across a broader range of the θ distribution. The conditional plots illuminate differences in the effectiveness of the various conditions that can sometimes be concealed by aggregated statistics being used for the study.

In addition to the measurement accuracy and precision measures, the study calculated the rate at which modules and panels were administered. Descriptive statistics for the administration are provided in the results section. Additionally important information for testing programs is the proportion of examinees routed between modules.

Understanding the rate of administration for each module allows for testing companies to monitor exposure rates of panels, modules, testlets, and stand-alone items. All such statistics of operational characteristics were extracted from the simulees audit trail. The audit trail is a record of the items and testlets administered to each simulee.

Chapter 4: Results

This study examined multistage tests (MST)s for four panel designs 1-5-5, 1-3-3, 1-5, and 1-3; two test lengths (long and short); three routing procedures approximate maximum information (AMI), stage-level defined population interval (SL-DPI), and module-level defined population interval (ML-DPI); and three local item dependence (LID) conditions ($\sigma^2_{(d_{(j)})} = (0.0, 0.8, \hat{\sigma}^2_{(d_{(j)})})$). The following design resulted in 60 conditions as seen in Table 1. This chapter presents the final panels assembled, the outcomes for evaluating the measurement accuracy and precision, and the module administration characteristics associated for each condition.

PANEL ASSEMBLY

The item pool for the study was a mixed-format item pool of both stand-alone dichotomous items and testlet-based dichotomously score items. In all, there were nine types of test units that items could be categorized into as seen in Table 2. Among the test unit types, 290 were dichotomous test units, 25 were test units with two items, 15 were test units with four items, 10 were test units with six items, 10 were test units with seven items, 10 were test units with nine items, 10 were test units with 12 items, 15 were test units with 13 items, and five were test units with 14 items. Figure 7 is an illustration of the information for the entire set of stand-alone dichotomous test units. Figure 8 is an illustration of the information for all the testlet based test unit types. Notable results from Figure 8 were the two item testlet based test unit information was bimodal and the test units with 13 testlet-based items had the largest amount of information.

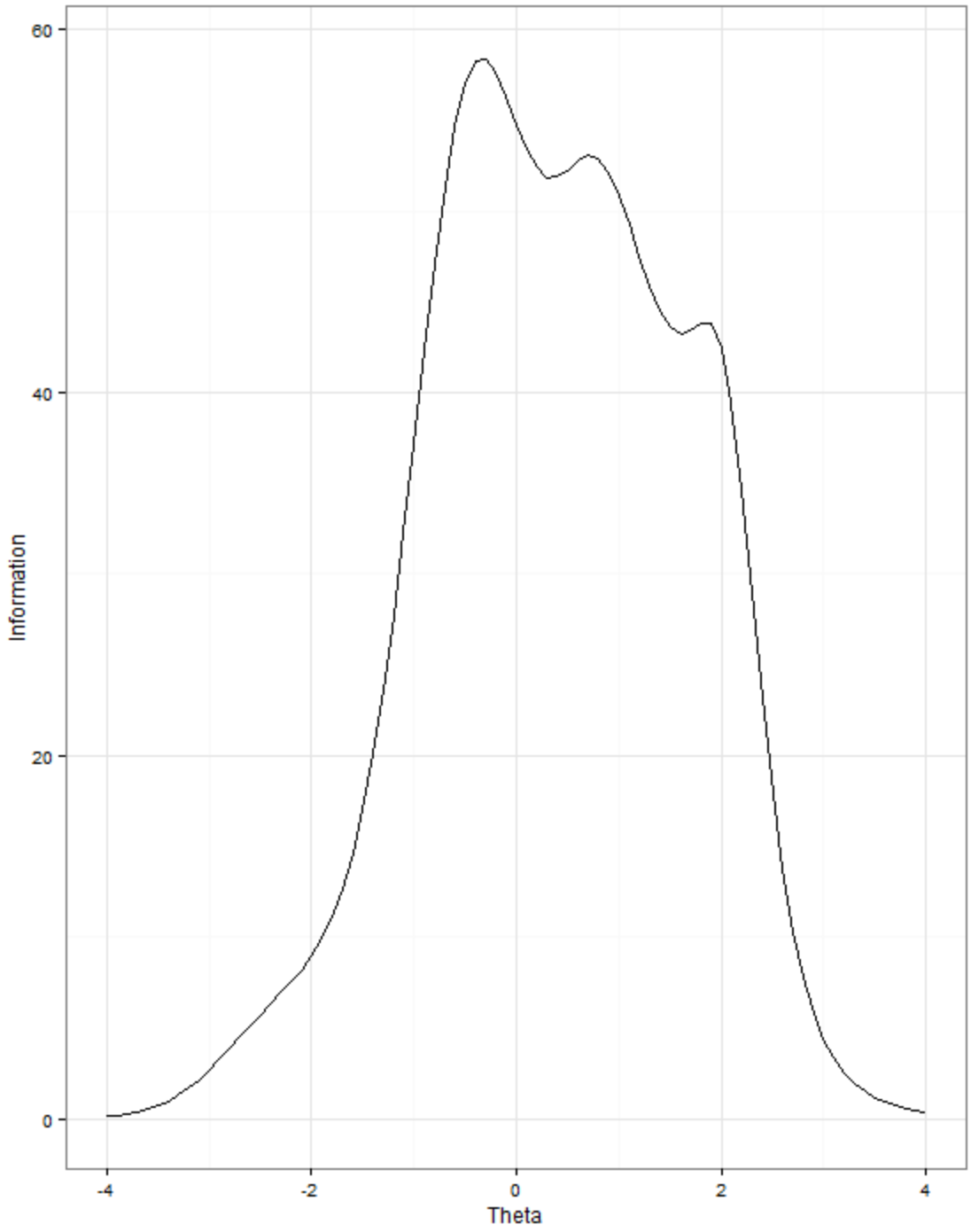

Figure 7. Information for the dichotomous stand-alone test units when $\gamma_{(d)} = 0$.

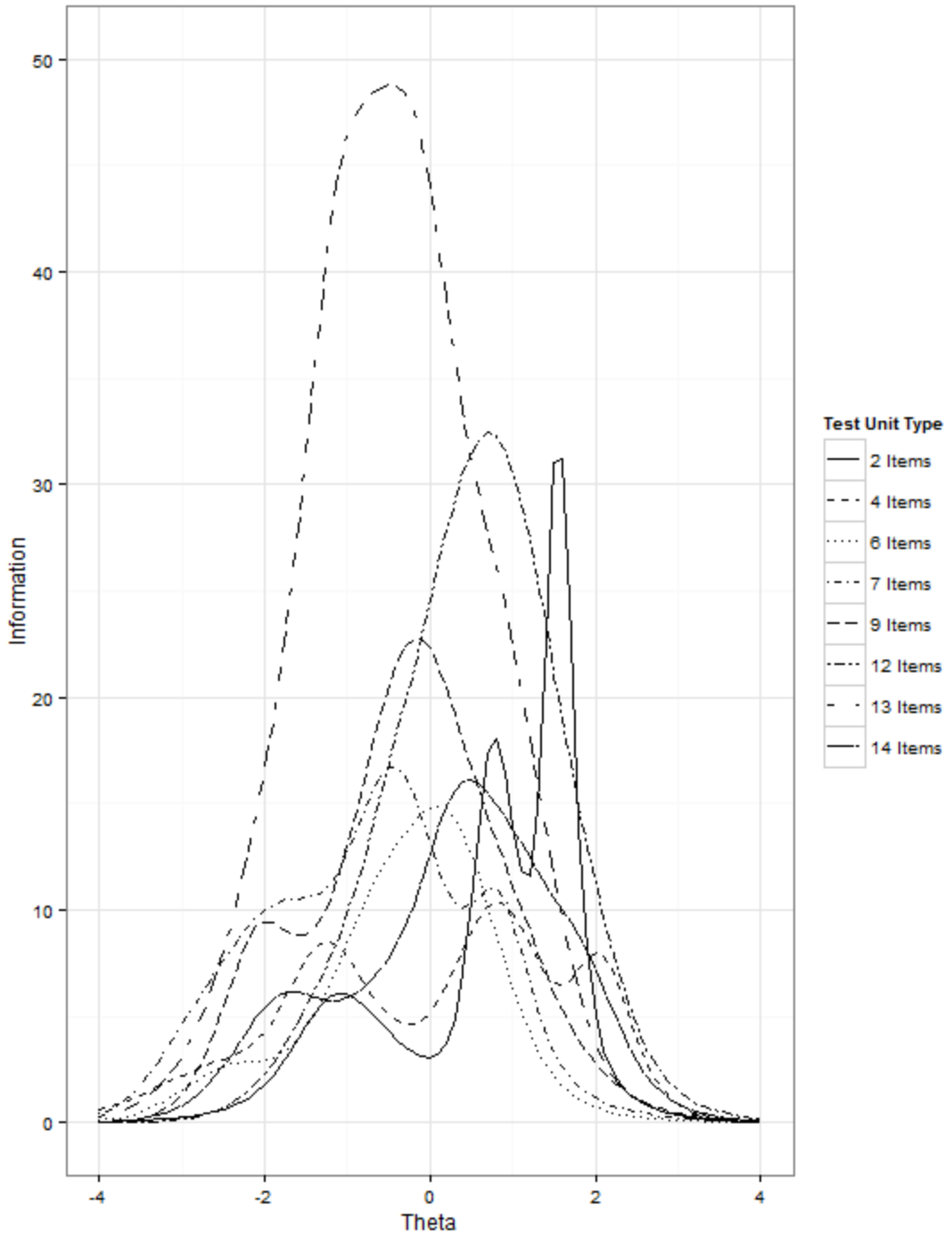

Figure 8. Information for all testlet-based test unit types when $\gamma_{(d)} = 0$.

MST panels were constructed from an automated test assembly (ATA) program using the R package lp_SolveAPI (Konis, 2013). To evaluate the adequacy of the ATA program, visual inspection of the relative target test information functions (TIFs) were examined based on targeted information across the ability range for each panel. For the sake of brevity, illustrations in this section are limited to Figures 9-16 for one three-stage panel and one two-stage panel. The remaining figures used in the visual analysis for the remaining panel designs can be found in Appendix A.

Recall from Table 5, the patterns used to inform the ATA algorithm for each test unit type at each stage and panel. Originally, the study set out to have no repeat items between all panels based on the specified test unit types. When this constraint was imposed on the ATA algorithm, it was unable to provide a solution for any panel design. Therefore, the constraint allowed a test unit to appear no more than twice, and only on different panels and at different stages. The panel design for the 1-5-5 long and short test length conditions was unable to meet these constraints for three panels. The 1-5-5 panel designs consistently identified multiple test units within a panel and used a test unit three times across panels. However, the ATA algorithm was able to optimize a solution for the 1-5-5 panel designs when using only two panels. Therefore, two panels were assembled for each of the 1-5-5 conditions and used for the simulation. Across all conditions, at most only three test units were used repeatedly for a given panel design.

Three-stage Panel Assembly

Figures 9-13 contain the relative target TIFs assembled for the short 1-5-5 panel design. The TIF plots for the short 1-5-5 panel design illustrate a three stage panel assembly. For the 1-5-5 panel design, five targeted theta values were specified, namely

$\theta_k = (-1.0, -0.5, 0.0, 0.5, 1.0)$. First, the routing module relative target TIFs at Stage 1 are discussed followed by Stage 2 and Stage 3 relative target TIF plots.

Figure 9 plots the TIFs for the routing stage for both panels. The routing stage was intended to be a uniform distribution across the theta range. The figure indicates that the information was consistent across for each θ_k with some deviation. Overall, the TIFs approximated a uniform distribution over the targeted ability range and were viewed as adequate for the simulation study.

Next the modules at each stage need to be inspected for uniformity at the peaks of each TIF for the targeted theta range. Figure 10 provides an illustration of the information across the targeted theta range for each panel at Stage 2. The information across the targeted thetas had peaked information functions occurring at each of the desired targets and was approximately uniform with respect to the peaks. Notice the bi-modal TIFs produced at the upper end of the ability distribution. This occurred from the use of test units for the two-item testlets. This pattern was found throughout the assembly process when test units with two-item testlets were used. Figure 11 provides an illustration of the information across the targeted theta range for each panel at Stage 3. The relative target TIFs were similar across the theta range all peaking uniformly at the targeted theta values.

Another aspect to the visual analysis was the ATA's ability to replicate relative target TIFs for each panel at each of the targeted difficulty levels, or each θ_k. Figures 12 and 13 provide an illustration of the TIF's for each difficulty level at Stage 2 and Stage 3 respectively. It was seen from the figures that each TIF was peaked at their targeted theta values and similarly constructed in the immediate surrounding areas for each panel. Again one should notice some Stage 2 TIFS were bimodal in Figure 10. Again, many of the TIFs were bimodal because of the use of the two-item testlet test units.

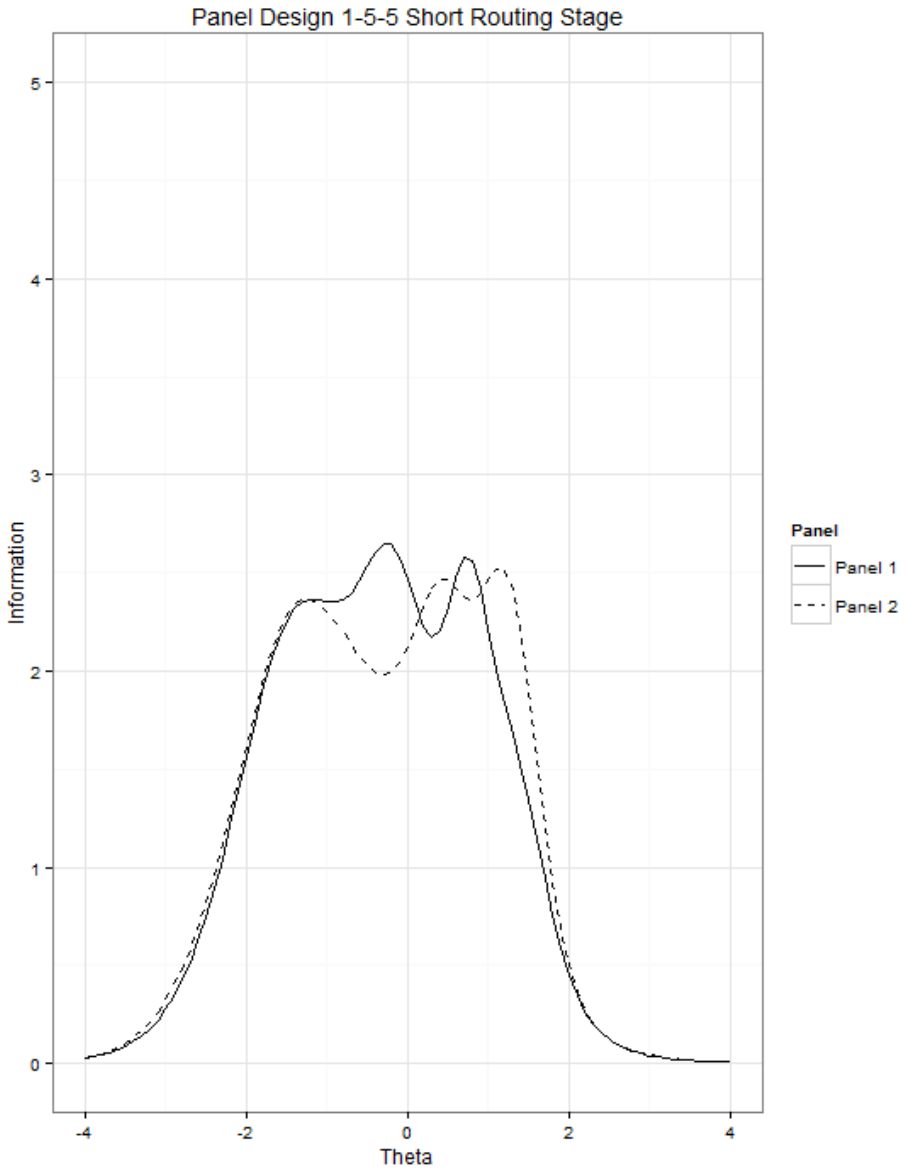

Figure 9. Stage 1 routing module relative target TIF plots for the 1-5-5 routing stage when $\gamma_{(d)} = 0$.

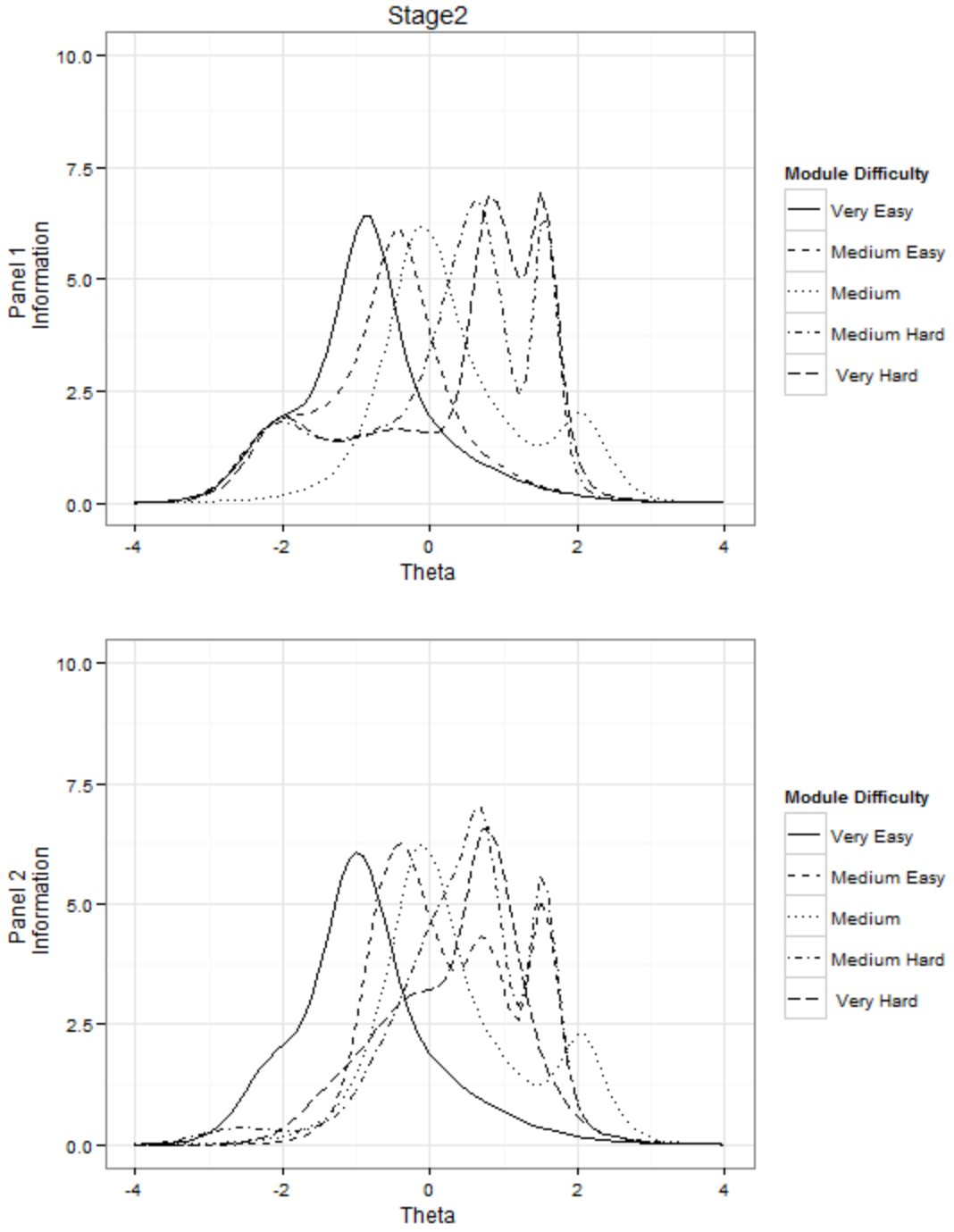

Figure 10. Stage 2 relative target TIFs for the short 1-5-5 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

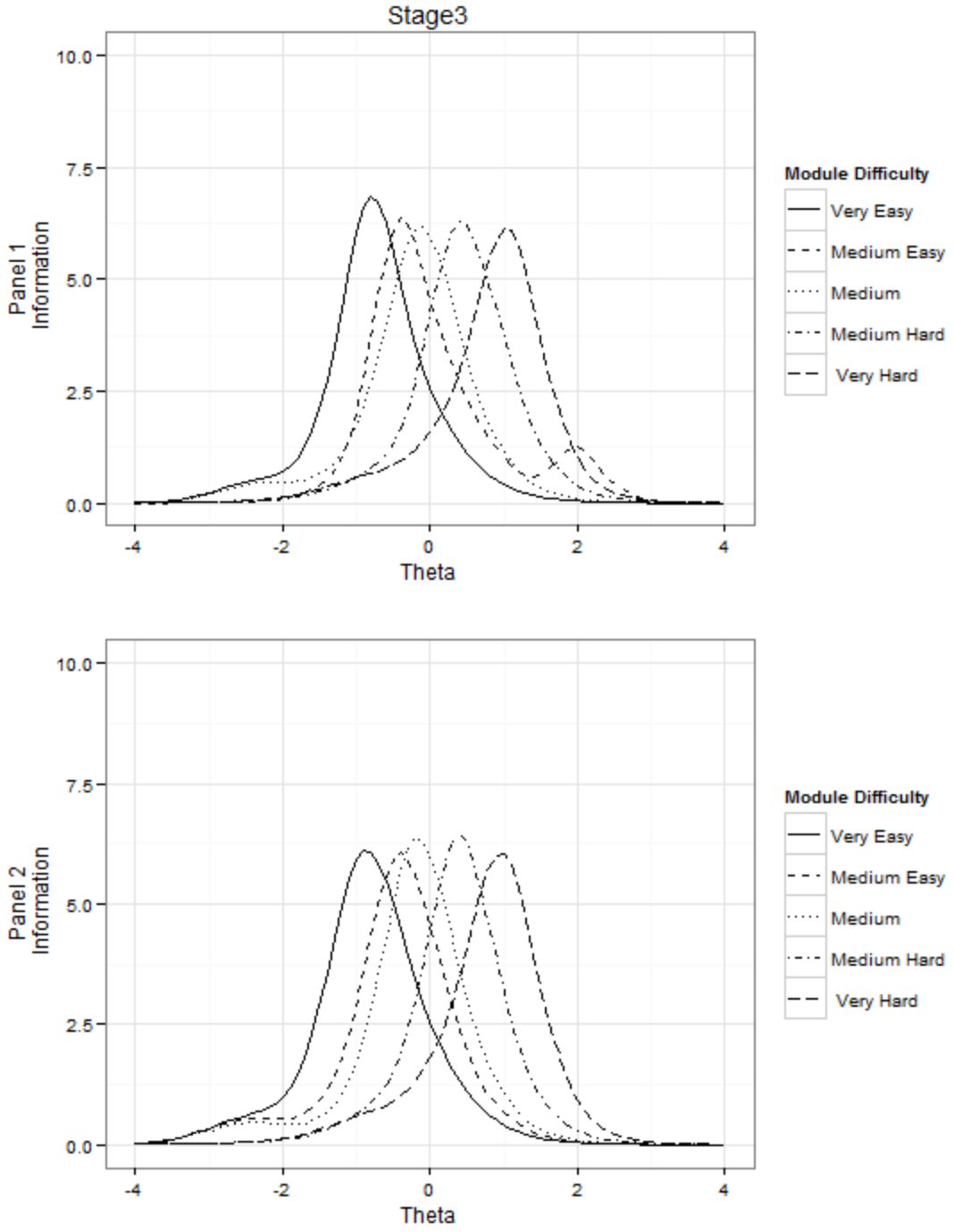

Figure 11. Stage 3 relative target TIFs for the short 1-5-5 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

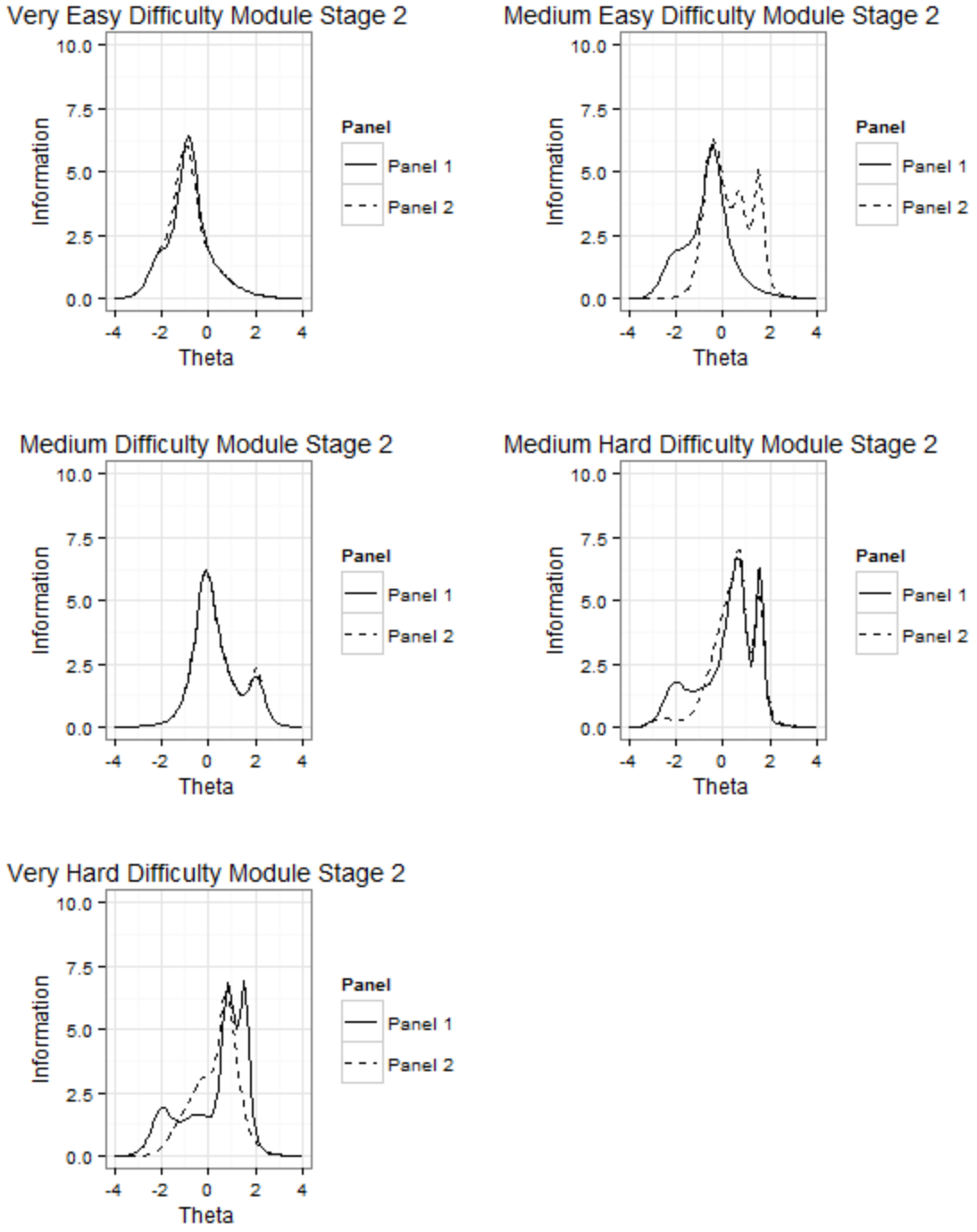

Figure 12. Stage 2 relative target TIFs for the short 1-5-5 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

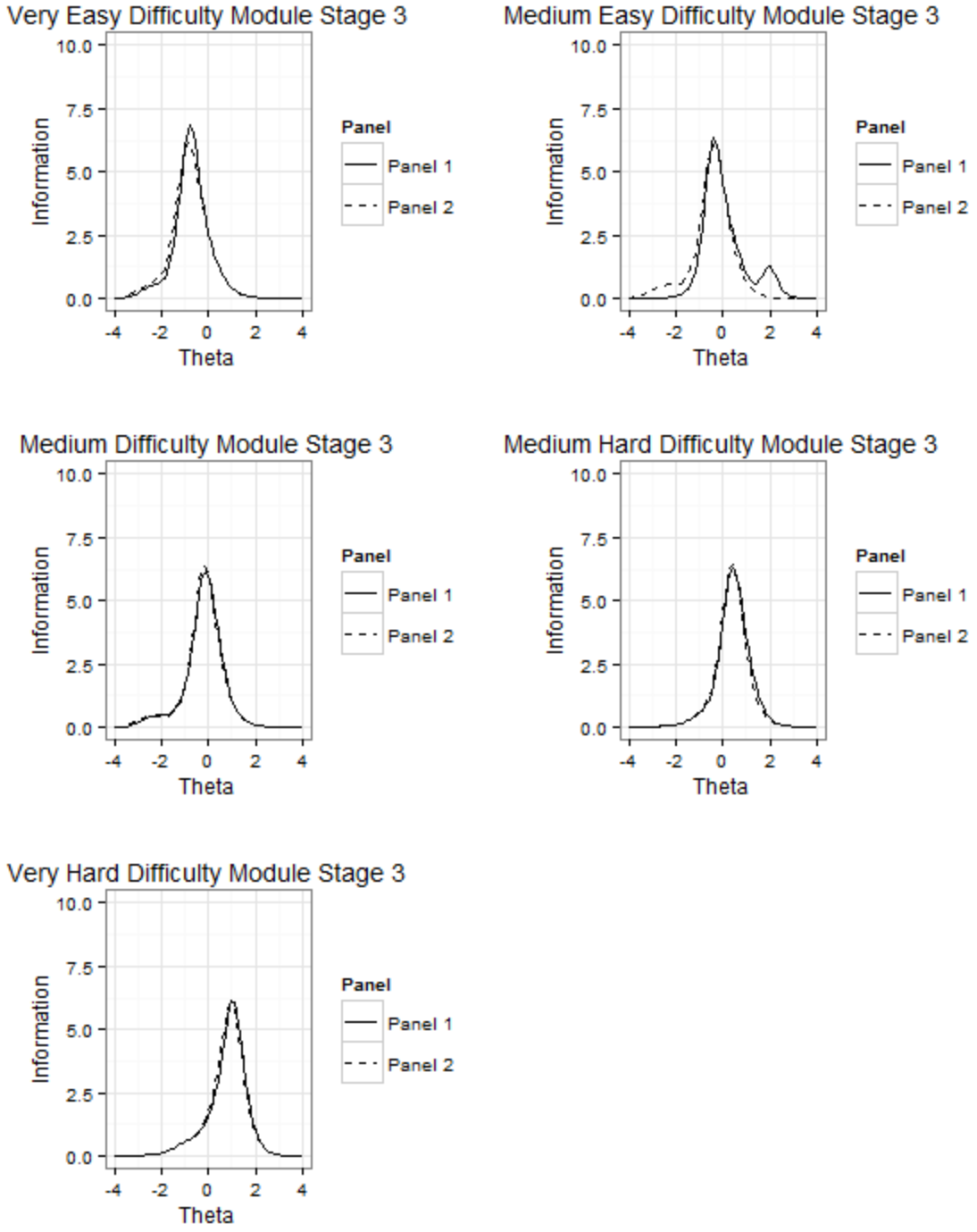

Figure 13. Stage 3 relative target TIFs for the short 1-5-5 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

Two-stage Panel Assembly

The two-stage panel assembly section provides the relative target TIFs for the 1-3 long panel design. First, the Stage 1 routing module TIFs were evaluated. Then, Stage 2 difficulty plots were inspected across the targeted theta range and at each difficulty. For the 1-3 panel design, three theta targets were used for the relative TIFs, namely $\theta_k = (-1.0, 0.0, 1.0)$.

Figure 14 plotted the Stage 1 routing module relative target TIFs for each of the three panels. The routing module assembled three modules that were approximately uniform across the specified targeted thetas. At each of the three targeted thetas, the information was very similar with some fluctuation between the target values. The information between -1.0 to 1.0 stays within a fairly small band of information and was considered adequate for the purposes of a uniform routing module.

At Stage 2, the three targeted thetas were specified to provide a uniform set of peaked relative target TIFs. Figure 15 provides the relative target TIFs across the targeted thetas. Approximately equivalent information was achieved for each of the difficulty levels. Similar to the three-stage assembly, one should note the bimodal TIFs especially prevalent in the harder modules. Again, this was where the test units with two-item testlets were administered.

Figure 16 provides the relative target TIFs at Stage 2 in order to inspect the replicability of a module across panels. The targeted difficulty TIFs resulted in similar amounts of information at each of the respective targeted thetas with some deviations beyond the general target area. Notice for the hard module that the information actually peaks slightly before the targeted theta, $\theta_k = 1.0$, but the information for all three panels was approximately equal at the targeted theta. Overall, similar amounts of information were achieved across the targeted theta range.

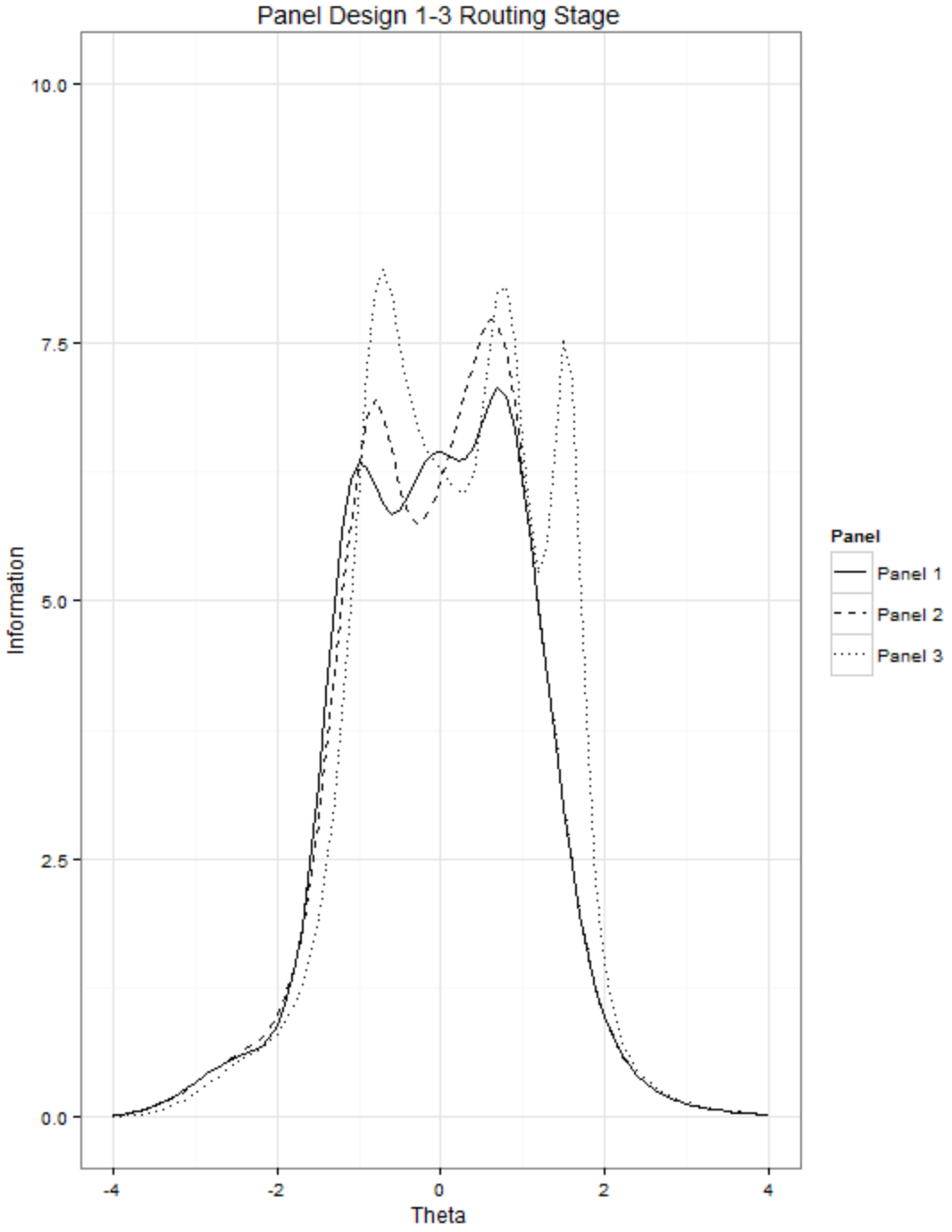

Figure 14. Stage 1 routing module relative target TIF plots for the long 1-3 routing stage when $\gamma_{(d)} = 0$.

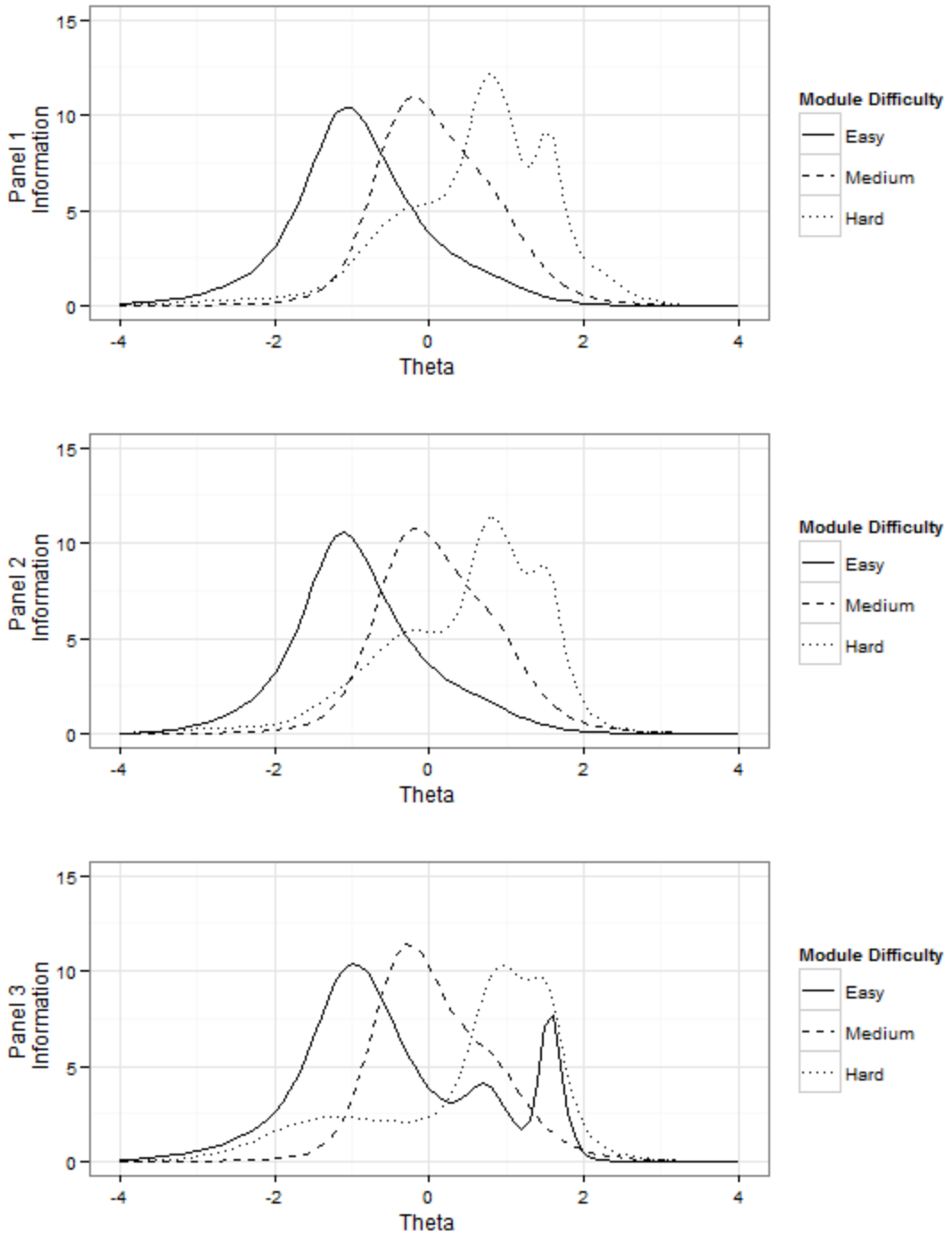

Figure 15. Stage 2 relative target TIFs for the long 1-3 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

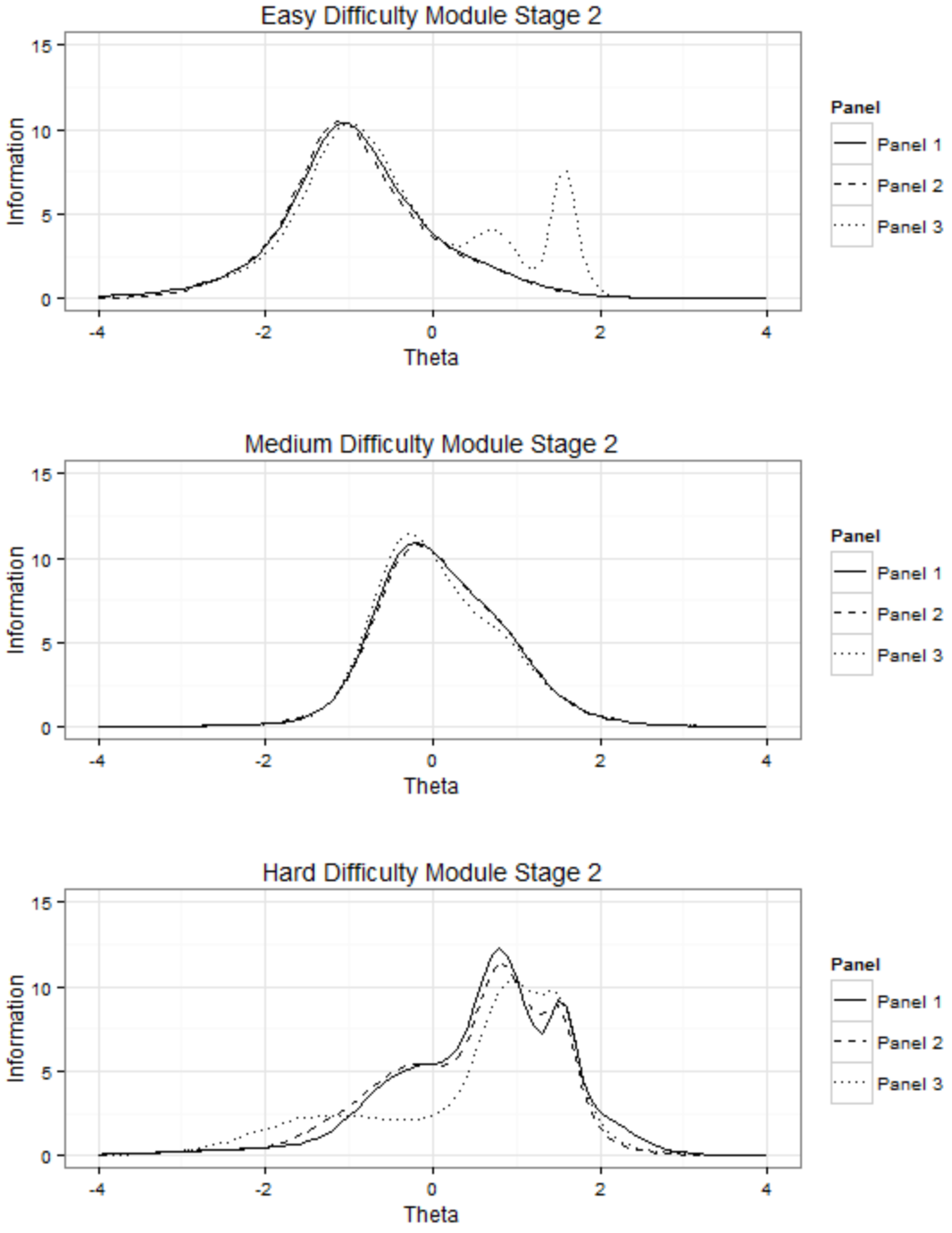

Figure 16. Stage 2 relative target TIFs for the long 1-3 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

MEASUREMENT ACCURACY AND PRECISION

The following section includes the dependent measures to evaluate the degree to which each testing condition recovered the known theta values. To evaluate the measurement accuracy and precision of the known theta recovery, the dependent measures for each sample included the mean estimated theta, mean standard error (SE), the Pearson product-moment correlation coefficient between the estimated and known theta values, bias, root mean squared error (RMSE), and average absolute difference (AAD). Then all of the statistics were averaged across the 100 replications. Additionally, conditional plots of the grand mean bias and SE were constructed to assess the testing conditions across the known theta scale.

Tables 7-10 contain the grand mean estimated theta and standard error (SE) for each panel design condition, i.e. 1-5-5, 1-3-3, 1-5, and 1-3. First the three-stage tables are presented, followed by the two-stage tables.

Table 7 provides the estimates for the 1-5-5 panel design conditions. The grand mean of the estimated thetas were reasonably close but slightly above the grand mean for the known thetas. As expected, there was an observed increase in the grand mean SE when the test length was decreased. The grand mean of the SEs for the long test length conditions ranged from 0.176 to 0.181 and the short test lengths ranged from 0.190 to 0.195. In addition, the grand mean of the SE for each routing procedures was very similar with the AMI consistently having the smallest SE. Across all conditions, the 1-5-5 panel design with the constant testlet effect LID condition, i.e. $\sigma^2_{d(j)} = 0.8$, resulted in the largest grand mean estimated theta. The grand mean of the SE's for each 1-5-5 condition were all similar in magnitude. The conditions that generated responses with a testlet effect of $\sigma^2_{d(j)} = 0.8$ consistently had the largest grand mean of the SEs, where the

conditions with no testlet effect, $\sigma^2_{d(j)} = 0$ and the item pool's estimated testlet effects, $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$, were similar.

Table 8 provides the descriptive statistics for the 1-3-3 panel design conditions. Similar to the 1-5-5 panel designs, the grand mean theta estimates were slightly above but close to the known thetas. As expected, the grand mean of the SE for the longer test length condition was smaller than the short test length. The range for the longer test lengths SEs were between 0.175 and 0.183, and the range for the short test lengths was between 0.192 and 0.200. All routing procedures produced similar grand mean SEs with the AMI routing having the smallest, and minimal differences being observed for the DPI procedures. The largest grand mean of the SE was recorded for the $\sigma^2_{d(j)} = 0.8$ LID conditions. The $\sigma^2_{d(j)} = 0.8$ LID condition had the largest grand mean of estimated theta ranging between 0.15 and 0.26.

Table 9 produced very similar descriptive statistic patterns for the 1-5 panel designs. The grand mean of the estimated thetas were all close but slightly above the grand mean of the known theta estimates ranging between 0.005 and 0.013. The grand mean of the SEs for the longer test lengths were smaller than the shorter test lengths. The range for the longer test lengths was 0196 to 0.200 and the range for the short test lengths was 0.214 to 0.220. Little to no difference was found between the grand mean of the SEs for each of the routing procedures for the 1-5 two stage tests. The $\sigma^2_{d(j)} = 0.8$ LID condition produced the largest grand mean of the SEs.

Table 10 provides the ability estimation descriptive statistics for 1-3 panel designs. The grand mean of the estimated thetas were all similar but slightly higher than the known theta estimates. The grand mean of the SEs were smallest for the longer test length conditions. Although differences were minimal, the AMI produced smaller grand mean of the SEs than did the DPI routing procedure. The $\sigma^2_{d(j)} = 0.8$ LID condition

producing the largest grand mean of the estimated thetas and grand mean of the SEs for each test length.

When comparing across panel designs, the grand mean of the theta estimates were all fairly similar. All were slightly higher than the grand mean for the known thetas. The descriptive statistics for the three-stage panel designs were very similarly to the two-stage tests. However, when examining the SEs for the three-stage test to the two-stage test designs, it was seen that the two-stage test produced higher standard errors than the three-stage tests.

Table 7. Estimated theta and Standard Error Descriptive Statistics for the 1-5-5 Panel Design.

Test Length	Routing Procedure	LID	Theta Estimate[a] Grand Mean (Min, Max)	SE Grand Mean (Min, Max)
Long	AMI	0.8	0.019 (-0.059, 0.094)	0.180 (0.176, 0.184)
		0.0	0.008 (-0.061, 0.086)	0.176 (0.172, 0.18)
		$\hat{\sigma}^2_{d(j)}$	0.013 (-0.083, 0.085)	0.176 (0.171, 0.18)
	ML-DPI	0.8	0.018 (-0.06, 0.085)	0.181 (0.178, 0.185)
		0.0	0.008 (-0.076, 0.089)	0.178 (0.173, 0.181)
		$\hat{\sigma}^2_{d(j)}$	0.011 (-0.071, 0.082)	0.178 (0.174, 0.182)
	SL_DPI	0.8	0.017 (-0.066, 0.078)	0.181 (0.177, 0.184)
		0.0	0.009 (-0.068, 0.091)	0.178 (0.173, 0.182)
		$\hat{\sigma}^2_{d(j)}$	0.013 (-0.091, 0.085)	0.178 (0.173, 0.181)
Short	AMI	0.8	0.015 (-0.082, 0.1)	0.195 (0.191, 0.199)
		0.0	0.009 (-0.063, 0.08)	0.190 (0.186, 0.193)
		$\hat{\sigma}^2_{d(j)}$	0.009 (-0.082, 0.085)	0.190 (0.187, 0.194)
	ML-DPI	0.8	0.015 (-0.073, 0.092)	0.195 (0.192, 0.199)
		0.0	0.009 (-0.066, 0.078)	0.191 (0.187, 0.194)
		$\hat{\sigma}^2_{d(j)}$	0.01 (-0.093, 0.08)	0.191 (0.187, 0.195)
	SL_DPI	0.8	0.015 (-0.071, 0.087)	0.195 (0.192, 0.198)
		0.0	0.009 (-0.063, 0.087)	0.191 (0.188, 0.194)
		$\hat{\sigma}^2_{d(j)}$	0.009 (-0.084, 0.075)	0.191 (0.187, 0.195)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. LID=local item dependence

[a] Known θ's: grand mean = -0.005, -0.001, and -0.003 for the $\sigma^2_{(d(j)} = (0.0, 0.8, \hat{\sigma}^2_{(d(j))})$

Table 8. Estimated Theta and Standard Error Descriptive Statistics for the 1-3-3 Panel Design.

Test Length	Routing Procedure	LID	Theta Estimate[a] Grand Mean (Min, Max)	SE Grand Mean (Min, Max)
Long	AMI	0.8	0.026 (-0.063, 0.102)	0.179 (0.176, 0.183)
		0.0	0.012 (-0.061, 0.083)	0.175 (0.172, 0.178)
		$\hat{\sigma}^2_{d(j)}$	0.018 (-0.07, 0.097)	0.176 (0.172, 0.18)
	ML-DPI	0.8	0.021 (-0.051, 0.091)	0.183 (0.18, 0.186)
		0.0	0.012 (-0.067, 0.089)	0.180 (0.178, 0.184)
		$\hat{\sigma}^2_{d(j)}$	0.016 (-0.085, 0.085)	0.180 (0.177, 0.184)
	SL_DPI	0.8	0.019 (-0.071, 0.087)	0.180 (0.177, 0.183)
		0.0	0 (-0.085, 0.072)	0.176 (0.173, 0.178)
		$\hat{\sigma}^2_{d(j)}$	0.015 (-0.071, 0.095)	0.177 (0.174, 0.181)
Short	AMI	0.8	0.022 (-0.047, 0.095)	0.198 (0.195, 0.201)
		0.0	0.01 (-0.053, 0.089)	0.194 (0.191, 0.197)
		$\hat{\sigma}^2_{d(j)}$	0.015 (-0.078, 0.087)	0.195 (0.19, 0.199)
	ML-DPI	0.8	0.015 (-0.072, 0.091)	0.200 (0.197, 0.203)
		0.0	0.01 (-0.054, 0.083)	0.196 (0.192, 0.199)
		$\hat{\sigma}^2_{d(j)}$	0.012 (-0.077, 0.087)	0.196 (0.193, 0.2)
	SL_DPI	0.8	0.015 (-0.073, 0.087)	0.197 (0.195, 0.201)
		0.0	0 (-0.068, 0.069)	0.192 (0.189, 0.195)
		$\hat{\sigma}^2_{d(j)}$	0.011 (-0.095, 0.072)	0.194 (0.191, 0.197)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. LID=local item dependence

[a] Known θ's: grand mean = -0.005, -0.001, and -0.003 for the $\sigma^2_{(d(j)} = (0.0, 0.8, \hat{\sigma}^2_{(d(j))})$

Table 9. Estimated Theta and Standard Error Descriptive Statistics for the 1-5 Panel Design.

Test Length	Routing Procedure	LID	Theta Estimate[a] Grand Mean (Min, Max)	SE Grand Mean (Min, Max)
Long	AMI	0.8	0.017 (-0.066, 0.085)	0.200 (0.196, 0.205)
		0.0	0.007 (-0.068, 0.08)	0.197 (0.192, 0.201)
		$\hat{\sigma}^2_{d(j)}$	0.011 (-0.068, 0.087)	0.196 (0.193, 0.2)
	DPI	0.8	0.012 (-0.067, 0.077)	0.200 (0.196, 0.205)
		0.0	0.007 (-0.073, 0.082)	0.197 (0.193, 0.201)
		$\hat{\sigma}^2_{d(j)}$	0.007 (-0.085, 0.093)	0.196 (0.193, 0.2)
Short	AMI	0.8	0.013 (-0.062, 0.084)	0.220 (0.216, 0.225)
		0.0	0.008 (-0.066, 0.082)	0.214 (0.21, 0.219)
		$\hat{\sigma}^2_{d(j)}$	0.009 (-0.086, 0.073)	0.214 (0.21, 0.217)
	DPI	0.8	0.009 (-0.079, 0.089)	0.220 (0.216, 0.225)
		0.0	0.005 (-0.074, 0.089)	0.214 (0.21, 0.22)
		$\hat{\sigma}^2_{d(j)}$	0.007 (-0.088, 0.079)	0.214 (0.211, 0.218)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. LID=local item dependence

[a] Known θ's: grand mean = -0.005, -0.001, and -0.003 for the $\sigma^2_{(d(j)} = (0.0, 0.8, \hat{\sigma}^2_{(d(j))})$

Table 10. Estimated Theta and Standard Error Descriptive Statistics for the 1-3 Panel Design.

Test Length	Routing Procedure	LID	Theta Estimate[a] Grand Mean (Min, Max)	SE Grand Mean (Min, Max)
Long	AMI	0.8	0.019 (-0.047, 0.082)	0.201 (0.198, 0.204)
		0.0	0.008 (-0.061, 0.079)	0.194 (0.191, 0.197)
		$\hat{\sigma}^2_{d(j)}$	0.011 (-0.08, 0.081)	0.197 (0.193, 0.201)
	DPI	0.8	0.015 (-0.064, 0.084)	0.203 (0.199, 0.206)
		0.0	0.008 (-0.06, 0.076)	0.199 (0.195, 0.202)
		$\hat{\sigma}^2_{d(j)}$	0.01 (-0.08, 0.087)	0.198 (0.195, 0.202)
Short	AMI	0.8	0.016 (-0.057, 0.075)	0.219 (0.216, 0.224)
		0.0	0.007 (-0.064, 0.079)	0.211 (0.206, 0.215)
		$\hat{\sigma}^2_{d(j)}$	0.008 (-0.084, 0.091)	0.215 (0.211, 0.219)
	DPI	0.8	0.013 (-0.061, 0.089)	0.223 (0.219, 0.227)
		0.0	0.006 (-0.075, 0.082)	0.216 (0.211, 0.221)
		$\hat{\sigma}^2_{d(j)}$	0.007 (-0.075, 0.078)	0.216 (0.212, 0.22)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. LID=local item dependence

[a] Known θ's: grand mean = -0.005, -0.001, and -0.003 for the $\sigma^2_{(d(j)} = (0.0, 0.8, \hat{\sigma}^2_{(d(j))})$

Tables 11-14 provide the summary statistics for the Pearson product moment correlation coefficient, bias, RMSE, and AAD for each panel design condition in the study. The patterns of results within each of the panel designs were very similar, therefore they are discussed collectively.

Table 11-14 provides the results for the measurement accuracy and precision measures for the 1-5-5, 1-3-3, 1-5, and 1-3 panel designs, respectively. The mean correlation coefficient for each condition was very high. As expected, the smallest mean correlation occurred in the short test lengths. But even in the short test length conditions the smallest mean correlation observed was 0.940 in the 1-3-3 panel design in Table 12. The range of mean correlations for all test lengths, routing procedures, and LID conditions was 0.943 to 0.961, 0.940 to 0.967, 0.945 to 0.967, and 0.942 to 0.967 for the 1-5-5, 1-3-3, 1-5, and 1-3 panel designs, respectively. When comparing the routing procedures very little difference was observed. While holding other variables constant, the AMI most frequently resulted in the highest mean correlation, but no practical difference was indicated as the largest observed difference between any routing procedure mean correlation was less than or equal to 0.001. The $\sigma^2_{d(j)} = 0.8$ constant large testlet effect consistently provided the smallest mean correlation regardless of test length, routing procedures, or panel design. The mean correlations ranged from 0.940 to 0.954, 0.952 to 0.967, and 0.950 to 0.964 for the LID conditions $\sigma^2_{d(j)} = (0.8, 0.0, \hat{\sigma}^2_{d(j)})$, respectively

The bias across the panel designs did not provide many consistent patterns. All conditions illustrate a small amount of positive bias but the magnitude was functionally zero, as the mean bias ranged between 0.005 and 0.026 across all conditions. The mean bias did not seem to be dictated by test length. Within a panel design, many instances produced a reduction in mean bias when routing procedure and LID condition were held

constant and test length was shortened, such as the case for the entire set of 1-3-3 panel design conditions. Another set of occurrences were the 1-5 and 1-3 panel designs for both AMI routing procedures when a testlet effect was present. When comparing routing procedure, no real difference was detected. The AMI did not consistently outperform the other routing procedures although the modules were being selected based on the highest amount of information. In some instances, the DPI outperformed the AMI procedures with respect to the mean bias, such as the 1-3-3 panel design in Table 12. The $\sigma^2_{d(j)} = 0.8$ LID condition typically produced the largest mean bias across the panel designs when holding constant test length and routing procedures. However, a couple of instances occurred when the $\sigma^2_{d(j)} = 0.8$ LID condition did not produce the highest mean bias, such as in Table 13 for the 1-5 panel design using the DPI procedure. One should note that the differences being discussed were minimal and the greatest difference in mean bias between any conditions was only 0.013.

The mean RMSE and mean AAD patterns were very similar across each set of panel designs. As expected, longer tests produced smaller mean RMSE and mean AAD than the shorter test length conditions. One should also notice that routing procedures show very little differences when all other conditions are held constant. For example, the range of mean RMSE for the 1-3-3 panel design for the longer test across all routing procedures was 0.287 to 0.292, 0.260-0.285, and 0.321 to 0.327 for the $\sigma^2_{(d(j)} = (0.0, 0.8, \hat{\sigma}^2_{(d(j))})$ LID conditions, respectively The largest mean RMSE and mean AAD was detected for the $\sigma^2_{d(j)} = 0.8$ LID condition, followed by the condition where $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$, and the smallest measures were present when no testlet effect was included in the response generation. The range for the RMSE was 0.322 to 0.337, 0.321 to 0.347, 0.301 to 0.327, and 0.304 to 0.335 for the 1-5-5, 1-3-3, 1-5, and 1-3 respectively for the $\sigma^2_{d(j)} = 0.8$ conditions. The range for the mean RMSE for the same respective panel

designs for the $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$ conditions were 0.285 to 0.310, 0.288 to 0.320, 0.266 to 0.290, and 0.268 to 0.295. Finally, the RMSE for same the respective panel designs under the $\sigma^2_{d(j)} = 0.0$ LID ranged from 0.278 to 0.307, 0.260 to 0.314, 0.255 to 0.278, and 0.257 to 0.279.

The differences across panels also needed to be evaluated. When the three-stage panel designs were compared to each other, very little differences were found for all dependent measures. The same results were seen when comparing the two-stage tests. Interestingly, when the three-stage tests (i.e. 1-5-5 and 1-3-3) were compared to the two-stage (1-5 and 1-3) panel designs some differences were found between the mean correlations, RMSE, and AAD. The mean correlation coefficients were consistently higher for the two-stage test designs than the three-stage test designs. For instance, when one compares the long test length conditions across the LID conditions, the three-stage mean correlations range from 0.946 to 0.948, 0.960 to 0.967, and 0.958 to 0.960 for the $\sigma^2_{d(j)} = (0.8, 0.0, \hat{\sigma}^2_{d(j)})$ LID conditions, respectively. Where the respective LID conditions for the two-stage test designs range from 0.952 to 0.954, 0.967 to 0.967, and 0.963 to 0.964.

When comparing panel designs, the mean RMSE and mean AAD produced similar patterns to the mean correlations. First, when looking at the three-stage tests we see that the mean RMSE and mean AAD were generally larger for the 1-3-3 panel design. For instance, when the AMI routing procedure was compared across the two panel designs for the long test length each LID condition was at least .002 higher for mean RMSE. However, the mean RMSE and mean AAD were smaller for the 1-3-3 panel design for the SL-DPI condition with no testlet effect LID condition. Where the mean RMSE for the 1-5-5 panel was 0.281 and 0.306 and for the 1-3-3 panel design 0.260 and 0.293 for the long and short length tests, respectively. The 1-3 mean RMSE and mean

Table 11. The Correlation Coefficient Between Known and Estimated Theta, Bias, RMSE, and AAD for the 1-5-5 Panel Designs.

Test Length	Routing Procedure	LID	Correlation Mean (Min, Max)	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	AAD Mean (Min, Max)
Long	AMI	0.8	0.947 (0.941, 0.954)	0.019 (-0.008, 0.043)	0.322 (0.299, 0.341)	0.253 (0.236, 0.271)
		0.0	0.962 (0.956, 0.966)	0.014 (-0.008, 0.03)	0.278 (0.266, 0.291)	0.218 (0.206, 0.231)
		$\hat{\sigma}^2_{d(j)}$	0.960 (0.952, 0.964)	0.015 (-0.008, 0.038)	0.285 (0.269, 0.301)	0.224 (0.209, 0.237)
	ML-DPI	0.8	0.947 (0.938, 0.954)	0.018 (-0.005, 0.044)	0.323 (0.303, 0.341)	0.254 (0.239, 0.265)
		0.0	0.961 (0.956, 0.965)	0.013 (-0.003, 0.034)	0.281 (0.265, 0.299)	0.221 (0.209, 0.236)
		$\hat{\sigma}^2_{d(j)}$	0.959 (0.951, 0.965)	0.014 (-0.01, 0.034)	0.287 (0.272, 0.307)	0.226 (0.214, 0.241)
	SL_DPI	0.8	0.947 (0.937, 0.955)	0.017 (-0.003, 0.04)	0.323 (0.302, 0.343)	0.254 (0.241, 0.269)
		0.0	0.961 (0.956, 0.967)	0.014 (-0.005, 0.039)	0.281 (0.268, 0.295)	0.221 (0.208, 0.233)
		$\hat{\sigma}^2_{d(j)}$	0.959 (0.954, 0.967)	0.015 (-0.002, 0.038)	0.286 (0.266, 0.301)	0.225 (0.208, 0.237)
Short	AMI	0.8	0.943 (0.936, 0.95)	0.015 (-0.007, 0.037)	0.336 (0.317, 0.353)	0.264 (0.25, 0.279)
		0.0	0.955 (0.947, 0.962)	0.015 (-0.005, 0.05)	0.304 (0.285, 0.322)	0.240 (0.223, 0.254)
		$\hat{\sigma}^2_{d(j)}$	0.953 (0.944, 0.959)	0.012 (-0.012, 0.034)	0.309 (0.29, 0.326)	0.244 (0.231, 0.262)
	ML-DPI	0.8	0.943 (0.936, 0.951)	0.016 (-0.006, 0.037)	0.337 (0.319, 0.361)	0.265 (0.25, 0.285)
		0.0	0.954 (0.945, 0.96)	0.015 (-0.003, 0.039)	0.307 (0.291, 0.327)	0.242 (0.23, 0.256)
		$\hat{\sigma}^2_{d(j)}$	0.952 (0.943, 0.96)	0.013 (-0.009, 0.039)	0.310 (0.293, 0.325)	0.245 (0.231, 0.258)
	SL_DPI	0.8	0.943 (0.933, 0.95)	0.015 (-0.012, 0.039)	0.336 (0.317, 0.356)	0.265 (0.253, 0.281)
		0.0	0.954 (0.948, 0.959)	0.015 (-0.006, 0.045)	0.306 (0.29, 0.325)	0.241 (0.227, 0.253)
		$\hat{\sigma}^2_{d(j)}$	0.953 (0.945, 0.959)	0.012 (-0.01, 0.035)	0.309 (0.29, 0.332)	0.244 (0.229, 0.259)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. AMI=approximate maximum information; ML-DPI=module-level defined population interval; SL-DPI=stage-level defined population interval; LID=local item dependence

Table 12. The Correlation Coefficient Between Known and Estimated Theta, Bias, RMSE, and AAD for the 1-3-3 Panel Designs.

Test Length	Routing Procedure	LID	Correlation Mean (Min, Max)	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	AAD Mean (Min, Max)
Long	AMI	0.8	0.947 (0.939, 0.955)	0.026 (-0.002, 0.061)	0.325 (0.308, 0.347)	0.255 (0.243, 0.274)
		0.0	0.961 (0.956, 0.968)	0.018 (-0.003, 0.045)	0.281 (0.263, 0.298)	0.222 (0.208, 0.238)
		$\hat{\sigma}^2_{d(j)}$	0.959 (0.953, 0.964)	0.021 (0.002, 0.039)	0.288 (0.275, 0.3)	0.227 (0.217, 0.236)
	ML-DPI	0.8	0.946 (0.936, 0.952)	0.021 (-0.007, 0.047)	0.327 (0.305, 0.347)	0.258 (0.241, 0.272)
		0.0	0.960 (0.955, 0.965)	0.017 (-0.003, 0.05)	0.285 (0.273, 0.306)	0.225 (0.213, 0.241)
		$\hat{\sigma}^2_{d(j)}$	0.958 (0.952, 0.962)	0.018 (0, 0.039)	0.292 (0.276, 0.309)	0.231 (0.217, 0.242)
	SL_DPI	0.8	0.948 (0.937, 0.957)	0.019 (-0.005, 0.042)	0.321 (0.3, 0.345)	0.253 (0.236, 0.273)
		0.0	0.967 (0.962, 0.971)	0.005 (-0.019, 0.021)	0.260 (0.239, 0.273)	0.205 (0.192, 0.217)
		$\hat{\sigma}^2_{d(j)}$	0.959 (0.951, 0.964)	0.018 (-0.014, 0.036)	0.287 (0.27, 0.306)	0.226 (0.215, 0.242)
Short	AMI	0.8	0.940 (0.932, 0.947)	0.022 (-0.002, 0.047)	0.347 (0.328, 0.362)	0.272 (0.255, 0.284)
		0.0	0.952 (0.944, 0.959)	0.016 (-0.011, 0.039)	0.314 (0.293, 0.33)	0.248 (0.229, 0.26)
		$\hat{\sigma}^2_{d(j)}$	0.950 (0.941, 0.956)	0.018 (-0.001, 0.036)	0.320 (0.298, 0.342)	0.252 (0.237, 0.27)
	ML-DPI	0.8	0.940 (0.93, 0.949)	0.015 (-0.009, 0.036)	0.344 (0.325, 0.373)	0.272 (0.257, 0.298)
		0.0	0.952 (0.944, 0.957)	0.015 (-0.008, 0.04)	0.314 (0.297, 0.338)	0.248 (0.234, 0.265)
		$\hat{\sigma}^2_{d(j)}$	0.950 (0.943, 0.956)	0.014 (0, 0.038)	0.318 (0.299, 0.342)	0.251 (0.235, 0.269)
	SL_DPI	0.8	0.941 (0.932, 0.948)	0.015 (-0.012, 0.048)	0.342 (0.32, 0.359)	0.270 (0.253, 0.285)
		0.0	0.958 (0.951, 0.963)	0.005 (-0.018, 0.025)	0.293 (0.277, 0.313)	0.232 (0.216, 0.251)
		$\hat{\sigma}^2_{d(j)}$	0.951 (0.942, 0.958)	0.013 (-0.012, 0.047)	0.314 (0.295, 0.334)	0.248 (0.233, 0.262)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. AMI=approximate maximum information; ML-DPI=module-level defined population interval; SL-DPI=stage-level defined population interval; LID=local item dependence

Table 13. The Correlation Coefficient Between Known and Estimated Theta, Bias, RMSE, and AAD for the 1-5 Panel Designs.

Note: All statistics were computed from across 100 replications; each replication contained 1,000

Test Length	Routing Procedure	LID	Correlation Mean (Min, Max)	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	AAD Mean (Min, Max)
Long	AMI	0.8	0.954 (0.946, 0.961)	0.017 (-0.007, 0.042)	0.301 (0.277, 0.327)	0.236 (0.218, 0.253)
		0.0	0.967 (0.962, 0.971)	0.012 (-0.004, 0.029)	0.255 (0.243, 0.271)	0.201 (0.191, 0.214)
		$\hat{\sigma}^2_{d(j)}$	0.964 (0.959, 0.969)	0.014 (-0.008, 0.034)	0.266 (0.251, 0.279)	0.209 (0.199, 0.22)
	DPI	0.8	0.953 (0.946, 0.962)	0.012 (-0.009, 0.035)	0.301 (0.282, 0.318)	0.237 (0.222, 0.248)
		0.0	0.967 (0.964, 0.972)	0.013 (-0.009, 0.032)	0.255 (0.241, 0.272)	0.201 (0.19, 0.211)
		$\hat{\sigma}^2_{d(j)}$	0.964 (0.957, 0.968)	0.009 (-0.008, 0.028)	0.266 (0.245, 0.282)	0.209 (0.194, 0.221)
Short	AMI	0.8	0.945 (0.938, 0.951)	0.014 (-0.009, 0.038)	0.327 (0.306, 0.347)	0.256 (0.241, 0.273)
		0.0	0.961 (0.955, 0.967)	0.014 (-0.008, 0.037)	0.277 (0.257, 0.295)	0.219 (0.202, 0.232)
		$\hat{\sigma}^2_{d(j)}$	0.958 (0.951, 0.963)	0.012 (-0.015, 0.033)	0.289 (0.272, 0.307)	0.228 (0.211, 0.24)
	DPI	0.8	0.945 (0.935, 0.953)	0.010 (-0.011, 0.037)	0.326 (0.301, 0.354)	0.256 (0.238, 0.271)
		0.0	0.961 (0.954, 0.965)	0.011 (-0.009, 0.034)	0.278 (0.263, 0.3)	0.219 (0.209, 0.235)
		$\hat{\sigma}^2_{d(j)}$	0.958 (0.952, 0.963)	0.009 (-0.007, 0.022)	0.290 (0.271, 0.308)	0.228 (0.214, 0.242)

observations. AMI=approximate maximum information; ML-DPI=module-level defined population interval; SL-DPI=stage-level defined population interval; LID=local item dependence

Table 14. The Correlation Coefficient Between Known and Estimated Theta, Bias, RMSE, and AAD for the 1-3 Panel Designs.

Test Length	Routing Procedure	LID	Correlation Mean (Min, Max)	Bias Mean (Min, Max)	RMSE Mean (Min, Max)	AAD Mean (Min, Max)
Long	AMI	0.8	0.953 (0.944, 0.959)	0.019 (-0.003, 0.045)	0.304 (0.279, 0.33)	0.238 (0.222, 0.257)
		0.0	0.967 (0.962, 0.972)	0.013 (-0.006, 0.04)	0.258 (0.242, 0.273)	0.202 (0.191, 0.215)
		$\hat{\sigma}^2_{d(j)}$	0.964 (0.959, 0.969)	0.014 (-0.008, 0.036)	0.268 (0.252, 0.284)	0.211 (0.2, 0.221)
	DPI	0.8	0.952 (0.945, 0.959)	0.015 (-0.007, 0.037)	0.305 (0.287, 0.32)	0.239 (0.228, 0.252)
		0.0	0.967 (0.962, 0.972)	0.013 (-0.009, 0.033)	0.257 (0.238, 0.271)	0.202 (0.187, 0.211)
		$\hat{\sigma}^2_{d(j)}$	0.963 (0.957, 0.968)	0.012 (-0.003, 0.029)	0.269 (0.256, 0.284)	0.212 (0.202, 0.225)
Short	AMI	0.8	0.943 (0.936, 0.95)	0.017 (-0.006, 0.044)	0.333 (0.311, 0.354)	0.262 (0.248, 0.275)
		0.0	0.961 (0.953, 0.965)	0.013 (-0.019, 0.037)	0.279 (0.261, 0.296)	0.219 (0.203, 0.233)
		$\hat{\sigma}^2_{d(j)}$	0.957 (0.951, 0.962)	0.011 (-0.011, 0.035)	0.293 (0.277, 0.311)	0.230 (0.217, 0.244)
	DPI	0.8	0.942 (0.931, 0.949)	0.014 (-0.009, 0.042)	0.335 (0.319, 0.357)	0.263 (0.25, 0.283)
		0.0	0.961 (0.955, 0.966)	0.011 (-0.007, 0.031)	0.278 (0.259, 0.292)	0.219 (0.203, 0.231)
		$\hat{\sigma}^2_{d(j)}$	0.956 (0.950, 0.961)	0.010 (-0.009, 0.032)	0.295 (0.277, 0.308)	0.232 (0.22, 0.243)

Note: All statistics were computed from across 100 replications; each replication contained 1,000 observations. AMI=approximate maximum information; DPI=defined population interval; LID=local item dependence

AAD were all greater than or approximately equal to the 1-5 panel designs, when holding the other conditions constant.

Interestingly, more pronounced differences in magnitude were seen when comparing the three-stage tests to the two-stage tests. When the long tests for both three-stage tests and two-stage tests were compared for all respective routing and LID conditions, the mean RMSE and mean AAD were smaller for the two-stage tests. The same result was found for the short length tests. The range for the three-stage long test length conditions for mean RMSE and AAD were 0.260 to 0.327 and 0.205 to 0.258, respectively. For the two-stage long test lengths the mean RMSE and AAD ranged from 0.255 to 0.304 and 0.201 to 0.238, respectively. The three-stage short conditions mean RMSE and mean AAD ranged from 0.293 to 0.347 and 0.232 to 0.272, respectively. Where the two-stage short length test mean RMSE and mean AAD ranged from 0.277 to 0.335 and 0.219 to 0.263, respectively.

Figures 17-21 illustrate the conditional plots across the three LID conditions' mean bias and grand mean of the SE conditional on theta. Figures 17-18 give the 1-3 panel design, short test length plots for the AMI and DPI routing procedures, respectively. Figures 19-21 give the 1-5-5 long test length plots for the AMI, ML-DPI, and SL-DPI routing procedures, respectively. These figures were presented as they represented the extremes of the panel designs for the study. Figure 17-21 shows that the three LID conditions performed similarly in terms of conditional mean bias and condition grand mean SE for the majority of the distribution. Differences start to occur in the MST designs for $-2.0 \geq \theta \geq 2.0$. In the extremes of the distribution we start to see that the condition with larger amounts of item dependency produce more bias. The LID condition with the constant large testlet effect, i.e. $\sigma^2_{d(j)} = 0.8$ produced the smallest SEs in the extremes of the distribution. Although the overall grand mean SEs were larger for

the $\sigma^2_{d(j)} = 0.8$ condition, it was seen in this study that the SEs may be underestimated in the extremes of the distribution. The condition using the real estimated testlet effects and the condition with no testlet effect were very similar. Over the set of the study conditions, the extremes for the negative θ's were typically less bias for the estimated testlet effects, $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$, while the positive θ's typically had less bias for the conditions with no testlet effect, $\sigma^2_{d(j)} = 0$. No discernable differences were seen in the conditional plots between the panel designs, test lengths, or routing procedures. The patterns discussed were observed throughout all the conditions investigated in the study. As such, the remainder of the conditional mean bias and conditional grand mean SE plots for the remaining manipulated variables can be found in Appendix B.

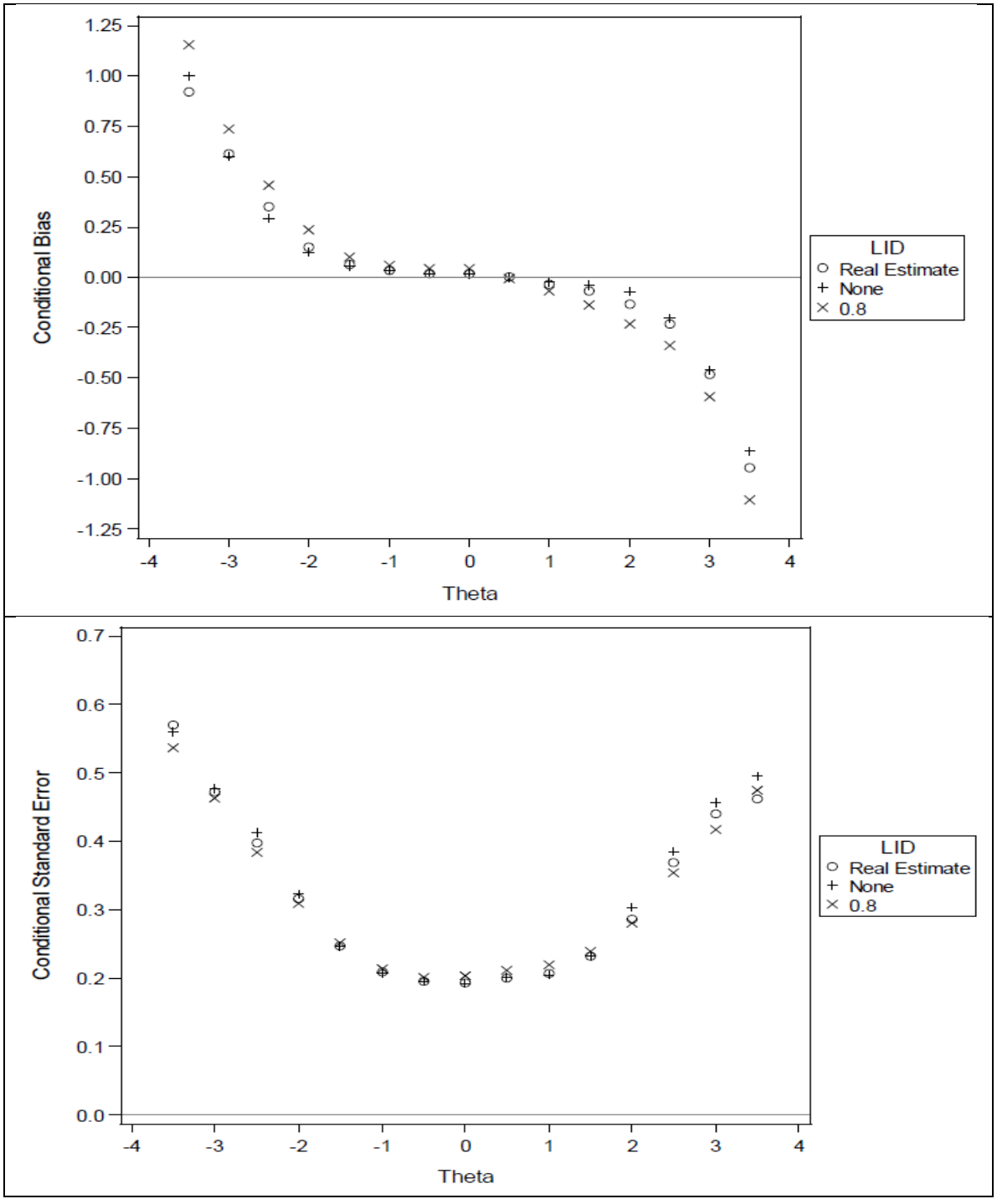

Figure 17. Conditional bias and standard error plots for the 1-3 panel design, short test length, AMI routing procedure across the LID conditions. *Note*: AMI=approximate maximum information; LID=local item dependence.

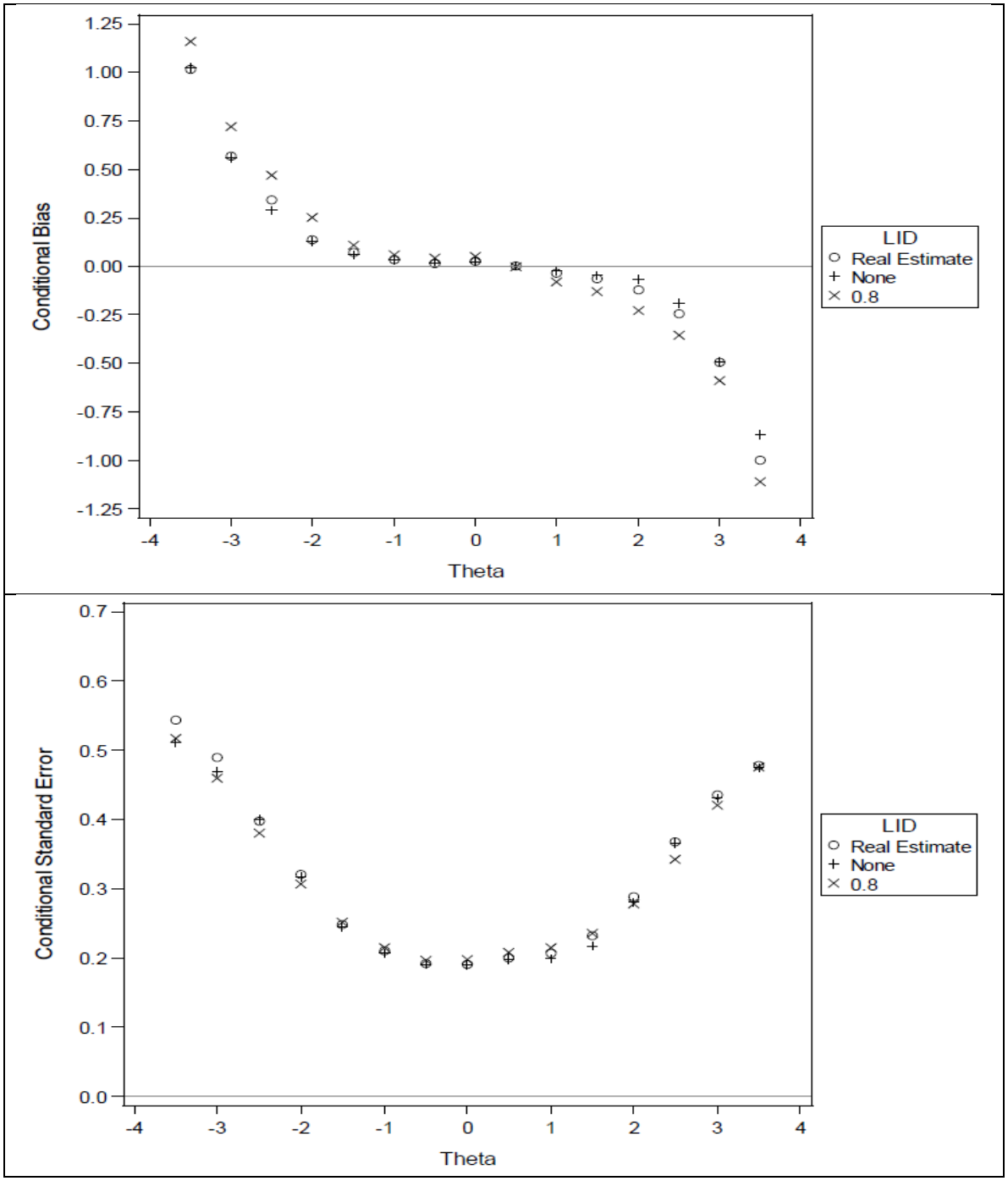

Figure 18. Conditional bias and standard error plots for the 1-3 panel design, short test length, DPI routing procedure across the LID conditions.
Note: DPI=defined population interval; LID=local item dependence.

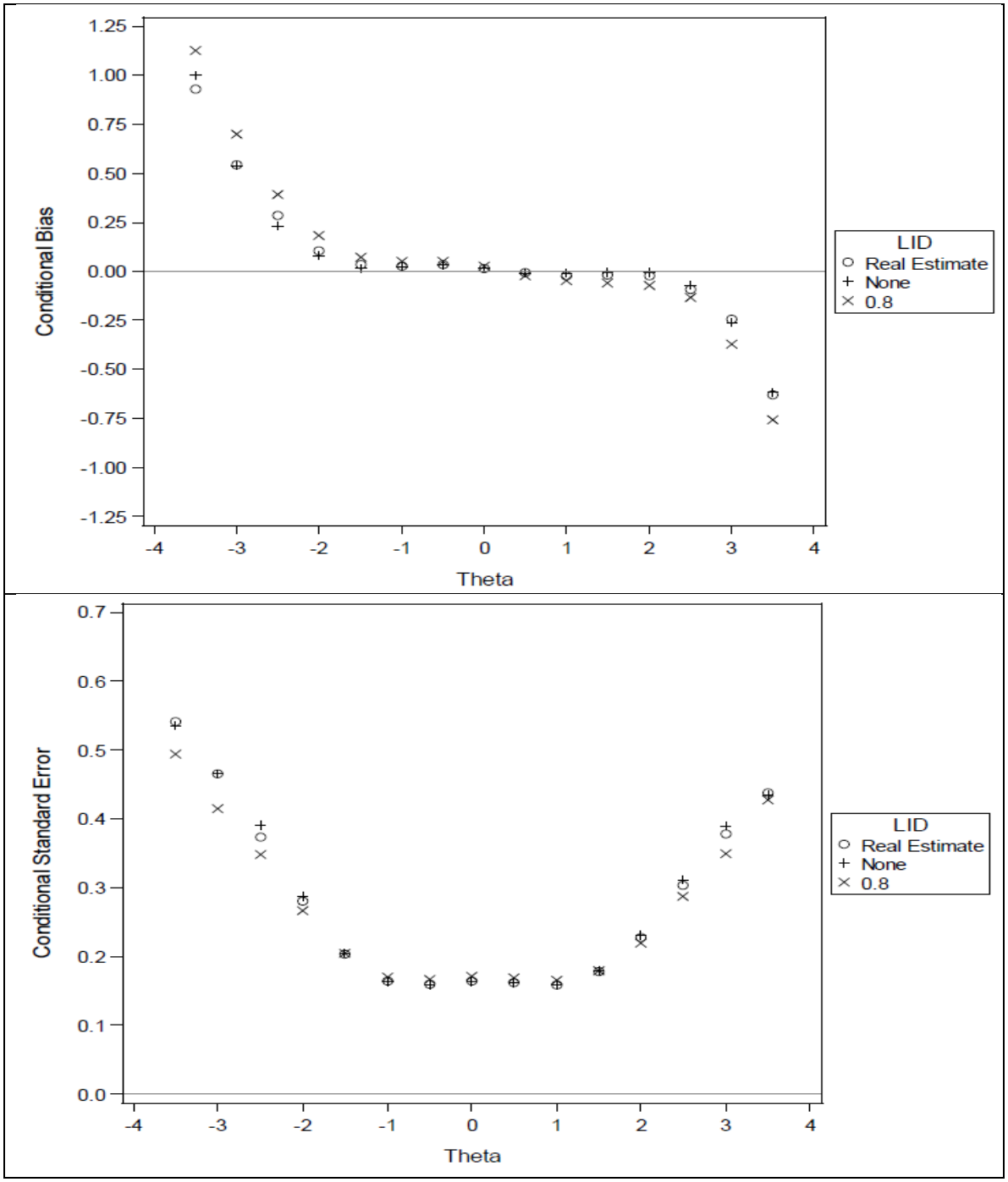

Figure 19. Conditional bias and standard error plots for the 1-5-5 panel design, long test length, AMI routing procedure across the LID conditions. *Note*: AMI=approximate maximum information; LID=local item dependence.

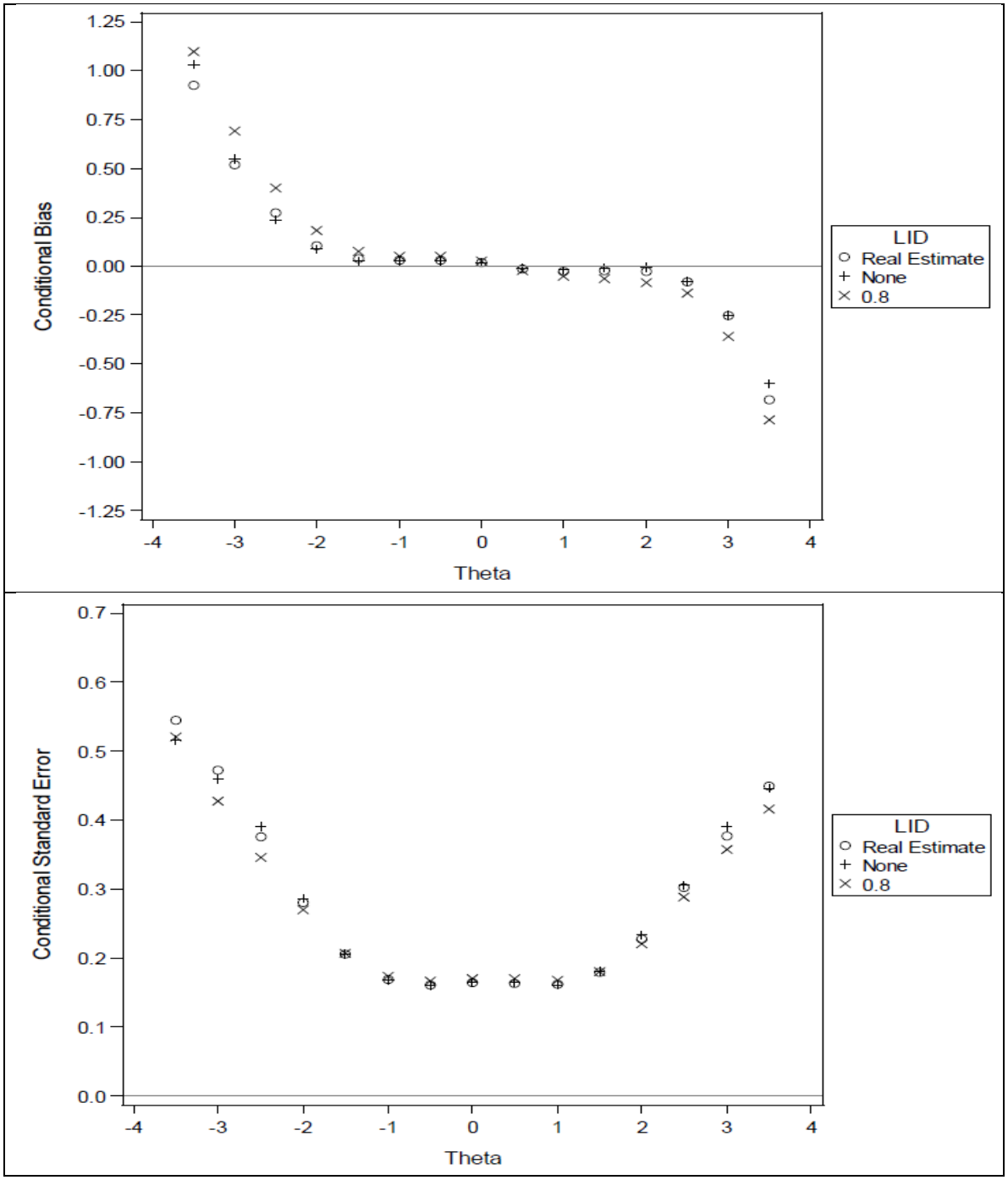

Figure 20. Conditional bias and standard error plots for the 1-5-5 panel design, long test length, ML-DPI routing procedure across the LID conditions. *Note*: ML- DPI=module-level defined population interval; LID=local item dependence.

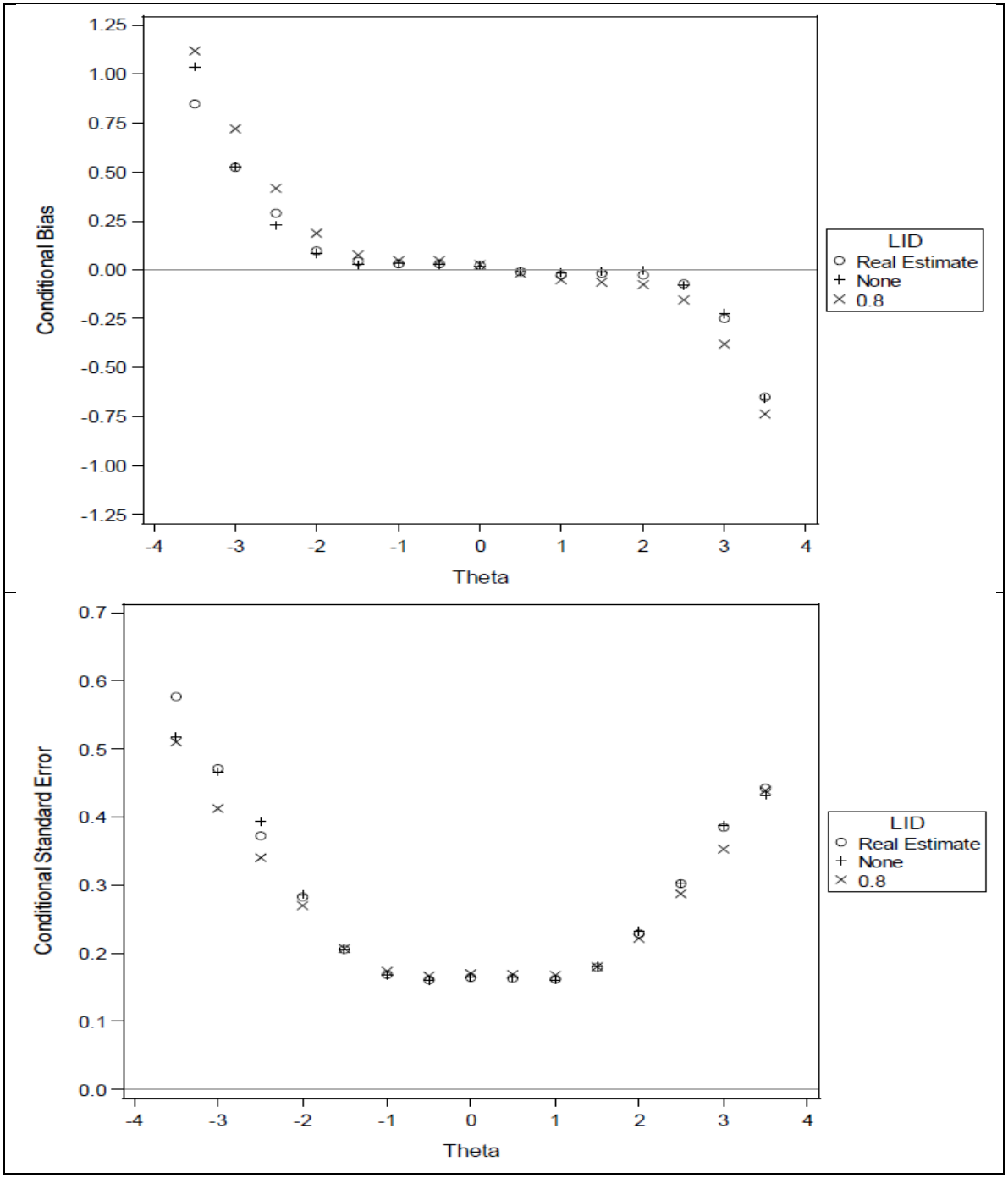

Figure 21. Conditional bias and standard error plots for the 1-5-5 panel design, long test length, SL-DPI routing procedure across the LID conditions.
Note: SL- DPI=stage-level defined population interval; LID=local item dependence.

Overall, three patterns emerged from the results. First, test length impacts the measurement precision of the assessment. The longer tests consistently produced smaller grand mean SEs and higher mean correlations, mean RMSE, and mean AAD. The second pattern was with regard to the amount of LID used in the generating responses. When a consistent testlet effect variance of 0.8 was used in the response generation, the measurement precision slightly decreased. This LID condition produced smaller mean correlations, higher RMSE, and higher AAD. Similarly, the condition that generated responses with the estimated testlet effect variances, $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$, performed better with respect to the larger testlet effect condition, $\sigma^2_{d(j)} = 0.8$ but illustrated some depreciation when comparing it to the condition where generated responses did not have a testlet effect present, $\sigma^2_{d(j)} = 0$. The third pattern occurred when comparing the three-stage tests to the two-stage tests. When comparing the respective test length, routing procedures, and LID conditions, the two-stage tests seemingly outperformed the three-stage tests. The two-stage tests resulted in higher correlations, lower RMSE, and lower AAD.

The study also attempted to identify an interaction that may be taking place with respect to the four manipulated variables. The only remote indication of an interaction occurred in the 1-3-3 panel designs for the SL-DPI routing procedure when no testlet effect was present. It was the only condition that decreased in RMSE when comparing the five module designs to the three module designs.

MODULE AND PANEL ROUTING PROPERTIES

Routing procedure properties were assessed by computing the frequency distribution of module administration for each manipulated condition. To compare the different routing procedures, the decision point for the AMI procedure needed to be

presented as the were only identified after panel assembly. Then to further evaluate the administration process, the average percentage of module administration was calculated over the 100 replications. The percentage of modules administered were an indication of the effectiveness of the desired module administration for each routing procedure and could enlighten any differences that may have occurred in the dependent variables.

Table 15 provides the decision points for the AMI routing procedure regarding the panel designs with five modules for the second and third stage. Table 16 provides the decision points for the AMI routing procedure regarding the panel designs with three modules for the second and third stage. Each panel assembled had unique decision points based on the intersection of the adjacent modules. When you look at the decision points across the panels for the respective pathways, the theta cut points were fairly similar, with the largest difference for a respective decision point being 0.2 on the theta scale.

Table 15. Approximate Maximum Information Theta Decision Points for the 1-5-5 and 1-5 Panel Designs.

		θ_{LL}	θ_{LM}	θ_{MU}	θ_{UU}
		1-5-5 Long			
Stage 2	Panel 1	-0.5	-0.3	0.3	1
	Panel 2	-0.5	-0.3	0.2	0.9
Stage 3	Panel 1	-1	-0.4	0	0.5
	Panel 2	-0.8	-0.5	0.2	0.5
		1-5-5 Short			
Stage 2	Panel 1	-0.7	-0.3	0.2	0.7
	Panel 2	-0.7	-0.3	0.1	0.7
Stage 3	Panel 1	-0.6	-0.3	0.1	0.7
	Panel 2	-0.7	-0.4	0.1	0.6
		1-5 Long			
Stage 2	Panel 1	-0.8	-0.4	0.2	0.7
	Panel 2	-1	-0.4	0.2	0.5
	Panel 3	-0.9	-0.4	0.1	0.5
		1-5 Short			
Stage 2	Panel 1	-0.8	-0.4	0.2	0.6
	Panel 2	-0.9	-0.4	0.2	0.6
	Panel 3	-1	-0.4	0.1	0.6

Table 16. Approximate Maximum Information Theta Decision Points for the 1-3-3 and 1-3 Panel Designs.

Panel Design		1-3-3				1-3			
Test Length		Long		Short		Long		Short	
		θ_L	θ_U	θ_L	θ_U	θ_L	θ_U	θ_L	θ_U
Stage 2	Panel 1	-0.6	0.5	-0.5	0.4	-0.7	0.4	-0.7	0.4
	Panel 2	-0.4	0.4	-0.5	0.5	-0.7	0.4	-0.7	0.4
	Panel 3	-0.4	0.4	-0.6	0.4	-0.7	0.5	-0.6	0.5
Stage 3	Panel 1	-0.7	0.3	-0.4	0.4	NA	NA	NA	NA
	Panel 2	-0.7	0.4	-0.7	0.2	NA	NA	NA	NA
	Panel 3	-0.8	0.2	-0.7	0.1	NA	NA	NA	NA

Table 17 provides the percentage of simulees for the 1-5-5 panel design, long test length, and estimated testlet variance LID condition that took each module. When examining the marginal percentage, it was seen the AMI procedure administered a large portion of the very easy module with 30.5% of simulees viewing this module at Stage 2. In Stage 3, we see that the distribution was pretty evenly distributed with the largest percentage of simulees taking the very hard module. Both DPI procedures performed similarly with marginal percentages at approximately 20% at both Stage 2 and Stage 3. The difference in the two procedures was elucidated when comparing how they rerouted simulees during administration between stages. The ML-DPI roughly rerouted equal proportions within a module at Stage 2 to an adjacent module at Stage 3. By contrast, the SL-DPI consistently retained the majority of simulees to take a module of similar difficulty at Stage 2 and Stage 3. Tables 18 and 19 are the additional tables describing the administration for the 1-5-5 panel design, long test length, for the responses generated with no testlet effect and a consistent testlet variance of 0.8. Similar patterns were observed as discussed above for the two remaining LID conditions.

Tables 20-22 provide the percentage of simulees taking each module for the short test length 1-5-5 panel design conditions. The results indicated that the short 1-5-5 tests performed similarly to that of the long test length administrations. The most noticeable difference occurred with the AMI procedure. The distribution of simulees administered to each module at each difficulty level was approximately uniform. However, when examining the Stage 3 module administration it seems to become less uniform more simulees seem to be administered to the very easy and very hard modules. Similar to the longer test length, the DPI procedures administer approximately administer 20% of each module, with the ML-DPI procedure rerouting more simulees than the SL-DPI routing procedure.

Table 17. Percentage of Module Administration for the 1-5-5 Panel Design, Long Test Length for the $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$ Local Item Dependency condition.

		Route 2					
	Route 1	VE	ME	M	MH	VH	**Total**
AMI	VE	17.3	13.2				**30.5**
	ME	0.5	2.6	4.5			**7.5**
	M		2.6	10.8	9.1		**22.5**
	MH			2.6	7.6	13.2	**23.4**
	VH				0.4	15.7	**16.1**
	Total	**17.8**	**18.3**	**17.9**	**17.1**	**28.9**	**100.0**
		VE	ME	M	MH	VH	**Total**
ML-DPI	VE	13.5	5.4				**18.9**
	ME	6.3	7.1	7.3			**20.7**
	M		7.0	7.0	7.5		**21.5**
	MH			6.3	7.9	6.4	**20.6**
	VH				5.0	13.4	**18.4**
	Total	**19.8**	**19.5**	**20.5**	**20.4**	**19.8**	**100.0**
		VE	ME	M	MH	VH	**Total**
SL-DPI	VE	15.4	3.6				**19.0**
	ME	3.8	11.4	5.4			**20.5**
	M		4.9	11.3	5.2		**21.4**
	MH			4.4	12.4	3.8	**20.6**
	VH				3.2	15.3	**18.5**
	Total	**19.2**	**19.9**	**21.0**	**20.9**	**19.1**	**100.0**

Note: All percentages were calculated across all 100 replications and 1,000 simulees per replication (i.e. N=100,000). AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard.

Table 18. Percentage of Module Administration for the 1-5-5 Panel Design, Long Test Length for the $\sigma^2_{d(j)} = 0.0$ Local Item Dependency Condition.

		Route 2					
	Route 1	VE	ME	M	MH	VH	**Total**
AMI	VE	17.6	12.4				**30.0**
	ME	0.4	2.8	4.4			**7.6**
	M		2.4	11.9	9.0		**23.3**
	MH			2.4	7.8	13.1	**23.3**
	VH				0.2	15.7	**15.9**
	Total	**18.0**	**17.5**	**18.7**	**17.0**	**28.7**	**100.0**
		VE	ME	M	MH	VH	**Total**
ML-DPI	VE	13.9	4.9				**18.8**
	ME	6.3	7.7	6.6			**20.6**
	M		7.0	7.4	7.3		**21.8**
	MH			6.2	8.2	6.1	**20.5**
	VH				4.6	13.9	**18.4**
	Total	**20.1**	**19.6**	**20.3**	**20.2**	**19.9**	**100.0**
		VE	ME	M	MH	VH	**Total**
SL-DPI	VE	15.7	3.1				**18.8**
	ME	3.7	12.0	4.7			**20.4**
	M		4.8	12.3	4.9		**21.9**
	MH			4.2	12.9	3.3	**20.5**
	VH				2.8	15.7	**18.5**
	Total	**19.3**	**19.9**	**21.2**	**20.6**	**19.0**	**100.0**

Note: All percentages were calculated across all 100 replications and 1,000 simulees per replication (i.e. N=100,000). AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard.

Table 19. Percentage of Module Administration for the 1-5-5 Panel Design, Long Test Length for the $\sigma^2_{d(j)} = 0.8$ Local Item Dependency Condition.

		Route 2					
	Route 1	VE	ME	M	MH	VH	**Total**
AMI	VE	16.1	15.8				**31.9**
	ME	0.6	2.2	4.9			**7.7**
	M		3.3	9.1	9.1		**21.5**
	MH			4.3	7.1	12.0	**23.3**
	VH				1.0	14.6	**15.6**
	Total	**16.7**	**21.2**	**18.2**	**17.2**	**26.5**	**100.0**
		VE	ME	M	MH	VH	**Total**
ML-DPI	VE	12.4	7.3				**19.7**
	ME	5.8	6.1	9.6			**21.6**
	M		7.1	5.2	7.8		**20.2**
	MH			7.9	6.2	6.4	**20.4**
	VH				6.5	11.8	**18.3**
	Total	**18.2**	**20.5**	**22.7**	**20.4**	**18.2**	**100.0**
		VE	ME	M	MH	VH	**Total**
SL-DPI	VE	14.1	5.4				**19.6**
	ME	3.8	9.8	8.0			**21.5**
	M		5.5	8.9	5.9		**20.3**
	MH			6.1	10.0	4.4	**20.5**
	VH				4.8	13.4	**18.2**
	Total	**17.9**	**20.7**	**23.0**	**20.7**	**17.8**	**100.0**

Note: All percentages were calculated across all 100 replications and 1,000 simulees per replication (i.e. N=100,000). AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard.

Table 20. Percentage of Modules Administered for the 1-5-5 Panel Design, Short Test Length for the $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$ Local Item Dependency Condition.

		Route 2					
	Route 1	VE	ME	M	MH	VH	**Total**
AMI	VE	17.6	4.9				**22.5**
	ME	5.4	3.6	5.9			**14.9**
	M		5.7	7.1	6.9		**19.7**
	MH			5.4	9.3	6.4	**21.1**
	VH				3.9	18.0	**21.9**
	Total	**23.0**	**14.2**	**18.4**	**20.1**	**24.4**	**100.0**
		VE	ME	M	MH	VH	**Total**
ML-DPI	VE	11.4	6.3				**17.7**
	ME	7.7	5.6	7.9			**21.2**
	M		8.7	5.5	8.2		**22.4**
	MH			7.4	6.0	7.4	**20.8**
	VH				5.6	12.3	**17.8**
	Total	**19.1**	**20.6**	**20.8**	**19.8**	**19.7**	**100.0**
		VE	ME	M	MH	VH	**Total**
SL-DPI	VE	13.0	4.8				**17.8**
	ME	5.5	9.1	6.4			**21.1**
	M		6.8	9.4	6.2		**22.4**
	MH			5.9	9.8	5.0	**20.8**
	VH				4.1	13.8	**17.9**
	Total	**18.5**	**20.7**	**21.7**	**20.1**	**18.9**	**100.0**

Note: All percentages were calculated across all 100 replications and 1,000 simulees per replication (i.e. N=100,000). AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard.

Table 21. Percentage of Modules Administered for the 1-5-5 Panel Design, Short Test Length for the $\sigma^2_{d(j)} = 0.0$ Local Item Dependency Condition.

		Route 2					
	Route 1	VE	ME	M	MH	VH	**Total**
AMI	VE	17.4	4.2				**21.6**
	ME	5.8	3.8	5.9			**15.5**
	M		5.6	7.5	7.1		**20.2**
	MH			5.2	9.7	6.2	**21.1**
	VH				3.5	18.2	**21.7**
	Total	**23.2**	**13.6**	**18.6**	**20.2**	**24.4**	**100.0**
		VE	ME	M	MH	VH	**Total**
ML-DPI	VE	11.6	5.5				**17.1**
	ME	7.9	6.1	7.7			**21.7**
	M		8.6	5.9	8.3		**22.9**
	MH			7.3	6.2	7.1	**20.7**
	VH				5.1	12.6	**17.8**
	Total	**19.4**	**20.2**	**21.0**	**19.7**	**19.8**	**100.0**
		VE	ME	M	MH	VH	**Total**
SL-DPI	VE	13.0	4.1				**17.1**
	ME	5.6	9.9	6.2			**21.7**
	M		6.6	10.1	6.3		**23.0**
	MH			5.7	10.1	4.9	**20.6**
	VH				3.6	14.0	**17.7**
	Total	**18.6**	**20.6**	**21.9**	**20.0**	**18.9**	**100.0**

Note: All percentages were calculated across all 100 replications and 1,000 simulees per replication (i.e. N=100,000). AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard.

Table 22. Percentage of Modules Administered for the 1-5-5 Panel Design, Short Test Length for the $\sigma^2_{d(j)} = 0.8$ Local Item Dependency Condition.

		Route 2					
	Route 1	VE	ME	M	MH	VH	Total
AMI	VE	17.0	8.3				**25.3**
	ME	4.4	2.6	5.9			**12.9**
	M		6.2	5.6	6.3		**18.2**
	MH			7.4	8.0	5.7	**21.0**
	VH				5.7	16.9	**22.6**
	Total	**21.5**	**17.1**	**18.9**	**20.0**	**22.6**	**100.0**
		VE	ME	M	MH	VH	Total
ML-DPI	VE	10.8	10.6	.			**21.4**
	ME	6.2	4.2	8.4			**18.7**
	M		8.9	4.3	7.4		**20.5**
	MH			9.4	5.1	6.2	**20.7**
	VH				7.2	11.4	**18.7**
	Total	**16.9**	**23.6**	**22.1**	**19.7**	**17.6**	**100.0**
		VE	ME	M	MH	VH	Total
SL-DPI	VE	12.7	8.7				**21.4**
	ME	4.5	7.0	7.0			**18.4**
	M		7.4	7.3	5.8		**20.5**
	MH			8.1	8.4	4.6	**21.0**
	VH				5.7	13.0	**18.6**
	Total	**17.2**	**23.1**	**22.4**	**19.9**	**17.5**	**100.0**

Note: All percentages were calculated across all 100 replications and 1,000 simulees per replication (i.e. N=100,000). AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard.

Tables 23 and 24 provide the percentage of simulees in 1-3-3 panel design for the long and short test length, respectively, taking each module during administration. At stage 2 the AMI procedure routed fairly similarly to both DPI procedures except that the AMI tended to under administer the easy modules for the long test lengths. The disparity between the AMI and the DPI was more prevalent at stage 3 where both the medium and hard modules were administered at a rate of roughly 40%. The DPI procedures consistently administered approximately one-third of each module across the theta range. Again differences between the ML-DPI and the SL-DPI were seen between stages when rerouting simulees. The ML-DPI showed a much higher rate of module reroute in difficulties, where the SL-DPI simulees primarily stayed in the module to which they were originally assigned in stage 2.

Table 25 provides the percentage of simulees' module administration for the two-stage conditions. The AMI tended to administer the medium difficulty level and very hard difficulty level at a higher rate than the other modules. However, it should be noted that range across all 1-5 conditions for percentage of module difficulty administered was 15.0% to 28.0%. The DPI administered module fairly uniformly across the theta range. The DPI module difficulty percentage administered ranged from 18.7% to 21.5%

Overall, the AMI procedure tended to deliver modules less uniformly than the two DPI procedures. However, at Stage 2 for the 1-5-5 short length tests, the AMI administered each module in a fairly uniform fashion with each module being administered close to 20% for each difficulty level. As expected, the DPI administered each module about the same amount for each design. The only difference was the amount of module reroutes taking place between stages. The ML-DPI rerouted more examinees than the SL-DPI procedure.

Table 23. Percentage of Modules Administered for the 1-3-3 Panel Design, Long Test Length for the $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}, 0.0, 0.8$ Local Item Dependency Condition.

		$\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$				$\sigma^2_{d(j)} = 0.0$				$\sigma^2_{d(j)} = 0.8$			
		Route 2				Route 2				Route 2			
	Route 1	Easy	Medium	Hard	Total	Easy	Medium	Hard	Total	Easy	Medium	Hard	Total
AMI	Easy	21.0	8.5		**29.4**	21.3	8.9		**30.2**	19.9	8.3		**28.2**
	Medium	1.2	27.3	7.6	**36.1**	1.1	27.8	7.6	**36.5**	1.6	27.4	8.6	**37.6**
	Hard		1.8	32.7	**34.5**		1.4	31.9	**33.3**		1.7	32.5	**34.2**
	Total	**22.1**	**37.6**	**40.3**	**100.0**	**22.4**	**38.2**	**39.5**	**100.0**	**21.4**	**37.5**	**41.1**	**100.0**
MLDPI	Easy	20.2	12.9		**33.1**	20.3	13.1		**33.4**	19.3	15.2		**34.5**
	Medium	12.5	7.8	12.5	**32.7**	12.6	8.6	13.2	**34.3**	12.7	6.1	13.0	**31.8**
	Hard		14.7	19.4	**34.1**		12.8	19.6	**32.3**		15.7	18.1	**33.7**
	Total	**32.7**	**35.4**	**31.9**	**100.0**	**32.9**	**34.4**	**32.7**	**100.0**	**32.0**	**37.0**	**31.0**	**100.0**
SLDPI	Easy	28.8	4.4		**33.2**	29.6	3.8		**33.4**	27.7	7.0		**34.7**
	Medium	4.2	23.6	5.0	**32.8**	4.2	25.6	4.5	**34.3**	4.8	20.8	5.9	**31.6**
	Hard		3.5	30.5	**34.1**		2.6	29.7	**32.3**		5.2	28.5	**33.7**
	Total	**33.0**	**31.5**	**35.5**	**100.0**	**33.8**	**32.1**	**34.2**	**100.0**	**32.6**	**33.0**	**34.4**	**100.0**

Note: All proportions were calculated based on 100 replications and 1,000 simulees per replication. AMI=approximate maximum information; MLDPI=module-level defined population interval; SLDPI=stage-level DPI

Table 24. Percentage of Modules Administered for the 1-3-3 Panel Design, Short Test Length for the $\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}, 0.0, 0.8$ Local Item Dependency Condition.

		$\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$				$\sigma^2_{d(j)} = 0.0$				$\sigma^2_{d(j)} = 0.8$			
		Route 2				Route 2				Route 2			
	Route 1	Easy	Medium	Hard	**Total**	Easy	Medium	Hard	**Total**	Easy	Medium	Hard	**Total**
AMI	Easy	19.7	7.4		**27.1**	19.8	6.9		**26.7**	16.8	9.0		**25.8**
	Medium	5.5	19.7	12.9	**38.2**	5.6	20.5	12.5	**38.6**	7.3	18.6	14.1	**39.9**
	Hard		6.0	28.8	**34.7**		6.0	28.8	**34.7**		6.8	27.5	**34.3**
	Total	**25.2**	**33.1**	**41.7**	**100.0**	**25.4**	**33.4**	**41.2**	**100.0**	**24.1**	**34.4**	**41.6**	**100.0**
ML-DPI	Easy	17.2	15.8		**33.0**	17.3	15.3	.	**32.6**	16.1	18.5		**34.6**
	Medium	15.1	6.3	13.7	**35.0**	15.4	7.0	13.2	**35.6**	15.2	4.9	12.6	**32.7**
	Hard		14.0	18.1	**32.1**	.	14.0	18.1	**32.1**		16.1	16.7	**32.8**
	Total	**32.2**	**36.0**	**31.8**		**32.7**	**36.2**	**31.3**	**100.0**	**31.3**	**39.5**	**29.3**	**100.0**
SL-DPI	Easy	25.7	7.0		**32.67**	26.0	6.5		**32.5**	25.1	9.6		**34.7**
	Medium	7.9	20.4	6.8	**35.18**	7.7	21.6	6.4	**35.6**	9.2	17.3	6.8	**33.2**
	Hard		5.1	27.0	**32.17**		4.8	27.1	**31.9**		6.8	25.3	**32.1**
	Total	**33.6**	**32.6**	**33.9**	**100.0**	**33.7**	**32.9**	**33.5**	**100.0**	**34.2**	**33.7**	**32.1**	**100.0**

Note: All proportions were calculated based on 100 replications and 1,000 simulees per replication. AMI=approximate maximum information; ML-DPI=module-level defined population interval; SL-DPI=stage-level DPI

Table 25. Percentage of Modules Administered for the 1-5 and 1-3 Panel Designs

1-5 Panel Design Long Test Length															
	$\sigma^2_{d(j)} = \hat{\sigma}^2_{d(j)}$					$\sigma^2_{d(j)} = 0.0$					$\sigma^2_{d(j)} = 0.8$				
Route	VE	ME	M	MH	VH	VE	ME	M	MH	VH	VE	ME	M	MH	VH
AMI	17.8	16.2	22.9	15.0	28.0	17.9	16.3	23.2	15.3	27.4	17.6	16.4	23.4	14.8	27.9
DPI	19.2	20.3	21.1	20.5	18.8	19.2	20.5	21.4	20.1	18.8	18.8	20.8	21.5	19.3	19.5
1-5 Panel Design Short Test Length															
AMI	17.9	16.1	23.1	16.5	26.5	17.6	16.2	23.4	16.7	26.2	18.2	16.9	22.9	15.4	26.7
DPI	18.9	20.6	21.3	20.5	18.7	18.8	20.8	21.5	20.4	18.6	19.4	21.2	20.8	19.1	19.5
		1-3 Panel Design Long Test Length													
		E	M	H			E	M	H			E	M	H	
AMI		23.0	43.9	33.1			23.3	43.8	32.8			21.7	44.5	33.8	
DPI		33.7	32.8	33.6			33.6	33.3	33.1			33.5	33.3	33.1	
1-3 Panel Design Short Test Length															
AMI		24.1	43.3	32.6			24.1	43.4	32.6			22.6	43.2	34.1	
DPI		33.6	33.0	33.4			33.6	33.4	33.0			33.7	32.8	33.4	

Note: All proportions were calculated based on 100 replications and 1,000 simulees per replication. AMI=approximate maximum information; DPI=defined population interval; VE=very easy; ME=medium easy; M=Medium; MH=medium hard; VH=very hard; E=easy; H=hard.

Chapter 5: Discussion

This chapter discusses the study's results. The study investigated the operation characteristics of MSTs with a mixed-format testlet based item pool under the three-parameter logistic testlet response theory (3PL-TRT) model. Included in the discussion are three main sections. First the research questions are addressed based on the results of the study. Then practical implications from the findings are described. Finally, limitations and future directions of research are discussed.

RESEARCH QUESTIONS

To fully explicate the findings with respect to the research questions, a review of the panel assembly outcomes needs to be discussed. The panel assembly does influence the outcomes from the dependent measure as the amount of information provided to an examinee is dictated by the items one receives. Unlike fully adaptive testing, multistage test (MST) panel assembly occurs prior to administration. A solid grasp on the approximate information for an examinee at various ability levels for a given testing route can then be approximated prior to administration through the panel assembly process.

Although using automated test assembly (ATA) for panel assembly was successful in providing parallel panels for each panel design, the panel designs for the 1-5-5 MSTs were only able to assemble two panels rather than the desired three. The remaining panel designs were all able to create three panels as desired for the mixed-format testlet-based item pool. These results were a function of the number of constraints placed in the ATA algorithm and the amount of test units in the item pool. The number of modules in 1-5-5 was greater than the remainder of the panel designs, creating

increased constraints on the number of test units used during assembly. The 1-5-5 panel designs were then too demanding on the ATA algorithm and only two of three panels could be created.

One of the main goals of the ATA program was to build modules that provided uniform information across a targeted ability range. Due to the additive nature of item information, testlet based information will generally contain more information across a broader range of the ability-level-spectrum (Murphy et al., 2010). This was supported by the evaluation of the panel assembly. When examining the relative target TIFs illustrated in Figure 9, Figure 10, and Figure 15, the information for a broad range of abilities provides very similar information across the target theta range, for this study the 1-5-5 and 1-5 panel designs targeted $\theta_k = (-1.0, -0.5, 0.0, 0.5, 1.0)$ and the 1-3-3 and 1-3 panel designs targeted $\theta_k = (-1.0, 0.0, 1.0)$. Across panel designs, similar information was able to be achieved across a broad range of abilities at each stage. The testlet based item pool not only supported the building of modules with uniform peaked TIFs, it also illustrated very minimal depreciation in information between adjacent targeted ability ranges.

How does panel design impact the measurement precision of an MST with a mixed-format testlet-based item pool?

The panel design conditions were chosen because the panel design is one of the first choices a testing program will have to make during test development. These conditions were assembled with respect to the original large-scale standardized assessment used to create the mixed-format testlet-based item pool.

The four panel designs performed very similarly with respect to the measurement accuracy under the 3PL-TRT model. Minimal differences were seen between estimated thetas and bias. The overall bias measures were functionally zero but consistently

exhibited a small positive bias. This result was present across all the manipulated variables. Minimal differences in the measurement precisions between the four panel designs with equal number of stages were found. The correlations between known and estimated theta, root mean squared error (RMSE), and average absolute deviation (AAD) for panel designs with equal number of stages were all very similar. However, a slight increase in measurement precision was observed from the reduction in RMSE and AAD, and increase in correlation coefficients when moving from a three-stage test design to a two-stage test design.

Previous research has found that routing stage information has impacted the precision of an MST administration (Galindo et al., 2013; Kim & Plake, 1993; Kim, 2010; Zenisky, 2004). The results seen in this study regarding the correlation, RMSE, and AAD may be an indication of the amount of information found in the routing stages. The panel designs were assembled to have equal total number of items administered, with approximately equal number of items administered at each stage. For instance, the long test length three-stage MST had between 18 and 19 items at each stage, and the long two-stage test had 27 or 28 items at each. The information was then increased at the routing stage for the two-stage test, which likely reduced the routing error that occurred between stages. As a result the correlations were increased and the RMSE and AAD decreased for two-stage tests.

When examining the bias conditional on abilities under the 3PL-TRT model, little bias was found across the bulk of the distribution. Towards the extremes of the distribution similar amounts of bias were detected across all panel designs. Additionally, minimal differences in the pattern of measurement error for simulees were detected across panel design for the range of abilities.

How does test length impact the measurement precision of an MST with a mixed-format testlet-based item pool?

One of the benefits to an adaptive test is providing similar precision for an ability estimate while being able to decrease the number of items administered when compared to a fixed-length test. A goal of the study was to push the boundaries of test length, while maintaining the overall test design with respect to the assessment which provided item parameters.

As expected, there were differences in the measurement precision between the longer (55 items) and shorter test length (44 items) conditions. The correlations were all reduced, while the SEs, RMSE, and AAD all increased for the shorter test length. When comparing the bias minimal differences in measurement accuracy were found between the long and short test lengths.

When comparing the test lengths across the range of ability distributions, the conditional bias was very similar for both test lengths. The conditional SEs were also similar for both test lengths with a slight increase in SEs for the shorter test lengths. Overall, there did not appear to be a substantial decrease in the measurement accuracy or precision for the shorter length tests.

How do various routing procedures impact the ability estimation of an MST with a mixed-format testlet-based item pool?

Three routing procedures were examined in the study, namely the approximate maximum information (AMI), module-level defined population interval (ML-DPI), and stage-level defined population interval (SL-DPI). Routing procedures are designed to dictate the proportion of module administration to examinees. Only minor differences in ability estimation was found between the three routing procedures for all the dependent

measures. Previous studies have shown that AMI procedure tends to be the routing procedure with better precision (see Kim et al.2013; Zenisky, 2004), however the current study found only negligible differences between the procedures.

The minimal differences found between the routing procedures may be partially explained by the nature of the mixed-format testlet-based item pool and the panel assembly process. Testlet-based items tend to provide more information across a broad range of abilities. The overlap of the information curves between the targeted theta's was very minimal across each of the panel designs. The information overlap may have contributed to all routing procedures having similar amounts of information for a wide range of abilities even when a module was not optimal with respect to the overall module information.

The main differences found between routing procedures primarily occurred with respect to module administration. As expected the AMI, tended to administer modules less uniformly than did either DPI procedure. Both DPI procedures administered modules that were approximately equally proportioned for the respective panel designs. For the case of the 1-5-5 and the 1-5 panel designs, the DPIs both administered approximately 20% of each module at Stage 2 and Stage 3. The 1-3-3 and 1-3 panel designs administered approximately 33% of modules at Stage 2 and Stage 3. The differences found between the two DPI procedures had more to do with the amount of reroutes produced during the administration process. As expected more simulees were rerouted to adjacent modules at Stage 3 for the ML-DPI, while the majority of simulees remained in the same module difficulty level from Stage 2 to Stage 3 for the SL-DPI. It should also be noted that some of the 1-5-5 and 1-5 AMI procedure were not dramatically different in proportion of modules administered to the DPI. Although it was less uniform across the

modules as the two DPI procedures, some instances only ranged between 15-25% of module administrations across the five modules. This was likely a result of the proximity of the targeted thetas and width of information for each module due to the nature of the item pool.

How does the magnitude of local item dependence (LID) effect the administration and ability estimation of an MST with a mixed-format testlet-based item pool?

Three LID conditions were used to generate item responses for the mixed-format testlet-based item pool. The three conditions represented no testlet effect, a constant large testlet effect, and estimated testlet effects from the item pool, or $\sigma^2_{d(j)} = (0.0, 0.8\hat{\sigma}^2_{d(j)})$. Investigating LID conditions was important as testlet-based items are common practice in standardize assessments.

The results of the current study suggest that, in the presence of a large testlet effect yield less measurement precision than an item pool with no testlet effect, or a small overall testlet effect. For the largest LID condition, $\sigma^2_{d(j)} = 0.8$, minimal overall bias was detected while the SEs, RMSEs, and AADs were all consistently larger, and the correlations were all consistently smaller than the other two LID conditions $\sigma^2_{d(j)} = 0.0, \hat{\sigma}^2_{d(j)}$.

Further exploration into the conditional bias and conditional SE plots showed that the bias was largest in the extremes of the ability distribution for all three LID conditions with $\sigma^2_{d(j)} = 0.8$ exhibiting the largest amount of bias. Overall, minimal bias was detected across the majority of the ability distribution for all LID conditions. The grand mean SE for the $\sigma^2_{d(j)} = 0.8$ LID condition was consistently largest. Interestingly, when examining the conditional SE plots, the SEs for $\sigma^2_{d(j)} = 0.8$ condition tended to be slightly smaller in the extremes of the distribution while being slightly larger towards the middle of the

distribution than the other two LID conditions. These results support previous research (Murphy et al., 2010; Sireci et al., 1991; Wainer, Bradlow, & Wang, 2007; Wainer & Thissen, 1996; Yen, 1993) that the discrimination may be inflated when larger amount of conditional independence is present leading to underestimated SE, especially in the portions of the distribution that are measured with less precision. However, the 3PL-TRT model performed fairly well with adequate amounts of measurement precision for under a variety of LID conditions.

How do panel design, test length, routing procedures, and LID, interact with respect to the accuracy and precision of ability estimation?

The study also sought to find any possible interactions that may be present between the different possible test designs. All the test designs appeared very similar with respect to the dependent measures. There was a slight improvement in measurement precision for the SL-DPI routing procedure when moving from the 1-5-5 panel design to the 1-3-3 panel design when no testlet effect was present. However, it was very slight and may provide little practical advantage to other conditions investigated in the study.

PRACTICAL IMPLICATIONS

Computer based testing is very prevalent in today's assessment environment. Although MSTs have been in the psychometric literature for many years, (Lord, 1971a), they have recently become more prevalent in large-scale standardized test settings. Therefore, it behooves the psychometric community to fully investigate such testing environments to understand the implications associated with implementing an MST. Additionally, many programs are implementing assessments with mixed-format testlet-based item pools. These item pools most likely exhibit some form of LID, whether it be a

relatively small amount, such as the real item pool, or a large amount of testlet dependencies, as were investigated in this study. Although some literature has addressed mixed-format MSTs, very few studies have actually examined the use of the 3PL-TRT model with parameters from a real dataset. All the manipulated conditions in this study, such as panel design, test length, routing procedures, and LID are all potential decision points that would arise during the test development process for an assessment. Therefore, the findings from this study contribute to the knowledge regarding the practical guidelines for programs considering MSTs and expand the current literature for programs considering mixed-format testlet-based item pools and the use of the 3PL-TRT model under realistic testing conditions.

First, this study demonstrates the use of 3PL-TRT model as an appropriate model to handle mixed-format testlet-based MSTs. Only one study has investigated a mixed-format testlet-based item pool assembling MSTs from an item pool that was completely simulated (Lu, 2010). This study expanded on previous research in two regards. First, the item pool was constructed from an existing test. Secondly, the MST was assembled to have mixed-format modules. The 3PL-TRT model effectively handled an MST that administers modules consisting of stand-alone items and testlet-based items. As supported by previous research, the 3PL-TRT appropriately handles testlet-based dependencies for varying magnitudes of a testlet effect. Minimal bias was detected. Although an increase in LID slightly depreciates the measurement precision, the 3PL-TRT model still provides adequate measurement accuracy and precision for the conditions investigated in this study.

Although the focus of the study is not ATA, the study also highlights some of the advantages and disadvantages of using an ATA branch and bound solver. The ATA

system was able to construct multiple panels with adequate amounts of information for each module to provide the desired levels of measurement precision for each of the conditions. The ATA found solutions for all panel designs and constructed uniformly peaked target TIFs for all but one of the panel designs. Given the composition of the test units in the item pool, only two panels for the 1-5-5 panel designs were able to be constructed. A testing company would likely need a larger test unit pool than the one used in this study should they wish to implement a larger panel design such as the 1-5-5. Overall, the ATA was highly successful and an efficient way to form the panels used in the study.

The study also demonstrated that the mixed-format testlet-based modules can provide a set of uniformly peaked modules across the targeted theta range. Because testlet-based target TIFs overlapped overs a broad range of abilities, the mixed-format administration provided fairly adequate measurement accuracy and precision for a majority of the simulee distribution. Using mixed-format item types increases the number of item types that can be administered while maintaining a high level of measurement precision.

Clearly test length influences the measurement precision of an assessment. A longer test has more measurement precision than a shorter test. As expected, the current study supported this notion. Even though the measures for the short test lengths were considered less precise, both test lengths yielded high levels of precision and could be considered viable for use in a testing program.

Under similar conditions and item pools very little difference was found between overall measurement properties of the three routing procedures. So the use of a particular routing procedure depends on the desires of the testing program. If the testing program is

interested in providing the most precise measurement, then the AMI should be used. If exposure control is an interest than one of the DPI procedures should be used. With respect to the two DPI procedures, the main difference was not a difference in psychometric properties but rather the number of routing decision points needed to implement each DPI procedure, with the ML-DPI producing more distinct decision points.

Additionally, little difference was found between panel designs. This may suggest that the 1-3 or 1-3-3 may be the most useful panel design in terms of measurement properties. This conclusion is reached not because the measurement properties are better, but the assembly process and size of the item pool needed would be smaller. Testing organizations must also consider available resources when constructing an MST design. Item pool development can be an expensive process, which might hinder the use of a 1-5-5 panel design. If, however, module administration is a primary concern then the 1-5 and 1-5-5 panel designs may be of interest as they would help in reducing module exposure.

LIMITATIONS AND FUTURE RESEARCH

The findings in this study address the posed research questions. Simulation studies, by nature, can only generalize to the conditions investigated. As with any study, there are limitations that need to be addressed. Because investigating all possible scenarios in a simulation study quickly becomes unwieldy, the current section provides areas in which the future studies could address limitations present in the current study.

The study assembled panels to have equal items at each stage. This created a situation where the two-stage MSTs had considerably more items at each stage than the

three-stage items. This means more information was provided at a given stage for the the 1-3 and 1-5 panel designs. Specifically, more information was provided during the routing stage. Based on previous research, it can be hypothesized that the increased information at the routing stage increased the routing classification accuracy. Future studies should assemble MSTs with varying amounts of information at each stage in the mixed-format contexts. Controlling how much information occurs at each stage and comparing the results could help inform panel assembly guidelines.

The test lengths were determined in a fashion that maintained proportionality to item pool's administered versions of the originating test. Therefore, certain proportions of test units, i.e. standalone items and testlet-based items, were used to guide the assembly process. As such, only limited reductions in test lengths were permitted. This occurred Based on the proportions of test units within the mixed-format testlet-based item pool, because as test length reduction occurs eventually standalone items would be the only test unit utilized. Future studies could further reduce test lengths of an MST administering mixed-format MST and mixed-format modules by allowing a looser interpretation, or any combination of mixed-format assembly of test units could administered for a pool similar to the one used in the current study. This would provide further knowledge about the minimum test length requirement for mixed-format testlet-based MSTs.

The simulee responses were all generated with a normal distribution. Normality is a typical assumption made in testing programs, but does not reflect all possible realistic examinee distributions. It is a common occurrence for a normal distribution to represent the range of abilities at the beginning of the testing program, to then become negatively skewed over time. Previous studies using a purely testlet-based item pool under the 3PL-

TRT model have demonstrated that departures from normality can lead to over-exposure of certain modules (Keng, 2008). In addition, the two main types of routing procedures used for MSTs are the AMI and DPI. The DPI functions correctly well when the trait levels are normally distributed. T he impact on module administration would likely be influenced by the underlying distribution of examinees and should be investigated to understand the potential impacts on administration and security of the items. The use of skewed distribution should be compared under the 3PL-TRT model and with various routing procedures in future studies with a mixed-format testlet-based item pool.

Finally, it was noticed that some of the 1-5 and 1-5-5 panel designs administered modules under the AMI routing procedure very similarly to the DPI procedures. This is in part due to the target TIF overlap across the range of abilities. The ability range of the target TIF is also a function of the targeted thetas. The study only ranged the targeted thetas from -1 to 1 over equal increments. As a result, minimal differences were found in measurement precision when increasing the number of modules. Relatedly, this study also found minimal differences between the measurement precision of routing procedures. Future studies could investigate the range of targeted thetas for the relative target TIF functions. This might help clarify when and if using a 1-5-5 or 1-5 panel design might provide increased measurement precision over the more typical 1-3-3 panel design, and how targeted theta's impact the measurement precision when using various routing procedures for a mixed-format testlet-based item pool. .

Appendices

APPENDIX A: PANEL DESIGN RELATIVE TARGET TEST INFORMATION FUNCTIONS

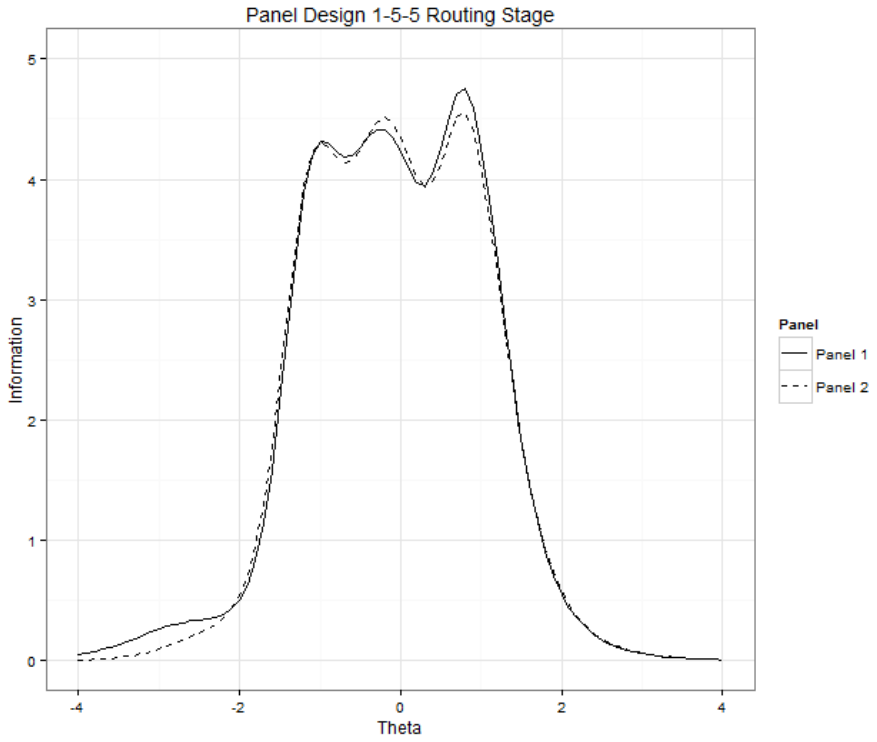

Figure A.1. Stage 1 routing module relative target TIF plots for the long 1-5-5 panel design when $\gamma_{(d)} = 0$.

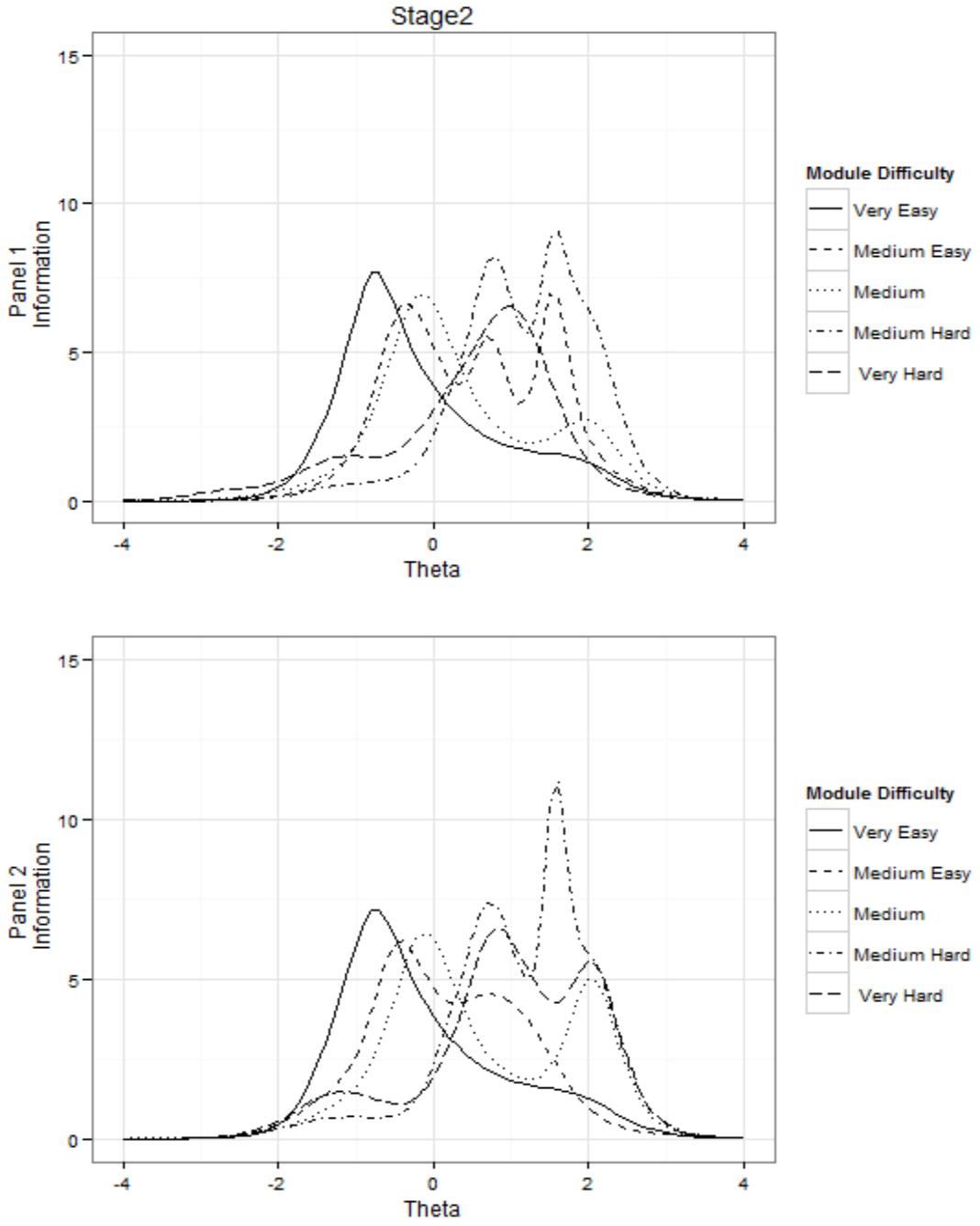

Figure A.2: Stage 2 relative target TIFs for the long 1-5-5 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

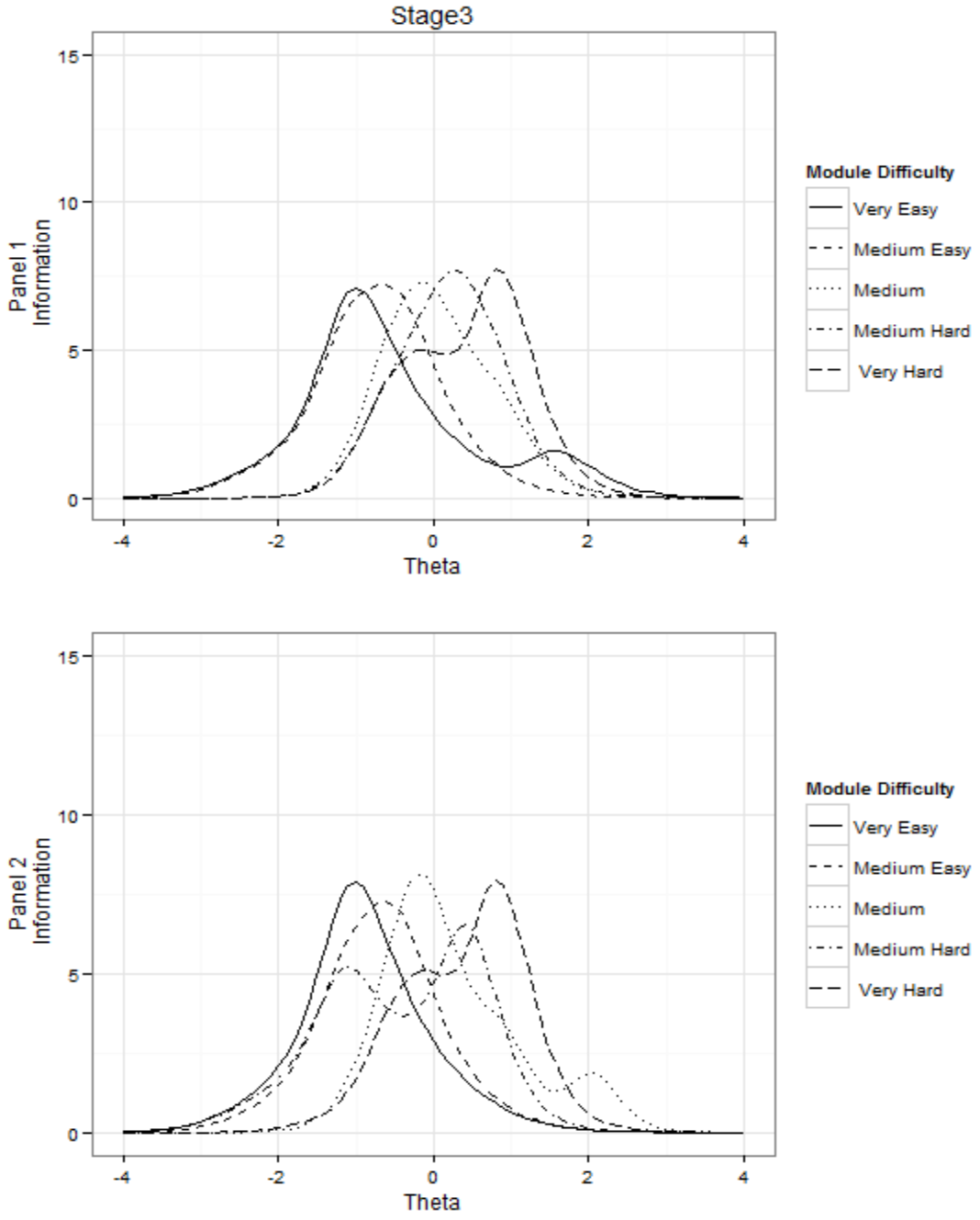

Figure A.3. Stage 2 relative target TIFs for the long 1-5-5 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

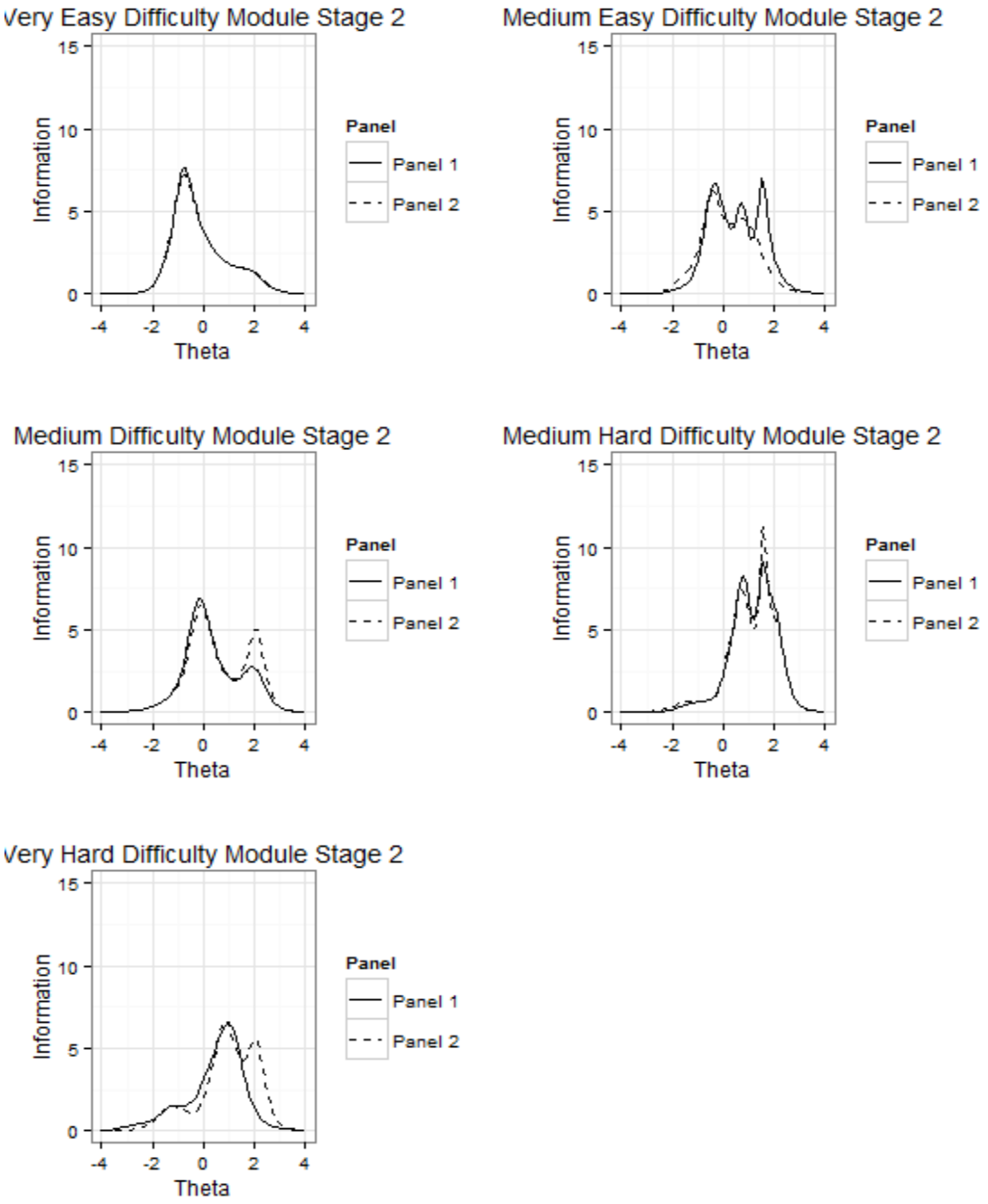

Figure A.4. Stage 2 relative target TIFs for the long 1-5-5 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

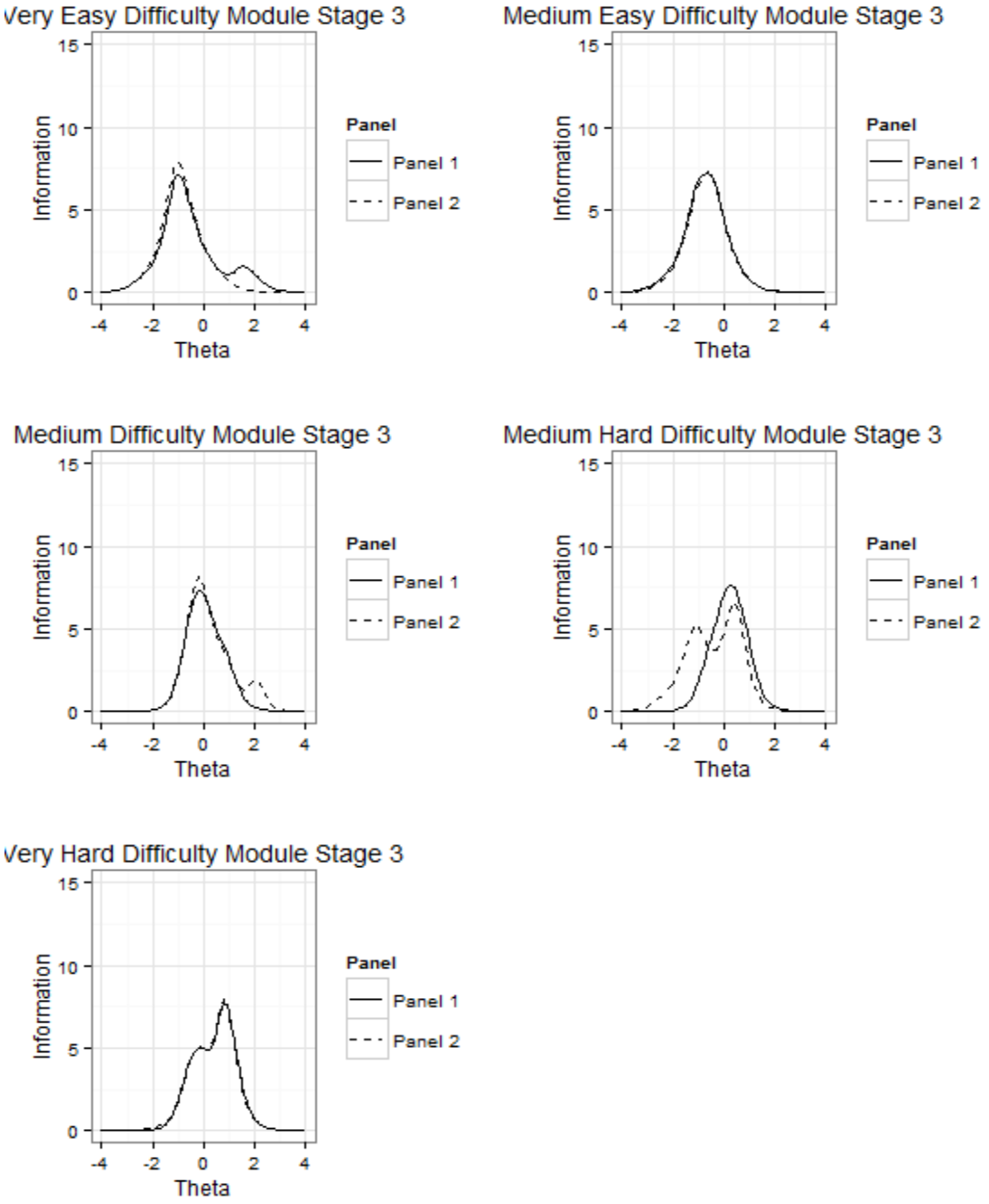

Figure A.5. Stage 3 relative target TIFs for the long 1-5-5 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

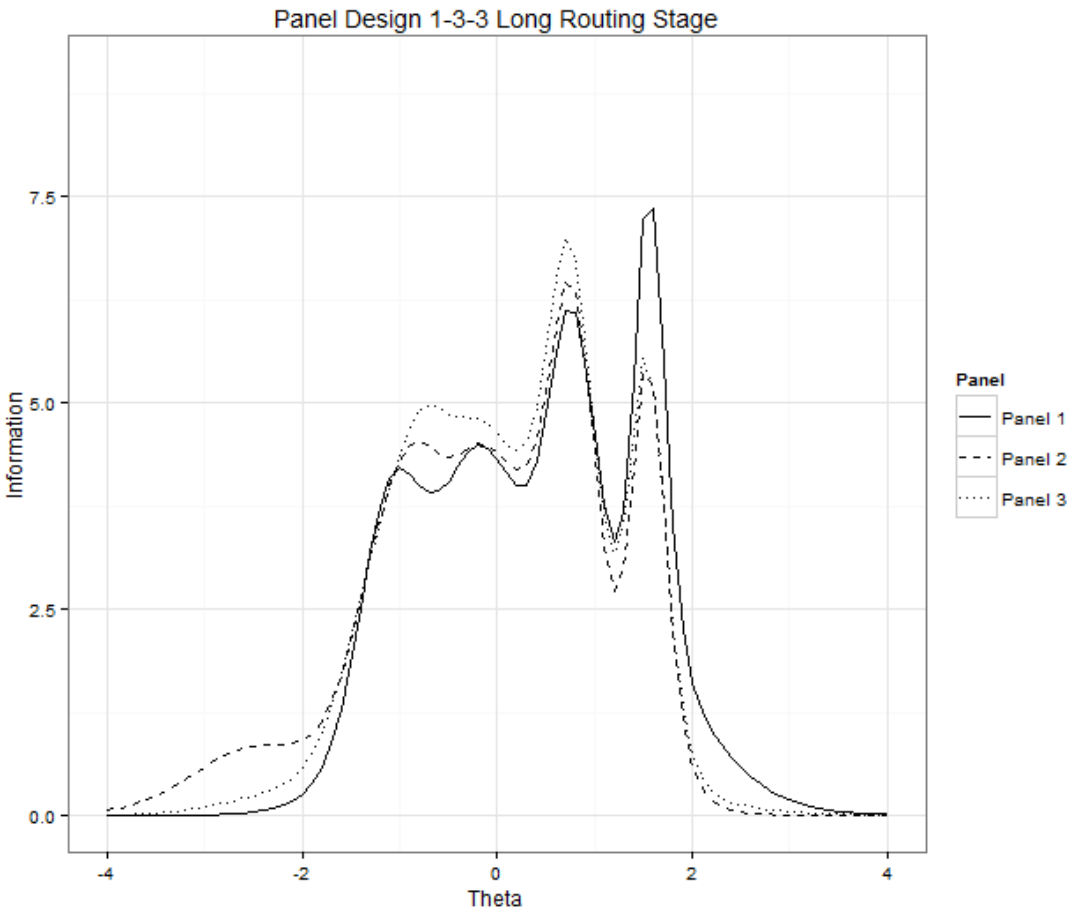

Figure A.6. Stage 1 routing module relative target TIF plots for the long 1-3-3 panel design when $\gamma_{(d)} = 0$.

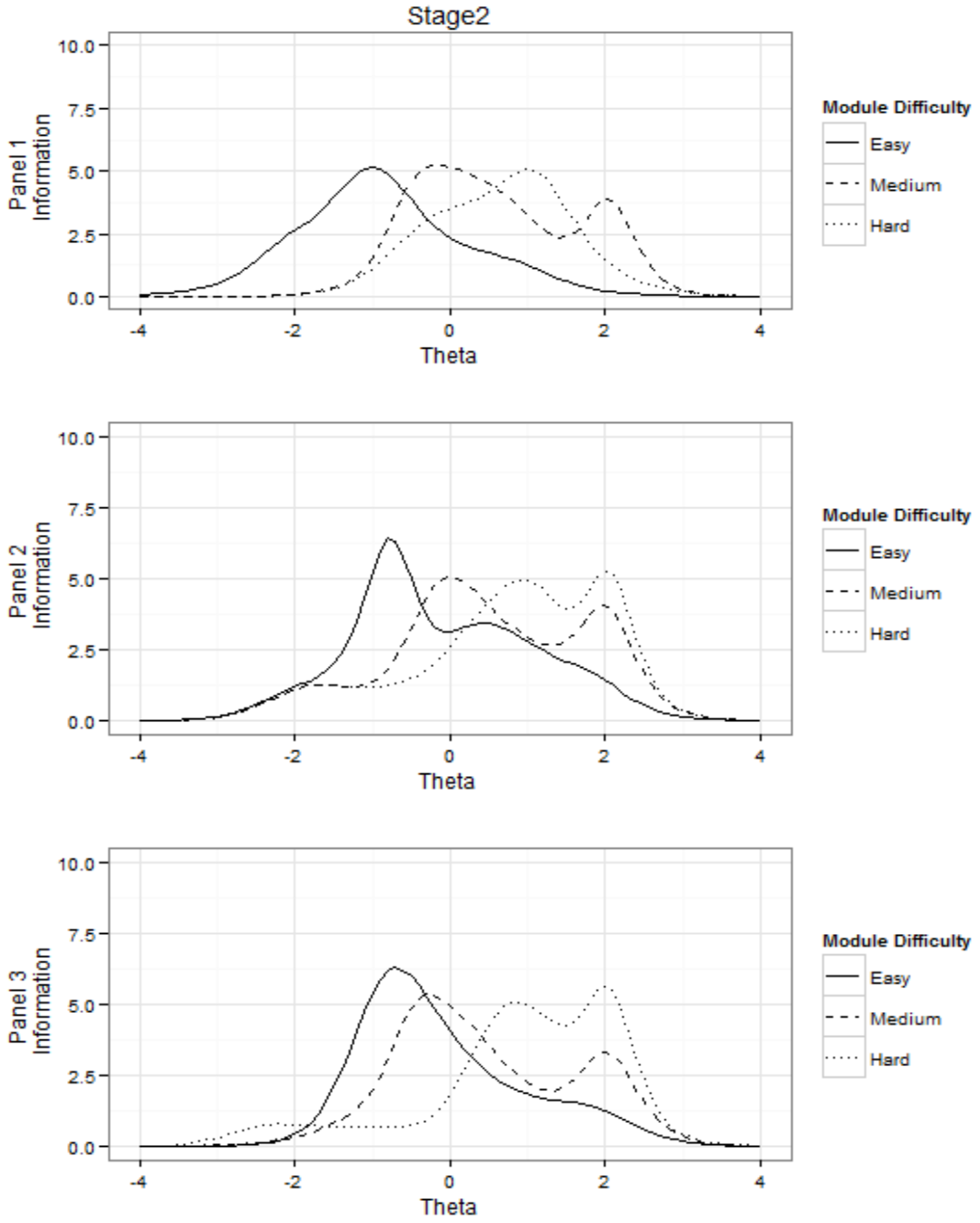

Figure A.7. Stage 2 relative target TIFs for the long 1-3-3 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

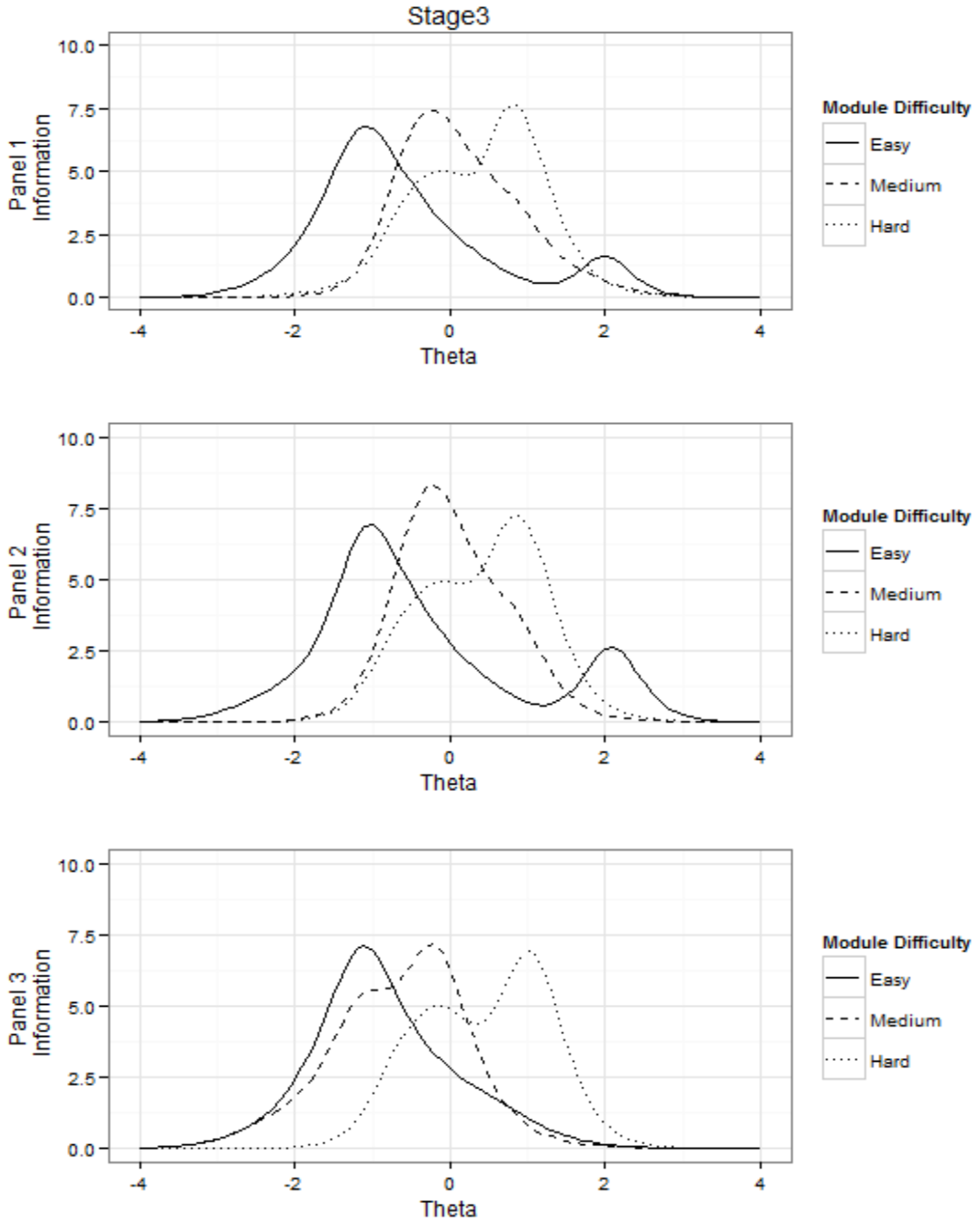

Figure A.8. Stage 3relative target TIFs for the long 1-3-3 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

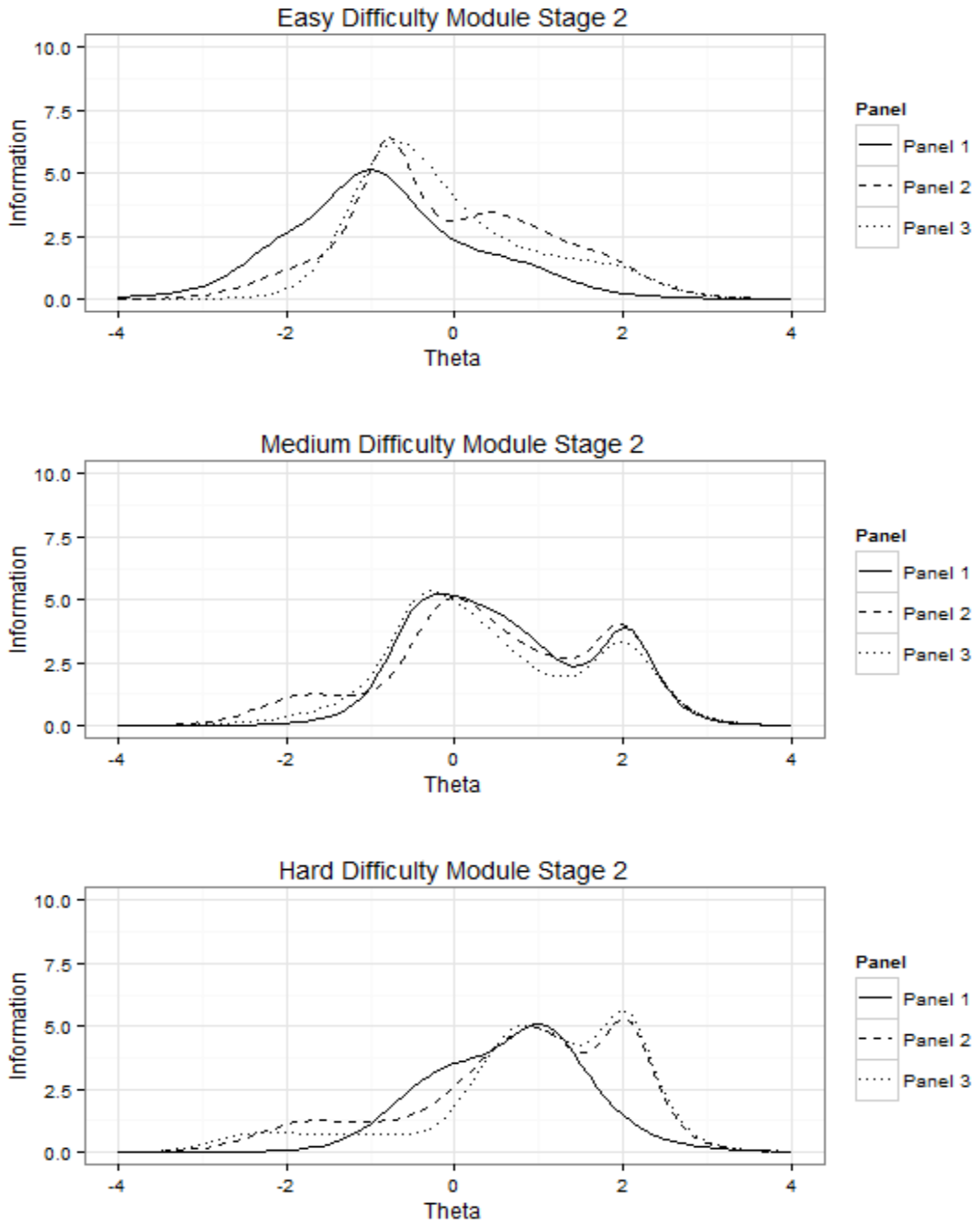

Figure A.9. Stage 2 relative target TIFs for the long 1-3-3 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

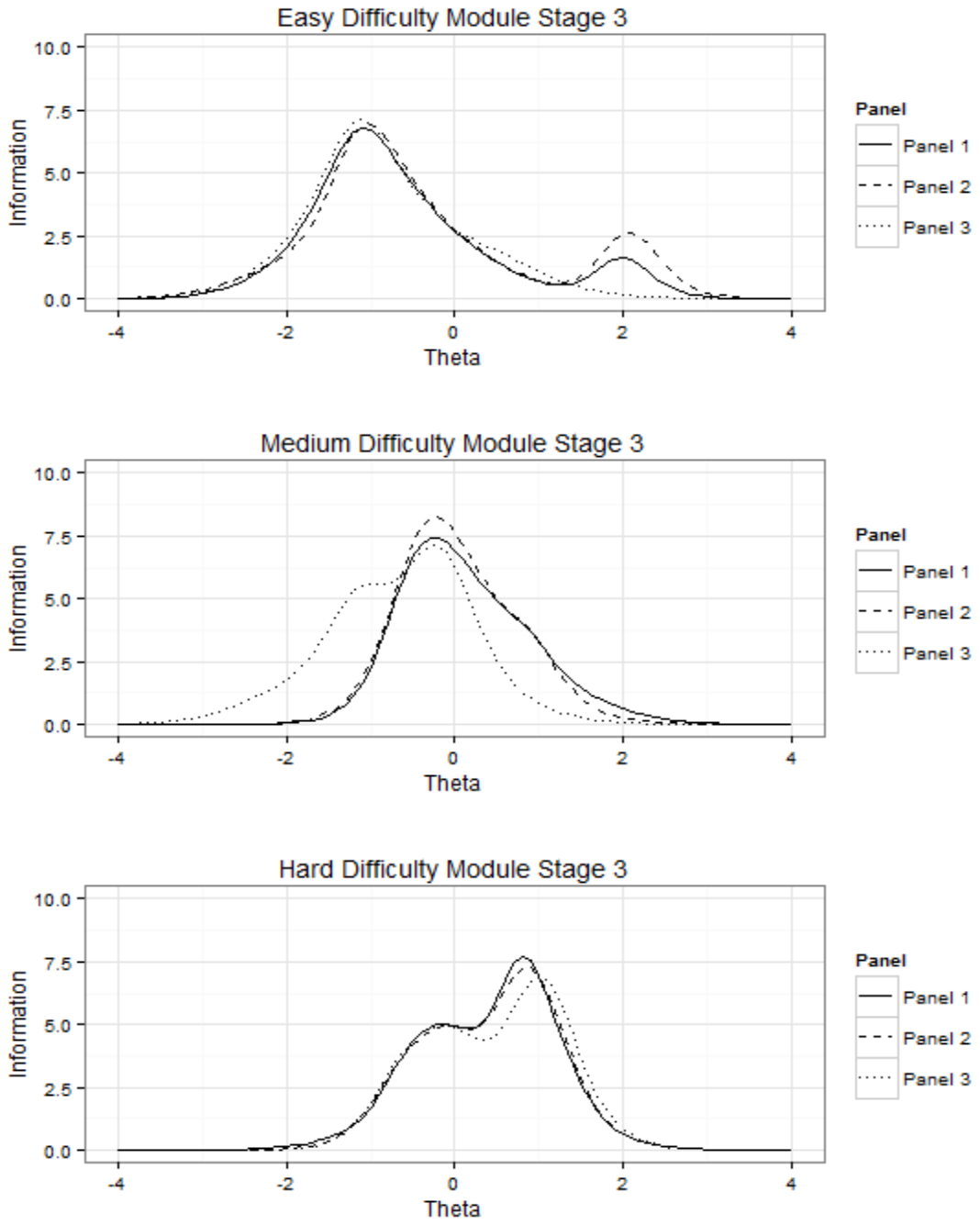

Figure A.10. Stage 3 relative target TIFs for the long 1-3-3 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

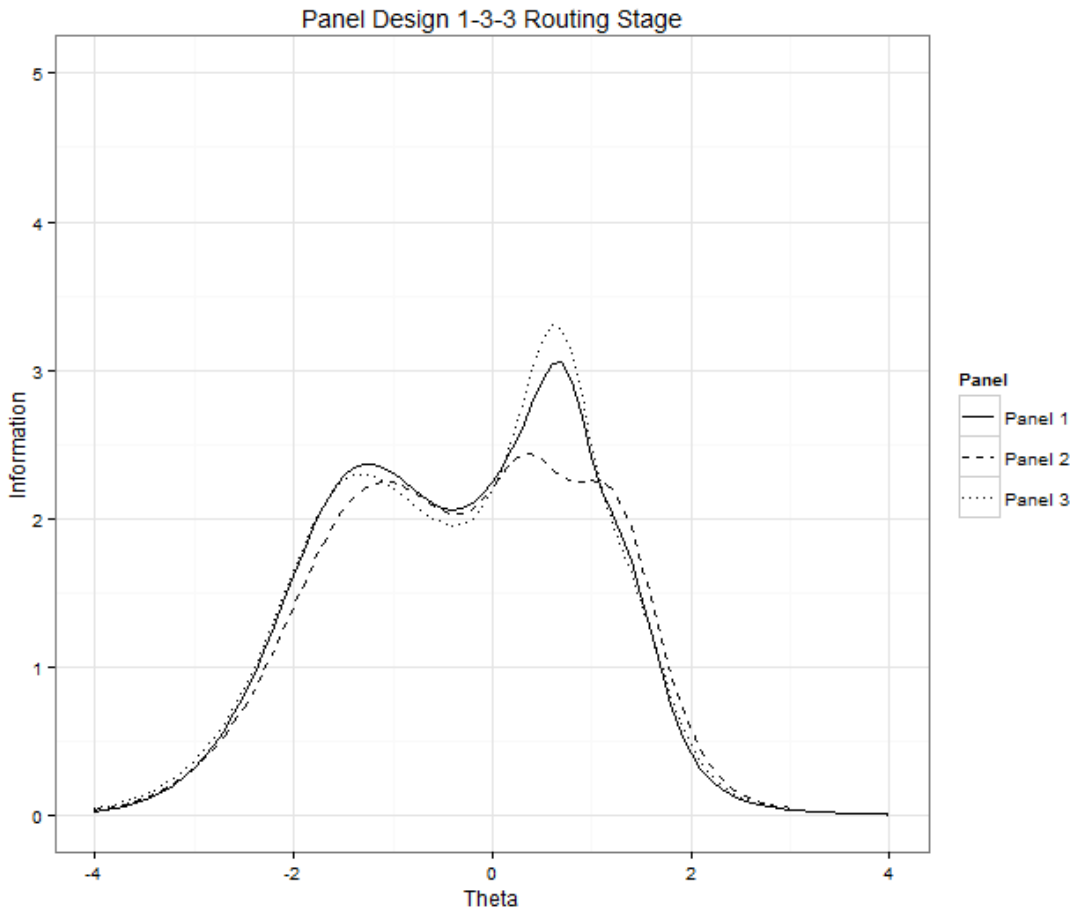

Figure A.11. Stage 1 routing module relative target TIF plots for the short 1-3-3 panel design when $\gamma_{(d)} = 0$.

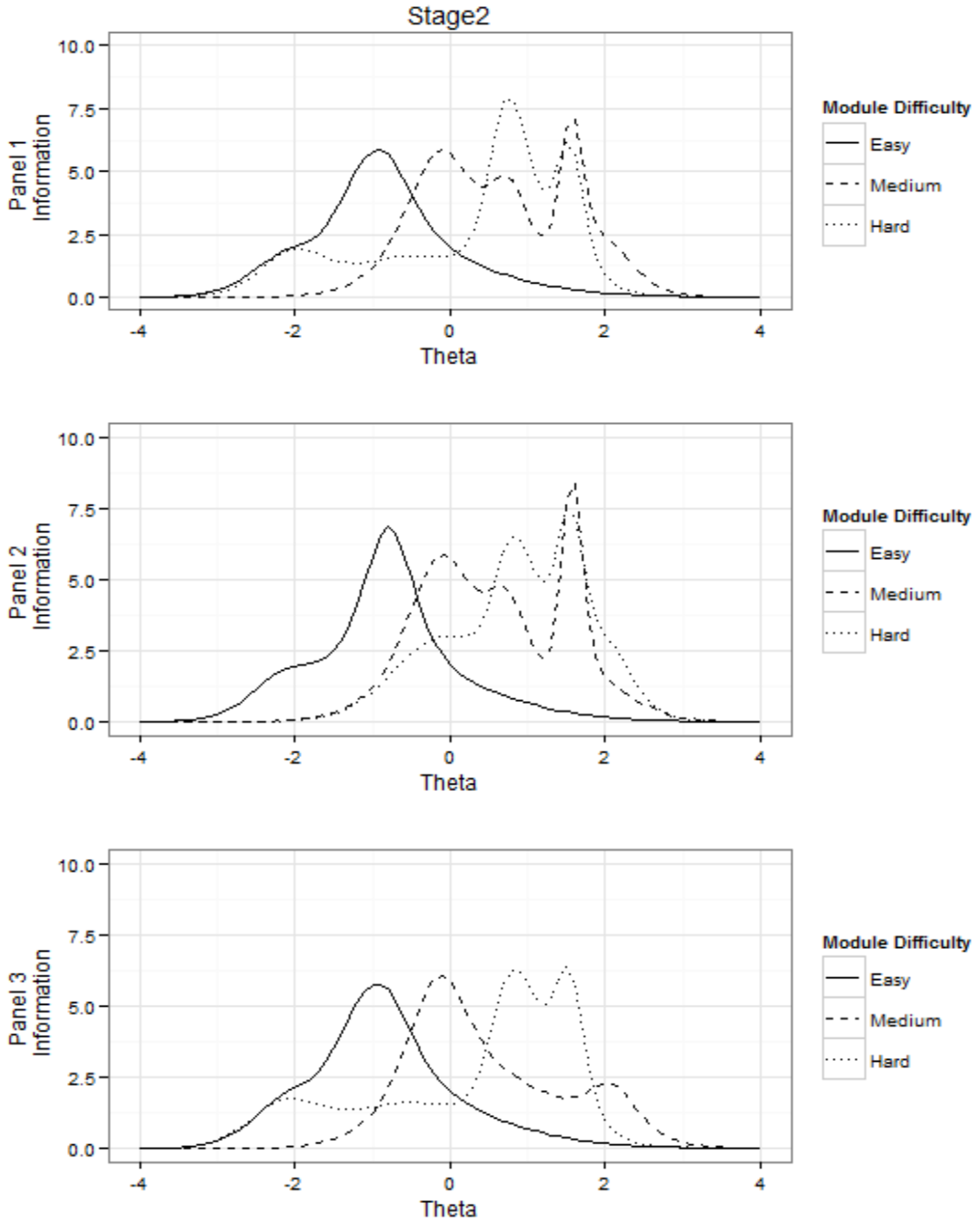

Figure A.12. Stage 2 relative target TIFs for the short 1-3-3 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

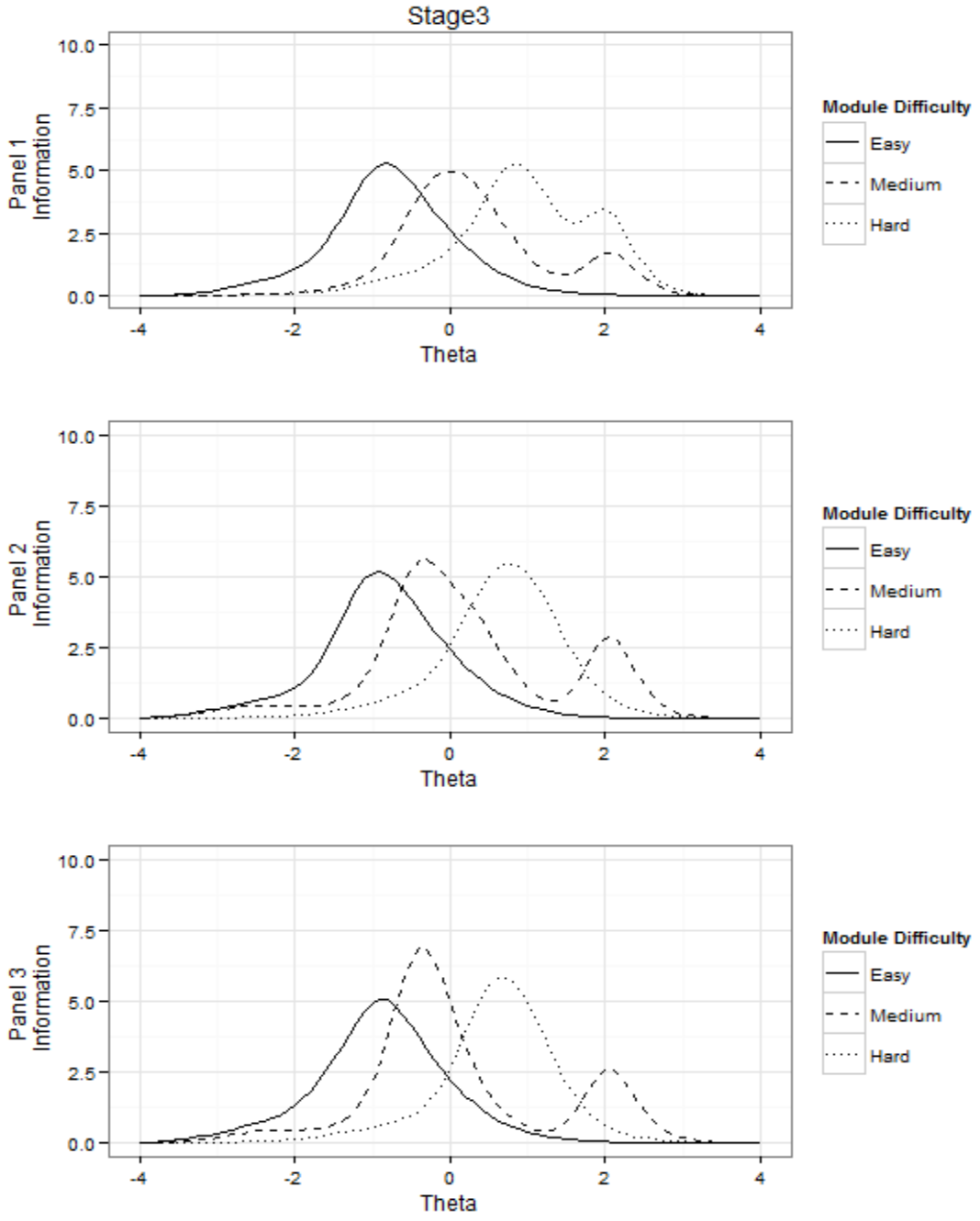

Figure A.13. Stage 3 relative target TIFs for the short 1-3-3 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

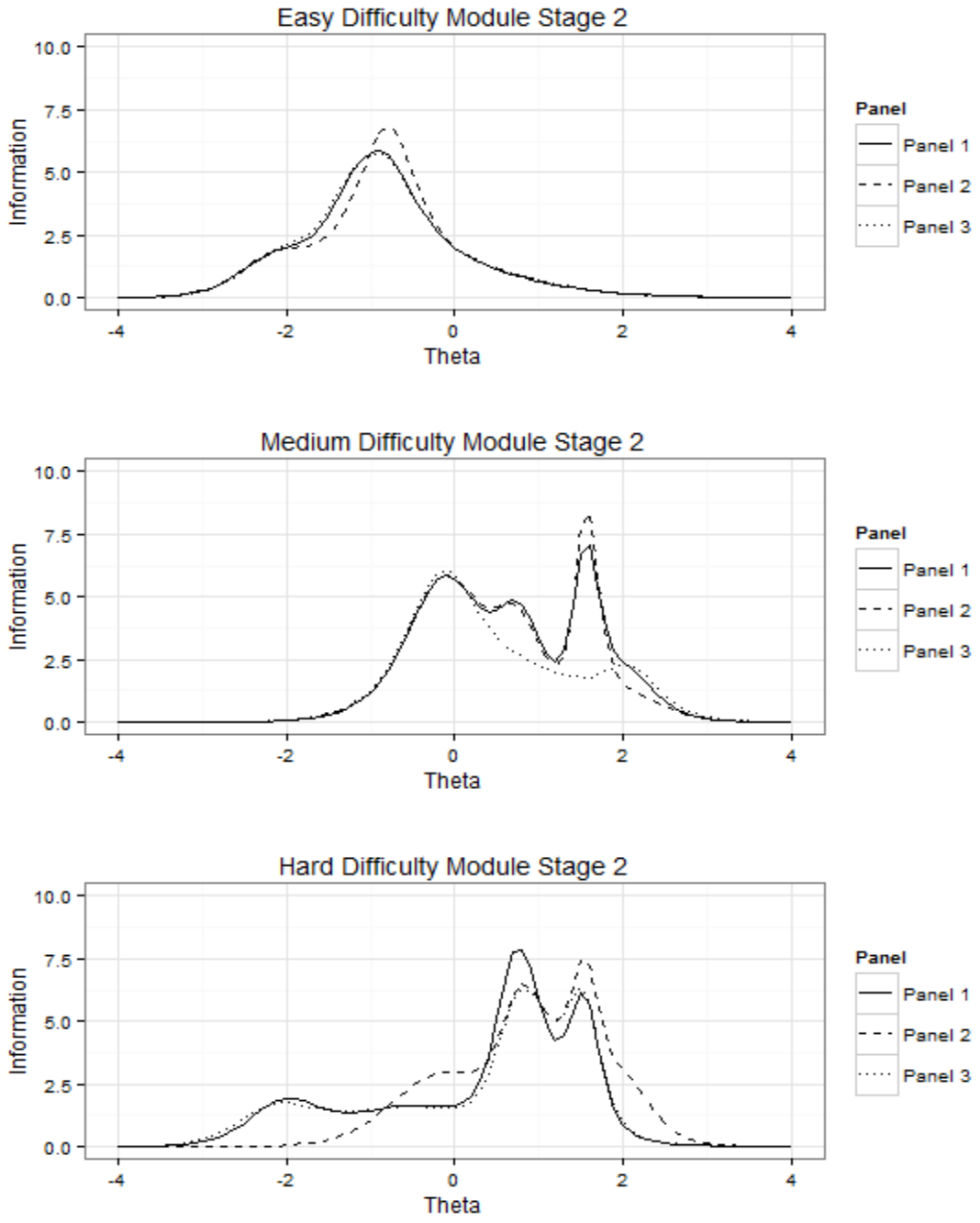

Figure A.14. Stage 2 relative target TIFs for the short 1-3-3 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

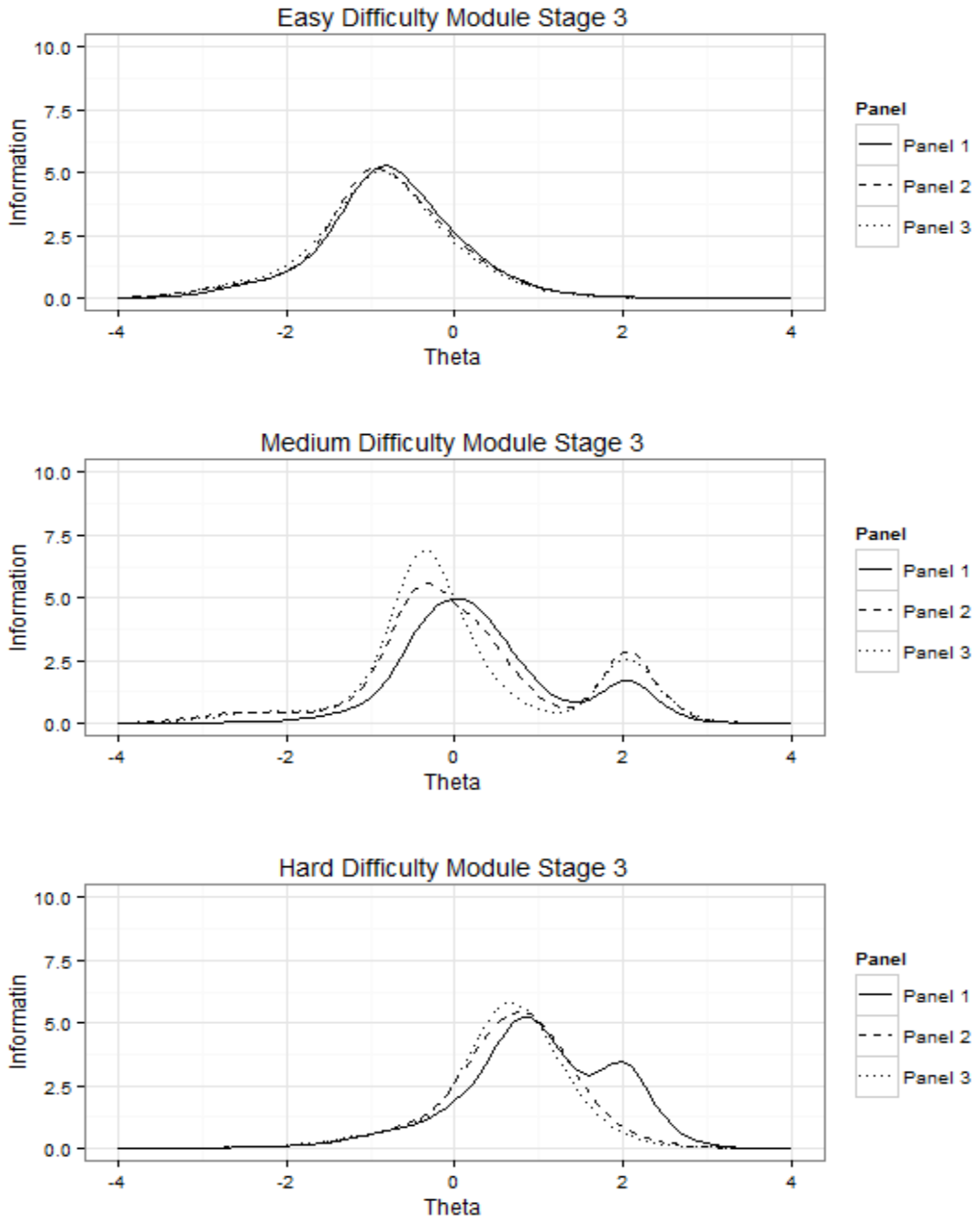

Figure A.15. Stage 3 relative target TIFs for the short 1-3-3 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

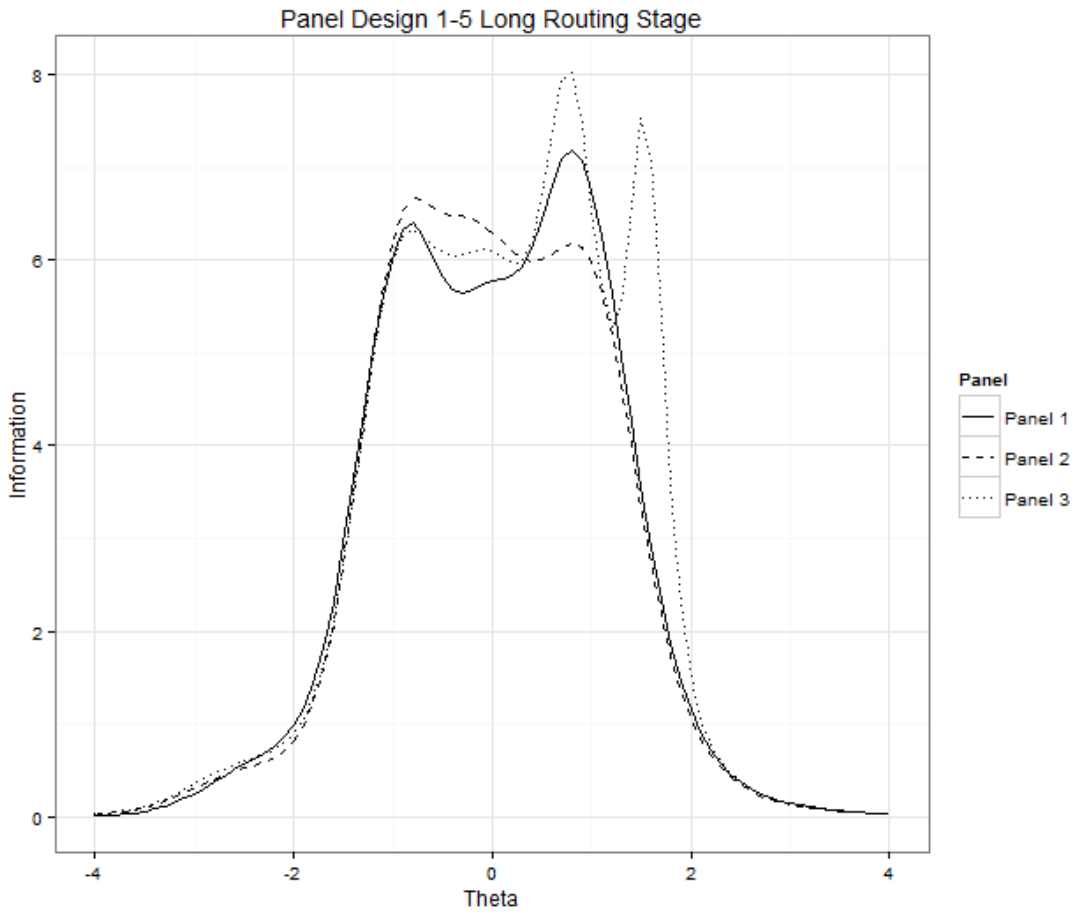

Figure A.16. Stage 1 routing module relative target TIF plots for the long 1-5 panel design when $\gamma_{(d)} = 0$.

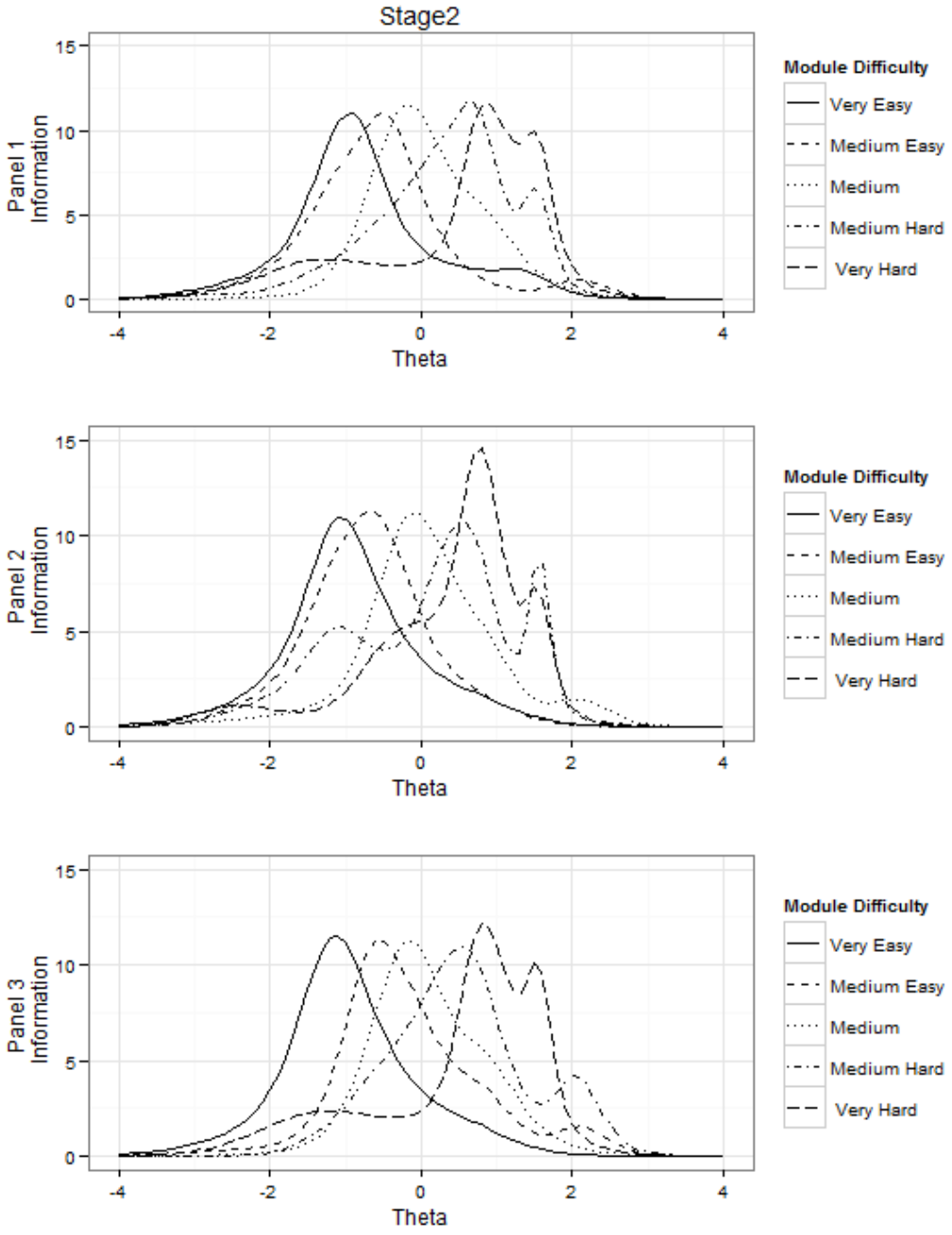

Figure A.17. Stage 2 relative target TIFs for the long 1-5 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

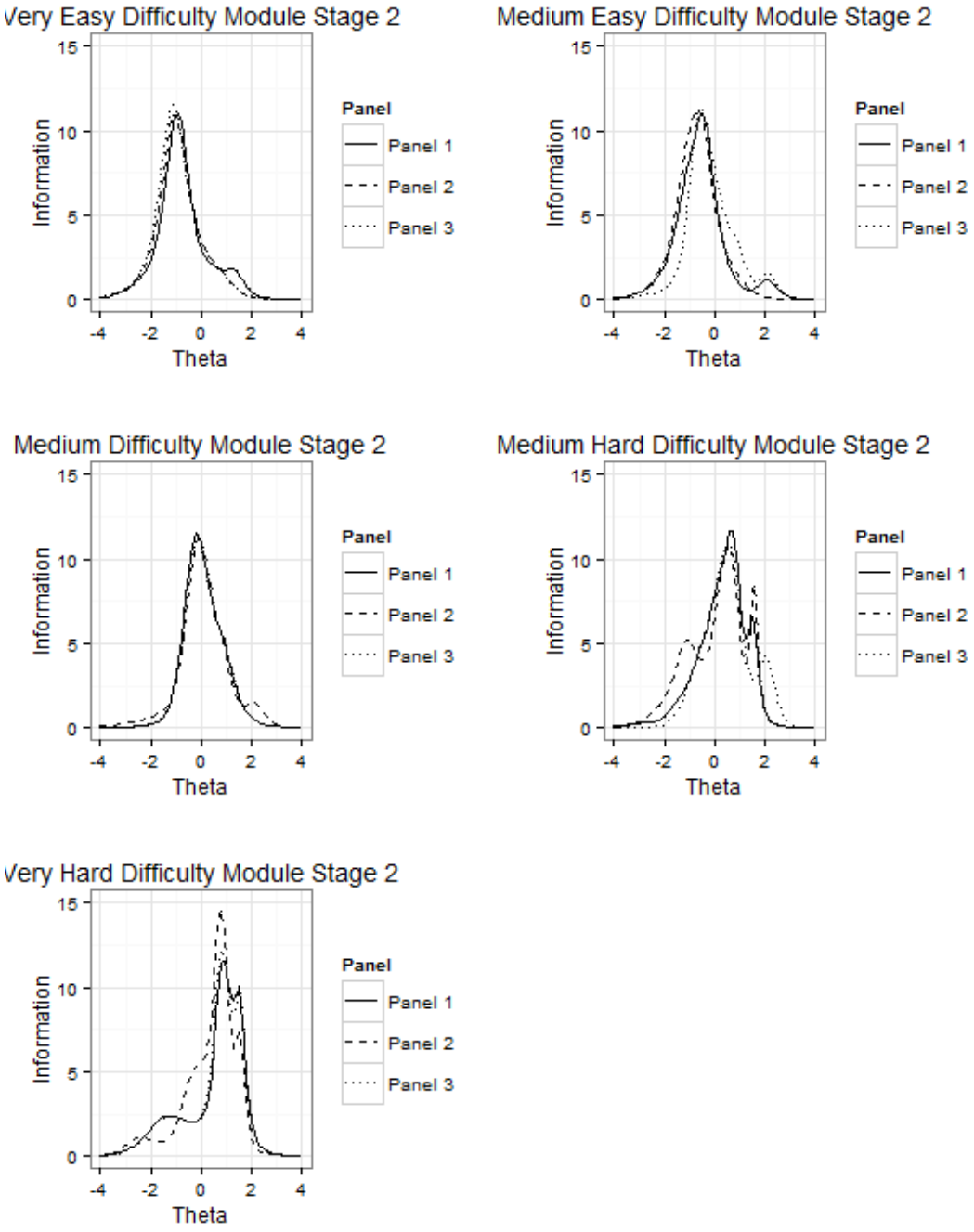

Figure A.18. Stage 2 relative target TIFs for the long 1-5 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

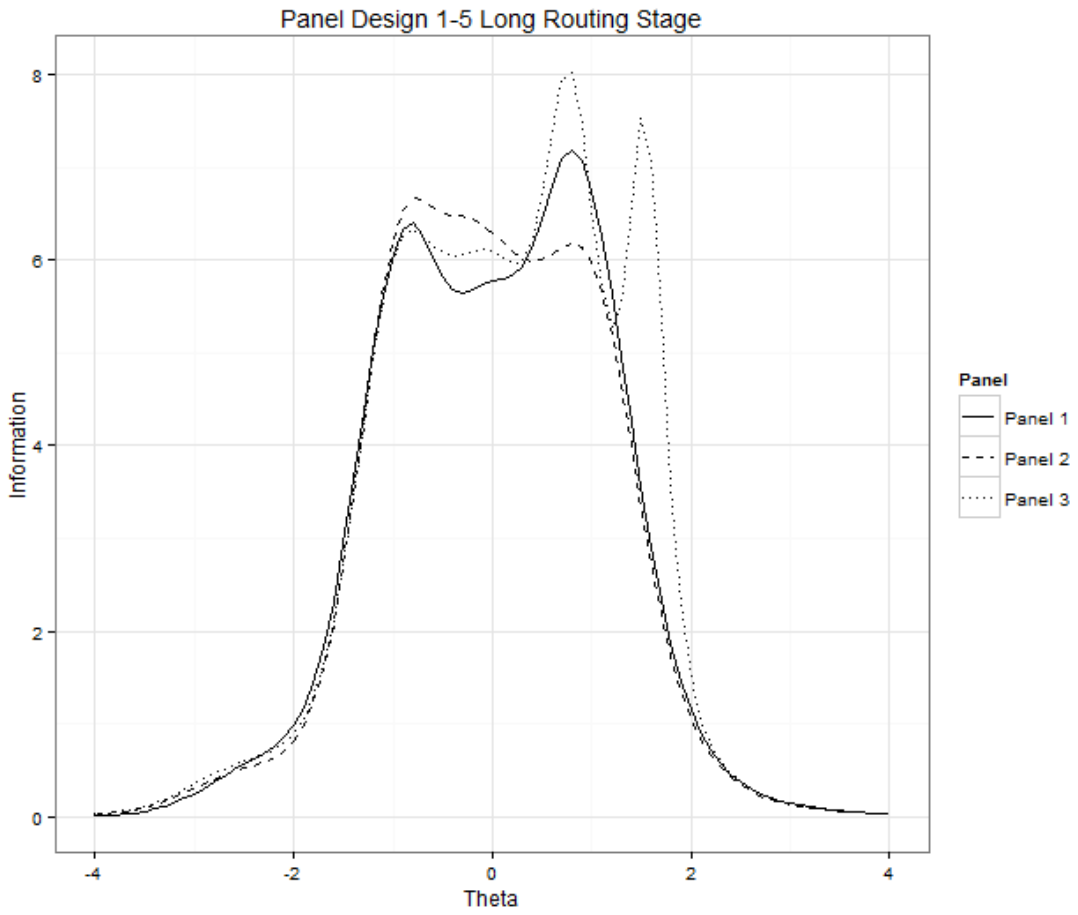

Figure A. 19. Stage 1 routing module relative target TIF plots for the short 1-5 panel design when $\gamma_{(d)} = 0$.

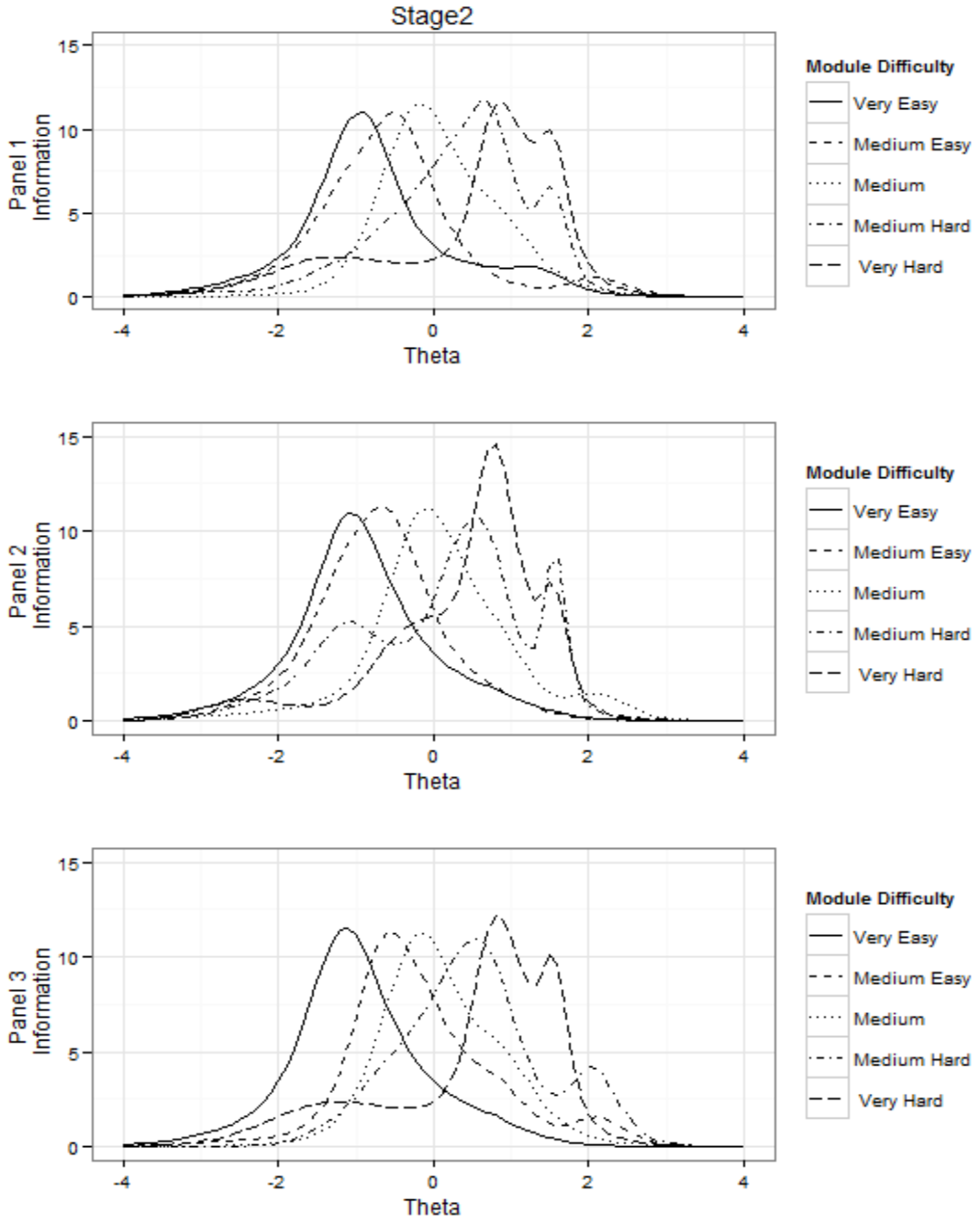

Figure A.20. Stage 2 relative target TIFs for the short 1-5 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

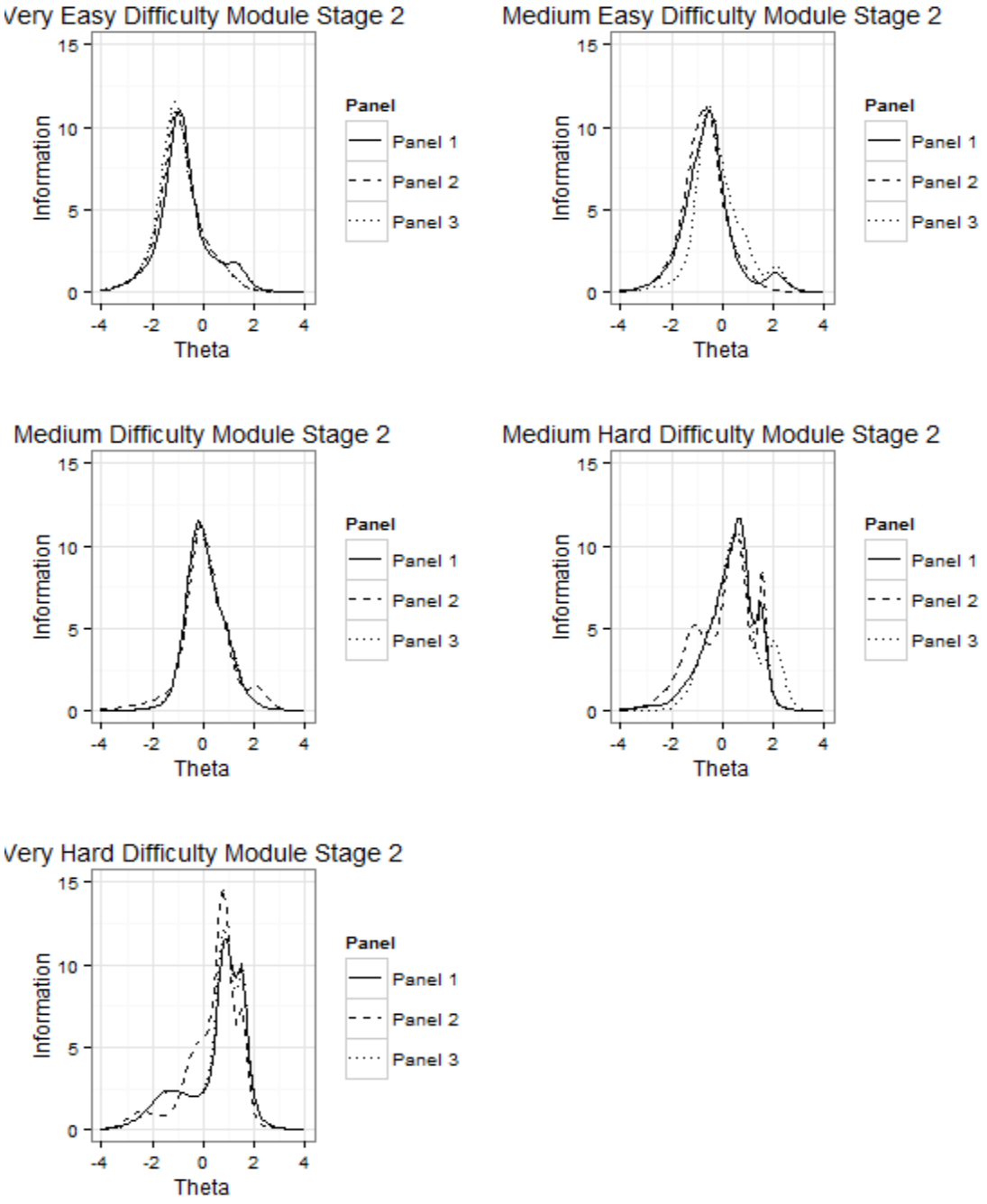

Figure A.21. Stage 2 relative target TIFs for the long 1-5 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

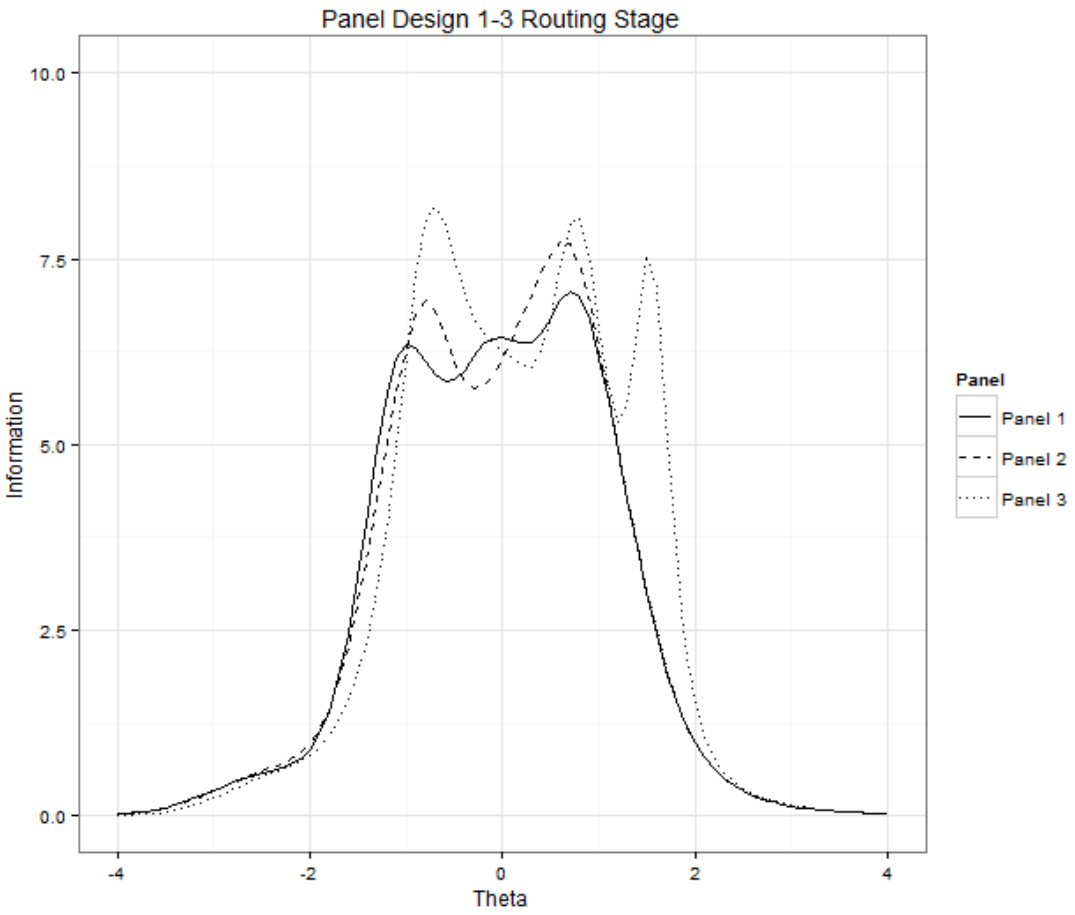

Figure A.22. Stage 1 routing module relative target TIF plots for the short 1-3 panel design when $\gamma_{(d)} = 0$.

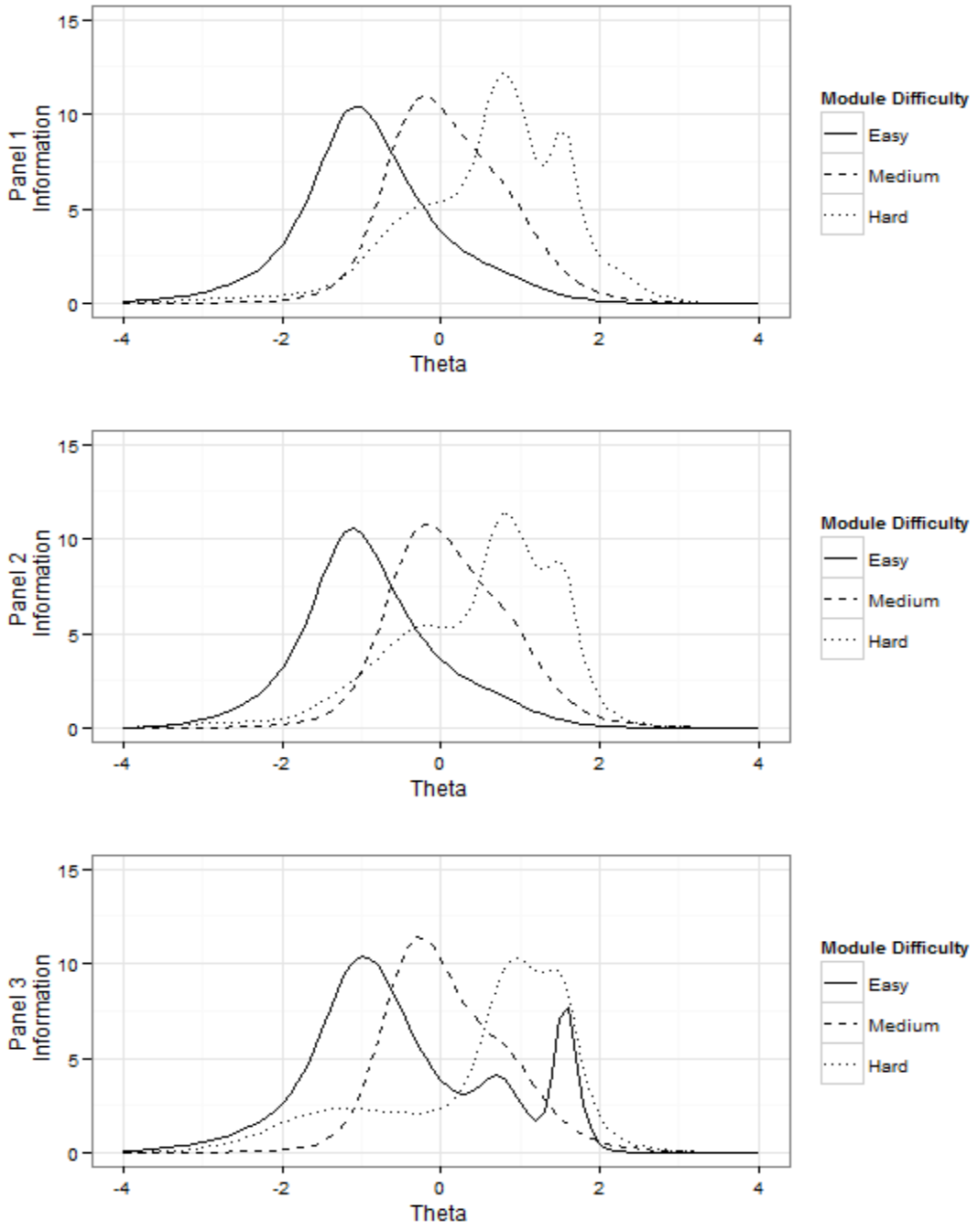

Figure A.23. Stage 2 relative target TIFs for the short 1-3 panel design across the targeted theta range when $\gamma_{(d)} = 0$.

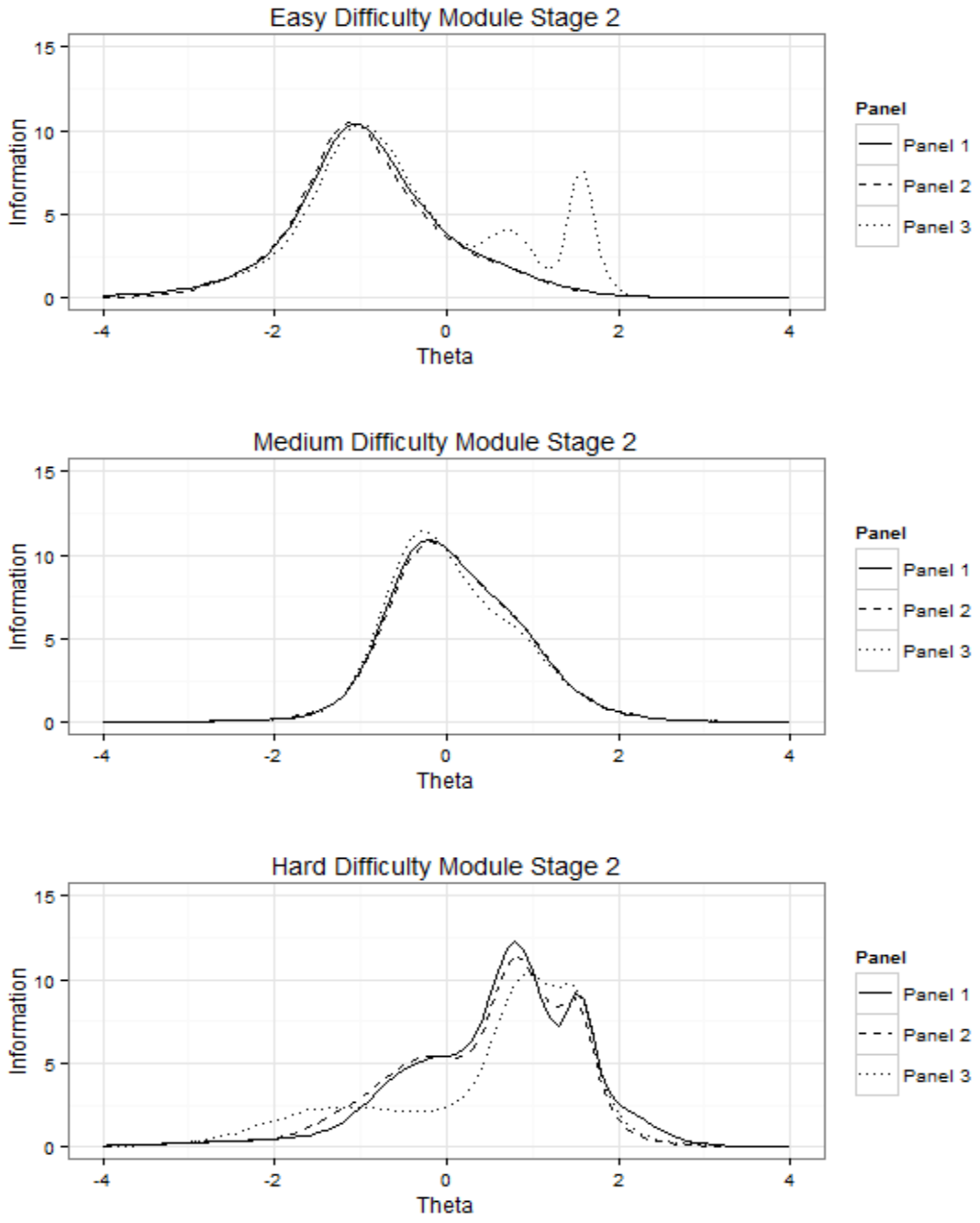

Figure A.24. Stage 2 relative target TIFs for the short 1-3 panel design at each targeted difficulty level when $\gamma_{(d)} = 0$.

APPENDIX B: CONDTIONAL BIAS AND CONDITIONAL STANDARD ERROR PLOTS.

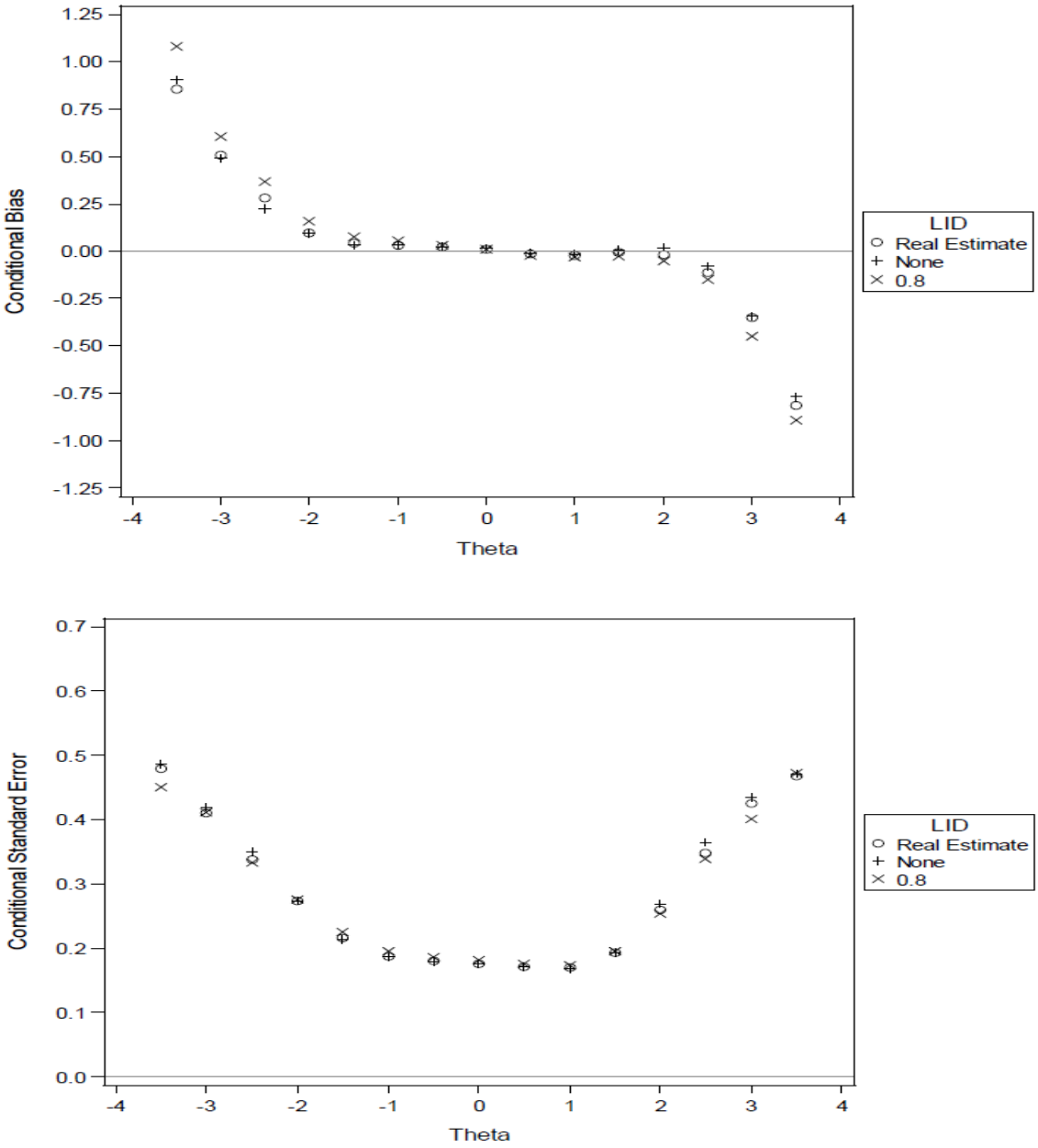

Figure B.1. Conditional bias and standard error plots for the 1-5-5 panel design, short test length, AMI routing procedure across the LID conditions.
Note: AMI=approximate maximum information; LID=local item dependence.

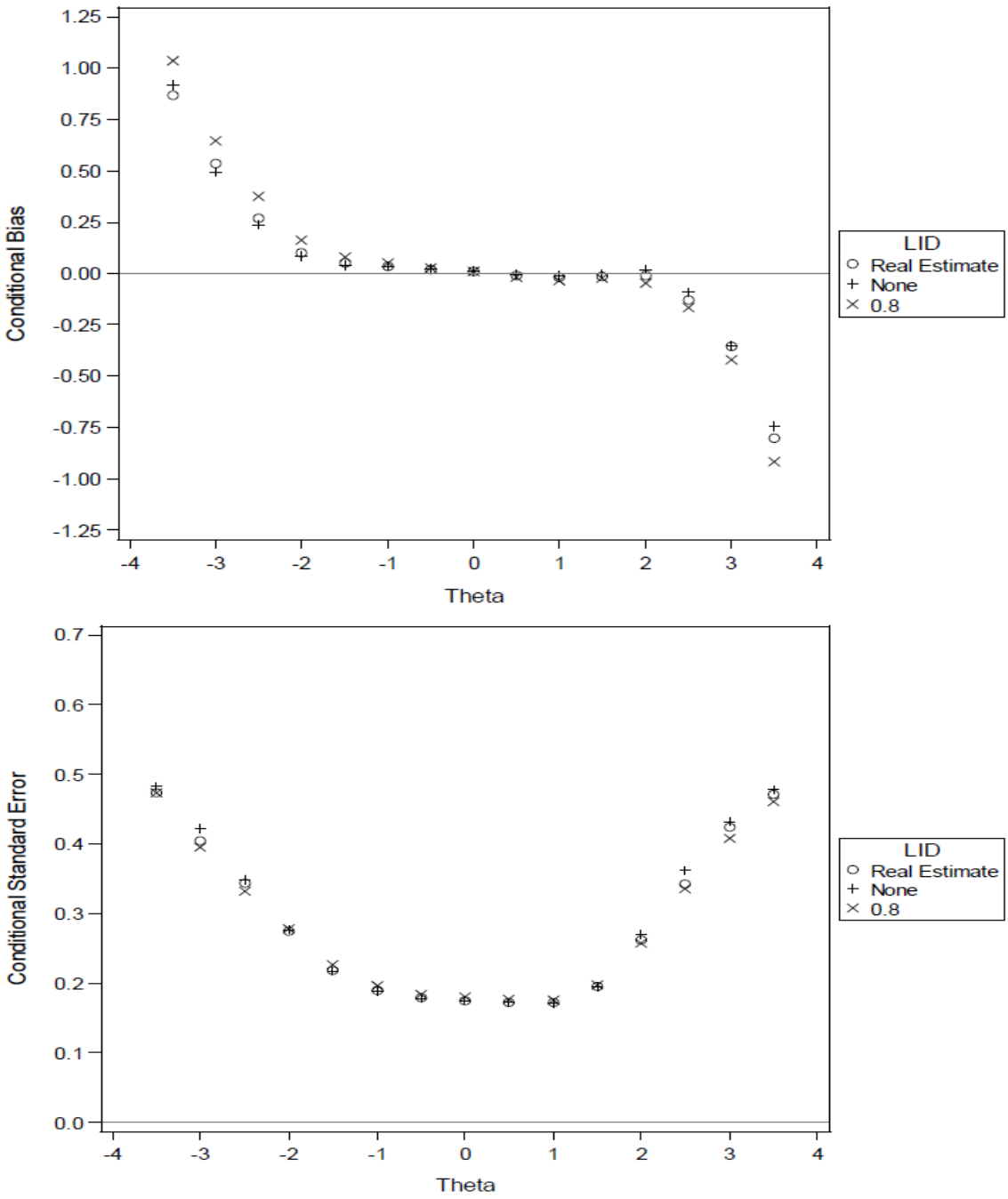

Figure B.2. Conditional bias and standard error plots for the 1-5-5 panel design, short test length, ML-DPI routing procedure across the LID conditions.
Note: ML-DPI=module-level defined population interval; LID=local item dependence.

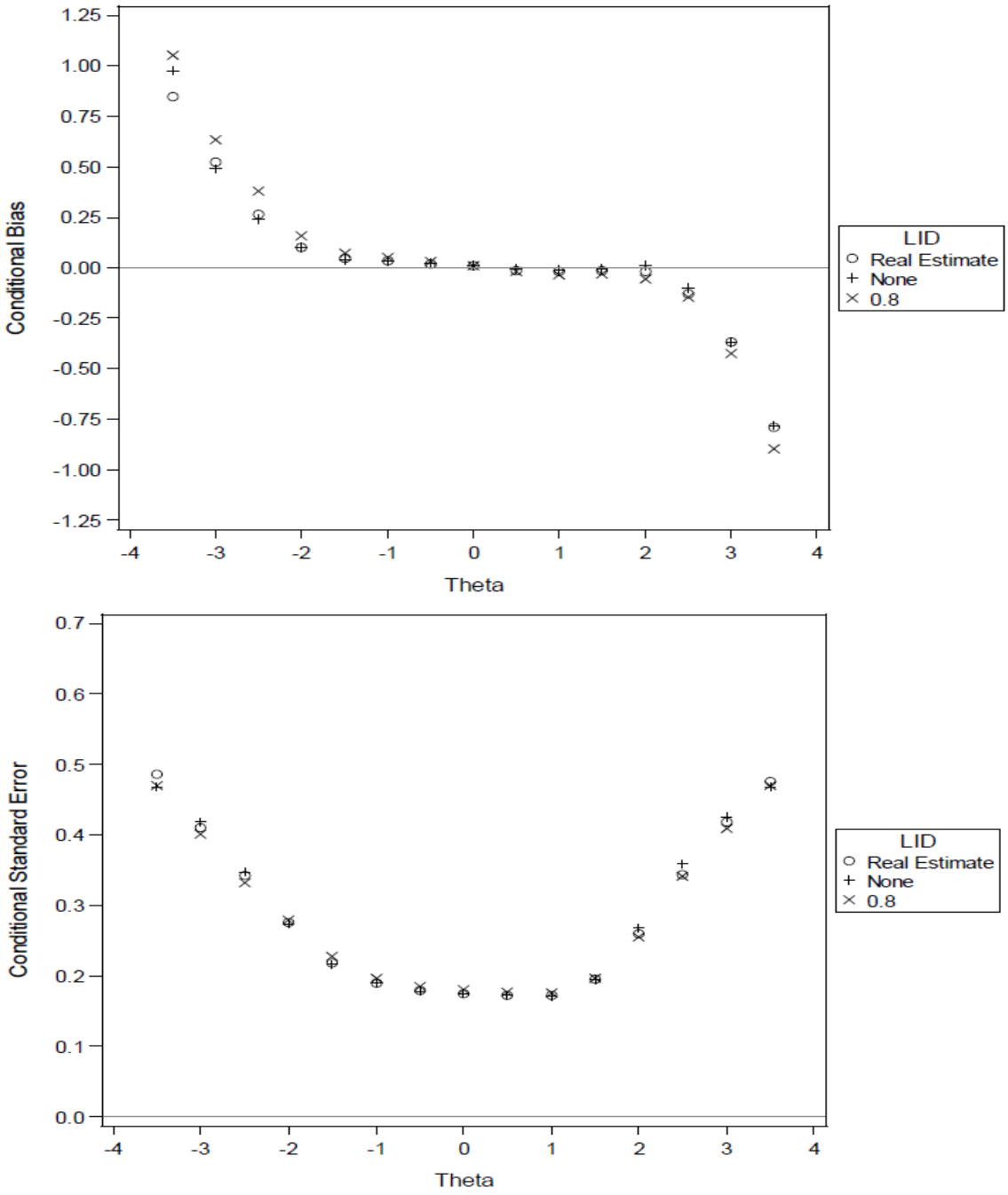

Figure B.3. Conditional bias and standard error plots for the 1-5-5 panel design, short test length, SL-DPI routing procedure across the LID conditions. *Note*: SL-DPI=stage-level defined population interval; LID=local item dependence.

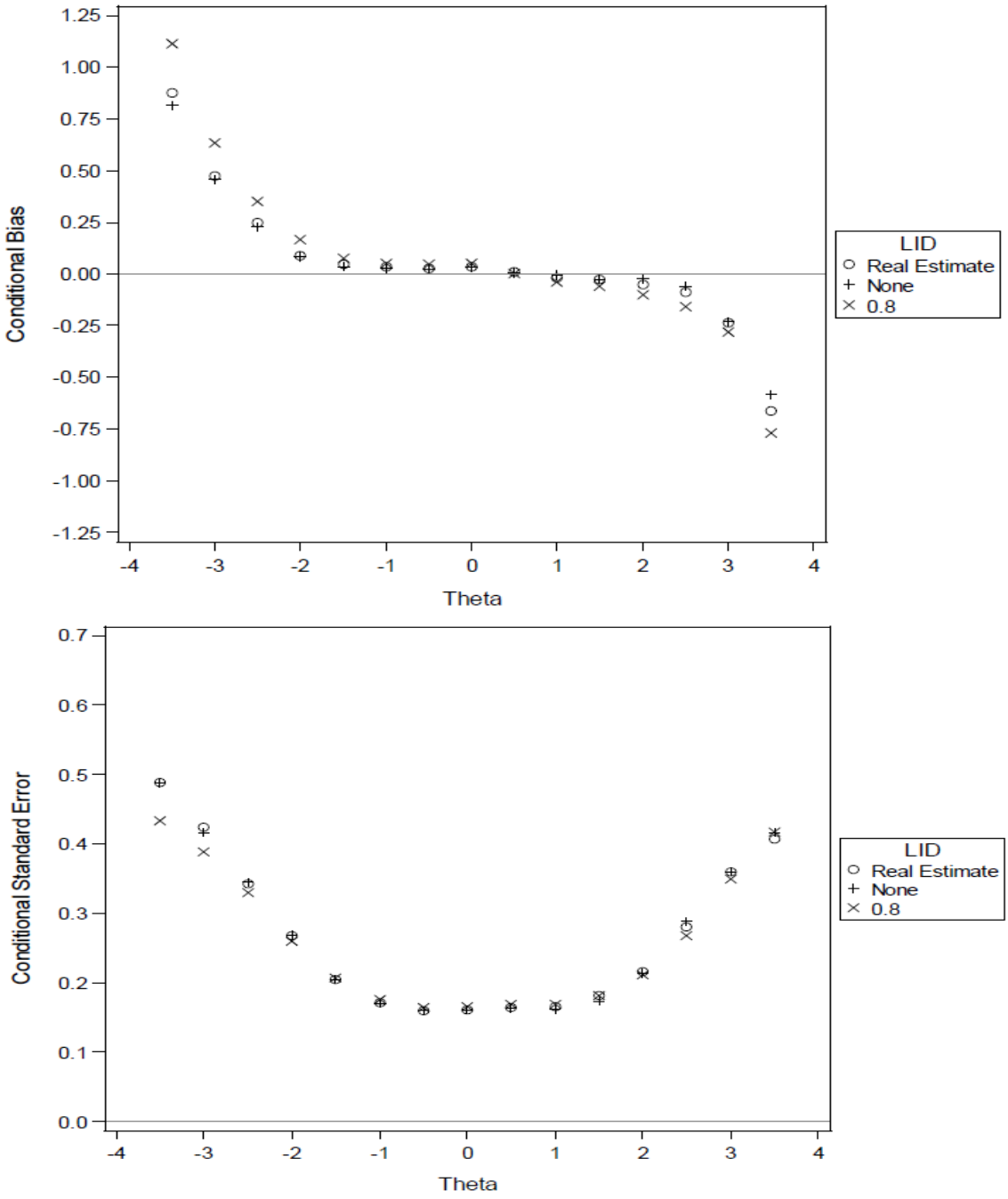

Figure B.4. Conditional bias and standard error plots for the 1-3-3 panel design, long test length, AMI routing procedure across the LID conditions. *Note*: AMI=approximate maximum information; LID=local item dependence.

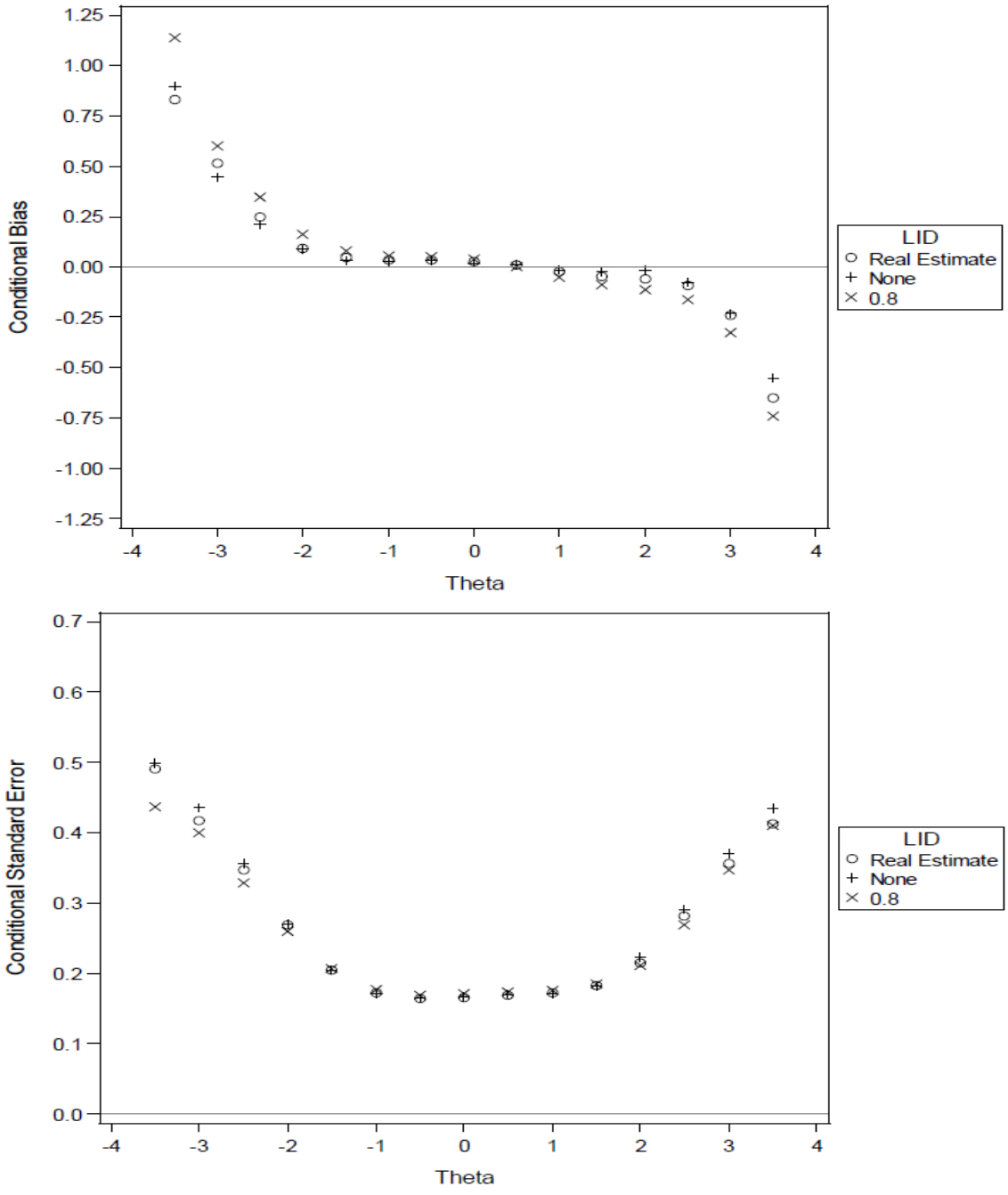

Figure B.5. Conditional bias and standard error plots for the 1-3-3 panel design, long test length, ML-DPI routing procedure across the LID conditions. *Note*: ML-DPI=module-level defined population interval; LID=local item dependence.

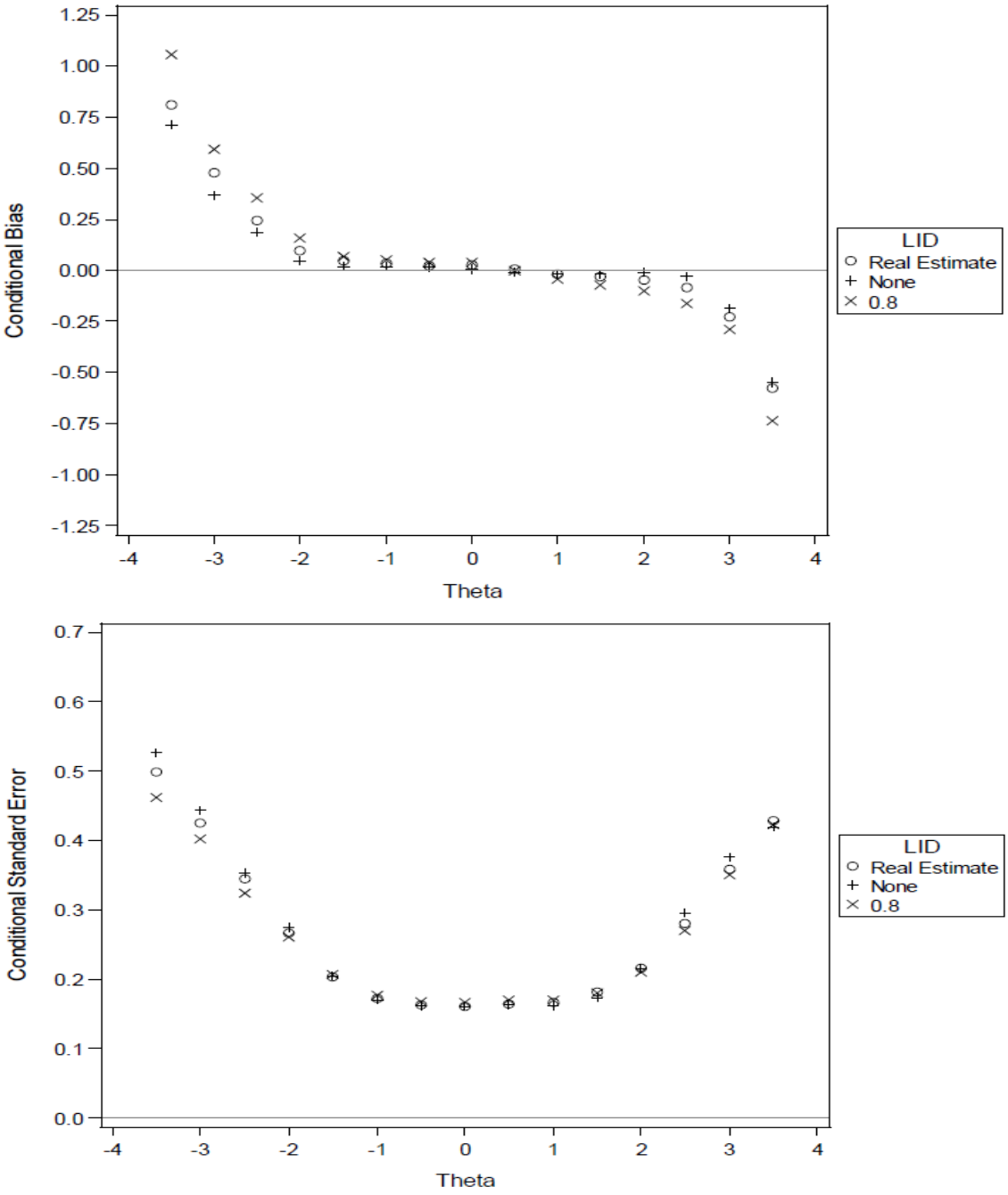

Figure B.6. Conditional bias and standard error plots for the 1-3-3 panel design, long test length, SL-DPI routing procedure across the LID conditions. *Note*: SL-DPI=stage-level defined population interval; LID=local item dependence.

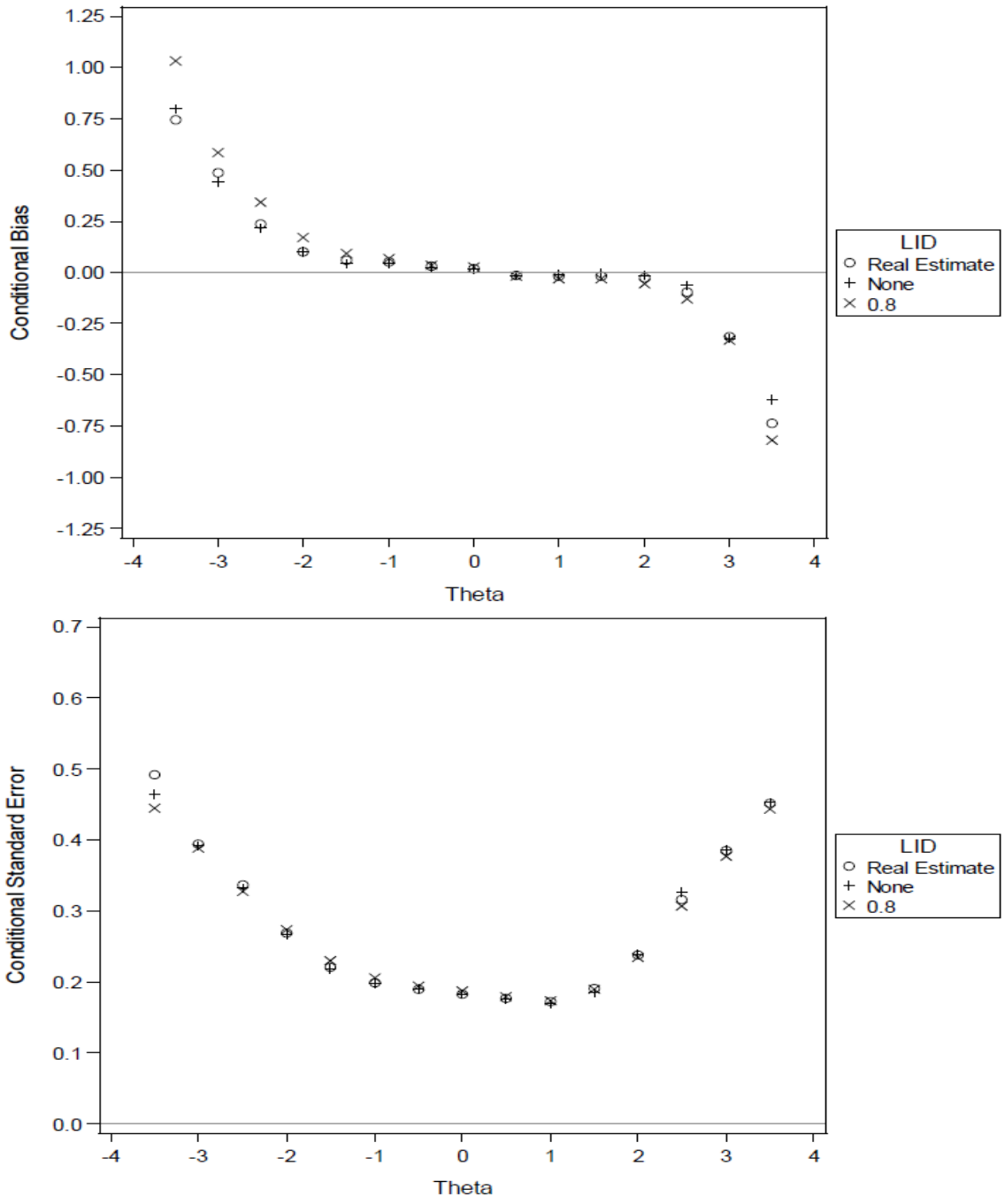

Figure B.7. Conditional bias and standard error plots for the 1-3-3 panel design, short test length, AMI routing procedure across the LID conditions. *Note*: AMI=approximate maximum information; LID=local item dependence.

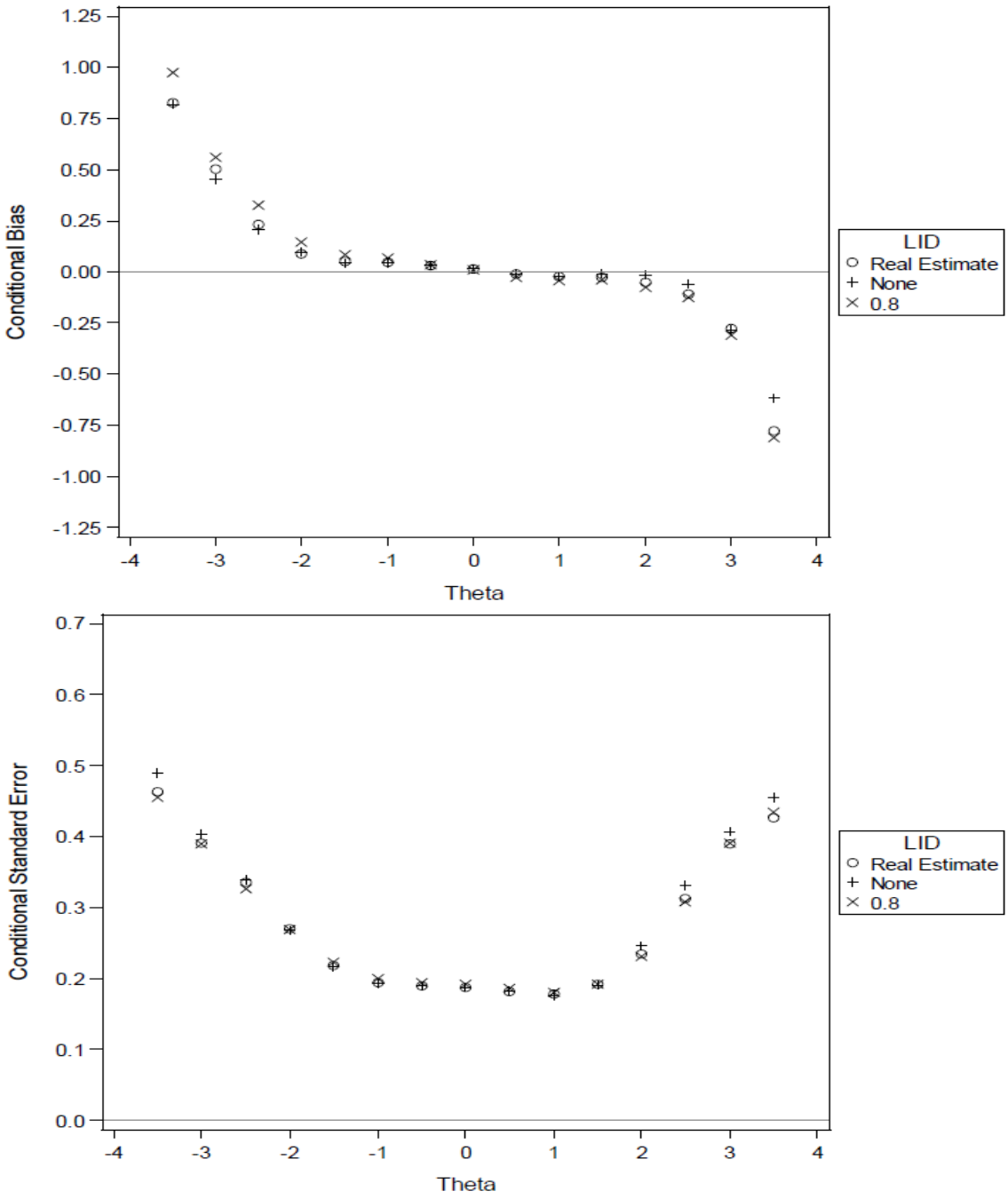

Figure B.8. Conditional bias and standard error plots for the 1-3-3 panel design, short test length, ML-DPI routing procedure across the LID conditions. *Note*: ML-DPI=module-level defined population interval; LID=local item dependence.

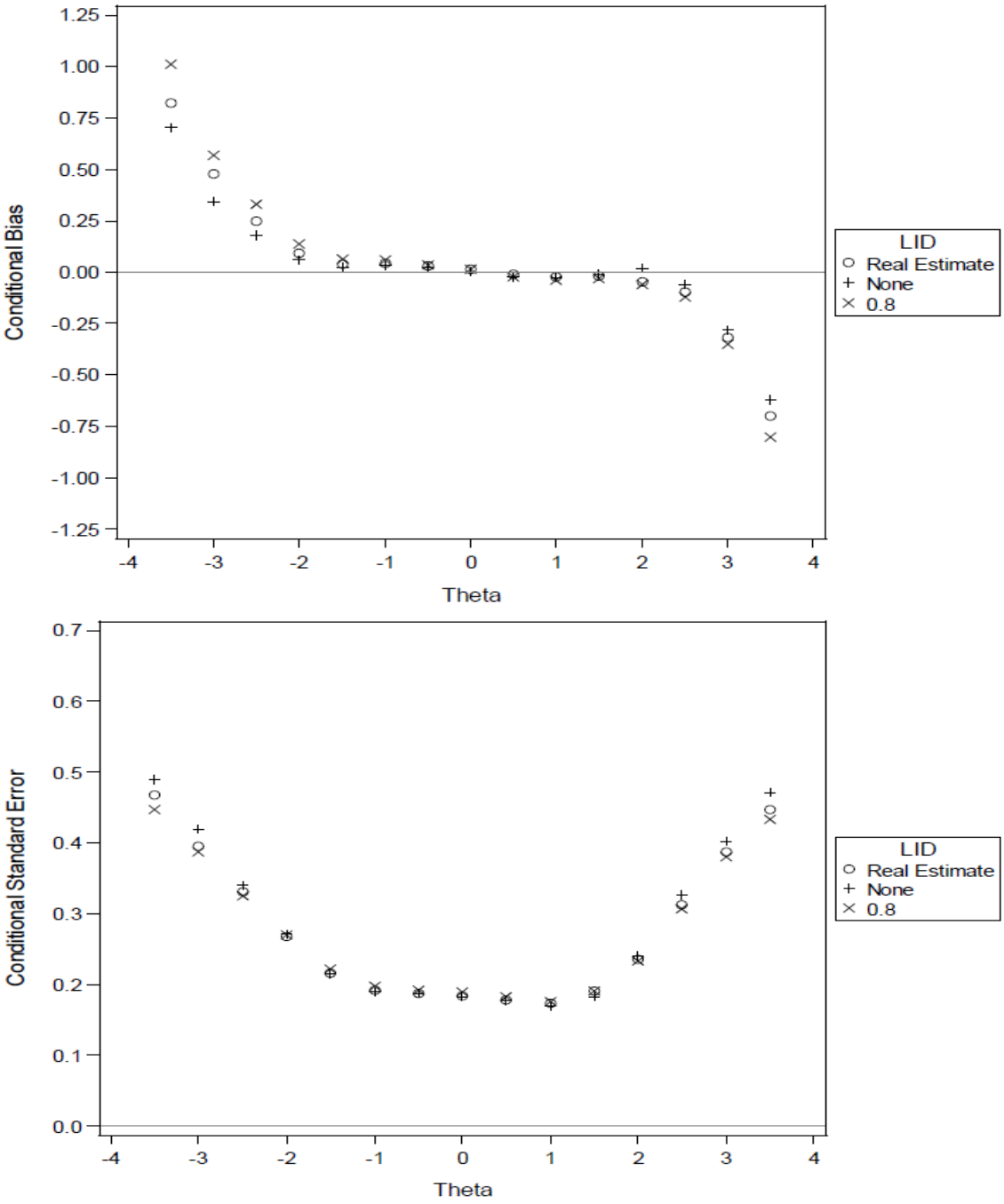

Figure B.9. Conditional bias and standard error plots for the 1-3-3 panel design, short test length, SL-DPI routing procedure across the LID conditions. *Note*: SL-DPI=stage-level defined population interval; LID=local item dependence.

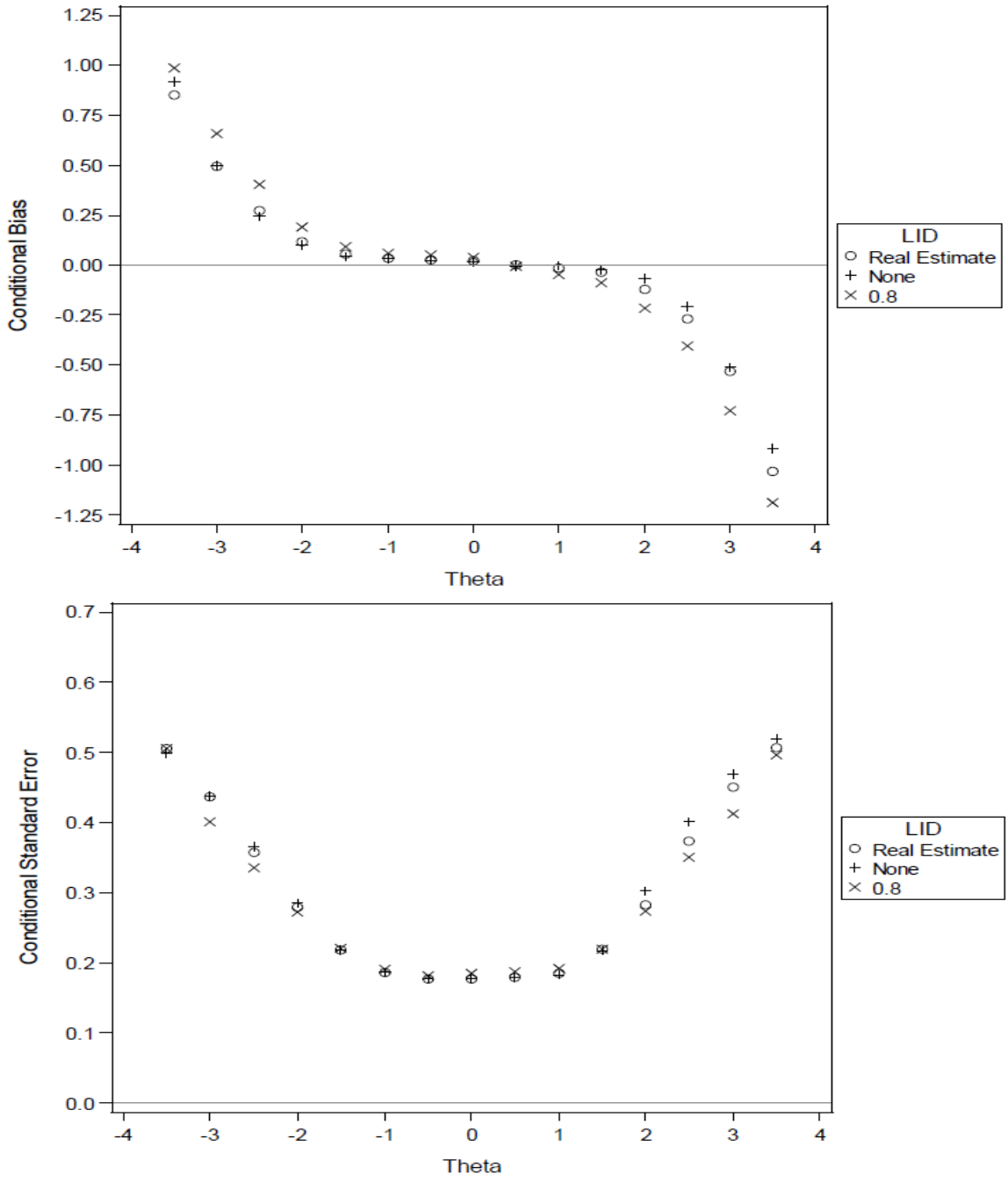

Figure B.10. Conditional bias and standard error plots for the 1-5 panel design, long test length, AMI routing procedure across the LID conditions.
Note: AMI=approximate maximum information; LID=local item dependence.

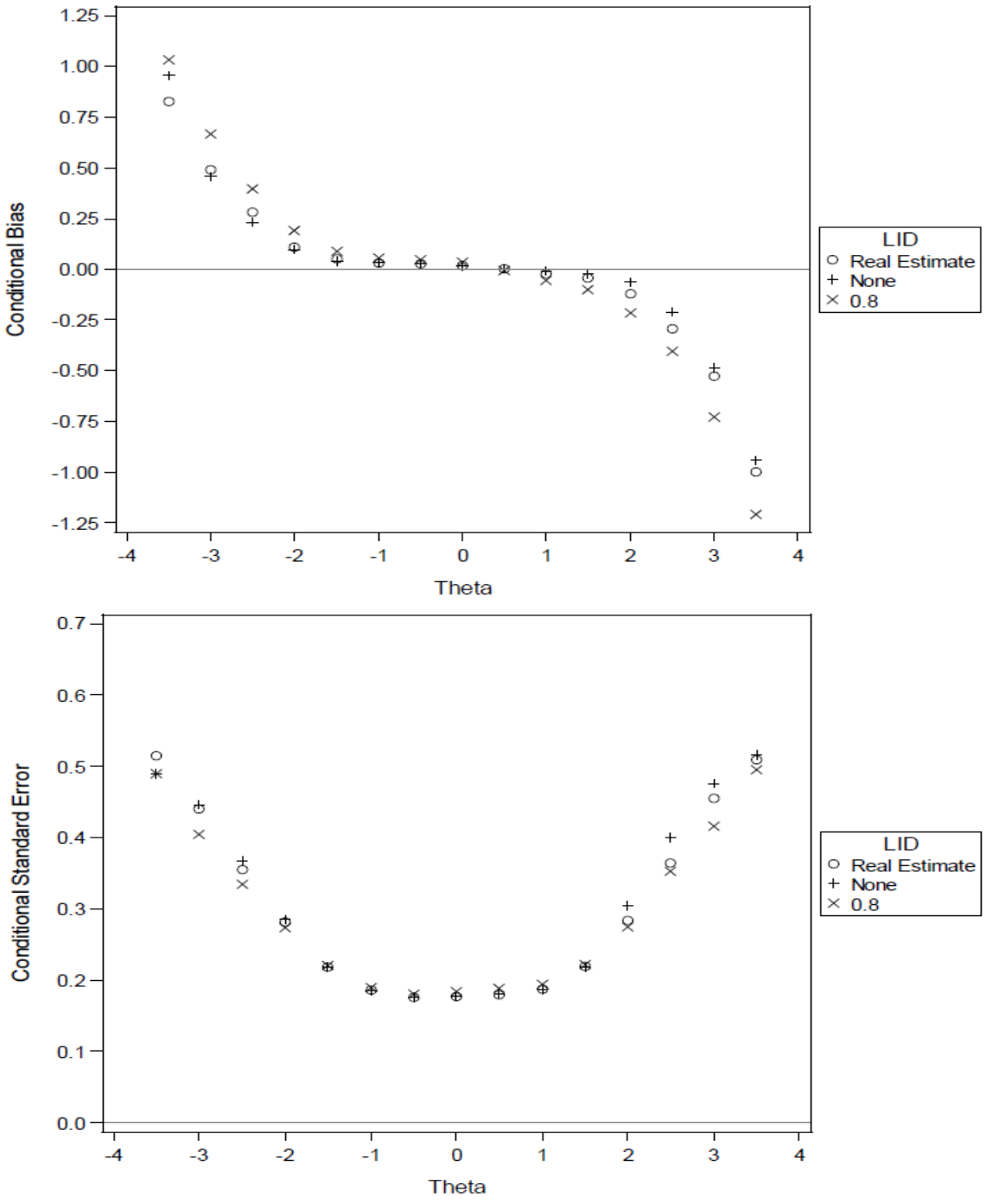

Figure B.11. Conditional bias and standard error plots for the 1-5 panel design, long test length, DPI routing procedure across the LID conditions.
Note: DPI=defined population interval; LID=local item dependence.

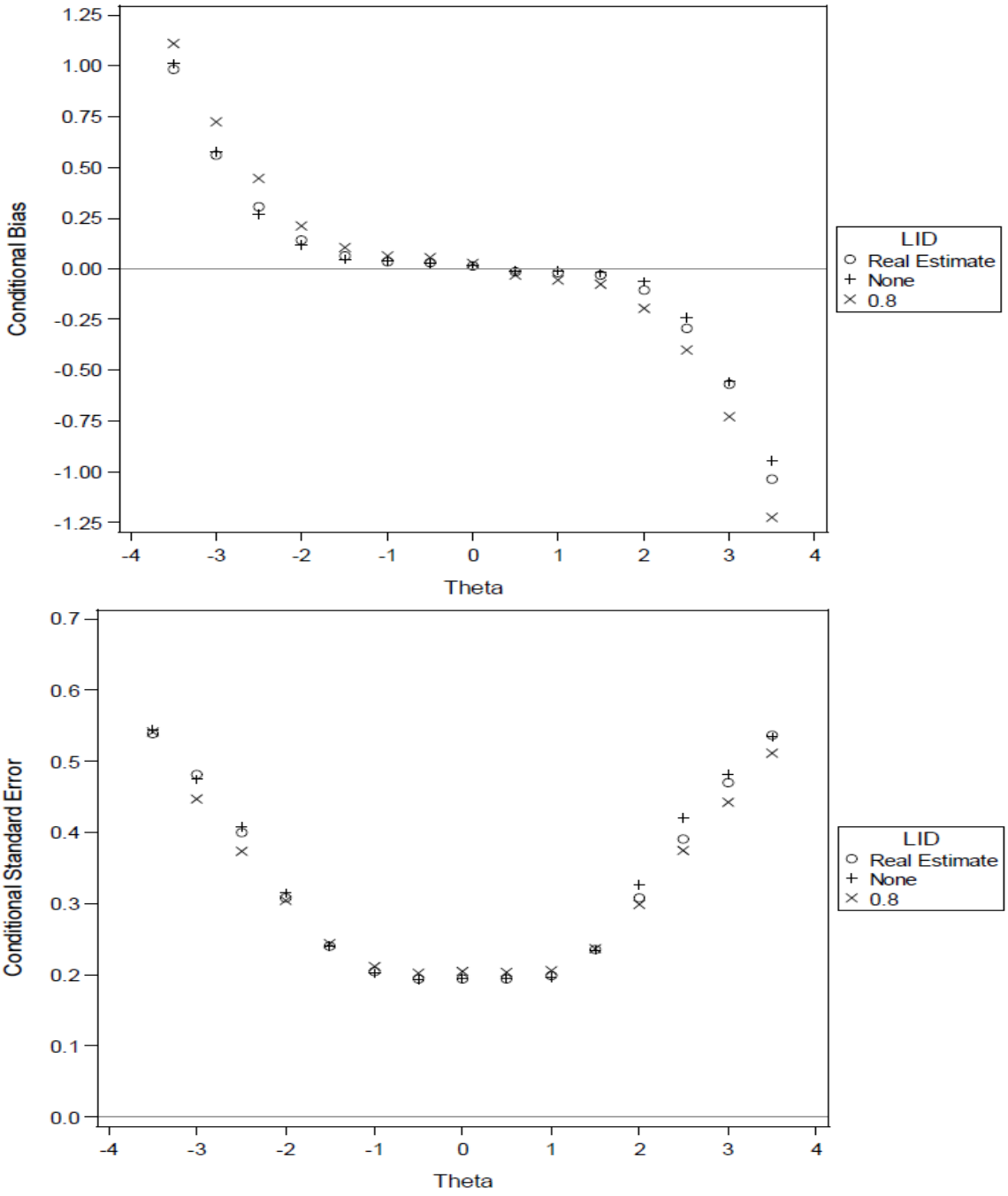

Figure B.12. Conditional bias and standard error plots for the 1-5 panel design, short test length, AMI routing procedure across the LID conditions.
Note: AMI=approximate maximum information; LID=local item dependence.

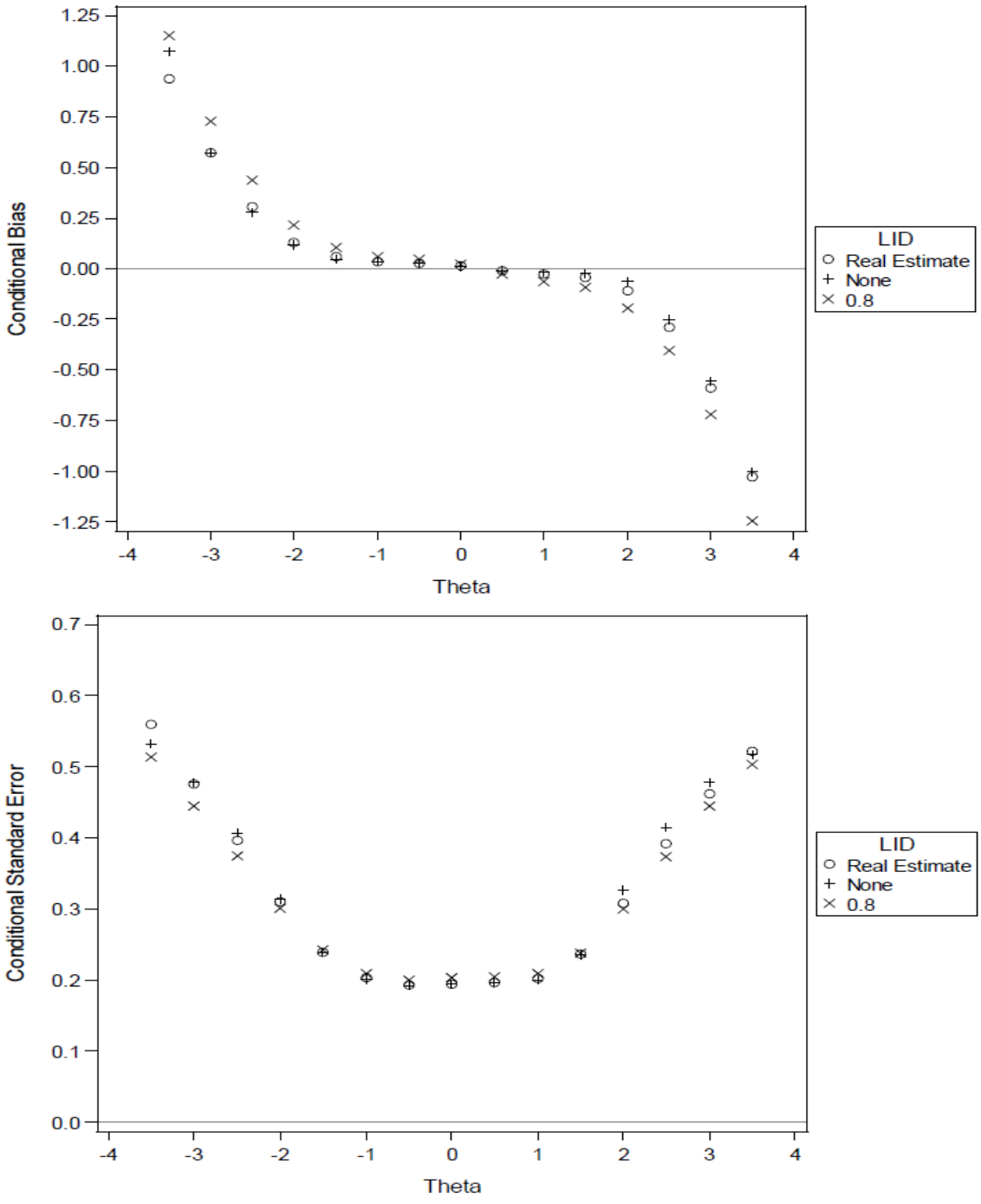

Figure B.13. Conditional bias and standard error plots for the 1-5 panel design, short test length, DPI routing procedure across the LID conditions.
Note: DPI=defined population interval; LID=local item dependence.

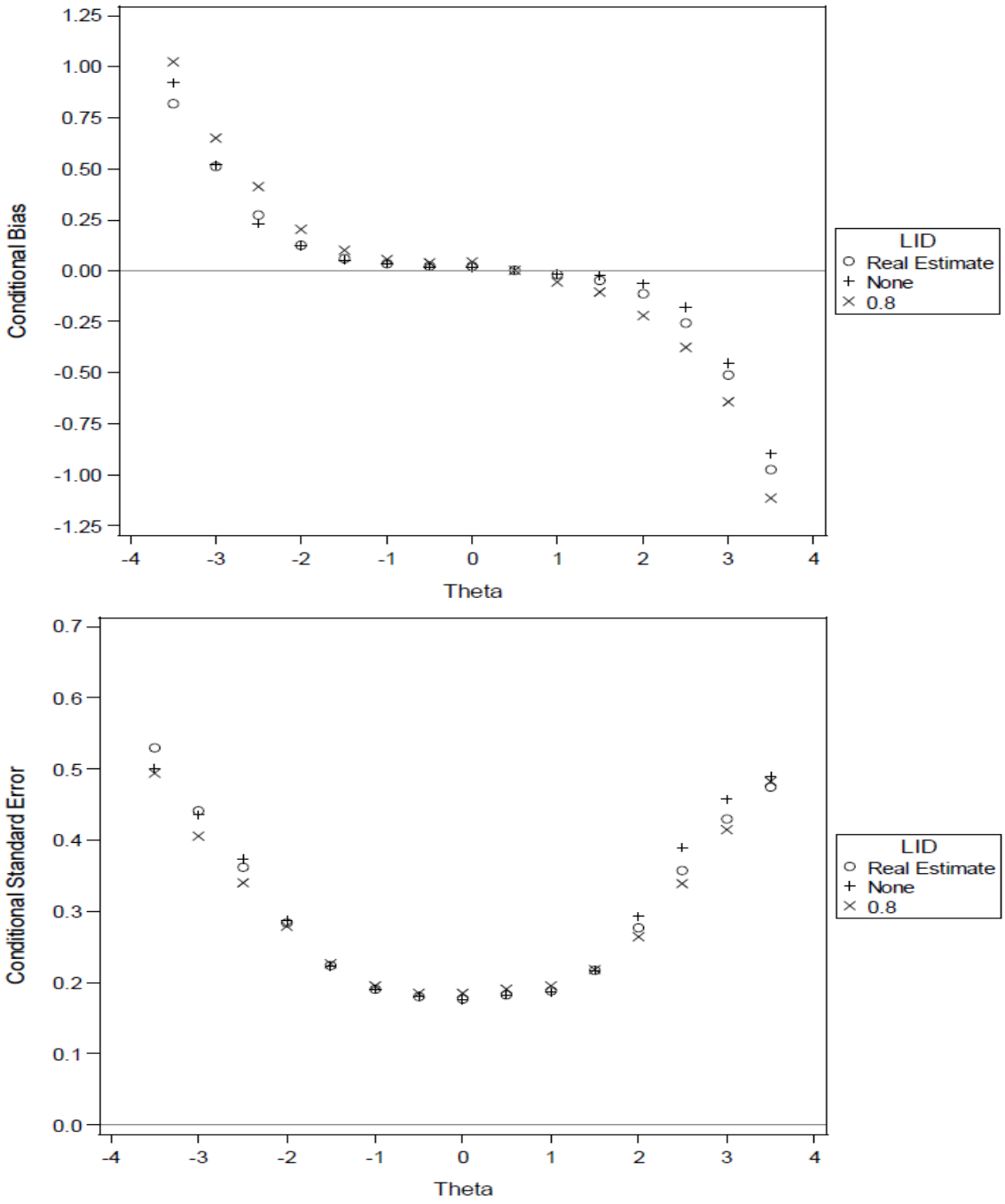

Figure B.14. Conditional bias and standard error plots for the 1-3 panel design, long test length, AMI routing procedure across the LID conditions.
Note: AMI=approximate maximum information; LID=local item dependence.

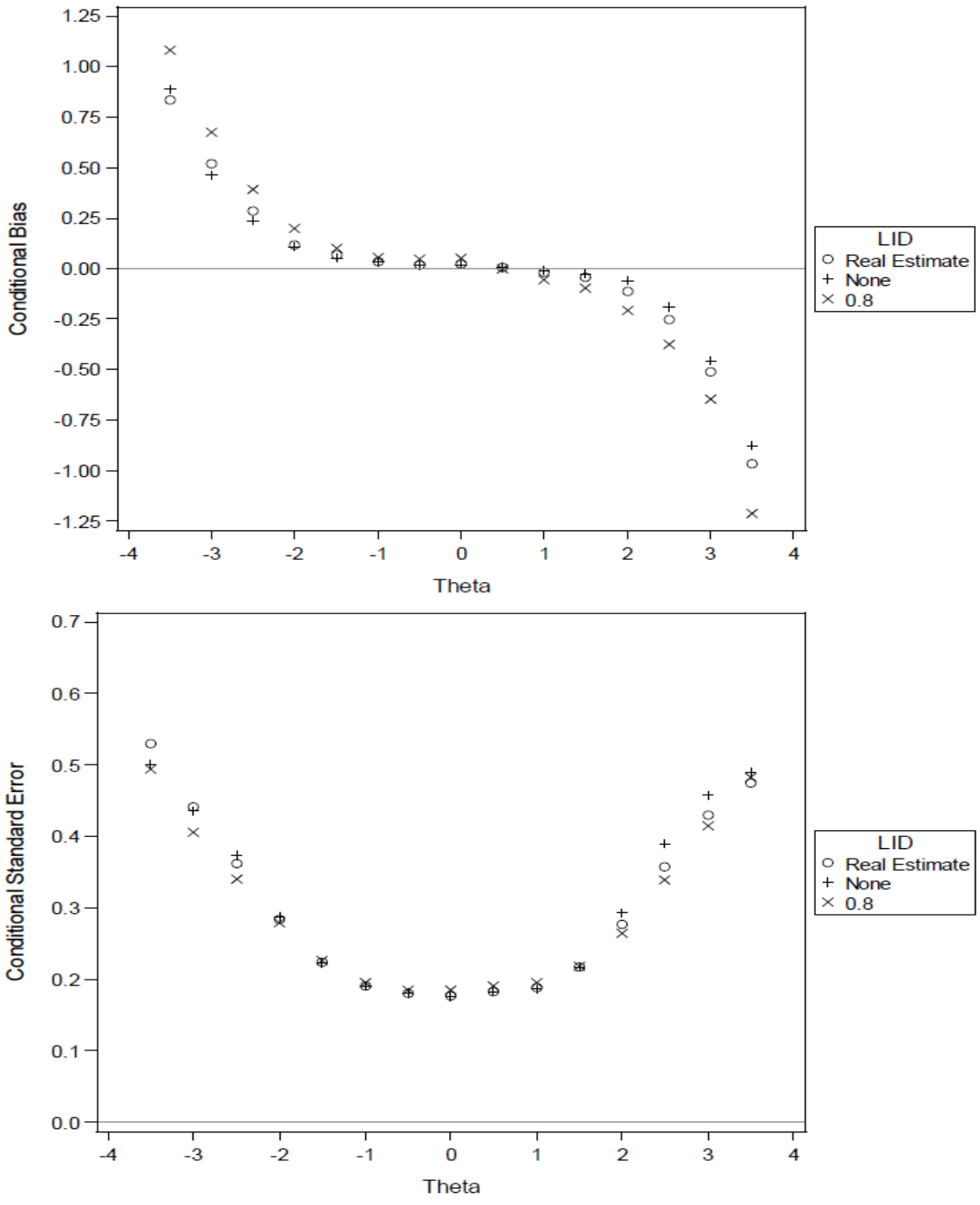

Figure B.15. Conditional bias and standard error plots for the 1-3 panel design, long test length, DPI routing procedure across the LID conditions.
Note: DPI=defined population interval; LID=local item dependence.

References

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. doi:10.1177/0146621604263652

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. doi:10.1177/014662168200600405

Bock, R., Zimowski, M., & Panel, N. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP*. Paolo Alto. Retrieved from http://iacat.org/node/1335

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3110732)

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

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing*, *5*(3), 319–330. doi:10.1207/s15327574ijt0503_8

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, *67*(1), 5–20. doi:10.1177/0013164406288162

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: Praeger Publishers.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Chen, L. L. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model*(Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 698453020)

Chuah, S. C., Drasgow, F., & Luecht, R. (2010). How big Is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, *19*(3), 241–255.

Cor, K., Alves, C., & Gierl, M. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Practical Assessment, Research & Evaluation*, *14*(14).

Cronbach, L. J., & Gleser, C. G. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test*. Princeton, NJ.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, *28*(3), 165–185. doi:10.1177/0146621604264133

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*(5), 335–356. doi:10.1177/0146621603256804

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

Diao, Q., & van der Linden, W. J. (2011a). Automated test assembly using lp_Solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398–409. doi:10.1177/0146621610392211

Diao, Q., & van der Linden, W. J. (2011b). Automated test assembly using lp_Solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398–409. doi:10.1177/0146621610392211

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport, CT: Praeger Publishers.

Edwards, M. C., Flora, D. B., & Thissen, D. (2012). Multistage computerized adaptive testing with uniform item exposure. *Applied Measurement in Education*, *25*(3), 118–141.

ETS. (2011). *GRE information and registration bulletin*. Princeton, NJ. Retrieved from www.ets.org/s/gre/pdf/gre_info_reg_bulletin.pdf

Galindo, J. L., Park, R., & Dodd, B. G. (2013). The effects of test structure, routing test length, and total test length on multistage testing using the 3PL-testlet response theory model. In *Annual Meeting of the National Council on Measurement in Education*. San Francisco, CA.

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

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, *5*(8). Retrieved from http://www.jtla.org

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

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.

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

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203–220. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15324818ame1903_3

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3315089)

Kim, H., & Plake, B. (1993). Monte Carlo simulation comparison of two-stage testing and computer adaptive testing. In *Annual Meeting of the National Council on Measurement in Education*. Atlanta, GA.

Kim, J. (2010). *A Comparison of Computer-Based Classification Testing Approaches Using Mixed-Format Tests with the Generalized Partial Credit Model* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3429006)

Kim, J., Chung, H., Dodd, B. G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement, 72*(4), 574–588. doi:10.1177/0013164411428977

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*. doi:10.3758/s13428-013-0316-3

Konis, K. (2013). lpSolveAPI, version 5.5.20. Retrieved from http://cran.r-project.org/web/packages/lpSolveAPI/index.html

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Lord, F. M. (1952). *A theory of test scores*. Richmond, VA. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf

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

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*(3), 147–151.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157–162. doi:10.1111/j.1745-3984.1986.tb00241.x

Lu, R. (2010). *Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3443478)

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

Luecht, R. M. (1998). CASTISEL. Philadelphia, PA: National Board of Medical Examiners.

Luecht, R. M. (2003). Exposure control using adapative multi-stage item bundles. In *Annual Meeting of the National Council on Measurement in Education*. Chicago, IL.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–249. doi:10.1111/j.1745-3984.1998.tb00537.x

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3328282)

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

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

Melican, G. J., Breithaupt, K., & Zhang, Y. (2010). Designing and implementing a multisatage adaptive test: The uniform CPA exam. In W. J. van der Linden & C. a. W. Glas (Eds.), *Elements of adaptive testing* (pp. 167–190). New York, NY: Springer.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*(6), 424–437. doi:10.1177/0146621609349804

Park, R., Kim, J., Dodd, B. G., & Chung, H. (2011). JPLEX: Java simplex implementation with branch-and-bound search for automated test assembly. *Applied Psychological Measurement*, *35*(8), 643–644. doi:10.1177/0146621610392912

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

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9950199)

Rasch, G. (1960). *Probabilistic models for some intelligence ad attaiment tests.* Chicago, IL: University of Chigo Press.

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

Rosenbaum, P. R. (1988). A note on item bundles. *Psychomtrika, 53,* 349-360.

Samejima, F. (1969). Estimation of latent trait ability using response pattern of graded scores. *Psychometrika*, (Psychometric Society Monograph No. 17).

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–354). Westport, CT: Praeger Publishers.

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

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. doi:10.1111/j.1745-3984.1991.tb00356.x

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

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

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika*, *50*(4), 411–420.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–133). Mahwah, NJ: Routledge.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260. doi:10.1111/j.1745-3984.1989.tb00331.x

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), 181–195. doi:10.1037//1082-989X.6.2.181

Van der Linden, W. J. (2005). *Linear models for optimal test assembly*. New York, NY: Springer.

Van Der Linden, W. J., Veldkamp, B. P., & Reese, L. M. (2000). An integer programming approach to item bank design. *Applied Psychological Measurement*, *24*(2), 139–150. doi:10.1177/01466210022031570

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*(4), 328–345. doi:10.1111/j.1745-3984.1998.tb00542.x

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *15*(1), 15–20. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1993.tb00519.x/full

Wainer, H. (1995). Precision and differential item functioning on a testlet-based tests: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157–187.

Wainer, H. (2000a). *Computerized adaptive testing: A primer* (2nd ed., pp. 1–315). New York, NY: Routledge.

Wainer, H. (2000b). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 1–22). Mahwah, NJ: Routledge.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing, theory, and practice* (pp. 246–270). Boston, MA: Kluwer-Nijhoff.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*(1), 1–14. doi:10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., Lewis, C., Kaplan, B. A., & Braswell, J. S. (1990). *An adaptive algebra test: A testlet-based, hierarchically-structured test with validity-based scoring*. Princeton, NJ.

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

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning : Definitions and detection. *Journal of Educational Measurement*, *28*(3), 197–219. doi:10.1111/j.1745-3984.1991.tb00354.x

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on Reliability? *Educational Measurement: Issues and Practice1*, *15*(1), 22–29.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *Users guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis*. Princeton, NJ.

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests*. Newtown, PA.

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*(1), 5–21. doi:10.1177/0013164403258393

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3136800)

Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In A. W. van der Linden, Wim, J.; Glas, Cees (Ed.), *Elements of adaptive testing* (2nd ed., pp. 355–372). New York, NY: Springer.

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* (Vol. 2012). Iowa City, IA.

Vita

Ian Fredrick Hembry was born in Red Oak, IA on March 19, 1982. He attended elementary, middle, and secondary school at the Nishna Valley Community School District in Hastings, IA. He earned a B.A. in Mathematics at the University of Iowa in Iowa City with a certification in Secondary Education. From 2005-2008, Ian was a high school mathematics instructor for the Clinton Community School District located in Clinton, IA. He then began his graduate studies in Iowa City, IA with the Department of Psychological and Quantitative Methods in the College of Education at the University of Iowa. At the University of Iowa, he worked as a research assistant with the Iowa Testing Programs and earned his Master's in Measurement and Statistics. Then in 2010 he began his doctoral studies in the College of Education at the University of Texas at Austin. During his time at the University of Texas, he held a psychometric internship with Pearson from 2011-2012, was a teaching assistant for multiple statistics courses, and was a research assistant with Dr. S. Natasha Beretvas and Dr. Barbara Dodd. He began working for Amplify Inc. as a Psychometrician in May 2014.

Permanent address: 325 Ivy Meadow Lane, Durham,NC 27707

Email: ian.hembry@gmail.com

This dissertation was typed by Ian Fredrick Hembry.