December 2017

Evaluating Item Selection Methods for Adaptive Tests with Complex Content Constraints

Logan Rome
University of Wisconsin-Milwaukee

EVALUATING ITEM SELECTION METHODS FOR ADAPTIVE TESTS WITH COMPLEX CONTENT CONSTRAINTS

by

Logan Rome

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

The University of Wisconsin-Milwaukee

December 2017

ABSTRACT

EVALUATING ITEM SELECTION METHODS FOR ADAPTIVE TESTS WITH COMPLEX CONTENT CONSTRAINTS

by

Logan Rome

The University of Wisconsin-Milwaukee, 2017
Under the Supervision of Professor Bo Zhang

Adaptive testing designs have become go-to methods for large-scale test administration due to their ability to provide more accurate scores with fewer items. In recent years, new designs have been introduced, such as on-the-fly multistage testing (OMST), that combine the advantages of the well-established computerized adaptive testing (CAT) and multistage testing (MST) designs. While adaptive testing has attracted a tremendous amount of research, most studies have used only one set of test specifications to constrain the content of the test. Through Monte Carlo simulation, this study evaluated the effectiveness of CAT, MST, and OMST under varying levels of test specification complexity. Specifically, the constrained item selection methods of the maximum priority index (MPI) and weighted penalty model (WPM) were examined in CAT and OMST while the normalized weighted absolute deviation heuristic (NWADH) was used to assemble MST forms. In addition to the complexity of the test specifications, the representation of each content category in the pool and on the test, size of the item pool, length of each stage, and number of preassembled MST difficulty levels were also varied. The performance of each test design was evaluated by three outcomes: content alignment, measurement precision, and test security. Results show that increasing the complexity of test specifications leads to worse content alignment across all test designs and item selection methods. The WPM item selection method performs better than the MPI and NWADH under

increased constraint complexity. Moreover, CAT and OMST provide higher measurement precision than MST, especially for the large item pool. Finally, CAT is the most secure among the three test designs and the security of MST benefits most from the larger item pool.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Over the last several decades, adaptive testing designs, such as computerized adaptive testing (CAT; Lord, 1971b) and multistage testing (MST; Lord, 1971a), have arisen as mainstream methods for large-scale test administration. These designs adjust the difficulty of the test to the ability of the examinee during test administration. Consequently, compared to the traditional paper-and-pencil linear tests, adaptive tests can provide more precise measurement with fewer items (Stocking, 1994). The traditional CAT is a fully-sequential adaptive design in that items are selected one-at-a-time and ability is estimated after each item. On the other hand, MST is a group-sequential adaptive design where sets of items, known as modules, are preassembled at target ability levels and the examinee is routed to the next module based on the ability estimate obtained from responses to the previous module(s).

Both CAT and MST have been successfully implemented in large-scale assessment. Over time, some notable drawbacks of each design have come to light. In CAT, early item responses lead to large changes in estimated ability, as little is initially known about the examinee. Later in the test, changes in estimated ability from one item to the next become smaller. This attribute of CAT makes it difficult for high-ability test takers to recover from early mistakes (Rulison & Loken, 2009). MST is less prone to this issue, as the initial ability estimate is delayed until after a set of items has been completed. As a tradeoff, final ability estimates in MST are often not as precise as those in CAT, as MST modules are designed to be of optimal difficulty only at a limited number of target ability levels (e.g., three levels at low, medium, and high ability). For an examinee whose ability falls between any two target levels (e.g., between low and medium),

difficulty of the modules will not be optimal, and subsequently, ability estimation will not be as accurate as in the CAT design.

To address these issues, researchers have continued to develop new adaptive testing designs. Han and Guo (2014) introduced MST by shaping (MST-S) while Zheng and Chang (2015) proposed "on-the-fly" MST (OMST). Both methods utilize a group-sequential design similar to MST, except that the items are selected during administration, as in CAT. Thus, MST-S and OMST represent a compromise between CAT and MST. These new methods still possess many of the advantages of MST but with the additional benefit that final ability estimates can be nearly as precise as CAT. While MST-S and OMST present a promising new direction for adaptive testing, they are relatively new, and more research needs to be done to determine their performance in various testing situations.

Together, CAT, MST, MST-S, and OMST present testing organizations with a myriad of options to achieve precise ability estimation efficiently. However, challenges still exist. For instance, inherent in adaptive testing is a large number of unique test forms. With as many as one unique form per examinee, ensuring that all test forms are equivalent in terms of content can be challenging. Wise, Kingsbury, and Webb (2015) contend that the degree of content alignment for an adaptive test is related to the extent that the test items (1) present an optimal challenge for the examinee, and (2) represent the desired content domain. With respect to the first goal, matching the difficulty of the test to the ability of the examinee is central to adaptive testing. This goal, on its own, can be met in CAT, MST-S, and OMST, and to a somewhat lesser extent in MST. The second goal can be readily accomplished when test forms are created and closely examined before administration, as in linear testing. The challenge in adaptive testing then becomes meeting both goals simultaneously in a test form that is created during administration.

The key to meeting the content alignment standards of adaptive testing lies in the item selection algorithm. Methods that consider item content, in addition to item statistical properties (i.e., information), have been developed for linear testing and preassembled MST (Swanson & Stocking, 1993; Luecht, 1998) as well as CAT (Cheng & Chang, 2009; Shin, Chien, Way, & Swanson, 2009) and OMST (Zheng & Chang, 2015). While these methods have been shown to be effective in many testing situations (He, Diao, & Hauser, 2014), they have not yet been studied for tests with complex content specifications. One example of such constraints comes from the Programme for International Student Assessment (PISA). Its mathematics test uses four indices – content, cognitive process, context, and format type – for each item (OECD, 2012). Each of these categories has 3 or 4 levels and the levels of each category are not exclusive (i.e., items from each content area could be of any cognitive process, context, and format type). Ensuring that each of the levels of each category is adequately represented on every test while also selecting items of optimal difficulty for the examinee can be extremely challenging.

Cheng and Chang (2009) introduced the maximum priority index (MPI) as an item selection method for CAT, which has since been extended to OMST (Zheng & Chang, 2015). The MPI calculates a priority index for each item in the pool based on item content and statistical characteristics. The item with the highest priority index is then selected for administration at each step. The weighted penalty model (WPM; Shin et al., 2009) and normalized weighted absolute deviation heuristic (NWADH; Luecht, 1998) use similar logic to consider both statistical and non-statistical attributes. The WPM also considers the prevalence of each content area in the item pool in order to account for the quality of the pool while the NWADH aims to assemble multiple test forms that are equivalent in terms of both content and statistical properties. So far, the WPM has only been applied to CAT while the NWADH has been used to

select items for both linear tests and MSTs. All three methods show potential for assembling tests with very complex content constraints, due to their ability to accommodate situations where items have multiple content indices.

While originally proposed as item selection methods for CAT, the MPI and WPM can be applied to OMST (as in Zheng & Chang, 2015). MST-S, on the other hand, does not use an index to select items; instead, a fixed number of items are randomly selected from each content area at each stage. This random selection process is repeated a predetermined number of times in order to achieve a desired level of measurement precision and item exposure control. So far, MST-S has only been studied for tests with simple test blueprints (Han & Guo, 2014), as the random nature of MST-S makes it challenging to consider multiple content indices at once.

The main goal of this study is to investigate the effectiveness of item selection methods for adaptive tests with varying levels of test specification complexity. The following five combinations of item selection method and test design will be studied: MPI and WPM for CAT, NWADH for MST, and MPI and WPM for on-the-fly MST. Evaluation of these methods will be based on the following three criteria: accuracy of ability estimation, satisfaction of test content constraints, and item exposure and test overlap rates. The importance of accurate ability estimation is self-evident. Many score-based decisions depend on the accuracy of latent trait measurement. Satisfaction of test constraints is directly related to the content validity of test scores. Violations of the constraints make the test scores invalid for the target construct and thus difficult to compare across examinees. Item exposure and test overlap rates are test security concerns. Overexposed items and high test overlap rates may result in a testing program that is vulnerable to compromised items due to question sharing between examinees. These three criteria are clearly related and tradeoffs will have to be made among them. For instance,

increasing ability estimation accuracy is likely to come at the expense of item exposure control, as the best items will be administered more frequently.

The effectiveness of the competing item selection methods may vary by the features of the item pool and test design; hence, these features will be closely examined in this study. Specifically, the size of the item pool, complexity of the test blueprint, representation of each content category in the item pool, number of items in each MST stage, and number of difficulty levels in each preassembled MST stage may all play a role.

CHAPTER 2

LITERATURE REVIEW

Item Response Theory

Item Response Theory (IRT) has been the dominant model in large-scale testing since at least the 1970s (Hambleton & Swaminathan, 1985). Different from Classical Test Theory (CTT), which uses number-correct scoring to produce scores that are dependent on the particular set of items included on the test (van der Linden, 1986), IRT focuses on modeling the response probabilities to individual items. Examinee abilities are scored based on the probability of the response pattern instead of the number of correct responses. IRT has many advantages over CTT, such as latent trait estimation that is not dependent on the test and item parameter calibration that is not dependent on the sample.

Dichotomous IRT models. The dichotomous IRT models aim to predict the probability of a correct response to an item. The three-parameter logistic (3PL) model, the most general form, can be expressed as (Birnbaum, in Lord & Novick, 1968):

$$P_{ij}\left(X_{ij}=1|\theta_i\right)=c_j+(1-c_j)\frac{1}{1+\exp\{-D\,a_j(\theta_i-b_j)\}} \quad (1)$$

The outcome, $P_{ij}(X_{ij}=1|\theta_i)$ or p_{ij}, is the conditional probability of a correct response to item j by examinee i ($X_{ij}=1$), given the examinee ability parameter, θ_i, and item parameters a_j, b_j, and c_j. D is a scaling constant used to approximate the normal ogive function by the logistic function, and is usually set equal to 1.702. The probability of an incorrect response, $P_{ij}(X_{ij}=0|\theta_i)$ or q_{ij}, is simply $1-p_{ij}$.

In Equation (1), the item difficulty parameter, b_j, represents the point on the ability (θ) continuum at which the probability of a correct response is 0.5. The item discrimination parameter, a_j, measures how well the item discriminates between examinees of different ability

levels and is related to the maximal slope of the item response function. The guessing parameter, c_j, represents the probability of a correct response for an examinee of very low ability (i.e., θ_i approaching $-\infty$) (de Ayala, 2009). If $c_j = 0$, the 3PL model reduces to the two-parameter logistic (2PL) model. Further nested IRT models are the one-parameter logistic (1PL) model, in which a_j is restricted to be equal across all items, and the Rasch model, a special case of the 1PL model where $a_j = 1$ for all items.

The item response function can be examined visually using the item characteristic curve (ICC). Figure 1 shows the ICCs for three example items.

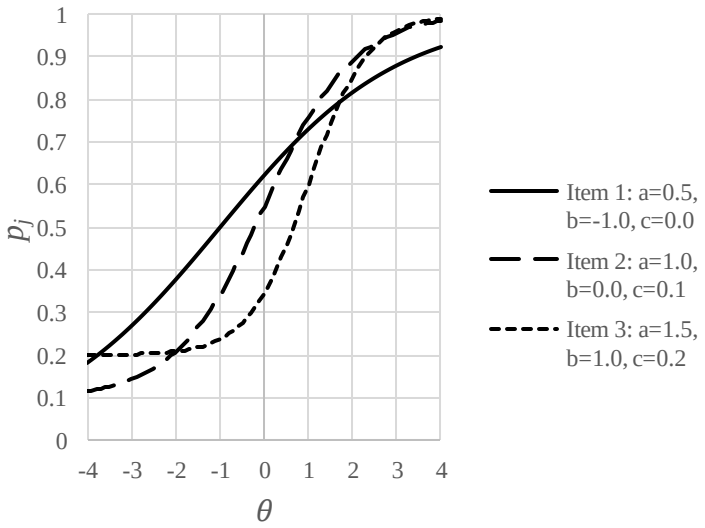

Figure 1. Item Characteristic Curves for three items with varying parameters.

Figure 1 demonstrates how the item parameters impact the predicted response probabilities. All three items differ in difficulty, as can be seen by the location on the θ continuum of the inflection points of the curves. Item 1 is the easiest item, so at $p_j = 0.5$ the curve for this item is further to the left than that of the other items. The items also differ in their discriminating power; this is evidenced by the slope of the curves (de Ayala, 2009). More discriminating items, or items with higher values of a_j, have ICCs that are steeper near the inflection point. Finally, differences in

the guessing parameters can be seen by examining the lower asymptote. The ICC for Item 3 is nearly flat around $p_j = c_j = 0.2$, meaning that even very low ability examinees have a nonzero chance of answering the item correctly by guessing. It should be noted that the presence of a nonzero guessing parameter shifts the entire ICC upward. Thus, the probability of a correct response at $\theta = b_j$ under the 3PL model is not 0.5, but instead can be computed by $\frac{(1+c_j)}{2}$.

Assumptions. IRT models carry strong assumptions. First, traditional IRT models assume unidimensionality, which states that all items measure only one latent trait. While it might seem impossible for this assumption to be met in practice, due to nuisance factors such as motivation or test-taking skill, this assumption does not need to be met strictly. Generally, it is instead required that there exists one "dominant" trait that accounts for test performance more so than any other trait (Hambleton & Swaminathan, 1985).

The second assumption is local independence, which requires that the response of an examinee to any given test item be independent of all other item responses in the test for any examinee (Birnbaum, in Lord & Novick, 1968). Local independence will be violated when the responses to two or more items are still related after accounting for the target ability. This may occur in situations where several items are related to a common stimulus or responses to later items are made based on responses to earlier items (de Ayala, 2009).

Another important assumption is monotonicity. This assumption requires that the ICC is monotonically increasing and somewhat S-shaped (Hambleton & Swaminathan, 1985). Monotonicity is important as it demonstrates that the latent trait is being measured by the item(s). If an examinee has a higher value of θ, they should have a higher probability of answering the item correctly.

In general, the form of the ICC should be close to what is specified by the model. The 3PL model thus provides the most relaxed assumptions; items may vary in their difficulty, discrimination, and guessing parameters. For the 2PL model, items may vary in difficulty and discrimination, but should possess a common guessing parameter of 0. Finally, the 1PL and Rasch models have the most stringent assumptions; items must have a guessing parameter of 0 and equal discrimination parameters.

When the above assumptions are met, IRT has the properties of sample-free calibration and test-free measurement. Sample-free calibration means that the values of the item parameters do not depend on the sample of examinees used to calibrate the parameters (Rupp & Zumbo, 2006). Thus, item parameters are invariant across test-takers. The property of test-free measurement indicates that examinee ability estimates do not depend on the particular set of items administered (Hambleton & Swaminathan, 1985). Therefore, unlike in CTT, examinees who respond to different sets of items can still be given comparable scores. These properties are extremely important in adaptive testing, where item parameters are treated as known and examinees typically see different test forms.

IRT scoring. Ability estimation can be accomplished using maximum likelihood (ML) methods. ML estimation aims to find the model parameters that are most likely to have produced the observed responses. Given local independence, the likelihood of a response pattern is simply the product of the conditional probability of each item response (Hambleton & Swaminathan, 1985):

$$P(u_1, u_2, \ldots, u_J|\theta) = \prod_{j=1}^{J} p_j^{u_j} q_j^{1-u_j} \quad (2)$$

where u_j is the response to item j and $p_j^{u_j} q_j^{1-u_j}$ is the Bernoulli distribution for the probability of an item response. For a correct response, $u_j = 1$ and $p_j^{u_j} q_j^{1-u_j}$ simplifies to p_j. For an incorrect response, $u_j = 0$ and $p_j^{u_j} q_j^{1-u_j}$ becomes q_j.

The probability in Equation (2) is conditional on θ, meaning each unique value of θ will result in a different likelihood for the response pattern. The ML ability estimate is the value of θ that maximizes the likelihood. One standard method for obtaining the estimate is to set the first derivative of the log of the likelihood function equal to zero and solve for θ. As the form of this derivative is irregular, numerical methods, such as the Newton-Raphson, are often applied (Hambleton & Swaminathan, 1985).

ML estimation can be enhanced by the Bayesian approach that utilizes prior information about the ability distribution in addition to the likelihood function of the response pattern. Specifically, Bayesian methods multiply the likelihood of the response pattern, given θ, by the prior distribution of θ to obtain the posterior density of θ. This is expressed as:

$$f(\theta|u) = f(u|\theta)f(\theta)/f(u) \tag{3}$$

Here $f(\theta|u)$ is the posterior density of θ, $f(\theta)$ is the prior distribution, $f(u)$ is the marginal distribution, and $f(u|\theta)$ is equivalent to $P(u_1, u_2, \ldots, u_J|\theta)$ in Equation (2). The estimated a posteriori (EAP) and maximum a posteriori (MAP) estimators, defined as the mean and mode of the posterior distribution, respectively, are commonly used ability estimators in IRT (Hambleton & Swaminathan, 1985).

Bayesian estimation has the distinct advantage of being able to provide an estimate no matter the response pattern. ML estimation will not find a solution if the response pattern is non-mixed (i.e., all 0s or all 1s), as the likelihood function will not have a maximum. A constant

concern with Bayesian estimation, however, is the accuracy of the prior information. In general, research has shown that differences between ML and Bayesian estimates are negligible when items are well-matched to examinee ability, as is the goal in adaptive testing (Wang & Vispoel, 1998; Kim, Moses, & Yoo, 2015).

Information and standard error. Under CTT, measurement accuracy is represented by reliability and the standard error of measurement at the test level. Thus, it is assumed that all examinees are measured to the same degree of accuracy, regardless of ability level. This is rarely true in practice. IRT, on the other hand, provides more localized estimates of measurement error in the form of test information and the standard error of estimate (Embretson, 1996).

Information can be calculated along the θ continuum at both the item and test levels by taking the second derivative of the likelihood function with respect to θ. The formulas for computing item and test information under the 3PL model, given estimated ability $\hat{\theta}_i$, are given in Equations (4) and (5), respectively.

$$I_j(\hat{\theta}_i) = a_j^2 \left[\frac{(p_{ij}-c_j)^2}{(1-c_j)^2}\right]\left[\frac{q_{ij}}{p_{ij}}\right] \quad (4)$$

$$I(\hat{\theta}_i) = \sum_{j=1}^{J} I_j(\hat{\theta}_i) \quad (5)$$

In Equation (4), item information, $I_j(\hat{\theta}_i)$, is calculated using the probabilities of a correct and incorrect response, p_{ij} and q_{ij}, and item parameters a_j and c_j. Equation (5) shows that the test information, $I(\hat{\theta}_i)$, is simply the sum of item information. Item and test information can also be examined visually, as shown in Figure 2.

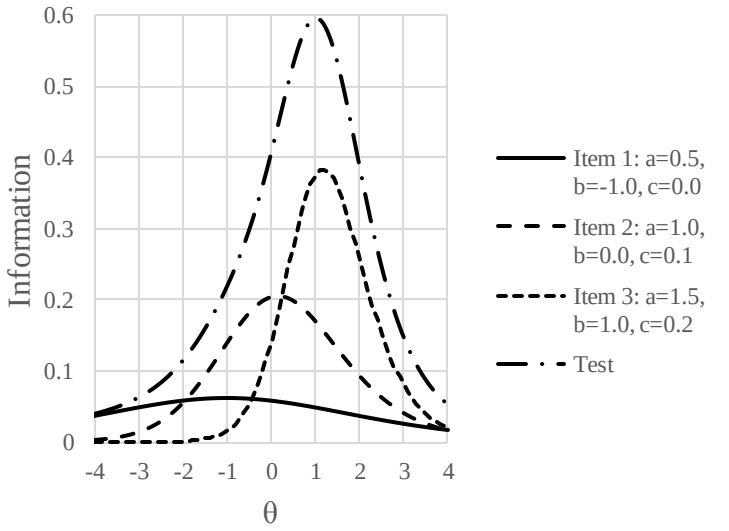

Figure 2. Item and test information curves for three items with varying parameters.

Figure 2 clearly shows that item information peaks near the item difficulty parameter while the discrimination parameter determines the amount of information. Guessing introduces noise into the measurement process, thus reducing information (de Ayala, 2009). Accordingly, one way to effectively increase test information is to add items with high discriminating power and difficulty near the examinee's ability level.

The IRT equivalent of the standard error of measurement is the standard error of estimate, $\sigma_e(\hat{\theta}_i)$. Much like the standard error of measurement, $\sigma_e(\hat{\theta}_i)$ represents the uncertainty associated with the ability estimate and can be used to build confidence intervals for $\hat{\theta}_i$. Equation (6) shows the relationship between test information and the standard error of estimate (Hambleton & Swaminathan, 1985).

$$\sigma_e(\hat{\theta}_i) = \frac{1}{\sqrt{I(\hat{\theta}_i)}} \tag{6}$$

As $\sigma_e(\hat{\theta}_i)$ is inversely related to test information, the more information that a test provides at θ, the more certain one is about the ability of examinees at θ. This relationship between measurement uncertainty and information is critical to item selection in adaptive testing.

Computerized Adaptive Testing

Computerized adaptive testing (CAT) aims to select and administer only the most appropriate items for each examinee (Parshall et al., 2002). CAT was originally conceptualized by Lord (1971b) as a method to tailor the test to the examinee by administering items whose difficulties are closely matched to examinee ability. Thanks to increases in computing power, a myriad of options are currently available for CAT administration. Unique design issues, such as the response model, item pool attributes, ability estimation and item selection methods, starting point, and stopping criterion, must be considered when developing a CAT (Weiss & Kingsbury, 1984).

In a CAT administration, items are selected sequentially in a process that can be described in the following steps:

1. Administer the first item from the item pool.
2. Estimate examinee ability based on all item responses.
3. Use the provisional ability estimate to select the best item from the item pool.
4. Administer the item selected in step 3.
5. Repeat steps 2 through 4 until a preset stopping criterion has been reached.

In step 1, the first item can be chosen using the mean of a proposed ability distribution (Mills & Stocking, 1996) or some known information about the examinee (Weiss & Kingsbury, 1984). One can also start the test by simply selecting a relatively easy item to reduce test anxiety (Wainer & Kiely, 1987). Both ML and Bayesian methods can then be used to estimate ability (Wang & Vispoel, 1998). Bayesian methods are typically used at least until a mixed response vector is obtained. In step 3, several algorithms exist for identifying the "best" item. Traditionally, when test specifications and item exposure are not of concern, the "best" item is

the item with the highest information at the provisional ability estimate. Algorithms that consider more than just item statistical properties will be discussed in great detail later. Finally, the stopping criterion can be a fixed number of items, which guarantees an equal test length for all test-takers, an acceptable standard error of estimate, which ensures equal measurement precision for all examinees (Weiss & Kingsubry, 1984), or simply a fixed testing time.

While extremely popular for large-scale testing (Chang, 2015), CAT has received its fair share of criticism. One disadvantage is that ability estimation may be inaccurate at early points in the test, when little information is known about the examinee. These errors in estimation are compounded by the fact that the item selection method depends on the provisional ability estimate. Another downside of CAT is the infeasibility of test form review. In testing, forms are typically reviewed to ensure that test specifications are met and that undesirable characteristics, such as item order or context effects, are not present (Wainer & Kiely, 1987). This is not possible in CAT, as forms are assembled during administration and most examinees will see very different sets of items, resulting in a large number of unique forms. Finally, examinees are not able to skip items or modify answers to earlier items (Parshall et al., 2002). The issues presented here arise because of the fully-sequential nature of CAT, and can be addressed by a group-sequential adaptive design.

Multistage Testing

Lord (1971a) proposed a two-stage testing design that has since been expanded upon by researchers (e.g., Wainer & Kiely, 1987; Kim & Plake, 1993) and become known as multistage testing (MST). MST adapts in stages, such that ability is estimated only after a set of items has been administered and the next set of items is chosen based on this estimate. In this sense, MST can be considered a compromise between CAT and linear testing, in which all examinees

respond to the same or equivalent test forms. MST utilizes the advantage of tailored testing, adjusting test difficulty to match examinee ability, while also allowing for test form review. These advantages have prompted some testing programs, such as the Graduate Record Examination (GRE), to move completely from CAT to MST (Zheng & Chang, 2014).

In MST, the item sets of differing difficulty levels within each stage are referred to as *modules*. The basic design of an MST can be simply described by the number of stages and the number of modules at each stage. Figure 3 shows a 1-3-3 MST design; that is, a 3-stage MST with one difficulty level in stage 1, and three difficulty levels in both stages 2 and 3.

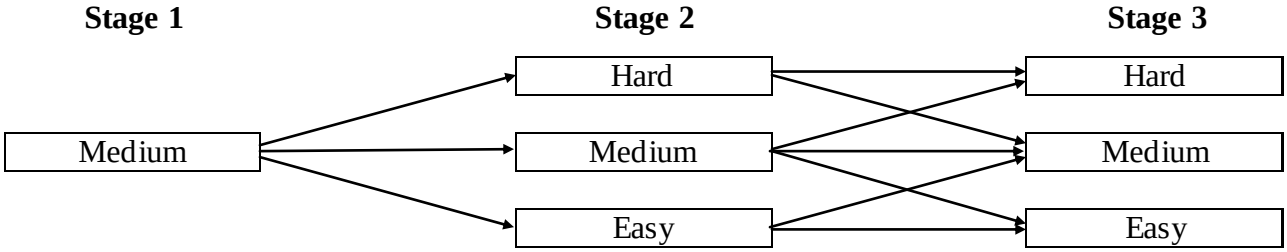

Figure 3. Three stage 1-3-3 MST design.

Each box in Figure 3 represents a module and the arrows show the possible routes that an examinee can take through the test. Some routes are not permitted; for example, there is no path moving from the hard module in stage 2 to the easy module in stage 3. Such a path would have indicated an aberrant response pattern. Each route in Figure 3 is called a *pathway*. Multiple parallel test forms are usually assembled for each pathway, where each form is called a *panel*. Typically, a panel is randomly selected before the first stage (Zheng, Nozawa, Gao, & Chang, 2012). This random assignment helps to ensure even exposure of items in the bank, thus increasing test security. However, since modules at the later stages are chosen based on the provisional ability estimate, examinees of similar ability assigned to the same panel will likely see the same items, increasing the test overlap rate.

MSTs usually begin with a module of medium difficulty, as shown in Figure 3 (stage 1). After the first module, also known as a *routing test*, examinees are assigned to the next module using either number-correct or IRT scoring (Weissman, Belov, & Armstrong, 2007). Typically, routing is accomplished by setting cut points, either by finding the point where the two adjacent module information curves cross (e.g., Zheng et al., 2012) or by using assumptions about the ability distribution to route a certain percentage of examinees to each module (e.g., Jodoin, Zenisky, & Hambleton, 2006). Using the crossing point of the module information curves often results in more precise measurement, as this is akin to choosing the most informative module for the examinee, while routing based on the ability distribution allows for better test security, as each module can be exposed to a set proportion of examinees. Examinees with $\hat{\theta}_i$ (or number-correct score) below the first cut point, θ_1, are routed to the easiest module while examinees with $\theta_1 \leq \hat{\theta}_i < \theta_2$ receive the second easiest module, and so on.

MST presents many advantages over both CAT and linear testing. Compared to CAT, provisional ability estimates are more accurate at early stages, as more items are administered between each estimation point. Second, MST forms can be preassembled and each possible pathway can be carefully reviewed with context and item order effects in mind (Wainer & Kiely, 1987). Third, the MST design allows examinees to skip and review items within a stage (Zheng et al., 2012). Finally, since MST is still adaptive, it provides more precise ability estimation than linear testing. Compared to CAT, one obvious disadvantage of MST lies in having fewer adaptation points. While CAT adapts after each item, MST adapts only after each stage. Also, as each module maximizes information at only one θ point, examinees whose abilities are far from the target abilities will receive modules that are not of ideal difficulty. This mismatch reduces the accuracy of final ability estimation.

On-the-fly MST. Two methods have been proposed to increase the measurement precision of MST: MST by shaping (MST-S; Han & Guo, 2014) and "on-the-fly" MST (OMST; Zheng & Chang, 2015). Like MST, MST-S and OMST are administered in stages and examinee ability is estimated only after the completion of each stage. However, in MST-S and OMST, there are no panels, no preassembled modules at fixed difficulty levels, and no routing rules. Instead, items within each stage are chosen during administration, based on the provisional ability estimate. MST-S accomplishes this by randomly selecting items iteratively for inclusion in the next stage. The set of items that minimizes the distance from the target information value is then chosen for administration. OMST, on the other hand, utilizes sequential item selection methods developed for CAT to build MST stages on-the-fly. Both methods have been shown to result in measurement precision close to that of CAT and considerably better than MST (Han & Guo, 2014; Zheng & Chang, 2015).

Test Specifications

Over the last two decades, educational policy, such as No Child Left Behind and Every Student Succeeds, has focused on holding schools accountable via assessments that are aligned to certain content standards. This alignment between educational standards and test content is critical to ensuring that inferences made from test scores are valid. Webb (2006) described four criteria that can be used to judge the alignment of an assessment. Categorical concurrence describes the degree to which topics or categories (e.g., algebra, geometry) within the broader category (e.g., mathematics) are represented both in the standards and on the test. Depth-of-knowledge relates to the cognitive demand, or complexity, of what students are required to do. Finally, range-of-knowledge and balance of representation refer to the span of knowledge required and the distribution of topics on the test, respectively. As they relate to the test content

specifications, the first two criteria describe what levels of content and complexity are to be assessed while the last two criteria define the distribution of these levels across the test. Evaluation of these criteria are based on judgments made by subject matter experts and are not the same as item statistical properties, which are usually based on the actual testing data.

Content specifications are one aspect of the overall test specifications, which may also include item format, context, or other traits important to the goal of measurement (Webb, 2006). The complexity of test specifications can vary greatly by assessment. For example, the test blueprint for the National Assessment of Educational Progress (NAEP) reading assessment includes two levels of passage type and three levels of cognitive targets (National Assessment Governing Board, 2015b). In comparison, the NAEP mathematics assessment specifies five levels of item content, three levels of cognitive complexity, and two levels of item format (National Assessment Governing Board, 2015a). Thus, the specifications for the mathematics assessment are much more complex than that of the reading assessment, even within the same testing program.

In adaptive tests, content alignment is defined by the agreement between the ability of the examinee and the difficulty of the test form, as well as the representation of the desired content domain (Wise et al., 2015). While the adaptive designs outlined previously were created with the intention of tailoring the test difficulty to match examinee ability, representation of the content domain is not inherent in these designs. That is to say, features like categorical concurrence, depth-of-knowledge, and balance of representation are not explicitly addressed. Additionally, further test specifications, such as item context or format, are also not automatically controlled by the test design. If test forms are to be created during administration, the item selection

algorithm will need to ensure that each test is aligned to the content standards and other criteria expressed by the test blueprint.

Item Selection in Adaptive Testing

The item selection algorithm arguably plays the most important role in adaptive testing. This algorithm must balance three elements: measurement precision, test specifications, and item exposure. Unfortunately, these often work against one another. For instance, high measurement precision requires selecting highly informative items, but repeated selection of those items will lead to their overexposure. Additionally, ignoring content specifications may result in tests that differ in content validity (Mills & Stocking, 1996; Wise et al., 2015). Thus, the goal of the item selection algorithm is threefold: to achieve maximum measurement precision, to satisfy test specifications, and to reduce item overexposure and test overlap.

In early versions of adaptive testing, item selection focused only on item information while content balancing was seen as a fairly simple problem. Kingsbury and Zara (1989) described a mathematics test where addition and subtraction problems were required to make up 30% of the test each while the remaining 40% was divided equally between multiplication and division items. The authors' proposed solution was simply to select the maximally informative item from the content category that was furthest from meeting its desired percentage. However, as outlined previously, test blueprints often categorize items by more than just the content area. When more than one categorical label is assigned to each item, the simple methods proposed in early CAT studies (e.g., Kingsbury & Zara, 1989) will not suffice.

Item selection in tests with many constraints is typically accomplished using one of two methods: 0-1 linear programming and heuristics. Linear programming methods attempt to maximize information across the test, subject to the test constraints; constructing the entire test

form at once. The shadow test approach (van der Linden & Reese, 1998) is an example of a 0-1 linear programming approach that can be applied to adaptive testing. Heuristic methods, on the other hand, build the test one-item-at-a-time by treating test construction as a series of local optimization problems (Zheng & Chang, 2014). Unlike linear programming methods, heuristics do not attempt to find a perfect solution, but they are generally faster, computationally simpler, and will at the very least minimize constraint violations. When the test blueprint is complex, heuristics can be very useful, as linear programming methods may encounter infeasibility issues, where no test is created because no perfect solution can be found (Cheng & Chang, 2009). Accordingly, this study focuses on heuristic methods.

Maximum priority index. The maximum priority index (Cheng & Chang, 2009) combines the statistical and non-statistical attributes of test items by multiplying item information by a value that measures the item's contribution toward meeting the test constraints. The item that maximizes this product is chosen for administration. The priority index for item j, PI_j, given the provisional ability estimate $\hat{\theta}_i$, can be calculated as:

$$PI_j = I_j(\hat{\theta}_i) \prod_{c=1}^{C} (w_c f_c)^{r_{jc}} \quad (7)$$

Here, each constraint is represented by c and is dummy coded such that r_{jc} is 1 when constraint c is relevant for item j and 0 otherwise. The weights, w_c, are part of the test blueprint and are assigned based on the importance of each constraint, with larger weights associated with major content areas. Finally, for each constraint, f_c measures the proportion of the constraint that still needs to be met. This is calculated by:

$$f_c = (X_c - x_c)/X_c \quad (8)$$

where X_c represents the number of items required from constraint category c and x_c is the number of items administered from c so far.

Oftentimes, the test blueprint will specify a lower (l_c) and upper (u_c) bound for each constraint. These bounds represent the minimum and maximum number of items allowed from category c. In these cases, the MPI requires a two-phase selection procedure, where phase one focuses on meeting the lower bounds and phase two tries not to exceed the upper bounds. In phase one, items may be selected from content area c until $x_c = l_c$, or $f_c = 0$. Once all content areas have satisfied their lower bounds (all $f_c = 0$), phase one ends. In phase two, items can be selected from content area c until $x_c = u_c$. The two-phase MPI ensures that all lower bounds will be met as long as the test length is sufficient and upper bounds will not be exceeded unless the test is too long (Cheng & Chang, 2009).

The MPI presents several options for controlling item exposure. Cheng and Chang (2009) suggested specifying a desired exposure rate as a constraint. To do this, X_c in Equation (8) is replaced by the desired exposure rate and x_c is updated to represent the current exposure rate, which is the number of times item j has been administered divided by the total number of tests. Items with current exposure rates higher than the desired rate will have negative values of f_c, making them unlikely to be selected. He et al. (2014) applied a randomesque method similar to that of McBride and Martin (1983) in which the administered item is chosen randomly from a group of items with the highest PI_j. All items in the group, including the unselected items, are then eliminated from the pool for the remainder of the test. Introducing randomness into the selection process ensures that the "best" item is not selected every time, reducing the chance of item overexposure.

The MPI is a very straightforward item selection method that considers all three aspects of item selection. Measurement precision is addressed through the presence of item information, deviations from the content constraints consider the test specifications, and exposure control can

be incorporated as outlined above. Cheng and Chang (2009) likened the MPI to a simple modification of the maximum information method that instead considers the overall "attractiveness" of the item in terms of both statistical and non-statistical properties. The content weights, w_c, can be adjusted to control the scale of the priority index. That is, larger weights can be used if test specifications are deemed to be more important than item information.

Weighted penalty model. Similar to the MPI, the weighted penalty model (Shin et al., 2009) assigns a unique penalty value to each item at each selection point. The item with the smallest penalty value is then selected for administration. The penalty value for item j, F_j, is calculated as:

$$F_j = w' F_j' + w'' F_j'' \tag{9}$$

where F_j' and F_j'' represent the standardized penalty values for item content and information, respectively. The weights associated with content and information, w' and w'', control the trade-off between non-statistical and statistical item properties and can be updated throughout the test. Shin et al. suggested changing the information weight throughout the test based on a logistic or quadratic function so that item selection focuses on meeting test constraints early in the test before giving larger weight to information near the end.

The constraint penalty value, F_j', is computed in five steps. The first step is to compute Pr_c, the proportion of items from category c that would be administered by the end of the test if items from c were selected in proportion to their prevalence in the remaining item pool. This can be written as:

$$Pr_c = [x_c + Prv_c(J - x)]/J \tag{10}$$

Here, x and x_c are the number of items administered so far and the number of items administered from constraint category c so far, respectively. The prevalence, Prv_c, is the proportion of items from c in the complete item pool and J is the test length (Shin et al., 2009).

Next, the difference between the projected proportion of items to be administered from c and the midpoint of the lower and upper bounds is calculated by:

$$D_c = Pr_c - m_c \tag{11}$$

where m_c is the midpoint between the lower and upper bounds. For the WPM, the lower and upper bounds, l_c and u_c, are expressed as proportions; thus, m_c represents the midpoint between the lowest and highest acceptable *proportion* of test items from category c. The deviation from the midpoint, D_c, is then used in one of Equations (12) through (14) to calculate P_c, the penalty value specific to category c.

$$\text{If } Pr_c < l_c \text{ then } P_c = \frac{1}{2(l_c - m_c)} D_c^2 + \frac{l_c - m_c}{2} \tag{12}$$

$$\text{If } Pr_c \geq u_c \text{ then } P_c = \frac{1}{2(u_c - m_c)} D_c^2 + \frac{u_c - m_c}{2} \tag{13}$$

$$\text{If } l_c \leq Pr_c < u_c \text{ then } P_c = D_c \tag{14}$$

Shin, Chien, and Way (2012) defined P_c as the quadratic distance between the number of items administered from c so far and the midpoint of the lower and upper bounds of c. It can be seen that when $Pr_c < l_c$, P_c will be lower, as $l_c - m_c$ is always negative. On the other hand, categories with $Pr_c \geq u_c$ will have positive values of P_c. Thus, penalty values are lower for items from categories that may not meet their lower bounds and higher for items from categories that may exceed their upper bounds.

In the third step, the total content penalty value for item j is calculated using the content-specific penalties and weights, P_c and w_c, respectively, and the dummy code for item j belonging to category c, r_{jc}:

$$P_j = \sum_{c=1}^{C} P_c w_c r_{jc} \tag{15}$$

Here, P_j is the unstandardized content penalty value and consists of the product of the content penalty value and its associated weight, summed across all content categories relevant to item j. The weights, w_c, are defined by the test blueprint, as in the MPI (Equation (7)). The final step for computing F_j' is to standardize the penalty value using the minimum ($min(P_j)$) and maximum ($max(P_j)$) content penalty values across all items remaining in the pool (Shin et al., 2009):

$$F_j' = \frac{P_j - min(P_j)}{max(P_j) - min(P_j)} \tag{16}$$

Similarly, the standardized information penalty value, F_j'', is calculated using the information at $\hat{\theta}_i$ for item j and the maximum information at $\hat{\theta}_i$ across all items remaining in the pool, $max(I_j(\hat{\theta}_i))$:

$$F_j'' = -\left(\frac{I_j(\hat{\theta}_i)}{max(I_j(\hat{\theta}_i))}\right)^2 \tag{17}$$

Note that the standardized information value is multiplied by negative one. Thus, items with more information will have smaller (negative with larger absolute value) penalty values, making them more likely to be selected (Shin et al., 2009). After computing F_j' and F_j'', these values are substituted into Equation (9) to determine F_j. The item that minimizes F_j is chosen for administration.

The techniques used to control item exposure with the WPM are similar to those applied to the MPI. Shin et al. (2009) recommended using a conditional randomesque procedure where the item is randomly selected from a group of items and the group size varies based on the ability

estimate. Conditioning the item selection group size on ability, as proposed by Kingsbury and Zara (1989), accounts for the fact that the item pool may have more items available at some difficulty levels than others. An unconditional randomesque procedure (i.e., McBride & Martin, 1983) can also be used, or the desired exposure rate could be specified as a constraint, as recommended for the MPI by Cheng and Chang (2009).

The WPM is not as a simple and straightforward as the MPI. There are, however, some notable advantages to this method. First, the WPM uses the prevalence of the content area in the item pool to project the number of items that will be administered. It is this projection, rather than the current deviation from the bounds, that determines which items will be given more preference. This helps account for any differences in the representation of each content category in the item pool. Second, since the WPM uses the quadratic distance from the midpoint (D_c^2 in Equations (12) and (13)), a relatively larger penalty is assigned to items that violate content constraints and more preference is given to items that do not (Shin et al. 2012). Finally, the WPM standardizes both the content and information penalty values. Thus, these values are on a similar scale and the content and information weights, w' and w'', can more readily be manipulated to control the tradeoff between statistical and non-statistical attributes.

Both the MPI and WPM have performed well in CATs with simple constraints. They are capable of meeting test specifications and minimizing item overexposure with minimal sacrifices in measurement precision. He et al. (2014) showed that the WPM was able to meet constraints more consistently than the MPI. The two methods did not differ in measurement precision or exposure control. Computationally, the MPI is much simpler than the WPM. However, the WPM possesses many theoretical advantages over the MPI, as discussed above.

MST Module Assembly

The group-sequential nature of MST lends itself to several options for module assembly. Preassembly of MST forms allows for each module and pathway to be reviewed by content area experts and test specialists at the possible expense of measurement precision, as items may not be of optimal difficulty for all examinees. On-the-fly assembly, on the other hand, prioritizes measurement precision over test form review and thus relies more heavily on the item selection algorithm. Item selection methods for both preassembled and on-the-fly assembled MST are discussed next.

Preassembled MST. When modules are preassembled, MST design aspects, such as the number of stages, the number of difficulty levels within each stage, and the number of parallel panels, will greatly impact measurement precision and item exposure rates. In general, greater precision can be achieved by including more stages, or adaptation points. Designs with more than two stages are recommended, as this gives the test an opportunity to recover from any inappropriate routings that may occur after the initial stage (Zheng & Chang, 2014). Measurement precision can be greatly impacted by the number of difficulty levels within each stage. Research has shown that a maximum of four difficulty levels at the final stage is desired while three difficulty levels will usually suffice (Armstrong, Jones, Koppel, & Pashley, 2004). Aspects such as the number of items included in each stage can also impact measurement precision. Longer routing tests achieve more accurate routing with the tradeoff of decreased precision by including fewer items in later stages when more is known about the examinee (Kim & Plake, 1993). Finally, it can easily be seen that the number of stages, difficulty levels in each stage, and panels will all impact item exposure and test overlap rates; more stages, difficulty levels, and panels will lead to lower item exposure and test overlap. These features of the MST,

which need to be decided on before assembling the test forms, can be just as important as the item selection method.

Once the details of the MST design have been established, the test assembly method has three goals: (1) to make the information functions of different modules within a stage distinct enough to provide appropriate adaptation; (2) to make the information functions of corresponding pathways similar across all panels; and (3) to meet all test specifications in every pathway and panel (Zheng, Wang, Culbertson, & Chang, 2014). The first two goals concern the target information function (TIF) of the modules and panels, respectively, while the third goal considers the test blueprint. These goals can be met using either a *bottom-up* or *top-down* approach (Luecht & Nungester, 1998). In the bottom-up approach, module-level TIFs and constraints are specified and each module is assembled individually to meet these criteria. Thus, any combination of modules should result in a pathway that meets the test-wide TIF and content constraints. Top-down assembly, on the other hand, focuses only on test-wide TIFs and constraints and attempts to meet these criteria in each pathway and panel. While top-down assembly may be easier when constraints are specified at the test level, bottom-up assembly allows for modules to be mixed and matched, resulting in lower test overlap rates. Since heuristic item selection approaches focus on local optimization, these methods often utilize a bottom-up strategy (Zheng et al., 2012).

Normalized weighted absolute deviation heuristic. While the MPI and the WPM have not been applied to MST preassembly, the normalized weighted absolute deviation heuristic (Luecht, 1998), a similar method, has been used successfully (Zheng et al., 2012). The NWADH can be used to select items one-at-a-time for inclusion in a stage (Zheng & Chang, 2015). The k^{th} item selected for inclusion is the item that maximizes:

$$D_j = w_t e_{jt} + \sum_{c=1}^{C} w_c e_{jc} r_{jc} \quad (18)$$

where D_j is the normalized weighted absolute deviation. The normalized absolute deviation from the TIF is represented by e_{jt} while each e_{jc} represents the normalized absolute deviation from constraint c. Once again, r_{jc} is a dummy code representing whether or not item j belongs to c. The weights, w_t and w_c, are specified for the TIF and each content constraint, respectively (Luecht, 1998).

At each selection point, the normalized absolute deviation from the TIF is calculated for every item remaining in the pool as:

$$e_{jt} = 1 - \frac{d_{jt}}{\sum_{j \epsilon R_{k-1}} d_{jt}} \quad (19)$$

where $j \epsilon R_{k-1}$ represents all items in the pool except for the $k-1$ items already included on the test. The d_{jt} are computed as:

$$d_{jt} = \left| \frac{T(\theta) - \sum_{l=1}^{k-1} I_l(\theta)}{J-k+1} - I_j(\theta) \right| \quad (20)$$

where $T(\theta)$ represents the target test information at θ. Subtracted from T is the sum of the item information at θ for the $k-1$ items included on the test so far. The denominator is the number of items remaining on the test, where J is the total test length. Thus, d_{jt} represents the absolute deviation for item j from the average information required over the remaining test items (Luecht, 1998).

To compute e_{jc}, the normalized absolute deviation from constraint c, Luecht (1998) suggested assigning weights to each constraint, W_c, such that:

$$\text{if } x_c \geq u_c \text{ then } W_c = 0 \quad (21)$$

$$\text{if } l_c \leq x_c < u_c \text{ then } W_c = 1 \quad (22)$$

$$\text{if } x_c < l_c \text{ then } W_c = 2 \quad (23)$$

Note that these weights are not the same as the w_c used in Equation (18). Items that have not yet met their lower bounds will be given more weight than those that have and items that have met their upper bounds will be given no weight. A complement to W_c, $\underline{W}_c$, is then computed as:

$$\underline{W}_c = max\ (W_c) - \frac{1}{C}\sum_{c=1}^{C} W_c \tag{24}$$

where $max\ (W_c)$ is the maximum W_c out of the C constraints. Finally, the e_{jc} used in Equation (18) are calculated by first finding d_{jc}:

$$d_{jc} = r_{jc}W_c + (1 + r_{jc})\underline{W}_c \tag{25}$$

then normalizing to e_{jc} by:

$$e_{jc} = \frac{d_{jc}}{\sum_{j \epsilon R_{k-1}} d_{jc}} \tag{26}$$

The NWADH is similar to the MPI and the WPM in that item information and deviations from the content constraints are combined into one index. The differences between the NWADH and the other two heuristics reflect the differences between the goals of CAT item selection and those of MST assembly. First, rather than focusing on information at a provisional estimate, the NWADH computes information at predetermined target θs. This allows for adaptation in the preassembled test form. Second, the NWADH aims to minimize deviations from the target information function instead of simply maximizing information. This needs to be done to ensure that all modules within a stage and across panels have similar information functions. Finally, when the bottom-up assembly approach is used, the NWADH focuses on meeting content constraints at the module level, rather than across the entire test. Meeting constraints within each module helps to ensure that each pathway meets the test constraints.

The NWADH has been successfully applied to both linear (Luecht, 1998) and MST (Zheng et al., 2012; Zheng & Chang, 2015) assembly. Both MST studies used bottom-up

assembly. Zheng et al. (2012) found that backward assembly of MST modules, where later modules are assembled first, led to higher classification accuracy compared to forward assembly. This was attributed to the fact that the later stages are more complex, in that there are more modules of differing difficulty levels. Assembling these modules may require access to the full item pool. If assembled later, when the pool has shrunk considerably, estimation accuracy may suffer. Zheng and Chang (2015) found that MSTs assembled using the NWADH led to lower measurement precision than CAT and OMST with item selection via the MPI. This, however, is likely an effect of differences in the test design, rather than a deficiency in the test assembly heuristic.

On-the-fly MST. When MST stages are built on-the-fly, CAT item selection methods can be used to choose items for each stage. Zheng and Chang (2015) demonstrated this with the MPI. In their method, items are added to each stage one-at-a-time. Thus, the formulas are the same as those outlined in Equations (8) and (9); however, item information only needs to be calculated once for each item at each stage, since the entire stage is based on the same provisional ability estimate. The content constraint deviations (f_c), on the other hand, must be updated after each selection, due to the change in the number of items administered from the content area(s).

After a stage of items is selected, an item replacement step can be added where test specifications for the stage are evaluated and items are replaced as needed. Zheng and Chang (2015) outlined the steps for item replacement when a lower bound violation exists for the stage-specific content constraints. These steps are as follows:

1. For every lower bound violation of constraint c_1, identify a constraint c_2 that is above its lower bound.

2. Replace a randomly selected item from c_2 in the current set with the item from c_1 with maximum information at $\hat{\theta}_i$ in the item pool.
3. Evaluate the constraints for the current set of items and repeat steps 1 and 2 until a set has been found that meets all constraints.

By utilizing an item replacement step, it is guaranteed that every test will meet the constraints, provided that meeting all constraints is possible given the test or stage length, test blueprint, and item pool characteristics. Zheng and Chang reported zero constraint violations in their study, but their test specifications were very simple and the authors did not report how often the item replacement step was needed.

The MPI has been shown to result in similar measurement precision when applied to OMST, compared to CAT, while also minimizing constraint violations (Zheng & Chang, 2015). Research on OMST, however, has been limited to tests with very simple constraints. Zheng and Chang's (2015) study only required that one item be administered per stage from each of eight content areas. As each stage included 15 items, these constraints could be met fairly easily. It is not clear how item selection will work when each item belongs to several categories that must be constrained. Additionally, situations where constraints are specified only at the test level, where the number of constraints may exceed the number of items in each stage, have yet to be discussed. Finally, other heuristic item selection methods, such as the WPM, have not yet been applied to OMST.

MST by shaping. Han and Guo (2014) proposed MST by shaping, a different method for assembling MST stages during administration. This method aims to create stages that will help meet the test specifications and TIF. Specifically, after a stage of items has been completed, the provisional ability estimate and current test information at this estimate are calculated. Next, the

difference between the current information and the TIF is used to develop a *TIF mold*, which represents the ideal information function for the next stage. In the item selection step, the number of items required from each category is determined from the test specifications and items are randomly selected in accordance with these constraints.

After the initial set of items is selected, the difference between the information function for the current set of selected items and the TIF mold is calculated as:

$$A_s = \left| I(\theta)^* - \tau_{\theta_s} \right| \tag{27}$$

Here, $I(\theta)^*$ is the information at θ for the currently selected items and τ_{θ_s} is the information at θ for the TIF mold for stage s. After calculating A_s for the current set, the first item in the set is replaced with another random selection from the same content area and A_s is recalculated. If the new item leads to a decrease in A_s, this item is kept in the stage. If not, the new item is discarded from the pool for the current stage and the initial item is kept in the stage. This random item replacement is repeated for each item in the stage, and then the entire process is repeated for a fixed number of iterations. The set of items that comprises the stage after the final iteration is then administered to the examinee (Han & Guo, 2014).

MST-S attempts to meet test constraints by selecting the appropriate number of items from each content area in each stage. The tradeoff between measurement precision and item exposure is controlled by the number of iterations in the item selection, or shaping, process. For example, if the item selection process does not iterate, the resulting test will be a random set of items from each content area. Thus, measurement precision will be poor but item exposure will be ideal. On the other hand, if the number of iterations is 100, the resulting stages will likely feature items of near optimal difficulty for the examinee. However, the randomness of the

selection process will be greatly reduced and items with high a_j (i.e., high information) may be selected too frequently.

Results of Han and Guo's (2014) initial study on MST-S are promising. When the shaping process iterated only three times, the standard error of the resulting ability estimates were comparable to those in a preassembled MST. As the number of iterations increased, results approached the precision levels of CAT. In terms of item exposure, MST-S had more even exposure rates than both preassembled MST and CAT when up to six iterations were used in the shaping process. Predictably, as the number of iterations increased to 100, items with high a_j became overexposed. This overexposure, however, was still not as extreme as in CAT with maximum information item selection. MST-S is still very new; Han and Guo's study is the lone demonstration of this method. While this method is ideal for controlling the tradeoff between measurement precision and item exposure, MST-S can only be applied to tests with simple content constraints. When items are classified on more than one variable, the random selection required by MST-S does not allow for the consideration of multiple indices for each item. Thus, this method will not be examined in this study.

Research Questions

The main purpose of this study is to evaluate the performance of heuristic item selection methods on adaptive tests with various levels of test specification complexity. Specifically, the following research questions are of interest:

1. How does on-the-fly MST compare to preassembled MST and CAT on tests with complex constraints?
2. How do the different heuristic item selection methods compare within and between adaptive testing designs with complex constraints?

The test designs to be compared will be: CAT with MPI item selection, CAT with WPM item selection, MST preassembled with the NWADH, OMST with MPI item selection, and OMST with WPM item selection.

CHAPTER 3

METHODOLOGY

A Monte Carlo simulation study was conducted to evaluate the effectiveness of the aforementioned item selection methods and adaptive testing designs. Complexity of the test specifications, representation of each content category, item pool size, the number of items in each stage, and number of difficulty levels in the preassembled MSTs were varied to simulate typical testing conditions. The outcomes of interest were alignment with test specifications, measurement precision, and item exposure and test overlap rates.

The simulation began by randomly generating examinee abilities, item parameters, and item content categorizations. Next, MST forms were preassembled from the item pool. Responses were then generated for each examinee on each of the testing designs. Each test contained 36 items. Final ability estimates were recorded, along with the items administered on each test and their corresponding content categorizations. This information was used to calculate root mean square error (RMSE) and bias for ability estimation, the number and type of constraint violations, a general index of content alignment, and item exposure and test overlap rates.

Item Pool Construction

Previous MST studies have typically used a fixed item pool of moderate size, ranging from 420 to 600 items (Routo, Patsula, Manfred, & Rizavi, 2003; Zheng et al., 2012; Han & Guo, 2014). In this study, the item pool size was varied at two levels – 360 and 720 items – to represent a small and large pool. To make the item pools realistic, item parameters were generated using real item pools. As described in Table 1, the means and standard deviations were based on the pool used in Zheng et al. (2012) while the distributions followed Edwards, Flora, and Thissen (2012). In both studies, the item parameters came from operational tests.

Table 1 Means, standard deviations, and distributions for item parameter generation

Parameter	Mean	Standard Deviation	Distribution
a	1.196	0.329	Lognormal
b	0.060	1.430	Normal
c	0.153	0.072	Logit-normal

For the a and c parameters, the means listed in Table 1 represent the means of the parameters after transforming to the appropriate distribution. Compared to the mean of 1 and standard deviation of 0.5 often used in simulation studies, the a parameters used in this study are larger and more centered. In other words, the items are better. This is due to the fact that adaptive tests are very dependent on the quality of the items. The same case can be argued for the c parameter distribution. Similar to most IRT simulation studies, the b parameter distribution has a mean close to 0, but the standard deviation in this study is larger than 1. This distribution has a wide spread in order to better cover the entire θ distribution. Examinee abilities were generated from a standard normal distribution. Sample size was set at 1,000. This size and distribution are similar to those used in Kim, Chung, Dodd, and Park (2012) and Zheng et al. (2012).

Test specifications. Test specification complexity was based on real large-scale test blueprints. The complexity of the test specifications can be summarized by the number of indices associated with each item. Each item is characterized by one, two, three, or four categories for the baseline, simple, medium, and complex specifications, respectively. The baseline condition was based on the blueprint described by Kingsbury and Zara (1989) and is shown in Table 2.

Table 2 Test blueprint for the baseline content constraint condition

Constraint Category	Level	% of items in pool
Content	Addition	30%
	Subtraction	30%
	Multiplication	20%
	Division	20%

The simple blueprint, shown in Table 3, simulates the one used for the NAEP 12th grade reading test (National Assessment Governing Board, 2015b).

Table 3 Test blueprint for the simple content constraint condition

Constraint Category	Level	% of items in pool
Passage type	Literary	30%
	Informational	70%
Cognitive targets	Locate/recall	20%
	Integrate/interpret	45%
	Critique/evaluate	35%

The medium complexity blueprint simulates the one used in the NAEP 12th grade mathematics test (National Assessment Governing Board, 2015a). This blueprint is shown in Table 4.

Table 4 Test blueprint for the medium content constraint condition

Constraint Category	Level	% of items in pool
Content	Number properties and operations	10%
	Measurement	15%
	Geometry	15%
	Data analyses, statistics, and probability	25%
	Algebra	35%
Complexity	Low	25%
	Moderate	50%
	High	25%
Format	Multiple choice	50%
	Constructed response	50%

Table 5 shows the most complex test blueprint condition. These specifications are akin to those used in the PISA mathematics assessment (OECD, 2012).

Table 5 Test blueprint for the complex content constraint condition

Constraint Category	Level	% of items in pool
Content	Change and relationships	27%
	Quantity	26%
	Space and shape	24%
	Uncertainty and data	23%
Cognitive process	Formulate	25%
	Employ	44%
	Interpret	31%
Context	Occupational	23%
	Personal	29%
	Public	25%
	Scientific	23%

Format	Simple multiple choice	27%
	Complex multiple choice	11%
	Constructed response (expert)	35%
	Constructed response (manual)	27%

In addition to the four levels of test blueprint complexity, the percentage of items in the pool from each category was also varied. The percentages listed in Tables 2 through 5 represent the realistic case, where categories vary in their representation in the item pool and on the test. This may present challenges in item selection, as categories with low representation in the pool may not have a wide variety of item difficulties to choose from. The realistic condition was contrasted with an even condition where each category was represented equally in the item pool and on the test. For example, in the even condition, each content area from Table 5 would make up 25% of the item pool.

Each item in the pool was assigned to each category randomly, using the percentages in Tables 2 through 5 for the realistic conditions and the average percentage for the even conditions. Thus, items from each content area could belong to any type of cognitive process, item format, etc. Also inherent in random assignment is that no relationship is assumed between the item parameters and item content categorizations.

The exact number of items required from each content category is simply the product of the total number of items on the test (36 in this case) and the proportion of items required by the given content area. As this product does not always produce a whole number, lower and upper bounds were set for each content constraint, as is often done in practice. For instance, the first row of Table 5 specifies that 27% of test items should be "Change and relationship" items. The desired number of items from this category is 9.72 (36x0.27), so the lower and upper bounds were set to 9 and 10, respectively.

Test Design

Test length was fixed at 36 items, a moderate test length in adaptive testing (Rotou, 2003; Edwards et al., 2012). Each MST and OMST consisted of 3 stages, as is common in research and practice (Hendrickson, 2007). The number of items in each stage was varied at three levels, as shown in Table 6.

Table 6 Number of items in each stage across conditions

Stage length	Items per stage
Equal	12, 12, 12
Decreasing	15, 12, 9
Increasing	9, 12, 15

MST Preassembly

After the item pool was generated, MST modules were preassembled using the NWADH. The MST module design was varied at two levels – 1-3-3 and 1-4-4 – to examine how the number of difficulty levels impacts the outcome variables. Each module was assembled using a bottom-up, backwards assembly approach. That is, modules were assembled one at a time starting with the final stage modules. Target difficulty values were set at $\theta = -1, 0,$ and 1 and $\theta = -1.5, -0.5, 0.5,$ and 1.5 for the 1-3-3 and 1-4-4 designs, respectively. The number of panels was determined based on the test design and size of the item pool. The 1-3-3 design yields a total of 7 modules with an average of 12 items per module. To exhaust the item pool to the fullest extent possible in this design, 4 and 8 panels were created for the 360 and 720 item pool conditions, respectively. For the 1-4-4 design, 3 and 6 panels were created for the two different item pool sizes.

A preliminary simulation study was conducted in order to determine appropriate TIFs for each condition. The NWADH was used to assemble MST modules in the preliminary simulation and the item pool and test specification conditions were the same as those in the final study.

Deviation from the TIF, d_{jt} in Equation (20), was replaced by the item information at the target θ. Thus, module assembly focused on maximizing information, rather than minimizing the deviation from the TIF. After each MST assembly, information at the target θ was averaged across all modules within each stage. Assembly was replicated 100 times for each condition and stage-specific module information was averaged across all replications of each condition. These average information values were used as the TIFs for MST preassembly in the final study. This method for developing TIFs was described by Zheng et al. (2014).

Item Response Generation

Unique item responses were generated for each testing design in each replication, using the same examinee abilities and item pool. For each item, the probability of a correct response, given the "true" examinee θ and the item parameters, was calculated based on the 3PL IRT model shown in Equation (1). This probability was then compared to a random number from the standard uniform distribution. If the probability was greater than or equal to the random number, the response was marked as correct; otherwise the response was scored as incorrect. After each stage, or each item for the CATs, an EAP ability estimate was calculated, as shown in Equation (3). EAP was chosen for this study as it always finds a solution and performs similarly to ML estimation in adaptive tests (Wang & Vispoel, 1998). Each testing format was simulated as outlined in the following sections.

CAT simulation. Two CATs were simulated in each condition, differing only in the item selection method: MPI or WPM. The purpose of the CAT simulations was twofold: (1) to compare the MPI and WPM item selection methods in a CAT with varying levels of constraint complexity; and (2) to set a baseline for the other conditions, as CAT has been well researched

and implemented in a number of testing programs. It is also known that CAT is able to achieve better measurement precision than MST and OMST under the baseline test blueprint condition.

The starting θ estimate was randomly drawn from a uniform distribution between -0.5 and 0.5. This is consistent with using the average ability as a starting estimate (Mills & Stocking, 1996), but item overexposure is reduced by ensuring that the initial "best" items are not the same for every examinee. To further control item exposure, each item chosen for administration was randomly selected from the five items with the highest priority index or lowest penalty value. The four items not selected for administration were eliminated from the pool for the remainder of the current test. This is similar to the method proposed by McBride and Martin (1983) and used with the MPI and WPM by He et al. (2014).

Early in the test, little is known about examinee ability, while the content specifications are well known. Hence, Shin et al. (2009) recommended giving more weight to item content at this stage and increasing the information weight throughout the test. However, He et al. (2014) compared various weighting schemes for the WPM in CAT and found that item exposure and content coverage results were best when constant weights of 6 and 2 for content and information penalties were used across the test. Thus, these weights were adopted for w' and w'' in this study. In their examination of the MPI, Cheng and Chang (2009) used content area-specific weights (w_c) ranging from 0.5 to 20, with an average weight of ~8.4. The same w_c were used with the WPM by Shin et al. (2009). He et al. examined both the MPI and WPM and used content-specific weights ranging from 1 to 11 with an average weight of ~6.7. In the current study, all constraints were treated as equally important and were assigned constraint-specific weights of 8, similar to the average value used in previous studies.

Preassembled MST simulation. Each preassembled MST simulation began by randomly selecting a module from the parallel forms of the stage 1 modules. After the stage 1 responses were simulated, ability was estimated and the next module was chosen based on this estimate. Specifically, cut values were defined as the midpoint between the target θs of two adjacent modules (e.g., if the target θs were 0 and 1, examinees with $\hat{\theta}_i \geq 0.5$ were assigned to the more difficult module), as in van der Linden and Diao (2014). Examinees did not follow panels, but were instead assigned randomly to a module of appropriate difficulty at each stage. Since MSTs were assembled via a bottom-up procedure, modules could theoretically be mixed and matched to form parallel pathways. This routing method was thought to help minimize test overlap by introducing randomness at each routing point.

On-the-fly MST simulation. Two OMSTs were simulated at each stage length in every condition, based on the method described by Zheng and Chang (2015). Items were selected via either the MPI or WPM. As with the CAT simulations, OMSTs started by randomly generating a number between -0.5 and 0.5 as the initial ability estimate. Items in the first stage were chosen based on this estimate. Item exposure was controlled by randomly selecting an item from the five items with the highest priority index or lowest penalty value at each item selection point. The MPI and WPM utilized the same weights as in the CAT design: constraint-specific weights were set to 8 and, for the WPM, w' and w'' were fixed across the test at 6 and 2, respectively. No item replacement phase was used for OMST, as this study focused on evaluating the item selection indices; the replacement phase serves as a correction for inadequate initial selection and would thus make comparison difficult.

Summary. The simulation design varied the size of the item pool (2 levels), complexity of the content constraints in the test blueprint (4 levels), and representation of each category in

the item pool (2 levels) for a total of (2x4x2) 16 conditions. In each condition, MSTs varied in the number of difficulty levels (2 levels) while OMSTs varied in the item selection method (2 levels). Both of these designs were simulated at each stage length (3 levels). Thus, 6 MSTs and OMSTs were simulated in each of the 16 conditions. Two CAT designs, varying in the item selection method, were simulated per condition. All conditions were replicated 100 times, with a new item pool and examinee abilities generated for each replication. Results were averaged within each test design and condition.

Analyses

Content coverage. Content alignment is a key focus of this study and was measured in two ways. Descriptively, the average number of constraint violations per test, $\bar{V}$, was calculated as in Cheng and Chang (2009):

$$\bar{V} = \frac{\sum_{i=1}^{I} V_i}{I} \tag{30}$$

Here, V_i is the number of constraints violated on examinee i's test and I is the number of examinees. Of particular interest in this study was the average number of lower ($\bar{V}_l$) and upper ($\bar{V}_u$) bound violations. These rates were also calculated using Equation (30), where V_i was replaced with the number of lower and upper bound violations on examinee i's test.

A second more general measure was the content alignment index proposed by Wise et al. (2015). CA_i measures the deviation from the test specifications for examinee i's test and is computed as:

$$CA_i = 1 - \frac{\sum_{c=1}^{C} |x_c - X_c|}{J} \tag{31}$$

where x_c is the number of items actually administered from content area c, X_c is the number of items required to be administered by the test specifications, and J is the length of the test. A CA_i of 1 represents perfect content alignment while lower values indicate the degree of misalignment.

Since this study used lower and upper bounds, rather than fixed constraints, deviation from 1 was expected. The degree of content misalignment for a given test design was examined relative to other designs. CA_i was averaged across all examinees to find the average content alignment, CA.

Due to the lack of a known minimum value, the scale of the CA index is not clear. To make CA more interpretable, it was transformed to CA' using Equation (32).

$$CA' = \frac{CA - min(CA)}{max(CA) - min(CA)} \tag{32}$$

Here, $min(CA)$ and $max(CA)$ are the minimum and maximum CA across all tests in the simulation. Thus, CA' has a minimum of 0, representing the worst content alignment among all tests, while the maximum value is 1, indicating the best alignment in the simulation.

Measurement precision. Measurement precision was investigated by calculating root mean square error (RMSE) and bias for final ability estimates. RMSE provides a relative measure of the amount of error in ability estimation. Bias, on the other hand, is used to examine whether or not there exists any systematic error in ability estimation. These two statistics were calculated as:

$$RMSE_{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^{I}(\hat{\theta}_i - \theta_i)^2}{I}} \tag{28}$$

$$Bias_{\hat{\theta}} = \frac{\sum_{i=1}^{I}(\hat{\theta}_i - \theta_i)}{I} \tag{29}$$

where $\hat{\theta}_i$ is the estimated ability for examinee i, θ_i is the "true" ability, and I is the total number of examinees.

Item exposure and test overlap. As with content coverage, a combination of descriptive measures and overall indices was used to measure item exposure and test overlap. Item exposure counts the number of times an item is administered across all examinees. An ideal exposure rate for items from a given item pool, $\overline{er_j}$, is defined as the test length divided by the number of items

in the pool (Chang & Ying, 1999). Moyer, Galindo, and Dodd (2012) defined overexposure as an exposure rate greater than 0.30. Given that the ideal exposure rate in their study was ~0.1, overexposure in this study was defined as 3 times $\overline{er}_j$ for the given condition. Descriptively, the proportion of overexposed items was calculated for each test along with the proportion of unused items, as the latter indicates underutilization of the item pool.

Moreover, the χ^2 statistic described by Chang and Ying (1999) was also reported. This statistic measures the similarity of the observed and ideal exposure rates across all items and can be written as:

$$\chi^2 = \frac{\sum_{j=1}^{N_p}(er_j - \overline{er}_j)^2}{\overline{er}_j} \tag{33}$$

where er_j is the exposure rate for item j and N_p is the number of items in the pool. Lower values indicate more even item exposure across the pool.

The test overlap rate can be calculated by counting the number of overlapping items for each pair of examinees and averaging across all pairs. As the number of possible pairs increases exponentially with increasing sample size, this calculation can be extremely tedious. Chen, Ankenmann, and Spray (2003) showed that, as the number of examinees increases, the test overlap rate, $\overline{T}$, approaches:

$$\overline{T} = \frac{\sum_{j=1}^{N_p} ER_j(ER_j - 1)}{JI(I-1)} \tag{34}$$

where ER_j represents the number of tests on which item j appears, and N_p, J, and I represent the number of items in the pool, the test length, and the number of examinees, respectively. In general, low test overlap rates are desired as higher overlap may indicate weakened test security. The item exposure statistics and test overlap rate were calculated for each simulated test in each replication.

CHAPTER 4

RESULTS

Results are presented in the following order. First, content alignment results are presented, as content specifications are the main interest of the current study. Next, measurement precision is examined. Finally, the results on test security are given by item pool usage and test overlap. For each outcome measure, ANOVAs were conducted to examine the effects of the design variables. Because the sample sizes are so large in this study, effect size (η^2) was reported in place of the ANOVA F tests and p-values. All ANOVAs were significant at an α-level of 0.05 unless stated otherwise. The effect size guidelines outlined by Cohen (1988) were used to interpret the size of the effects. Specifically, η^2 of 0.01, 0.06, and 0.14 were used to define small, medium, and large effects, respectively. A summary is included at the end of each section to highlight the major findings.

Content Alignment

ANOVAs were conducted on the standardized content alignment index, CA', as this was the most general measure of content alignment. For some ANOVAs, the within-group distribution of CA' deviated from normality. However, the homogeneity of variance assumption was always met. Since ANOVA is robust against non-normality (Maxwell & Delaney, 2004) and effects sizes, rather than p-values, were of interest, the analyses were deemed appropriate. First, the effects of test design (CAT, MST, and OMST) and item selection method (MPI, WPM, and NWADH) were examined. Test design and item selection method were found to explain 8 and 9 percent of the variability, respectively, in CA'. These are both considered medium effect sizes. Tables 7 and 8 show the mean CA', as well as the average number of lower and upper bound constraint violations per test, $\bar{V}_l$ and $\bar{V}_u$, by test design and item selection method.

Table 7 Mean content alignment and lower and upper bound violations by test design

Test Design	CA'	$\bar{V}_l$	$\bar{V}_u$
CAT	0.86	0.18	0.21
MST	0.78	1.43	1.36
OMST	0.86	0.18	0.20

Table 8 Mean content alignment and lower and upper bound violations by item selection method

Test Design	CA'	$\bar{V}_l$	$\bar{V}_u$
MPI	0.844	0.356	0.400
WPM	0.875	0.001	0.006
NWADH	0.780	1.429	1.361

Among the test designs, shown in Table 7, CAT and OMST performed similarly and considerably better than MST. For the item selection methods, Table 8 shows that the WPM performed the best, followed by the MPI, with the NWADH in a distant third. As a follow-up comparison between the MPI and WPM, *Cohen's D* showed an effect size of 0.30. Thus, the average CA' for the WPM was almost one-third of a standard deviation greater than that of the MPI. This is considered a small effect (Cohen, 1988); however, the number of constraint violations displays a clear advantage for the WPM.

The poor content alignment of the MST design and NWADH item selection method was further investigated by counting the number of lower and upper bound violations for each preassembled MST module. On average, there were 0.21 and 0.19 lower and upper bound violations per module. As each examinee was administered three modules, this should result in 0.63 and 0.57 violations per test. While these numbers are still higher than those for the MPI and WPM, they are not as large as the results for MST in Table 7. This discrepancy comes from the fact that lower and upper bounds were module-specific and it was possible to meet the constraints in each module but violate constraints at the test level. For instance, if 10 to 11 items were required from content area c, each module would require 3 to 4 items. An examinee who receives three modules with 3 items from c would violate a lower bound constraint at the test

level, while an examinee who receives three modules with 4 items from c would violate an upper bound. Hence, while the NWADH did not perform as well as the MPI or WPM, additional deviation from the constraints appeared to be inherent in the MST design.

ANOVAs were conducted to examine the effects of item pool size, test specification complexity, and content representation. Item pool size and content representation accounted for a small amount of the variability in CA'; just 2 and 1 percent, respectively. On the other hand, test specification complexity had an η^2 of 0.55; a large effect. Figure 4 shows the average number of total violations across item selection methods and item pool sizes.

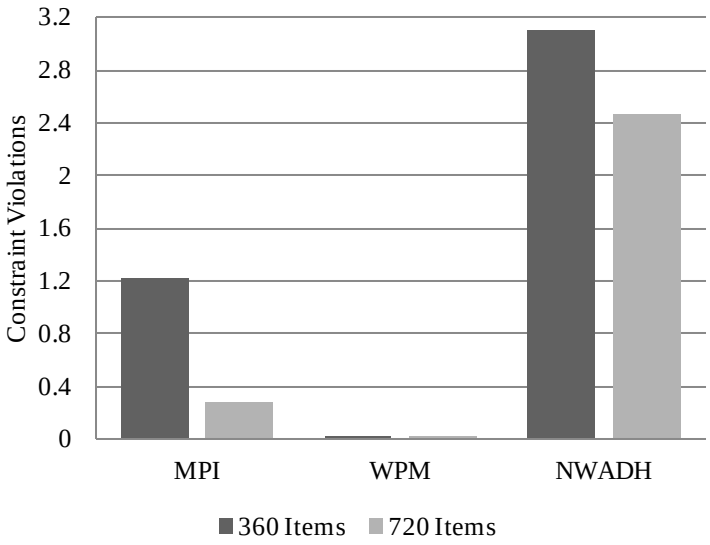

Figure 4. Average number of constraint violations by item selection method and item pool size.

Not surprisingly, access to a larger pool resulted in fewer constraint violations for all selection methods, as there were more items to choose from at each selection point. The larger pool size seemed to especially benefit the MPI, which saw a four-fold decrease in total constraint violations from the 360 to 720 item pools. The effect looked smaller for MSTs (with the NWADH). This can be explained by the increase in the number of parallel panels that coincided with increasing item pool size. The panels assembled later had a similar number of items to

select from in both pool size conditions. For the WPM, the effect of pool size was barely noticeable, as very few violations were committed using either pool.

Table 9 shows the average number of lower and upper bound violations across the selection methods and levels of test specification complexity, separated by the two levels of content representation: realistic and even.

Table 9 Average number of constraint violations by item selection method, constraint complexity, and content representation

Realistic Representation						
Constraint Complexity	$\bar{V}_l$			$\bar{V}_u$		
	MPI	**WPM**	**NWADH**	**MPI**	**WPM**	**NWADH**
Baseline	0.001	0.000	0.660	0.000	0.000	0.642
Simple	0.006	0.000	1.041	0.005	0.000	0.801
Medium	0.104	0.001	1.706	0.107	0.000	1.727
Complex	0.889	0.003	2.700	1.011	0.021	2.893
Even Representation						
Baseline	0.000	0.000	0.559	0.000	0.000	0.523
Simple	0.001	0.000	0.601	0.001	0.000	0.659
Medium	0.177	0.000	1.776	0.173	0.000	1.133
Complex	1.671	0.001	2.393	1.905	0.024	2.512

As expected, increasing complexity of the test specifications resulted in more constraint violations for all selection methods. This increase was most prominent for the NWADH, followed by the MPI. The WPM was highly robust to the increasing complexity; although, it seemed to be more prone to upper than lower bound violations.

The effects of content representation were intriguing. They were dependent on the selection method and the complexity of the constraints. The MPI performed better in the baseline and simple constraint conditions when content categories were evenly distributed and better in the medium and complex conditions when the distribution was realistic. The opposite was true for the NWADH. While it is not clear why this pattern emerged, it is worth more investigation in the future. The WPM had very few constraint violations regardless of test specification complexity or content representation.

Next, the effect of stage length was examined for MSTs and OMSTs. An ANOVA revealed that stage length accounted for 6 percent of the variability in CA', a medium effect. Table 10 shows the average number of lower and upper bound violations across item selection methods and stage lengths.

Table 10 Average number of constraint violations by item selection method and stage length

Stage Length	$\bar{V}_l$			$\bar{V}_u$		
	MPI	**WPM**	**NWADH**	**MPI**	**WPM**	**NWADH**
Equal	0.355	0.001	1.203	0.400	0.006	1.177
Decreasing	0.349	0.001	1.067	0.394	0.006	1.015
Increasing	0.358	0.001	2.019	0.402	0.005	1.893

For the MPI and WPM, stage length had little effect on content alignment. For the NWADH, more constraint violations were committed when the stage length was increasing. One possible explanation is the backwards assembly used to build MST modules. Because modules were assembled in reverse stage order, the routing modules, assembled last, faced a highly depleted item pool. The design with increasing stage length had a shorter routing stage. In general, shorter modules are harder to assemble, as there is less room for error. Thus, shorter routing stages, coupled with a depleted item pool, resulted in more constraint violations.

For MSTs, the module design (1-3-3 or 1-4-4) had a very small impact on content alignment. Module design accounted for less than 1 percent of the variability in CA'. On average, 2.94 and 2.64 total constraint violations were committed for the 1-3-3 and 1-4-4 designs, respectively. The fact that the simpler design resulted in slightly more violations may come as a surprise. However, this can be explained by the fact that the 1-3-3 design featured more parallel panels, making assembly slightly more difficult and increasing the opportunity for violations.

Summary. Examination of the content alignment results revealed that the CAT item selection methods, the MPI and WPM, did a better job of meeting content constraints than the NWADH. The WPM performed best of all. It committed very few violations even under the

most complex test specification conditions. For preassembled MSTs, extra deviation from the test constraints appeared to be introduced by randomly selecting panels at each stage, due to the flexible module-specific constraints. There were no apparent differences in content alignment between CAT and OMST.

The complexity of the test specifications played an important role in content alignment. All selection methods deteriorated when content specifications became more complex. Item pool size had a medium impact on content alignment, with better alignment for larger pools. Content alignment was also impacted by the length of each MST stage such that fewer violations were committed when the earlier stages were longer. This effect likely resulted from the backwards assembly method used to build MST modules and thus was not present in OMST.

Measurement Precision

Separate ANOVAs were conducted using RMSE and bias as outcome variables. RMSE was not normally distributed within test design and item selection method groups. However, the variances were equal, so the ANOVAs were examined. Both the normality and homogeneity of variance assumptions were met for all ANOVAs with bias as the outcome. The first sets of ANOVAs found that test design and item selection method each accounted for 92 percent of the variability in RMSE and 23 percent of the variability in bias. Thus, test design and item selection method had large effects on both measures of precision. Table 11 shows the average RMSE and bias across test designs and item selection methods.

Table 11 RMSE and bias by test design and item selection method

	CAT		OMST		MST
Outcome	**MPI**	**WPM**	**MPI**	**WPM**	**NWADH**
RMSE	0.219	0.223	0.226	0.230	0.420
Bias	-0.001	-0.001	-0.001	-0.000	-0.013

Overall, MST was the least precise while CAT was slightly more precise than OMST. Within both CAT and OMST designs, the MPI resulted in slightly lower RMSE than the WPM. In general, RMSE and bias were both quite low, particularly for the CAT and OMST designs.

To further examine the differences between CAT and OMST and their item selection methods, *Cohen's D* was computed for the pairwise comparisons of CAT vs. OMST and MPI vs. WPM. The effect sizes for the comparison of CAT and OMST were 0.32 and 0.03 for RMSE and bias, respectively. For the MPI vs. WPM comparison, the effect sizes were 0.21 and 0.01. Thus, RMSE was about one-third of a standard deviation higher for OMST than for CAT and one-fifth of a standard deviation higher for the WPM than the MPI. These are both small effects (Cohen, 1988). Differences in bias between the methods were negligible.

Given the large differences in measurement precision between MST and the other test designs, another set of ANOVAs was conducted to examine the effects of test design with CAT and OMST grouped together. These ANOVAs had η^2 of 0.92 and 0.23 for RMSE and bias, respectively, indicating that nearly all of the variance in measurement precision between test designs was accounted for by differences between MST and the other two designs. Therefore, for the remaining simulation conditions, MST measurement precision results were examined separately from those of CAT and OMST.

Separate ANOVAs for the effect of item pool size on measurement precision revealed differential effects for MST, compared to CAT and OMST combined. For MST, item pool size did not affect RMSE; even with a sample size of over 9,000, the ANOVA was not significant. However, for CAT and OMST together, item pool size explained approximately 74 percent of variability in RMSE, a very large effect. The effects on bias were minimal for all test designs. Figure 5 shows the effects of item pool size on RMSE across test designs.

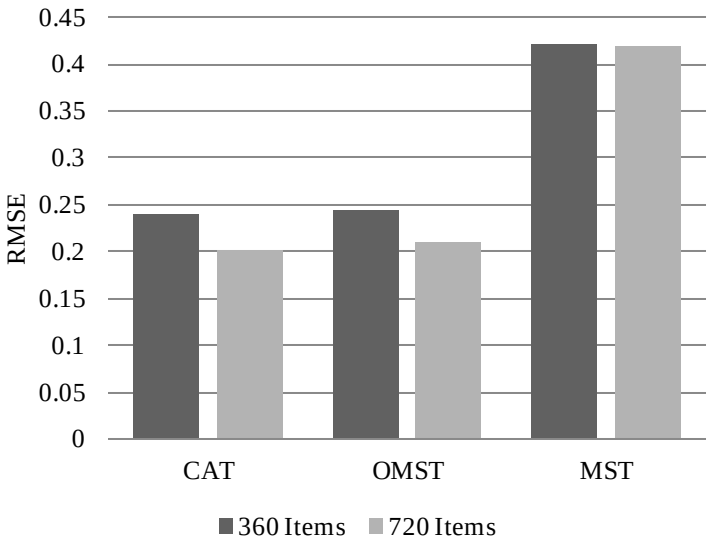

Figure 5. RMSE by test design and item pool size.

As expected, the larger item pool led to lower RMSE for CAT and OMST, as the item selection method had more items to choose from. For MST, the increase in item pool size appeared to be negated by the increase in the number of panels created. That is, because MSTs utilized most of the item pool by creating as many parallel modules as possible, the number of modules increased but the *quality* of the modules did not increase with the size of the pool. This hypothesis can be examined by looking at the information targets obtained from the preliminary simulation and used to assemble MST modules in the final simulation. The average module information targets across stages and item pool size conditions are shown in Table 12.

Table 12 Average information target by stage and item pool size

	Item Pool Size	
Stage	**360 Items**	**720 Items**
1	0.63	0.60
2	1.88	1.90
3	4.05	4.09
Test	6.56	6.59

Table 12 shows that target information was quite similar at each stage for the two item pool conditions. As measurement precision is a direct function of the target information, RMSE and bias were also similar for MSTs of varying pool sizes.

Similar to item pool size, complexity of test specifications showed differential effects for MST, compared to CAT and OMST. Complexity of test specifications accounted for 10 percent of the variability in RMSE for CAT and OMST, a moderate effect, but less than 1 percent of the variability in RMSE for MST. These effects are shown in Figure 6.

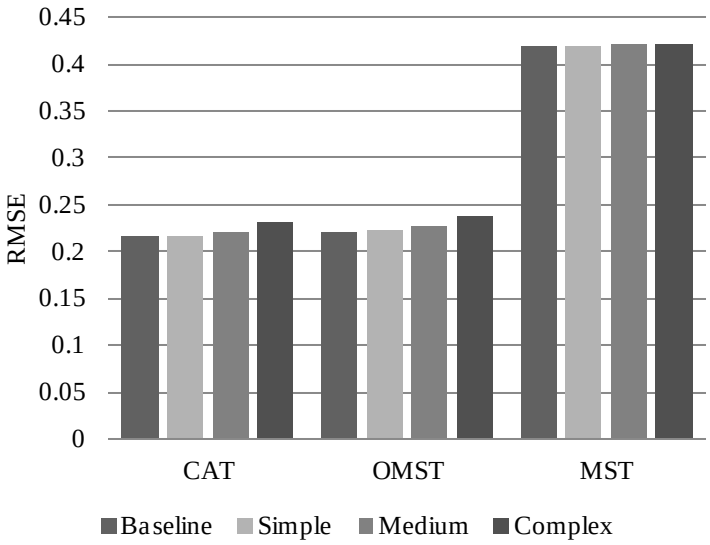

Figure 6. RMSE by test design and test specification complexity.

As expected, increasing constraint complexity resulted in higher RMSE for CAT and OMST. More complex constraints limit item selection, which in turn lowers measurement precision. Surprisingly, this was not observed for MST. One possible reason is that MSTs often failed to meet their content constraints. Thus, item selection appeared to have focused more on precision than on balancing precision and content alignment.

Separate ANOVAs for MST and CAT and OMST together revealed that bias did not differ significantly by content complexity while neither RMSE nor bias was affected by content representation. Another set of ANOVAs revealed differences between MST and OMST in the

effects of stage length on measurement precision. Stage length accounted for approximately 87 and 1 percent of the variability in RMSE and bias, respectively, for MST. For OMST, the effect of stage length was very small for RMSE and non-significant for bias.

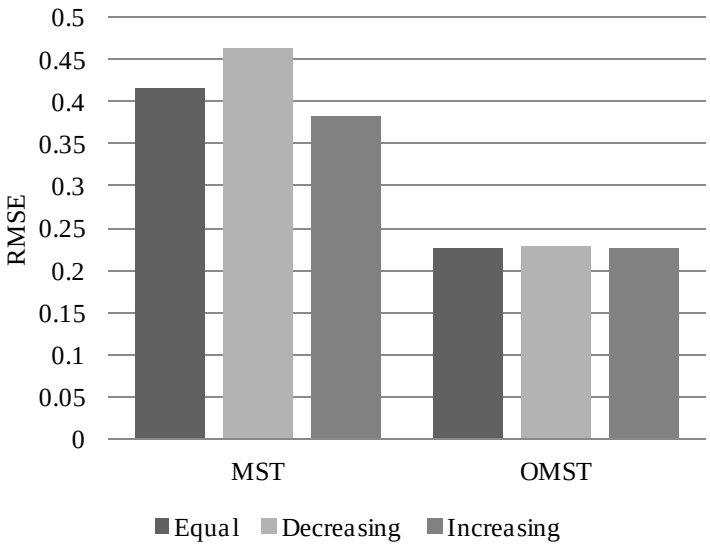

Figure 7. RMSE by test design and stage length.

Figure 7 shows the effects of stage length on RMSE separately for MST and OMST. For MST, the highest precision was achieved when a shorter routing stage was used. The order of performance was reversed from the content alignment results (Table 7), providing more evidence of a tradeoff between precision and alignment. The absence of an effect for OMST is consistent with this explanation, as stage length did not appear to affect either content alignment or measurement precision.

For preassembled MSTs, the effect of module design was examined through ANOVAs on RMSE and bias. Module design accounted for 1 and 3 percent of the variability in RMSE and bias, respectively. These are both small effects. Surprisingly, the 1-4-4 design had slightly higher RMSE and bias (0.42 and -0.02) than the simpler 1-3-3 design (0.42 and -0.01). The direction of this effect once again points to the trend of better content alignment leading to worse measurement precision.

Considering the results of both content alignment and measurement precision, it is clear that the WPM outperformed the MPI. To further investigate this finding, the PI_j and F_j values of selected items were inspected for two examinees: one from a CAT with simple constraints and one of similar ability from a CAT with complex constraints. Table 13 shows the PI_j and F_j for the selected items at five positions in the test.

Table 13 Selected $\boldsymbol{PI_j}$ and $\boldsymbol{F_j}$ values throughout the test for two average ability examinees

	Simple Constraints		**Complex Constraints**	
Item Number	$\boldsymbol{PI_j}$	$\boldsymbol{F_j}$	$\boldsymbol{PI_j}$	$\boldsymbol{F_j}$
1	87.90	-0.29	2536.01	-0.08
9	9.20	-1.11	426.44	-0.43
18	2.88	-0.92	43.30	-0.11
27	0.47	-0.90	1.40	-0.01
36	0.07	-0.80	0.01	-0.23

For the MPI, the values in Table 13 demonstrate an unclear scale for the PI_j statistic. Because the content priority values are not standardized, extremely large PI_j occur at the beginning of the test when the deviations from the bounds are large, while very small values are seen at the end of the test. This effect is amplified under complex constraints, as several content indices and their associated weights are multiplied together. The WPM, however, standardizes the information and content penalty values based on the minimum and maximum values found in the item pool. Therefore, F_j for all items at a given selection point are on a similar scale and selection considers the desirability of administering item *j relative to other items remaining in the pool*. Thus, while the MPI performs worse as content complexity increases, the WPM handles increasing complexity well.

While the WPM clearly possessed an advantage in content alignment, it performed slightly worse than the MPI in terms of RMSE. This could be explained by the tradeoff between

content alignment and measurement precision. To examine this hypothesis, Table 14 shows the average RMSE for the MPI and WPM across test specification conditions.

Table 14 RMSE for the MPI and WPM by test specification complexity

Constraint Complexity	MPI	WPM
Baseline	0.220	0.220
Simple	0.221	0.221
Medium	0.225	0.228
Complex	0.230	0.244

In the baseline and simple constraint conditions, no difference was observed between the two methods. As constraints became more complex, slight advantages in RMSE were evident for the MPI. Recall that the MPI was prone to more constraint violations under the more complex test specification conditions. Hence, the additional measurement precision achieved by the MPI came at the cost of content alignment. Given the importance of content validity, the slightly lower RMSE associated with the MPI in the medium and complex constraint conditions should not be taken to indicate an advantage of the MPI over the WPM.

Summary. The examination of measurement precision results revealed very small bias across all conditions, with few notable effects. However, many simulation factors had sizeable effects on RMSE. RMSE was most impacted by the test design, with MST displaying lower precision than CAT and OMST. CAT was slightly more precise than OMST. These results are consistent with those of Zheng and Chang (2015). Within CAT and OMST, the MPI and WPM performed similarly.

Item pool size and content complexity impacted measurement precision for CAT and OMST such that larger item pools and simpler content constraints resulted in better precision. For MST, measurement precision was affected by the stage length. Shorter routing stages resulted in the lowest RMSE. Finally, the module design of MSTs had a slight impact on both RMSE and bias, with the 1-3-3 design outperforming the 1-4-4 design.

Test Security

Test security was examined through a series of ANOVAs with the item exposure χ^2 statistic and the test overlap statistic, $\bar{T}$. Within-group distributions of both statistics deviated slightly from normality but the within-group variances were equal for all analyses. Therefore, examination of effect sizes were deemed appropriate. Test design accounted for 16 and 6 percent of the variability in item exposure χ^2 and test overlap, respectively, while item selection method accounted for 15 and 4 percent. Both of the effects on the χ^2 statistic are large, while the effects on test overlap are small to moderate. Much of the variability accounted for by the item selection method appeared to be due to differences in test design. When only CAT and OMST were considered, selection method accounted for less than 1 percent of the variability in item exposure and test overlap. This is not surprising, given that the MPI and WPM utilized the same method for exposure control in this study. Table 15 shows the average χ^2 and $\bar{T}$ statistics, as well as the proportion of overexposed and unused items across test designs.

Table 15 Average item exposure χ^2, test overlap rate, and proportion of overexposed and unused items by test design

Test Design	χ^2	$\bar{T}$	Overexposed	Unused
CAT	0.002	0.104	0.007	0.021
MST	0.004	0.127	0.048	0.084
OMST	0.003	0.117	0.033	0.050

CAT was more secure than both MST and OMST in every measure. Although CAT and OMST utilized the same item selection and exposure control methods, OMST had fewer ability estimation points; each stage of items was based on one ability estimate. The initial stage was selected based on a similar estimate (between -0.5 and 0.5) for all examinees, causing items with medium difficulties and high discriminations to become overexposed. The effects of stage length for OMST, discussed below, support this hypothesis. The high overlap rates and large proportions of overexposed and unused items are not surprising for MST. Items not included in

any module had no chance of being administered while items in a given routing module were administered to as many as 1 of every 3 examinees. These results reiterate that test security is often a major disadvantage for MST.

Because test security measures clearly differed between test designs, the effects of the remaining simulation variables were examined separately by test design. Separate sets of ANOVAs revealed that item pool size accounted for 62, 82, and 28 percent of the variability in item exposure χ^2 and 96, 91, and 82 percent of the variability in test overlap for CAT, MST, and OMST, respectively. These are all quite large effects. Figures 8 and 9 show the average χ^2 and test overlap, respectively, across the three test designs and two item pool sizes.

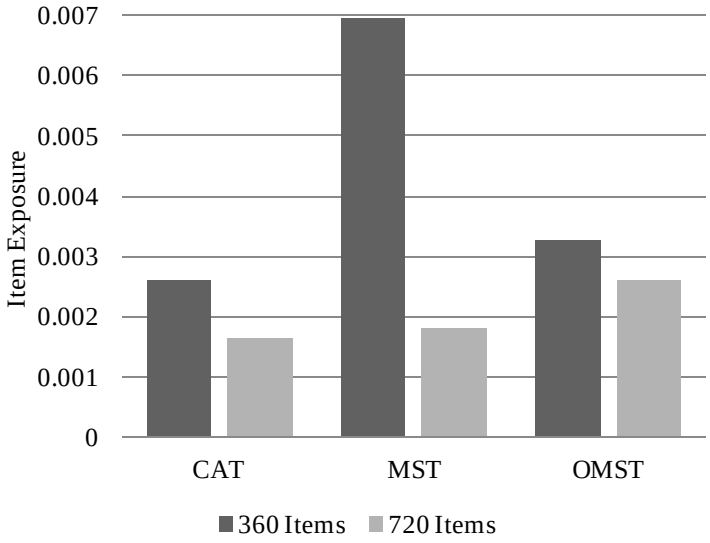

Figure 8. Average item exposure $\boldsymbol{\chi^2}$ by test design and item pool size.

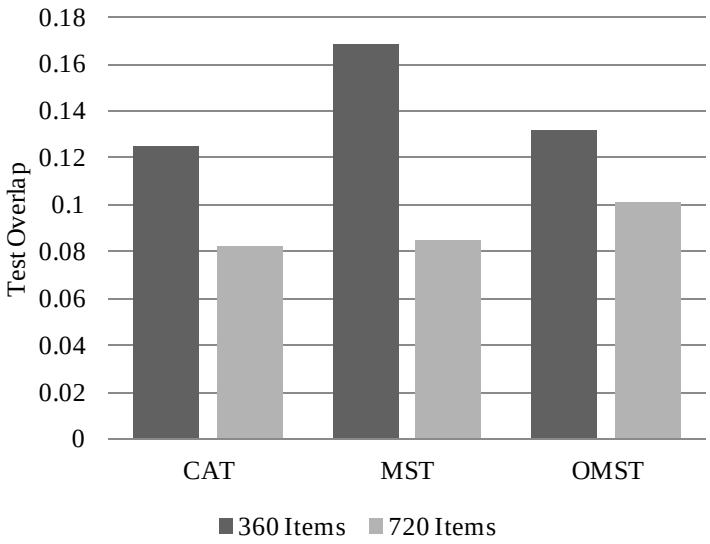

Figure 9. Average test overlap rate by test design and item pool size.

When the pool contained 360 items, MST had higher item exposure χ^2 and test overlap rates than the other designs, indicating worse security. However, when the pool contained 720 items, MST was nearly as secure as CAT. This is likely attributable to the large number of panels (6 or 8 parallel modules for the 1-4-4 and 1-3-3 designs, respectively) and the fact that examinees were randomly assigned a module at each stage, rather than following a set panel. OMST consistently had worse exposure distributions and test overlap rates than CAT and, surprisingly, was also worse than MST when the item pool was large. The increase from 3 or 4 parallel modules in the 360 item pool to 6 or 8 parallel modules in the 720 item pool appeared to be enough to push MST ahead of OMST in terms of test security.

ANOVAs were also conducted to examine the impact of content complexity on test security. Content complexity explained 19 and 42 percent of the variability in item exposure χ^2 and 2 and 11 percent of the variability in test overlap rates for CAT and OMST, respectively. The effects of content complexity on test security measures were non-significant for MSTs. Figure 10 shows the average item exposure χ^2 by content complexity and test design.

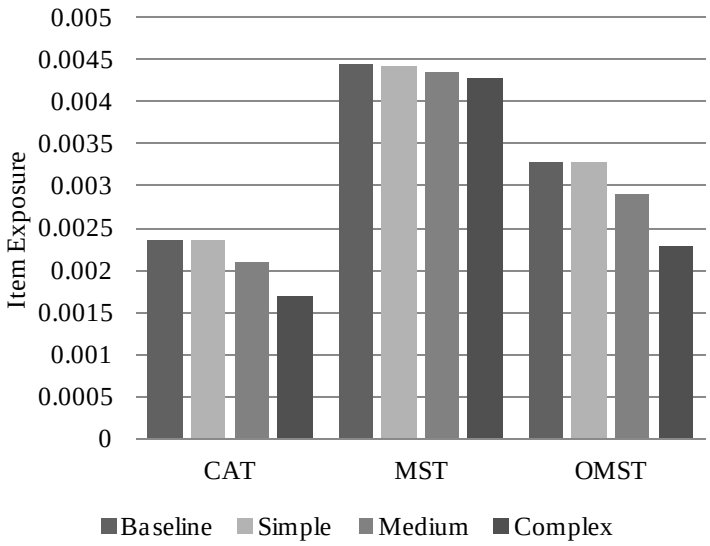

Figure 10. Average item exposure χ^2 by test design and test specification complexity.

Item pool usage consistently improved as the test specifications increased in complexity. This makes sense intuitively. When constraints are simple, items with desirable statistical properties (i.e., high discrimination) may be selected too frequently. However, when constraints are complex, items with low information may still be desirable due to the need to fulfill content constraints. The effect of content complexity was not present in MSTs. This is because MSTs used the same number of items and assigned examinees to modules the same way regardless of the test specifications.

Another set of ANOVAs revealed that content representation did not significantly affect either item exposure or test overlap for any test design. Stage length had very small effects ($\eta^2 \approx 0.01$) on the two outcomes for MST. For OMST, however, stage length explained 10 percent of the variability in item exposure χ^2 and 3 percent of the variability in test overlap rate, a medium and small effect, respectively. Figure 11 shows the average proportion of overexposed and unused items across stage length conditions for OMST.

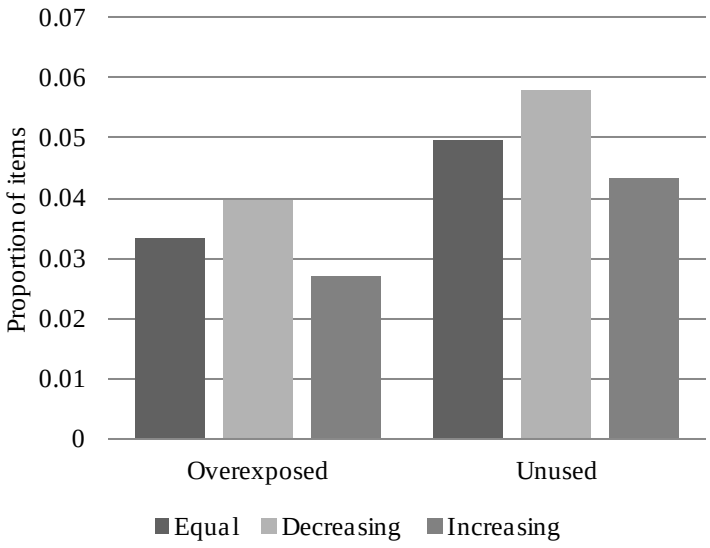

Figure 11. Proportion of overexposed and unused items across stage lengths for OMST.

Figure 11 supports the earlier hypothesis that test security is worse in OMST than in CAT because the entire first stage is based on a similar ability estimate for all examinees. All test security indices, including χ^2 and $\bar{T}$, seemed to favor OMSTs with shorter first stages, indicating that the length of the initial OMST stage impacts test security. This was also true for MST, but the effect was much smaller. For MSTs, each parallel version of the first stage is exposed to more examinees than the later stages, as there is only one stage 1 difficulty level. Thus, shortening the first stage leads to slight improvements in test security measures.

A final set of ANOVAs revealed that module design had a moderate effect on both general measures of test security. Specifically, module design explained 10 and 6 percent of the variability in item exposure χ^2 and test overlap rate, respectively. This effect was such that the 1-3-3 design was more secure than the 1-4-4 design. This can be explained by the difference in the number of total modules created for each design. The 1-3-3 design featured 28 and 56 total modules in the 360 and 720 item pool conditions, while the 1-4-4 design contained 27 and 54 modules. Thus, fewer items from the pool were required for the 1-4-4 design, resulting in more unused items and higher exposure rates for those items that were included.

Summary. The test security results demonstrated the advantages of CAT. Despite utilizing the same item selection and exposure control techniques, CAT performed better than OMST in every measure of test security. By selecting items one-at-a-time and updating the ability estimate after each item, CAT used a wider variety of items than the multistage designs.

Item pool size had a large effect on item exposure and test overlap across all test designs. The larger item pool was consistently associated with more even item exposure and lower test overlap. This effect was largest for MSTs, where increasing the number of parallel panels improved test security considerably. Increasing complexity of test specifications helped to improve test security in CATs and OMSTs, as a wider variety of items were required to meet content constraints. Stage length impacted all measures of test security for OMSTs. Longer initial stages resulted in repeated selection of the same items, increasing test overlap and skewing the item exposure distribution. For MST, module design had a moderate effect on test security, due to differences in the number of items required for the 1-3-3 and 1-4-4 designs.

CHAPTER 5

DISCUSSION

This study evaluated item selection methods for three adaptive testing designs with varying levels of content constraint complexity. Specifically, computerized adaptive, multistage, and on-the-fly multistage tests were studied. The normalized weighted absolute deviation heuristic was used to assemble MST modules while the maximum priority index and weighted penalty model were used to select items for CAT and OMST. For all tests, the complexity of the test specifications, representation of each content category, and size of the item pool were varied at 4, 2, and 2 levels, respectively. For the multistage designs, the length of each stage was varied at 3 levels: equal, decreasing, and increasing. Finally, the number of preassembled difficulty levels at each MST stage was manipulated by studying both 1-3-3 and 1-4-4 designs. All tests were evaluated based on measures of content alignment, measurement precision, and test security. The results were investigated by looking at ANOVA effect sizes and further exploring the effects descriptively. A discussion of the key findings and practical implications follows.

Content Alignment

The content alignment index of Wise et al. (2015) was computed as a general measure of content alignment and the average number of lower and upper bound violations per test were examined as a descriptive measure. The complexity of the test blueprint has a large effect on content alignment. As the number of content categories associated with each item increases from 1 to 4, the content alignment index lowers (indicating worse alignment) and the number of violations increases. This effect is present across all test designs and item selection methods. Content representation, on the other hand, does not impact content alignment in a consistent way. The effect appears to depend on the item selection method used and the complexity of the

constraints. Finally, a larger item pool consistently results in fewer constraint violations than a smaller pool. Thus, test specification complexity and item pool size are key factors in the ability to create content-aligned tests.

The test design and item selection method also have a considerable impact on content alignment. CAT and OMST consistently perform better than MST. The poor alignment of MST results from a combination of the NWADH, which performs worse than the other item selection methods, and the MST module selection method. When modules have flexible content constraints, randomly selecting a module of the desired difficulty at each stage can sometimes lead to test-level constraint violations even when no module-level violations are committed. Among the item selection methods, the WPM performs the best, which is consistent with previous research (He et al., 2014) under the simple constraint conditions. The advantage of the WPM over the MPI actually grows with increasing constraint complexity. This is likely due to a combination of three characteristics of the WPM: the standardization of content and information values, the summing, rather than multiplying, of these values, and the consideration of the number of items available from each content area in the pool.

The length of each MST stage has a small effect on content alignment, such that fewer violations are committed when the earlier stages are longer. When MST modules are assembled backwards, as in this study, the stage 1 modules have access to fewer items from the pool and are thus more difficult to assemble. This effect is most extreme for designs with a short routing test. Finally, the number of difficulty levels in preassembled MSTs has a small impact on content alignment. This only occurs because of differences in the number of panels that can be created. In this study, the 1-3-3 design allows for more panels and is thus more difficult to assemble than the 1-4-4 design, resulting in more constraint violations.

Measurement Precision

Much of the variability in measurement precision is accounted for by test design. MST is much less precise than CAT and OMST. This occurs because CAT and OMST select items specifically for each examinee while MST modules are preassembled at limited fixed difficulty levels. CAT is slightly more precise than OMST. The advantage of CAT comes from the fact that it updates the ability estimate at each item selection point, making each selected item optimal for assessing the examinee's ability. OMST, on the other hand, updates the ability estimate only after a set of items has been administered. The MPI appears to have a slight advantage over the WPM. However, this difference is explained by the MPI's inability to meet complex content constraints. As content alignment helps to validate the interpretations of test scores, this improved measurement precision has little importance when coupled with the deteriorating content alignment.

For CAT and OMST, the effect of item pool size is noticeable. Measurement precision increases with the larger item pool, as more items are available for selection. For MST, however, this effect is absent. This occurred in this study because the increase in the item pool size coincided with an increase in the number of MST panels assembled. Thus, the number of modules increased while the quality of the modules stayed the same. The effect of test specification complexity on measurement precision is also somewhat large for CAT and OMST, but negligible for MST. CAT and OMST see increased RMSE and bias as content constraints become more complex and item selection attempts to balance precision and content alignment. For MST, however, measurement precision does not suffer as a result of increasing content complexity because the content constraints are often not met. Therefore, MST assembly appears to focus more on precision than content alignment.

Stage length and module design appear to impact the measurement precision of MSTs. Both of these results provide evidence of a tradeoff between content alignment and measurement precision, as the conditions with the best alignment result in the least precise measurement. Specifically, tests with increasing stage lengths and tests with 1-3-3 module designs displayed the most precise ability estimation in this study.

Test Security

Test security was examined via the item exposure χ^2 statistic outlined by Chang and Ying (1999), the average test overlap rate, and the proportion of overexposed and unused items. One major finding is that CAT consistently outperforms both MST and OMST. Because CAT updates the ability estimate after each item, there is opportunity for greater variability in item selection and response patterns. The advantage of CAT is so great that, despite utilizing the same item selection and exposure control methods as CAT, OMST still has test security results that are much closer to those of MST than to those of CAT. There is no difference between the MPI and WPM selection methods.

For all three test designs, the larger item pool is associated with a more secure test in every measure. This effect is particularly prominent for MST in this study, as the number of parallel panels doubled from the 360 to the 720 item pool conditions. For CAT and OMST, item pool usage and test overlap improve as test specification complexity increases. This occurs because more complex constraints require a wider variety of items and items with undesirable statistical properties are more likely to be used, as they may help satisfy content constraints. The effect on test specification complexity, however, is absent for MSTs, since MSTs use a set number of items regardless of the test specifications.

Stage length has a large effect on test security measures for OMST. This helps explain the test security differences between CAT and OMST. OMSTs are less secure because the initial stage is created based on a similar ability estimate for each examinee. Longer initial OMST stages generally lead to higher exposure χ^2 and test overlap rates, indicating that the length of the first stage is critical to the test security results of OMST. Longer MST routing stages are also associated with slightly worse security. Because only one difficulty level is used for the first stage, the probability of an examinee seeing a given stage 1 module is greater than that of a given stage 2 or 3 module. Thus, longer routing stages are associated with slightly worse test security. Finally, the 1-3-3 MST design in this study was more secure than the 1-4-4 design. This occurred because of the difference in the number of panels. If the number of panels were held equal between the two designs, the 1-4-4 condition would be more secure, as this design would require more modules overall.

Conclusions

The results of this study have significant implications for testing programs that currently use, or are considering adopting, an adaptive testing design. One major contribution is the comparison of CAT, MST, and OMST under varying levels of test specification complexity. While previous research has often utilized one specific test blueprint, this study varied the content constraints to provide more general guidelines for practical application. There are clearly some major differences between the adaptive testing designs. CAT and OMST appear better than MST in just about every measurable way. Advantages in measurement precision are inherent in the CAT and OMST designs, while the MPI and WPM item selection methods help create a large advantage in terms of content alignment. An advantage of MST that could not be measured in the simulation, however, is that modules can be reviewed ahead of time and necessary changes

can be made before administration. Thus, while CAT and OMST outperform MST in simulation, MST may still have an important place in practice.

OMST holds up well against the well-established CAT design. While CAT is slightly more precise than OMST, this difference, as well as differences in content alignment, is quite small. However, CAT has a noticeable advantage over OMST in test security, which may be reduced by using a shorter first stage in OMST. OMST has its advantages over CAT, such as delaying the first ability estimate and allowing examinees to move freely between items within a stage. But, if test security is a high priority, other designs or alternative methods for selecting the initial stage should be considered.

Another key contribution of this study is the comparison of the CAT and OMST item selection methods, the MPI and WPM, under varying levels of content complexity. While the two methods are comparable when the test specifications are relatively simple, the WPM is able to meet complex constraints more consistently without sacrificing measurement precision. Generally, the WPM should be recommended over the MPI, especially when three or more content categories are associated with an item. This was the case in this study despite the fact that the WPM did not place items into color groups, as recommended by Shin et al. (2009). Placing items into groups would have provided additional protection against constraint violations; but it appeared as though it was not entirely necessary.

This study provides some general recommendations for implementing adaptive testing. Before deciding on a test design, the testing organization must carefully consider the importance of content alignment, measurement precision, and test security. If content alignment is of great importance, the complexity of the test specifications should be closely examined. Either CAT or OMST should be generally preferred over MST, particularly as constraint complexity increases.

If precise measurement is desired, examinee-specific item selection with many adaptation points is essential. That is to say, CAT should be preferred, with OMST as a close second option. CAT should also be preferred for testing programs concerned with test security. However, it should be kept in mind that test security is highly dependent on the size of the item pool. No test design or item selection method can make up for a pool lacking in quality items.

Once a decision has been made regarding the test design, additional steps may be taken to get the most out of the selected design. For CAT and OMST, the choice of item selection method is of most importance when content constraints are complex. Specifically, the WPM should be used whenever three or more content indices are associated with an item. Additionally, the content weights of the WPM and MPI can be manipulated to achieve greater control between content alignment and measurement precision. For MST, methods for selecting modules that consider the content of each module may need to be implemented in order to meet the test specifications. Additionally, the number of difficulty levels and parallel panels is crucial to the content alignment, precision, and security of MSTs. Finally, if test security is of high priority, alternative methods for creating the first OMST stage may be considered. For instance, the starting ability estimate could utilize information from prior administrations or other test scores for the given examinee. If no such information is available, the initial estimate could come from a wider range of abilities.

Limitations and Future Directions

There are a number of limitations in this study. One major limitation is that item pools were randomly generated and there was no relationship between item parameters and content categories. In practice, they are likely to be related. For instance, in the NAEP mathematics test, items from the high cognitive complexity category are likely to have higher item difficulties than

those from the low complexity category. Another disadvantage to working with a randomly generated pool is the correspondence between content representation in the pool and on the test. In this simulation, there was an approximately one to one relationship between the proportion of items from each content category in the pool and the proportion on the test. In practice, this is hardly the case. One pool is often used to design different tests. The correspondence between the item pool and the test can be quite disproportional.

Because this study examined varying test blueprints and item pools, the item selection methods could not be finely tuned to match the requirements of each test. The content weights of the MPI, WPM, and NWADH were kept constant. In practice, if a testing organization is interested in using the MPI, they may consider manipulating the size of the content weights to potentially achievement better content alignment. Finally, the randomization technique used to increase test security in this study was very straightforward and is certainly not the only method available. Other exposure control methods may be utilized based on the security needs of the testing program.

The findings from this study point to a number of lines of future research. One can continue to examine the performance of the relatively new OMST, compared to the well researched CAT and MST designs. For instance, this study shows that randomly selecting the initial ability estimate between -0.5 and 0.5 does not provide enough variability to alleviate item overexposure concerns. Future research can be conducted on the development of a more optimal initial OMST stage. Additionally, the MST-S design of Han and Guo (2014) was described in detail, but preliminary simulation results indicated that this method works well only when content constraints are simple. It may be interesting to look into how to apply MST-S for tests with complex specifications. Finally, another promising line of research is to investigate new

testing designs that combine CAT and MST. A good example of such research is the hybrid design of Wang, Lin, Chang, and Douglas (2016) that starts off as an MST before morphing into a CAT. Designs like these may increase in practical relevance and importance as more testing programs move their assessments online and embrace the advantages of computerized testing and adaptive testing designs.

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.

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

Belov, D. I. & Armstrong, R. D. (2008). A Monte Carlo approach to the design, assembly, and evaluation of multistage adaptive tests. *Applied Psychological Measurement, 32*(2), 119 – 137.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrica, 80*(1), 1 – 20.

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

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369 – 383.

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*(6), 467 – 482.

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

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

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341 – 349.

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

Han, K. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 119 – 134). New York, NY: CRC Press.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement, 74*(4), 677 – 696.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issue 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.

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.

Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement, 52*(1), 70 – 79.

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

Kingsbury, G. G., & Zara, A. R. (1989). Procedure for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359 – 375.

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

Lord, F. M. (1971b). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*(1), 3 – 31.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison – Wesley.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*(3), 224 – 236.

Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229 – 249.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

McBride, J., & Martin, J. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223 – 226). New York: Academic Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4), 287 – 304.

National Assessment Governing Board (2015a). *Mathematics Framework for the 2015 National*

Assessment of Educational Progress. Washington, DC: National Assessment Governing Board.

National Assessment Governing Board (2015b). *Reading Framework for the 2015 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

OECD (2012). *PISA 2012 Technical Report*. OECD, Paris.

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

Rotou, O., Patsula, L., Manfred, S., & Rizavi, S. (2003, April). *Comparison of multistage-tests with computerized adaptive and paper & pencil tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83 – 101.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63 – 84.

Shin, D., Chien, Y., Way, D., & Swanson, L. (2009). Weighted penalty model for content balancing in CATS. Retrieved from http://www.pearsonedmeasurement.com/research/research.htm

Shin, D., Chien, Y., & Way, D. (2012, April). *A comparison of three content balancing methods for fixed and variable length computerized adaptive tests*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Vancouver, British Columbia, Canada.

Stocking, M. L. (1994). *Three Practical Issues for Modern Adaptive Testing Item Pools*. (Report No. ETS-RR-94-5). Princeton, NJ: ETS.

Swanson, M. L., & Stocking, L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151 – 166.

van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. *Applied Psychological Measurement, 10*(4), 325 – 332.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*(3), 283 – 302.

van der Linden, W. J. & Diao, Q. (2014). Using a universal shadow-test assembler with

multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 119 – 134). New York, NY: CRC Press.

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

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

Wang, S., Lin, H., Chang, H.-H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement, 53*(1), 45 – 62.

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

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155 – 180). Mahwah, NJ: Lawrence Erlbaum.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361 – 375.

Weissman, A., Belov, D. I., Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests* (Law School Admission Council Research Report No. 07-05). Newtown, PA: Law School Admission Council, Inc.

Wise, S. L., Kingsbury, G. G., & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practice, 34*(4), 41 – 48.

Zheng, Y., & Chang, H.-H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. In Y. Chang, & H.-H. Chang (Eds.), *Advancing methodologies to support both summative and formative assessments* (Chapter 2). Charlotte, NC: Information Age Publishing.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement, 39*(2), 104 – 118.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H.-H. (2012). Multistage adaptive testing for a large-scale classification test: The designs, automated heuristic assembly, and comparison with other testing modes. *ACT Research Report*.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 119 – 134). New York, NY: CRC Press.

Logan Rome

Curriculum Vitae

Education

PhD in Educational Statistics & Measurement — August 2017
University of Wisconsin-Milwaukee
Dissertation: Evaluating Item Selection Methods for Adaptive Tests with Complex Content Constraints
Committee: Bo Zhang, Razia Azen, Michael Brondino, Cindy Walker

Certificate in Applied Data Analysis Using SAS — May 2016
University of Wisconsin-Milwaukee

BS in Psychology and Criminal Justice, cum laude — May 2013
University of Wisconsin-Oshkosh

Professional Experience

Curriculum Associates — August 2017 – present
Research Scientist

AMTC & Associates — August 2016 – present
Statistical Consultant

Curriculum Associates — April 2017 – August 2017
Psychometric and Research Intern

Consulting Office for Research & Evaluation (UWM) — August 2013 – May 2017
Research Assistant

University of Wisconsin-Milwaukee — January 2015 – December 2015
Teaching Assistant

Publications

Rome, L. & Zhang, B. *Investigating the effects of differential item functioning on proficiency classification* (In press).

Rome, L., Azen, R., & Zhang, B. *Detecting quadratic item position effects with a multilevel model* (Revised and resubmitted).

Rome, L., Cançado, L., Azen, R., & Zhang, B. *DIF detection methods in large-scale assessments* (In progress).

Technical Reports

Rome, L. & Walker, C. M. (2015). Year 5 evaluation of JA BizTown. *Junior Achievement Technical Report*.

Rome, L. & Walker, C. M. (2015). Year 5 evaluation of JA Finance Park. *Junior Achievement Technical Report*.

Presentations

Rome, L. & Zhang, B. (2017, April). *Investigating the effects of differential item functioning on proficiency classification*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Rome, L., Azen, R., & Zhang, B. (2016, April). *Detecting nonlinear item position effects with a multilevel model*. Poster presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Rome, L., Cançado, L., Azen, R., & Zhang, B. (2015, April). *Ability estimation and DIF detection in large-scale assessments*. Poster presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Professional Service

National Council on Measurement in Education
Reviewer – Graduate Student Poster Session

Educational Psychology Student Association – University of Wisconsin-Milwaukee
Co-President