A Comparison of Adaptive Mastery Testing Using Testlets With the 3-Parameter Logistic Model

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

BY

Melody Sue Jacobs-Cassuto

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

David J. Weiss, Adviser

MAY, 2005

ABSTRACT

Adaptive mastery testing is computerized testing that is tailored to the examinee and measures the mastery of some skill or set of skills. It grew out of two branches of measurement: criterion-referenced testing and item response theory (IRT). Traditionally in adaptive testing, branching occurs after every item based on some sort of selection/algorithm procedure, in the case of IRT using maximum item information at the most updated ability estimate. Later research was interested in branching or adapting after a testlet was administered, where testlets are defined as subgroups or "bundles" of items shorter in length than a test, that are interconnected through some common stimulus beyond the trait being measured. There has been little research on conditions involving multiple adaptive testlet selection procedures using maximum information item selection with a 3-parameter logistic model.

In this simulation study, four main effects and their interactions were investigated in a completely crossed design: three selection methods [traditional CAT (TCAT), maximum item information preserving testlet structure (MII-T), and maximum testlet information (MTI)], 7 levels of θ (-3, -2, … , 2, 3), 3 testlet lengths (2-item, 3-item and 5-item) and 2 item banks (600 items and 900 items). All data were simulated in a manner that reflected real data conditions based on a foreign language reading comprehension test. Three classes of criteria were evaluated: (a) recovery of θ, (b) quality of classification decisions, and (c) efficiency. The recovery of θ was evaluated through (a) correlations, (b) root mean square error (RMSE) and (c) bias. Quality of classification decisions was evaluated through (a) classification percentages, (b) average decision error, and (c) percentage of simulees for whom a 95% confidence decision could not be made. Efficiency was evaluated through (a) ϕ correlations (b) test length, and (c) the ratio of precision divided by test length.

Results indicated that θ level was by far the most influential condition, having the largest effect size (η^2), except for the correlation results. When θ level was collapsed for correlations, selection method emerged as the most influential condition. Testlet length had a smaller, more moderate amount of influence, and pool size had very little influence.

Estimation of θ was most precise near the mastery cutoff with TCAT and 2-item testlets being the two conditions with the highest precision. TCAT had a lower percentage of simulees for whom a 95% confidence decision could not be made than the two testlet selection methods. When a 95% confidence classification could not be made, TCAT had a shorter mean test length than the other two methods. The two testlet selection methods were more accurate than TCAT in terms of classification decisions. TCAT had a tendency to overidentify masters, which lowered its rate of correctly identifying nonmasters and increased its error rate of misidentifying masters. For all three selection methods, accuracy of classification increased with increased distance from the cutoff. Implications for future research involve: (a) improving estimation away from the mastery cutoff, particularly the extremes of the θ continuum to allow for equally accurate estimation along the entire θ continuum; and (b) modifying the construction of testlets to include a wider range of constructs, content balancing, and degree of the stakes (high vs. low) of the tests administered. Types of future research should include: manipulating initial θ for estimation, varying the cutoff criterion, extending the range of item/testlet difficulty of the item bank, varying the width of the testlets, expanding maximum testlet length, and balancing content/item-type within testlets.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

CHAPTER 2: METHODS

CHAPTER 3: RESULTS

List of Tables

Appendix Tables

List of Figures

Appendix Figures

CHAPTER 1: INTRODUCTION

Overview

The history of testing can be traced to about 2000 BCE in China where civil servants were tested to ascertain whether or not they were fit for their duties (Kaplan & Sacuzzo, 1997). Today, there are a multitude of tests designed to measure knowledge/skills in many content areas (ability or achievement) and professions (licensing/certification) as well as analytical tools for various clinical and counseling situations. When tests are used to make decisions, particularly high-stakes decisions, correct classification of individuals is of utmost importance. In the past, all examinees have been administered the same set of items in conventional paper-and-pencil tests and scores have been based on total number of items answered correctly. Unfortunately, this can result in long tests with fatigue having an influence on performance, since examinees are administered all items regardless of difficulty level. This leads to the question: How can we test as efficiently as possible with the greatest degree of precision? A solution to this problem, when tests are used for the purpose of classification, is adaptive mastery testing.

Adaptive mastery testing (Kingsbury & Weiss, 1980, 1981, 1983; Weiss & Kingsbury, 1984) is computerized testing that is tailored to the examinee (adaptive) and measures the mastery of some skill or set of skills. It is mainly applicable to achievement type testing where mastery is relatively easy to operationalize. It grew out of two branches of measurement: criterion-referenced testing and item response theory (IRT). Criterion-referenced testing measures achievement "in terms of an absolute standard of quality," as opposed to norm-referenced testing which measures achievement "upon a relative standard" (Glaser, 1963, p. 519). IRT is a type of measurement that takes not only the characteristics of the items into consideration but the pattern of responses as well (i.e., the interaction of ability and item parameters). It also allows for persons and items to be placed on the same scale. The remainder of this chapter reviews computerized adaptive mastery testing using models that either incorporate principles of IRT or models whose performance was compared to an IRT model. The review focuses on the more general 3-parameter logistic model and is limited to models using dichotomous scoring.

TEST THEORIES

Classical Test Theory (CTT)

Classical test theory (CTT), sometimes known as true and error score theory, states that any observed score can be broken down into two parts: true score and random error. True score is the fixed amount of the trait or attribute being measured: it is an error-free measurement, and it can never be directly observed. An observed score is an indirect, observed measurement of the trait or attribute and is associated with some degree of error.

Statistics associated with CTT. Four statistics are associated with CTT: item difficulty, item discrimination, reliability, and standard error of measurement (SEM). Item difficulty is defined as the proportion of examinees endorsing the correct or keyed response. Item discrimination is usually defined as the point-biserial correlation between the item response and total number-correct score. Reliability is conceptually defined as the proportion of observed score variance that is true score variance. Standard error of measurement (SEM) is the standard deviation of observed scores around the true score (Anastasi, 1988; Crocker & Algina, 1984).

Because these statistics are based on group performance and total number-correct scores, the main drawback of CTT is that estimates are group dependent and test dependent (Weiss & Yoes, 1991). For example, suppose an arithmetic test was administered to a group of 1st graders and a group of 6th graders. The difficulty estimates (proportion correct) of the items for the 1st graders would be much lower than those of the 6th graders. Likewise, if a math test were given to two different groups of 6th graders, but one group was administered a set of "easy" items and the other a set of "difficult" items, the total number-correct score, and in turn reliability and SEM, would be different for the two groups. This dissatisfaction with CTT was an impetus leading to the development of IRT, which is an item-driven model rather than a test-driven model and allows parameter estimates to be obtained independent of the examinees and the items used in a particular test (Weiss, 1985; Weiss & Yoes, 1991).

Criterion-Referenced Testing

Criterion-referenced testing became popular in the late 1960s and was in response to a trend in education toward individualized instruction. Jaeger (1987), in a review of educational measurement, maintained that the major advantage of criterion-referenced testing is its capabilities for diagnostic interpretation and tailoring instruction to individual needs. The origins of criterion-referenced testing are steeped in CTT and due to its nature, certain measurement problems became apparent. Hambleton and Novick (1973) stated two major problems: bandwidth-fidelity and reduced variability of scores. The bandwidth-fidelity paradox involves a trade-off between precision (fidelity) and range of measurement (bandwidth). Tests can be developed either to obtain precise information about a small range of ability or less precise information about a wider range of ability (Hambleton & Novick, 1973). This is a problem for any type of test based on CTT. The second problem, reduced variability of scores, is specific to criterion-referenced tests. Hambleton and Novick (1973) stated that criterion-referenced tests, unlike norm-referenced tests, are not explicitly constructed to maximize test variance. Scores tend to be more homogeneous, making it more difficult to order examinees along an ability or achievement continuum. According to Hambleton and Novick, this trend toward homogeneous scores is a result of a common background such as instruction or curriculum. In turn, reduced variability can lead to lower reliability (Anastasi, 1988, p.131). As a result, reliability using CTT may appear to be poor even when the test is performing well. When IRT models are combined with methods of adaptive testing, this is no longer an issue -- it is possible to obtain equally precise information about a wide range of abilities and item parameters independent of person parameters (Weiss, 1985; Weiss & Yoes, 1991).

Adaptive Testing

Adaptive testing is different from conventional testing in that the examinees do not all receive the same fixed set of items. An item is selected based on the scoring of the most recent response as well as the cumulative scoring (pattern of responses) up to that point in a particular test. It is considered a more efficient way of testing since instead of answering all the items, only those items that are considered appropriate (near the examinee’s trait level) are selected. Items that are too "easy" or too "difficult" for a given examinee are not administered. Adaptive testing has also been referred to as "tailored testing" (Lord, 1980).

The first adaptive test. The first adaptive test was developed in 1905. Alfred Binet was appointed by the French ministry of education to test and identify mentally deficient school-age children in France (Kaplan & Succuzzo, 1997). Binet’s test has all the characteristics of an adaptive test: "a mechanical branching strategy using a fixed branching rule, a variable entry point, and a variable termination criterion" (Weiss, 1985, p. 777).

While Binet's test is adaptive, it is still a norm-referenced test (age normed) (Anastasi, 1988; Kaplan & Succuzzo, 1997; Weiss, 1985) based on CTT scoring (percentage correct) with all the drawbacks of CTT (item parameters are sample-dependent and total score is item dependent). It was also individually administered and hand scored, restricting the amount of branching allowed and creating high costs in time and administration.

Computerized adaptive testing. Computerized adaptive testing (CAT) follows the same principles as Binet's adaptive test: variable starting point, item selection rule(s), and termination criterion. Administration by computer, however, increases the versatility of adaptive testing by (1) allowing for more complex mathematical models of measurement, (2) having little if no lag time for estimation of an examinee’s ability level between items, and (3) reducing administration costs and time due to the computer's great memory capacity and computational capabilities. The development of IRT ran parallel with the development of the computer and computerized testing.

IRT

History. IRT is based on a family of probabilistic models that relate latent traits or abilities that can not be directly measured and observable examinee test performance (Hambleton & Swaminathan, 1985). These models are characterized by S-shaped curves known as the item response function (IRF) and represent the probability of examinees endorsing a keyed response across varying levels of a particular trait denoted as θ (Weiss & Yoes, 1991). The exact beginnings of IRT are somewhat vague. Tucker in 1946 was the first to use the term item characteristic curve (ICC), an early term that recently has been replaced by IRF. Lawley in the mid 1940s related parameters in IRT to classical model parameters. Lazarsfeld in 1950 introduced the term "latent trait". The work of Lord in the early 1950s is considered the "birth" of IRT (see Hambleton & Swaminathan, 1985 for a more detailed history).

Assumptions. IRT differs from CTT in its assumptions and models. Weiss and Yoes (1991) and Baker (1985) list four assumptions of the dichotomous case of IRT:

(1) If the examinee knows the correct answer, s/he will answer it correctly.

(2) The probability of a correct/keyed response by an examinee is attributable to

his/her standing on some specific number (*k*) of latent traits or dimensions.

(3) For a fixed level of θ, an examinee's response to different items in a test are statistically independent (probability of the response pattern is equal to the product of the probability associated with each response); an assumption known as "local independence" (Hamilton & Swaminathan, 1985).

(4) The IRF is S-shaped and can be modeled as a cumulative normal ogive or logistic ogive.

The last assumption, the shape of the IRF, is the basis from which all dichotomous IRT models stem (Baker, 1985; Weiss & Yoes, 1991). Baker (1985) gives a succinct explanation. Every examinee is assumed to possess some degree of the underlying trait denoted as θ. For each level of θ there is a probability, P(θ), associated with it for an examinee to give a correct/keyed response. For examinees with low levels of θ, P(θ) is small, for examinees with high levels of θ, P(θ) is large. The resulting plot of θ along the *x*-axis and P(θ) along the *y*-axis is a non-linear, ogive or S-shaped curve known as the IRF (Lord, 1980; Weiss, 1985; Weiss & Yoes, 1991). This ogive indicates that for extreme levels of θ, there is a very low or very high probability of an examinee giving a correct/keyed response whereas for most of the θ range, there is a steady, but non-linear increase in probability of an examinee answering an item correctly.

In addition to relating probability of a correct/keyed response to trait level, the IRF is usually described by up to three item parameters: item difficulty, item discrimination, and pseudo-guessing. Item difficulty, *b*, is located at the center of the IRF where the slope is the steepest. A perpendicular dropped from this point on the IRF to the *x*-axis places difficulty and θ on the same scale (Lord, 1980; Weiss, 1985; Weiss & Yoes, 1991). Item discrimination, *a*, is proportional to the slope at the point of difficulty. It tells how well an item differentiates between two adjacent points (Lord, 1980; Weiss, 1985; Weiss & Yoes, 1991). If the item is very discriminating, reflecting a steep slope, two very different probabilities for a correct response for two adjacent levels of θ will result at its steepest point. If the item is not very discriminating, then the probabilities for a correct response for two contiguous levels of θ will be closer in value. Pseudo-guessing, *c*, is the probability of a correct response for a person at an extremely low θ whose response is based on uninformed guessing. This is not to be confused with informed guessing in which one or more alternatives is eliminated because they lack plausibility. Pseudo-guessing is guessing with the trait contributing no influence.

Mathematical models. One of two mathematical functions can be used to define the IRF: the cumulative normal ogive or the logistic function. Each function is actually a family of functions and can be further broken down into three models depending on the number of item parameters allowed to vary. From a mathematical perspective, the logistic function is easier to work with; the general equation for the probability of a correct response conditional on θ is

$$P_i(u_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 + c_\mathrm{i})\ [1 + \exp(-\mathrm{D}a_i(\theta\text{-}b_i)]^{-1} \qquad (1)$$

where $P_i(\mathrm{u}_{ij} = 1|\theta_j, \mathrm{a}_i, \mathrm{b}_i, \mathrm{c}_i)$ is the probability of a correct response given θ_j
a_i is the discrimination of item *i*
b_i is the difficulty of item *i*
c_i is the pseudo-guessing parameter of item *i*
θ_j is the ability level of examinee *j*
D is a constant ≅ 1.702 used to approximate the normal ogive function using the logistic function

(Weiss & Yoes, 1991).

IRT models. The one-parameter logistic model was developed by Danish mathematician Georg Rasch in the early 1960s. It is based on the assumptions that (1) all the items are equally discriminating with $a_i = 1.0$ (2) there is no pseudo-guessing (i.e., $c_i = 0$ for all items) and (3) only item difficulty is allowed to vary.

The two-parameter logistic model, like the one-parameter, fixes c_i at 0, but unlike the one-parameter model, both item difficulty and item discrimination are allowed to vary. The three-parameter logistic model allows all three item parameters to vary.

Test response function. Because items are assumed to be locally independent (i.e., the probability of answering an item does not influence the probability of answering other items), IRFs can be summed. In other words, for each level of θ, it is possible to sum the probabilities of a correct response across all the items in the test. A plot with θ along the *x*-axis by summed (or averaged) probabilities along the *y*-axis is known as a test response function (TRF; Baker, 1985; Hambleton & Swaminathan, 1985; and Weiss & Yoes, 1991). It is usually S-shaped and can be interpreted as a "sum" or "average" of the IRFs with "sum" or "average" parameters (Baker, 1985). In addition, the TRF provides a "translation" of scoring between the IRT θ scale and the CTT number- or proportion-correct scale (Kingsbury & Weiss, 1979). The summed probabilities on the *y*-axis are known as the expected number-correct score. When the summed probabilities are divided by the number of items this is known as an expected proportion-correct score (Weiss & Yoes, 1991). Thus, when θ is estimated, it is possible to convert back and forth between an IRT score (theta estimate) and a CTT score by dropping a perpendicular from the TRF to the *x*-axis (θ) and a perpendicular from the TRF to the proportion-correct *y*-axis (Kingsbury & Weiss, 1979; Weiss & Yoes, 1991, for examples, Appendix B). This ability to convert or translate from IRT scoring to CTT scoring is a crucial element in adaptive mastery testing (Kingsbury & Weiss, 1979) and will be discussed in further detail below.

Item information. Information was first defined by R.A. Fisher in the early 20s as the reciprocal of imprecision.

$$I = \frac{1}{\sigma^2{}_E} \tag{2}$$

Where $\sigma^2{}_E$ is the variance of the estimates around a particular parameter (Baker, 1985).

In IRT, item information indicates how well an item is performing (degree of precision) for various levels of θ. Item information for a particular θ level can be represented as:

$$I(\theta, U) = \sum P'^2_i / P_i Q_i \tag{3}$$

Where P'_i is the derivative of the item response function with respect to θ

U = the response 1 or 0 to item *i*

Q_i = Probability of an incorrect response to item *i*

(Hamilton & Swaminathan, p. 91, 1985).

The computational formula is:

$$I = \frac{D^2 a_i^2 Q_i}{P_i}\left[\left(P_i - c_i\right)^2 / \left(1 - c_i\right)^2\right] \tag{4}$$

(Hambleton & Swaminathan, p. 107, 1985).

Information varies with the slope of the IRF along the θ continuum. The location of the IRF where the slope is greatest or steepest, the item difficulty, is the maximum of the item information function (IIF). An item provides the most information at its difficulty level. The height of the maximum is related to discrimination (Weiss & Yoes, 1991). A more discriminating item will have a more peaked IIF than a less discriminating item, even if they are of the same difficulty. Thus, the difficulty of an item is the location on the θ continuum that provides the most information and more discriminating items provide more information than less discriminating items (Baker, 1985).

Test information function. Just as the case that the IRFs are additive to a TRF, the IIFs are additive to a test information function (TIF: Lord, 1980; Weiss, 1985; Weiss & Yoes, 1991). The TIF denotes how precisely the test measures over the range of θ. In general, increasing the number of items increases the amount of information (Baker, 1985).

Theta estimation (Maximum Likelihood Estimation: MLE). To better understand the MLE procedure, it is necessary to discuss joint probability, its relationship to the likelihood function, and how the adjustments to estimation are determined. Let P(U_i|θ) = probability of response U (1 or 0) to item i given a particular θ level. For a single item the probability of a response, given θ is:

$$\mathrm{P}(U_i|\theta) = P_i^{U_i} Q_i^{1-U_i} \qquad (5)$$

where P_i = probability of a correct response to item i
Q_i = probability of an incorrect response to item i

(Hambleton & Swaminathan, 1985).

For a set of n independent items the joint probability is:

$$\mathrm{P}(U_1, U_2, \ldots, U_n|\theta) = \mathrm{P}(U_1|\theta)\,\mathrm{P}(U_2|\theta),\ldots,\mathrm{P}(U_n|\theta) = \prod_{i=1}^{n} P_i^{U_i} Q_i^{1-U_i} \qquad (6)$$

(Hambleton & Swaminathan, 1985).

When the random variable U_i takes on a value of 1 or 0 (u_i), Equation 6 becomes a likelihood function (LF).

$$\mathrm{L}(\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_\mathrm{n}|\theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} \qquad (7)$$

(Hambleton & Swaminathan, 1985 p. 77).

The maximum of the LF is the estimate of θ. It is the value that has the highest product of probabilities given the observed response pattern (Hambleton & Swaminathan, 1985). The maximum for any function is the first derivative of the function set equal to 0. In the case of the LF, determining the derivative is quite complex mathematically.

The natural log of the LF, ln L(u_1, u_2, …, u_n|θ), is monotonically related to the LF and therefore reaches it maximum at the same θ value as does the LF (Hambleton & Swaminathan, 1985). Since it is much easier to define and there are special properties associated with multiplication of natural logs, the derivative of the natural log is used (see Equation 8).

$$\frac{d}{d\theta} \ln L(u|\theta) = 0 \qquad (8)$$

(Hambleton & Swaminathan, 1985). To solve Equation 8, an iterative procedure like Newton-Raphson is needed (Hambleton & Swaminathan, 1985).

Research Methods for Obtaining Response Data Sets

In prior literature on CAT, response data sets and person parameters were collected or generated in one of three ways: live testing, post-hoc simulation, or Monte Carlo simulation.

Live Testing

Live testing is the more traditional method of obtaining data. Examinees are administered items that may or may not have pre-calibrated item parameters. Based on response patterns and item parameter estimates, $\hat{\theta}$ is estimated for the examinees.

Post-Hoc Simulation

Post-hoc simulation is a modified version of live testing that combines traditional data collection and simulation. Like live testing, examinees are administered items and response patterns are recorded. Based on the item parameters and item responses scored 0-1, the test is "re-administered" post-hoc to the examinee adaptively as a simulation. The same set of items is administered but the order of administration is now based on maximum information at the most updated $\hat{\theta}$. For example, suppose an examinee has an initial $\hat{\theta}$ = 0.5. An item that has not yet been administered in the post-hoc simulation will be selected that maximizes information at that $\hat{\theta}$. Based on the 0-1 response from the live testing, $\hat{\theta}$ will be updated and the process repeated. This will continue until a pre-specified termination criterion has been reached.

Monte Carlo Simulation

Monte Carlo simulation is a pure simulation requiring no actual examinees. Response data are simulated based on item and person parameters. These item and person parameters are often based on calibrations from live testing or may be based on some theory. In a Monte Carlo simulation, two values are compared in order to simulate a 0-1 response matrix. The first value is randomly chosen from a uniform distribution ranging from 0 to 1. The second value is $P(\theta_j)$, the probability of a correct response to item *i* with θ level *j* (see Equation 1). If $P(\theta_j)$ is greater than the random number a response of 0 is denoted. If $P(\theta_j)$ is less than the random number a response of 1 is denoted. This is done across all items for all examines, yielding a simulated response data set. Once a simulated response data set has been obtained, $\hat{\theta}$ can be estimated using the item parameters and the response data set.

SEQUENTIAL-BASED MODELS OF MASTERY TESTING

SPRT-Based Models

SPRT

Wald (1947) was interested in quality control and developed a statistical method for testing large consignments of light bulbs using small samples, while at the same time maintaining a prespecified level of quality. Testing the elements (light bulbs) was done sequentially and decision-making regarding the outcome of each sampling was based on traditional hypothesis testing where the null and alternative hypotheses are:

H_o: $p = p_o$

H_1: $p = p_1$

where p = the proportion of defective bulbs in the population
p_o= upper limit permitted of defective light bulbs to accept the consignment
p_1 = lower limit permitted of defective light bulbs to reject the consignment.

Each stage of sampling is a Bernoulli trial (random sampling without replacement where each element is equally likely to be chosen and has either a defective or nondefective status). The probability of observing a specified number of defects in a certain sample size given H_o is true or given that H_1 is true follows the binomial probability density function.

Under H_o: $p = p_o$

$$p_{o_m} = p_o^{(m-W_m)}(1 - p_o)^{W_m} \tag{9}$$

Under H_1: $p = p_1$

$$p_{1_m} = p_1^{(m-W_m)}(1 - p_1)^{W_m} \tag{10}$$

where m = number of elements in the sample
W= number of defective elements (Kingsbury & Weiss, 1980).

The log of the ratio of these two probabilities is an indication of the strength or magnitude of the hypotheses (Kingsbury & Weiss, 1980). After each sampling, one of three possible decisions is made (accept H_o, reject H_o, or continue sampling) depending on the magnitude of the ratio (hence the name “sequential probability ratio testing”) and the values chosen for α (falsely rejecting a quality consignment) and β (falsely accepting an inadequate consignment).

If $\text{Log}\frac{P_{1_m}}{P_{o_m}} \geq \text{Log}\frac{1-\beta}{\alpha}$ then reject the consignment (assume H_1 is true) (11)

(Kingsbury & Weiss, 1980).

In other words, this ratio says that if the ratio of H_1 over H_0 is greater than the ratio of correctly rejecting over mistakenly rejecting, then consider the consignment inadequate.

If $\text{Log}\frac{P_{1_m}}{P_{o_m}} \leq \text{Log}\frac{\beta}{1-\alpha}$ then accept the consignment (assume H_0 is true) (12)

(Kingsbury & Weiss, 1980).

Likewise, this ratio says that if the ratio of H_1 over H_0 is less than the ratio of falsely accepting over correctly accepting then consider the consignment adequate.

If, however, $\text{Log}\frac{\beta}{1-\alpha} \leq \text{Log}\frac{P_{1_m}}{P_{o_m}} \leq \text{Log}\frac{1-\beta}{\alpha}$ then continue sampling (13)

(Kingsbury & Weiss, 1980).

An early educational application of the SPRT. Ferguson (1969) was the earliest to apply the SPRT model in an educational setting by replacing light bulbs with items and consignments with domains of learning. The purpose of his research was to create an individualized type of test to be used in conjunction with individualized instruction. The test was structured to reflect a hypothesized hierarchy of 18 objectives in addition and subtraction targeted to 3rd and 4th grade children. For each objective, there was an appropriate set of items. Within an objective, the SPRT model was employed to make a decision of mastery or nonmastery of that objective. Hierarchical branching was based on theory and expert opinion.

Like Wald's model, Ferguson's test within a particular objective employed Bernoulli type trials: two possible outcomes for each trial or item (correct or incorrect), the probability of each outcome was assumed constant over trials (i.e., difficulty and discrimination were fixed across items), and the outcome of any trial was assumed to be independent of other trials (i.e., administration of an earlier item does not influence the probability of correctly answering a later item). Different examinees received different sets of items and the tests themselves were of varying lengths (Ferguson, 1969). The null and alternative hypotheses were:

H_o: $p = p_o$

H_1: $p = p_1$

where p = proportion of items answered *incorrectly* for the population of items
p_o = proportion answered *incorrectly* by a master
p_1 = proportion answered *incorrectly* by a nonmaster (Ferguson, 1969).

Type I and Type II errors were:

α = Erroneously declaring a master a nonmaster
β = Erroneously declaring a nonmaster a master

Based on expert consultation, α was set at .10 and β was set at .20. For most of the objectives, $p_o = .15$ and $p_1 = .40$. If $p < p_o$ (the examinee answered less than 15% of the items incorrectly) a decision of mastery was made, if $p > p_1$ (the examinee answered more than 40% of the items incorrectly) a decision of nonmastery was made. If $p_o < p < p_1$ then testing continued.

For the null hypothesis the probability of a response pattern is:

$$p_o^{w_m}\left(1 - p_o\right)^{r_m} \qquad (14)$$

and for the alternative hypothesis the probability of response pattern is:

$$p_1^{w_m}\left(1-p_1\right)^{r_m} \tag{15}$$

where $\mathrm{w_m}$ = the number of items answered incorrectly in sample size *m*
$\mathrm{r_m}$ = the number of items answered correctly in sample size *m* (Ferguson, 1969)

The sequential probability ratio based on these two proportions was used and an acceptance value or criterion, a_{m}, and a rejection value, b_{m}, were computed based on α, β, $\mathrm{p_o}$, and $\mathrm{p_1}$. The acceptance and rejection numbers are related to cutoffs chosen. If the number of items in the sample that were incorrect exceeded the rejection value ($w_{\mathrm{m}} > b_{\mathrm{m}}$) then a decision of nonmastery was made. If the number of incorrect items in the sample was below the acceptable value ($w_{\mathrm{m}} < a_{\mathrm{m}}$) then a decision of mastery was made. If the number of incorrect items in the sample was between the acceptable and rejection value ($a_{\mathrm{m}} < w_{\mathrm{m}} < b_{\mathrm{m}}$) then testing continued. Once a decision was made regarding a particular objective, the test branched to a new objective.

The test was administered to 75 elementary students grades 1 through 6 in an experimental school emphasizing individualized prescribed instruction (IPI). There were three ability groups: low proficiency (n = 10), middle proficiency (n = 55), and high proficiency (n = 10). These groupings were made by the IPI coordinator. Exposure to the unit of learning was varied. Of the 75 students, 28 had not yet worked on the unit, 11 were working on it, and 36 had already completed the unit. The sequential portion of the test was done on computer and the untested objectives following the decision were completed using paper and pencil. The results showed a reduction of the number of objectives as well as the number of items within each objective prior to a mastery/nonmastery decision. In other words, fewer items were needed to make a mastery decision. The paper-and-pencil test consisted of 150 items with 18 objectives. The test using SPRT yielded means of 7.11 objectives and 52.12 items.

While Ferguson (1969) was successful in employing the SPRT model in an educational setting, three criticisms can be made: homogeneity of the items assumption, values chosen for α and β, and characteristics of the sample of examinees tested. Assuming that items have equal difficulty and equal discrimination is problematic in education. Although it might be appropriate to make this assumption for relatively simple and limited tasks such as addition and subtraction, research shows that it is violated for more complex tasks such as advanced algebra or calculus (Fischer, 1973; Medina-Diaz, 1992). The second concern is the method and values chosen for α and β. Type II error in this study (erroneously declaring a nonmaster a master) is a more serious error than Type I (erroneously declaring a master a nonmaster). The former can result in continued instruction at an inappropriate level or in other applications acceptance to a program or licensure of an unqualified candidate who does not meet the criteria, whereas the Type I error only results in unnecessary practice and time before retaking the test. Even though Type II is a more serious error, a greater magnitude (.20) was tolerated than for Type I (.10) in Ferguson’s study. The report only states that expert consultation was used to determine these values. A more quantitative approach such as was used in Lewis and Sheehan’s (1990) study, in which the values of α and β were determined through simulations and statistical tests, was needed. Lastly, Ferguson used 1st through 6th grade students with a test designed for 3rd and 4th grade students. It can be argued that by administering items of inappropriate level to part of the sample, variance was artificially increased. Despite these criticisms, however, Ferguson's study has been a stepping stone to other SPRT studies and in some cases modification of the model.

An example using SPRT. Ferguson's (1969) null and alternative hypotheses were concerned with the probability of an *incorrect* answer conditional on status whereas other studies hypothesized about the probability of a *correct* answer conditional on status (Frick, 1992, 1990, 1989; Kingsbury & Weiss, 1983, 1980; Lewis & Sheehan; 1990;

Reckase, 1983; Weiss & Kingsbury, 1984). This example will follow the latter studies. Hence, the null hypothesis and alternative hypotheses are H_o: $p = p_o$ and H_1: $p = p_1$ respectively where

p = proportion of items answered *correctly* for the population of items
p_o = proportion answered *correctly* by a nonmaster
p_1 = proportion answered *correctly* by a master.

Type I and Type II errors will then be

α = Erroneously declaring a nonmaster a master
β = Erroneously declaring a master a nonmaster.

Probabilities contingent on status will be

[P(C | M)] = Probability of a correct response conditional on mastery status
[P(C | N)] = Probability of a correct response conditional on non-mastery status
[P(I | M)] = Probability of an incorrect response conditional on mastery status
[P(I | N)] = Probability of an incorrect response conditional on non-mastery status

Assume that the following probabilities have been determined from prior data:

[P(C | M)] = 0.8	[P(I | M)] = 0.2
[P(C | N)] = 0.3	[P(I | N)] = 0.7.

It will be assumed that with no prior information the prior probabilities of status are equally likely [P(M) = P(N) = 0.5]. After each item is selected and scored, the probability of status [P(M) or P(N)] is updated by multiplying the prior probability status [P(M) or P(N)] × probability of a correct or incorrect response conditional on status [P(C | M), P(C | N), P(I | M) or P(I | N)] yielding a joint probability. This example is summarized in Table 1. For example, in Table 1, Item 1 was answered correctly. This yielded a joint probability of (.5 × .8) or .40 for a master and (.5 × .3) or .15 for a nonmaster. These two probabilities do not sum to 1. To make these joint probabilities sum to 1 they must be normalized. This is accomplished by dividing the joint probability for a particular status by the sum of the two joint probabilities. For Item 1, the sum of the two joint probabilities is .55. Hence, the posterior probability for a master is (.40/.55) or .73 and the posterior probability for a nonmaster is (.15/.55) or .27. Table 1 lists all the prior, joint, posterior probabilities for 6 items with the response pattern of (C, I, C, C, C, C).

Table 1
Example of SPRT with a 6-Item Test

Item	Status	Prior Probability of Status		Probability of Response		Joint Probability	Posterior Probability	Probability Ratio
1	M	.5	×	.8	=	.40	.73	
(correct)	N	.5	×	.3	=	.15	.27	2.703
						sum = .550	1.0	
2	M	.73	×	.2	=	.146	.44	
(incorrect)	N	.27	×	.7	=	.189	.56	.786
						sum = .335	1.0	
3	M	.44	×	.8	=	.352	.68	
(correct)	N	.56	×	.3	=	.168	.32	2.125
						sum = .520	1.0	
4	M	.68	×	.8	=	.544	.85	
(correct)	N	.32	×	.3	=	.096	.15	5.667
						sum = .640	1.0	
5	M	.85	×	.8	=	.68	.94	
(correct)	N	.15	×	.3	=	.045	.06	15.667
						sum = .725	1.0	
6	M	.94	×	.8	=	.752	.98	
(correct)	N	.06	×	.3	=	.018	.02	49
						sum = .770	1.0	

In addition to the posterior probabilities, the probability ratios are shown in Table 1. These values can be compared to the upper and lower bounds for nonmastery and mastery respectively. If, in the example, it is assumed that $\alpha = \beta = .025$, then the lower bound for the mastery threshold $[(1 - \beta)/\alpha] = (1 - .025/.025) = 39$ and the upper bound for nonmastery [$\beta/(1 - \alpha)$] = .025/(1 - .025) = .0256. Comparing the probability ratios to these upper and lower bounds in Table 1 shows that the first 5 items are between these two boundaries and testing must continue. Item 6 yields a probability ratio of 49 which is greater than the lower bound for mastery, indicating termination of testing with a decision of mastery status.

Robustness of the SPRT. The prior example assumes equality of item characteristics and subsequent use of average (expected) probabilities. This has been a main criticism of SPRT, so Frick (1989) was interested in testing the robustness of the SPRT model when item parameters were known to vary. His study consisted of two parts: a real-time conventional test and a post-hoc simulated SPRT test. Two different computer-based tests were used: the structure and syntax of digital authoring language (DAL) and knowledge of how computers functionally work (COM). The DAL test consisted of 97 items and was administered to 105 former and present graduate students who were either taking a course on DAL or had already taken the course. The COM test consisted of 85 items and was administered to 53 students (approximately two-thirds were graduate students and one-third were undergraduates) taking a course on computers in instruction. Items were of varying difficulty and varied in format (multiple-choice, dichotomous, and short answer). In addition, testing was done at different points of time during the course to ensure increased variability of scores due to varied exposure to the domain of learning. Examinees were administered the entire set of items on computer with items randomly selected. The score, the particular order of the items, and the response pattern were saved for later use in the post-hoc SPRT simulation. Both tests had high internal consistency reliability (α = .98 for DAL and α = .94 for COM). Mean score for DAL was 63% correct (SD = 24.6) and mean score for COM was 79% correct (SD = 13.6) (Frick, 1989).

The SPRT parameters were set *a priori* with .85 as the probability of a correct response, given that the examinee was a master, and .60 as the probability of an incorrect response for a nonmaster. The values for α and β were both .025. An SPRT test was created retroactively by selecting the items in the same order they were presented originally, examining the response pattern, computing the probability ratio, and making a decision of mastery, nonmastery, or continuation of testing.

Results from the post-hoc SPRT simulation indicated a substantial reduction in the number of items required to make a decision (Frick, 1989). For the DAL test, the mean number of items for a mastery decision was 19.1 and the mean number of items for a nonmastery decision was 17.4. For the COM test, the mean number of items for mastery and nonmastery decisions were 21.6 and 18.6, respectively.

Agreement of classification between SPRT and conventional testing was also compared. Mastery for the conventional test was determined to be the midway point between .60 and .85, or .725. Frick (1989) stated that agreement between SPRT decisions and conventional decisions was very high (.96 for DAL and .99 for COM).

Frick's (1989) study provides evidence that SPRT has a certain degree of robustness, at least when compared with conventional testing, but the exact degree is unknown. Frick did not manipulate different conditions in a systematic manner to determine what SPRT's capabilities really were. He warned that SPRT can be a problem if the representativeness of the sampled items is flawed. If a set of very difficult or very easy items was administered initially, the test can terminate too quickly with an incorrect decision. He suggested choosing very low decision error rates when item parameters are known to vary. Unfortunately, this study did not report data to substantiate either the warning or the suggestion. He also did not vary values for α and β, SPRT parameters, or classical cutoffs. All of these conditions need to be explored before coming to any conclusions regarding the robustness of the SPRT.

EXSPRT

Expert sequential probability ratio theory (EXSPRT) was Frick's (1992) modification of the SPRT model. He was interested in an alternative to IRT-based adaptive testing that did not require large sample sizes (250 - 1,000) for the calibration of item parameters, yet still took item characteristics into account. EXSPRT is based in artificial intelligence and the use of expert systems. An expert system is a type of decision-making instrument that consists of a set of questions and a knowledge base. An "inference engine" uses answers to the questions and the knowledge base to choose from a set of discrete alternatives. This goal of choosing from a few discrete alternatives rather than a continuum of alternatives in IRT-based item bank should yield similar results with a smaller data sample of examinees (Frick, 1992).

The basic EXSPRT model. To calibrate probabilities for masters and nonmasters, the entire pool of items is given to a representative sample of examinees where it is preferable that 50% are masters and 50% are nonmasters. A mastery cut-off is chosen and the group is divided up according to total test score and the cut-off. Probabilities are determined as follows:

$$P(C_i \mid M) = (nr_{im} + 1)/(nr_{im} + nw_{im} + 2) \qquad (16)$$

$$P(-C_i \mid M) = 1 - P(C_i \mid M) \qquad (17)$$

$$P(C_i \mid N) = (nr_{in} + 1)/(nr_{in} + nw_{in} + 2) \qquad (18)$$

$$P(-C_i \mid N) = 1 - P(C_i \mid N) \qquad (19)$$

where $P(C_i \mid M)$ = probability of a correct response to item *i* given that the examinee is a master,
$P(-C_i \mid M)$ = probability of an incorrect response to item *i* given that the examinee is a master,
$P(C_i \mid N)$ = probability of a correct response to item *i* given that the examinee is a nonmaster,
$P(-C_i \mid N)$ = probability of an incorrect response to item *i* given that the examinee

is a nonmaster,
nr_{im} = number of persons in the mastery group who answered item *i* correctly,
nw_{im}= number of persons in the mastery group who answered item i incorrectly,
nr_{in} = number of persons in the mastery group who answered item *i* correctly, and
nw_{in}= number of persons in the mastery group who answered item i incorrectly

(Frick, 1992).

A likelihood ratio is determined based on these probabilities:

$$LR = \frac{P_{om} \prod_{i=1}^{n} P(C_i | M)^s \left[1 - P(C_i | M)\right]^f}{P_{on} \prod_{i=1}^{n} P(C_i | N)^s \left[1 - P(C_i | N)\right]^f} \tag{20}$$

where P_{om} = prior probability that the examinee is a master,
P_{on} = prior probability that the examinee is a nonmaster,
$s = 1, f = 0$ if item *i* answered correctly,
$s = 0, f = 1$ if item *i* answered incorrectly, and
$s = 0, f = 0$ if item *i* not administered.

The stopping rule is based on the traditional SPRT criteria:

If $LR \geq (1 - \beta)/\alpha$, terminate and choose mastery, or

If $LR \leq \beta/(1 - \alpha)$, terminate and choose nonmastery,

otherwise select another question, update *LR*, and make a decision whether or not to continue based on the updated *LR*.

Item selection. Within EXSPRT there are two models based on item selection: Random item selection (EXSPRT-R) and intelligent item selection (EXSPRT-I). EXPSRT-R employs random item selection without replacement. It differs from SPRT in that the probabilities for each item are estimated individually based on responses by masters and nonmasters rather than an average probability across all the items. EXSPRT-I is modeled after Kingsbury and Weiss' (1980) maximum information search and selection (MISS) procedure. Items are chosen that best discriminate between masters and nonmasters and are least incompatible with the current estimate of achievement (Frick, 1992).

In the EXSPRT-I model, three item parameters are estimated for use in this "intelligent" search: item discrimination, item/examinee incompatibility, and item utility. Item discrimination is very similar to the item discrimination index used in CTT and is defined as:

$$D_i = P(C_i \mid M) - P(C_i \mid N) \tag{21}$$

where D_i = discrimination index for item *i* (Frick, 1992).

In other words the discrimination index is the probability of a master getting a correct response minus the probability of a nonmaster obtaining a correct response.

Item/examinee incompatibility is defined as:

$$I_{ij} = abs\{(1 - P(C_i)) - E(\Phi_j)\} \tag{22}$$

where $E(\Phi_j) = (nr_j + 1)/(nr_j + nw_j + 2)$ (estimated ability of examinee j)
$P(C_i) = (nr_i + 1)/(nr_i + nw_i + 2)$ (proportion-correct of item i irrespective of status)

(Frick, 1992).

Incompatibility is comparable to item difficulty in CTT. It is the discrepancy between ability and the probability of an incorrect response.

Item utility is defined as:

$$U_{ij} = D_{ij}/(I_{ij} + \delta) \tag{23}$$

where δ = some arbitrary small constant to avoid dividing by 0 (Frick, 1992).

Item utility is similar to item information in IRT and is the ratio of discrimination to incompatibility with person *j*'s achievement level. In EXSPRT-I, the next item chosen is the one that has the highest utility of those not yet chosen. Like information, utility will change during a test depending on examinee performance. Research using this model is discussed below in the comparison studies section.

A Sequential Procedure Using Bayesian Posterior Beta Distributions

Sequential testing using beta distributions has the same assumptions as SPRT: random sampling without replacement, independent observations, and equivalent treatment of items. The probability density function for the posterior beta distribution is:

$$\text{beta}\ (\Phi|\ s,f\,) = \frac{(s+f+1)!}{s!\,f\,!}\Phi^{s}(1-\Phi)^{f} \qquad (24)$$

where Φ is the estimated proportion correct

s is the number of successes

f is the number of failures (Frick, 1990).

This model assumes that the prior distribution is flat [beta(Φ|0,0)] and *s* and *f* are integers greater than or equal to zero. A distribution is formed from sets with the same number of successes and failures as Φ and is allowed to vary from 0 to 1. For example, when Φ = .85 with 3 successes and 2 failures, the posterior beta density is:

$$\text{beta}(.85|3,2) = \frac{(3+2+1)!}{3!2!}.85^{3}(1–.85)^{2} = .82906875$$

For Φ = .90 with 3 successes and 2 failures the posterior beta density is:

$$\text{beta}(.90|3,2) = \frac{(3+2+1)!}{3!2!}.90^{3}(1–.90)^{2} = .4374$$

To find a range such as the prob (Φ ≥ .85) numerical integration is required (see Frick, 1990, p. 488-489 for a more detailed explanation using Simpson’s rule).

Both Novick and Lewis (1974) and Tennyson, Christensen, and Park (1984) used Bayesian posterior beta distributions in their research. Novick and Lewis (1974) were interested in prescribing test length while minimizing losses due to false advancements and false retentions.

Tennyson, Christensen, and Park (1984) were interested in assessing concept attainment by adapting amount of instruction. Performance was measured by the number of interrogatory examples answered correctly in concept learning.

In both cases, the Bayesian posterior beta distributions were used to estimate the probability that Φ, an estimate of true proportion-correct score, is greater than or equal to a prespecified cutoff (Φ_c) given a particular response pattern of successes and failures and used the rule:

$$\text{If} \qquad \frac{\text{prob}\left(\Phi \geq \Phi_c | s, f\right)}{\text{prob}\left(\Phi \langle \Phi_c | s, f\right)} \geq \frac{a}{b}\text{, then advance the student} \qquad (25)$$

where *a* is the loss due to false advancement
b is the loss due to false retention (Novick & Lewis, 1974).

Since the beta distribution makes the same assumptions as SPRT, the same criticisms of equal treatment of the items and average probabilities can be made.

Comparison of the SPRT, Beta Model and Binomial Model

Both SPRT and sequential testing using beta distributions are similar to the binomial model. There is random selection without replacement and each event is equally likely to occur. The binomial model can be expressed as:

$$\binom{n}{r} p^r (1-p)^{n-r} \qquad (26)$$

where n = total number of events,
r = number of successes,
p = probability of a success occurring, and
$\binom{n}{r} = \frac{n!}{r!(n-r)!}$ (number of ways of obtaining r success from n events)

(Ross, 1984).

Frick (1990) demonstrated that the SPRT probability ratio is the same as the binomial model where $\binom{n}{r}$ has been canceled out in the numerator and denominator.

$$PR = \frac{P_m^s(1-P_m)^f}{P_n^s(1-P)^f} \tag{27}$$

where s = number of successes,
f = number of failures,
P_m = probability of success for a master, and
P_n = probability of success for a nonmaster (Frick, 1990).

The beta model uses Bayesian posterior beta distributions to calculate the probability that Φ (estimated true proportion correct) is greater than or equal to cut-off Φ_c (Frick, 1990). The probability density function for the posterior beta distribution is:

$$\text{beta}(\Phi \mid s, f) = \frac{(s+f+1)!}{s!\,f!}\Phi^s(1-\Phi)^f \tag{28}$$

(Frick, 1990).

This model assumes a flat prior distribution [beta (Φ| 0,0)] and *s* and *f* are positive integers greater than or equal to 0. In order to better compare across models, Frick (1990) suggested using a probability ratio of the posterior beta distribution to SPRT criteria:

$$\text{IF } \frac{\text{prob}(\Phi \geq \Phi_c|s, f)}{\text{prob}(\Phi < \Phi_c|s, f)} \geq \frac{(1-\beta)}{\alpha} \qquad \text{advance student} \tag{29a}$$

$$\text{IF } \frac{\text{prob}(\Phi \geq \Phi_c|s, f)}{\text{prob}(\Phi < \Phi_c|s, f)} \leq \frac{\beta}{(1-\alpha)} \qquad \text{retain student} \tag{29b}$$

ELSE continue testing.

ADAPTIVE MODELS OF MASTERY TESTING

Adaptive Mastery Testing

Kingsbury and Weiss (1979, 1980, 1981, 1983) were interested in combining computerized adaptive testing (CAT) and criterion-referenced testing into "adaptive mastery testing" (AMT). In CAT, IRT is used to determine the probability of a correct response, update $\hat{\theta}$ based on the cumulative pattern of responses using either MLE or Bayesian estimation, and select new items based on maximum information. In CAT, an initial $\hat{\theta}$ is used as a starting point. An item is administered. If the item is answered correctly, a more difficult item is administered. If the item is answered incorrectly, an easier item is administered. Each time an item is administered, $\hat{\theta}$ is updated based on the item parameters of the items administered and the cumulative pattern of response. Selection of the next item is based on maximum item information. The item with parameters that provide the most information relative to the current $\hat{\theta}$ is selected next. In AMT, termination occurs when a confidence interval around $\hat{\theta}$ is completely above or completely below some pre-selected mastery criterion; a 95% confidence interval has typically been used, but the size of the interval can be varied. Adaptive mastery testing is similar to sequential testing in the sense that items are administered one at a time, and based on the response a decision is made to either terminate the test with a status category (mastery or nonmastery) or to continue testing. What differs, however, is the treatment of item characteristics, the item selection algorithm, the scoring procedure, and the criteria for termination of the test. Items are *not* treated as replicates of each other. They vary in degree of difficulty, discrimination, and guessing. These characteristics are taken into account in determining the probability of a correct response formula (see Equation 1). Items are selected based on the MISS strategy. Kingsbury and Weiss (1979) were interested in examining performance of trainees on a computer-managed Weapon

Mechanics course at Lowry Air Force Base. Instruction was self-paced and consisted of 13 blocks of instruction. Multiple testing was done within each block. For this study, two tests were chosen from two different blocks with responses from 200 examinees for each test. The conventional length for the first test was 30 items and for the second test 50 items. All items were of multiple-choice format with five alternatives each. The study itself was a post-hoc simulation in which performance on a traditional full-length conventional test (real data) was compared to a retroactive performance (simulation) of the same items with the same responses had the test been adaptive.

There were three components of the study: item parameter estimation, item fit analyses, and post-hoc simulation of an adaptive test. Classical item parameters were used as the initial parameter estimates for IRT parameters. Then an ancillary correction procedure was used to obtain more precise estimates. For θ estimates, number-correct score was the initial estimate and then scores were replaced with Bayesian modal estimates. Items were culled in the first stage using Urry's item parameterization method (Urry, 1977). Items were excluded that met these conditions in the first stage: (1) a_is less than .80, (2) b_is less than -4.0 or greater than +4.0, and (3) c_is greater than .30. Once the parameters were estimated, dimensionality was examined through factor analysis and parameter estimates obtained in this study were compared to the range of parameter estimates in other 3- parameter model studies that demonstrated increased test efficiency. The final stage consisted of a simulated adaptive mastery test (AMT). Each examinee began with a prior achievement estimate of 0 and a prior variance of 1.0. Items were selected that maximized information for the trait estimate. Responses were the responses from the conventional test. Each time an item was selected, based on the response, $\hat{\theta}$ was updated, a 95% Bayesian confidence interval around the estimate was computed, and a new item was selected that maximized the information. Testing continued until a 95% Bayesian confidence interval no longer included the mastery level. Three levels of mastery,converted from the TRF to the θ scale, were included: *P* = .7, .8, and .9.

Evaluation of the study used three criteria: mean number of items, mean information, and correlation of decisions made between conventional and AMT procedures. This evaluation was made across four groups: (1) the total group of trainees, (2) declared masters, (3) declared nonmasters, and (4) trainees for which the AMT procedure made full confidence decisions (the 95% confidence interval around $\hat{\theta}$ was completely above or below the cutoff). Comparisons were made for the four groups across the three mastery levels and two types of testing (conventional and adaptive). For all four groups, there was a reduction in the mean test length using the AMT procedure. For the 25-item paper-and-pencil test, the mean test length for AMT version ranged from 7.0 to 20.8 items depending on the cutoff level and degree of confidence. For the 38-item paper-and-pencil test, the AMT mean test length ranged from 7.2 to 27.7. Mean information either increased or remained the same when using the AMT procedure. Correspondence between decisions was evaluated using the phi correlation. Across both tests and all mastery levels, the two procedures agreed for 95.9% of the trainees. It was slightly higher for the longer test and slightly lower for the shorter test. The AMT information function closely approximated the conventional test information function in the mastery level region but was lower than the conventional information function further away from the mastery level region (Kingsbury & Weiss, 1979).

The results provide strong evidence that adaptive mastery testing “works”. In other words, it agreed with conventional decisions while at the same time reduced the mean test length, making it a more efficient testing instrument. There are a few areas in the Kingsbury and Weiss (1979) study that are worth examining more closely. One is the issue of cut-offs. Kingsbury and Weiss did not investigate whether or not there were ceiling or floor effects with cutoffs. They did mention that it took more items to make a mastery decision as the mastery level became higher. Kingsbury and Weiss said it was because the cutoff falls above the steepest slope of the TCC. It is unfortunate that the high-confidence group was not further broken down into masters and nonmasters and this issue examined in a more systematic manner using phi correlations. As it had been alluded to in the study, it would seem that it is much more difficult (due to a ceiling effect) to make a full confidence mastery decision (ceiling effect) when the cut-off is .9 than when the cutoff is .7. Likewise, there is less room for a confidence interval (floor effect) to make a nonmastery decision at the .7 level than at the .9 level. Had the high-confidence group been further broken down, these effects could have been examined for mean test length and accuracy of decisions. .

A second issue that was not addressed are guidelines for suitable parameters in an adaptive mastery testing situation. Kingsbury and Weiss followed Urry's (1977) guidelines (a_is greater than .80, b_is widely and evenly distributed between -2.0 and +2.0, and c_is less than .30). Their b_is, though, were sparsely represented above 1.0. Kingsbury and Weiss (1979) appeared not to be concerned about the distribution of the b_is since their purpose was mastery testing as opposed to wide-range ability testing. Their discriminations, on the other hand, follow Urry's (1977) suggestion of being greater than .80, yet achievement type tests typically have lower discriminations than wide-range ability tests (Kingsbury & Zara, 1991). Urry's (1977) guidelines were designed for more wide-range ability testing. Guidelines concerning ranges of item parameters in an adaptive mastery setting that yield the most accurate decisions and/or shortest tests in an adaptive mastery setting need to be investigated more thoroughly.

Bayesian-Based Techniques

Bayes' theorem, discussed earlier in the section on parameter estimation, can be used in several stages of the adaptive mastery procedure. It can be used to estimate θ, select the next item or classify the examinees as masters/nonmasters.

For Bayesian estimation, each examinee has a prior trait estimate, prior variance of that estimate, and a confidence interval associated with that estimate. An item is selected and answered. Based on the response and the prior, a posterior estimate of the mean and variance is obtained. For each new item administered, the posterior from the previous items becomes the new prior and the trait estimate is updated again based on the response and the prior distribution values. Item selection is based on which item will reduce the posterior variance the most (Kingsbury & Zara, 1989).

Misclassification is probably the most prominent concern in adaptive mastery testing, especially in high-stakes tests such as licensing exams. Bayesian decision theory attempts to minimize the error of misclassification through use of a risk function and the cost of taking another observation. Risk is defined as the expected loss (relationship between the trait and decision) given a particular decision (Reckase, 1979). The termination rule is based on this risk. If the expected risk after administering another item plus the cost of the additional item is less than the risk prior to taking the observation, then the testing continues. If this expected risk plus the cost of an additional item is greater, then testing should stop (Reckase, 1979). Classification can be made by determining whether the Bayesian confidence is located completely above or below the mastery cutoff (Kingsbury & Weiss, 1979, 1980, 1983; Weiss & Kingsbury, 1984).

Reckase's (1979) simplified example. Reckase illustrated a simplified example of a Bayesian sequential decision process using two θ levels (θ_j = -.8 and θ_j = +.8) and to simplify the arithmetic, the one-parameter logistic model. The following is a summary of that example. The first step is the establishment of a loss function table (Table 2).

Table 2
Loss Function

	Decision	
θ_j	d_1	d_2
+.8	0	15
-.8	25	0

where d_1 = the decision to classify as a master (above the criterion score of 0.0)
d_2 = the decision to classify as a nonmaster (below the criterion score of 0.0)

(Reckase, 1979).

In Reckase's example the loss values of erroneously classifying an examinee as nonmaster (15) and of erroneously classifying an examinee as master (25) were set arbitrarily and units were not reported. Likewise, it was an arbitrary supposition in this example that a randomly selected person with ability .8 is .6 and that a randomly selected person with ability -.8 is .4.

The second step is to determine the risk (expected loss) associated with a particular decision when no observations have been made. Prior information about the θ distribution of the examinees is needed. In this example, $P(\theta = +.8) = .6$ and $P(\theta = -.8) = .4$. Using Bayes theorem for conditional probability,

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \tag{30}$$

where A and B are events,

the expected loss if a mastery decision (d_1) with no observations is:

$$\begin{aligned} E(\text{loss}|\text{d}_1) &= P(\theta_1)\,\ell\,(\text{d}_1|\theta_1) + P(\theta_2)\,\ell\,(\text{d}_1|\theta_2) \\ &= P(\theta = -.8)\,\ell\,(\text{d}_1|\theta = -.8) + P(\theta = +.8)\,\ell\,(\text{d}_1|\theta = +.8) \\ &= .4 \times 25 + .6 \times 0 \\ &= 10 \end{aligned} \tag{31}$$

Where $P(\theta_\text{i})$ = prior probability of θ_i
$\ell\,(\text{d}_\text{j}|\theta_\text{i})$ = the loss from making decision d_j when θ_i is true

Likewise, the expected loss of making a nonmastery decision (d_2) with no observations is

$$\begin{aligned} E(\text{loss}|\text{d}_2) &= P(\theta_1)\,\ell\,(\text{d}_2|\theta_1) + P(\theta_2)\,\ell\,(\text{d}_2|\theta_2) \\ &= P(\theta = -.8)\,\ell\,(\text{d}_2|\theta = -.8) + P(\theta = +.8)\,\ell\,(\text{d}_2|\theta = +.8) \\ &= .4 \times 0 + .6 \times 15 \\ &= 9 \end{aligned} \tag{32}$$

Where $\ell\,(\text{d}_\text{j}|\theta_\text{i})$ = the loss from making decision d_j when θ_i is false

Since the expected loss associated with a nonmastery decision is lower than the expected loss associated with a mastery decision, a nonmastery decision should be made.

The expected loss associated with one observation plus the cost of an observation needs to be determined in order to compare it with the loss associated with no observations. To find this expected loss, the Bayesian posterior probability is needed and is expressed in Equation 33.

$$P\left(\theta_a|x\right) = \frac{P\left(x|\theta_a\right)P\left(\theta_a\right)}{\sum_{a=1}^{2} P\left(x|\theta_a\right)P\left(\theta_a\right)} \tag{33}$$

where a = examinee 1 to N
x = response 1 or 0

The probability that an examinee with $\theta = +.8$ would elicit a correct response is:

$$P(.8|x=1) = \frac{P(1|.8)\,P(.8)}{P(1|.8)\,P(.8) + P(1|-.8)\,P(-.8)} \tag{34}$$

In this simplified example using the 1-parameter logistic model, P(1|.8) is

$$P(1\,|\,.8) = \frac{\exp(.8-0)}{1+\exp(.8-0)} = .69 \qquad (35)$$

Likewise $P(1|\text{-}.8) = .31$. Thus the posterior probability of a +.8 given a correct response is

$$P(.8|x=1) = \frac{(.69)(.6)}{(.69)(.6)+(.31)(.4)} = .77$$

The other posterior probabilities are listed in Table 3.

Table 3
Posterior Probabilities

Posterior Probability	Value
P(-.8| *x* = 1)	.23
P(+.8| *x* = 0)	.40
P(-.8| *x* = 0)	.60

The expected losses using the posterior probabilities are listed in Table 4.

Table 4
Expected Losses

Correct Response		**Incorrect Response**	
d_1	d_2	d_1	d_2
.23 × 25 + .77 × 0 = 5.75	.77 × 15 + .23 × 0 = 11.55	.60 × 25 + .40 × 0 = 15.00	.40 × 15 + .60 × 0 = 6

The expected overall risk given a response uses the smallest expected loss values from Table 4 and is:

$$\begin{aligned} E(\text{risk} \mid \text{responses}) &= E(\text{loss} \mid 1)P(1) + E(\text{loss} \mid 0)P(0) \\ &= 5.75 \times .538 + 6 \times .462 \\ &= 5.87 \end{aligned} \tag{36}$$

where $P(1) = P(1 | .8)P(.8) + P(1 | -.8)P(-.8) = .538$
$P(0) = 1 - P(1) = .462$

If cost, which is somewhat arbitrary, is assumed to be 1, the expected loss after taking a response is 6.87. Since this is less than the expected loss with no responses, 9, another item should be administered.

Criticisms of the Bayesian approach. Several criticisms have been made about the application of Bayesian theory to mastery testing, especially in the area of scoring (Weiss & McBride, 1984; Samejima, 1980). Bayesian estimates tend to regress toward the prior θ distribution (Baker, 1987; Weiss & Kingsbury, 1984) and the accuracy of Bayesian estimates can be compromised if the assumed prior is incorrect (Weiss & McBride, 1984). Samejima (1980) argued that priors (either multiple group priors or individual priors based on past performance) can result in two examinees with the exact same response vector receiving different scores. The test can be prematurely terminated if there are start-up problems (examinee answers early items incorrectly due to reasons other than ability level, such as nervousness or inexperience with computerized testing; Frick, 1990). Kingsbury and Zara (1989) indicated that while Bayesian scoring may reduce variance on a group level it can be quite problematic for individual scores. In addition, the risk and cost of administering another item are very difficult to determine, especially in an educational setting (Reckase, 1979). Despite these limitations, Bayesian estimation has some advantages over maximum likelihood estimation. First and foremost, it allows for all correct or all incorrect response vectors or estimation from a single-item test. MLE requires at least one correct and one incorrect response in order make an estimate, and all correct or all incorrect scores are not usable by the MLE procedure (Baker, 1987). Depending on the θ level of an examinee, the mastery cutoff, and the distribution of item parameters in an item bank, an all correct score is still realistic in an adaptive mastery type setting. A second advantage of Bayesian estimation is that infinite estimates of θ do not occur (Lord, 1984). Because of this advantage, some research has used a Bayesian decision theory model (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992) or Bayesian confidence intervals (Kingsbury & Weiss; 1980, 1981, 1983).

IRT-Based SPRT

Reckase (1979) combined IRT and SPRT in a study that used the SPRT model but allowed the probability of a correct response (IRT-based) to vary. In other words, probability of a correct response and item selection were based on CAT, but the actual mastery decision used SPRT rather than a $\hat{\theta}$ confidence interval. Reckase was interested in examining how SPRT performed with regard to (1) nonrandom item sampling, (2) varying region of indifference boundaries, and (3) effects of guessing on classification accuracy with a 1-parameter model.

Method/Design. This study consisted of nine conditions: three models (1-parameter, 3-parameter, and 1-parameter using 3-parameter based responses) crossed with three areas of indifference: ± .3, ± .8, and ± 1.0 (region where the decision to continue testing will be made).

The item parameters for the 1-parameter and 3-parameter tests came from an existing pool of 72 vocabulary items with difficulty approximately normally distributed. The distribution of a_i and c_i were not specified. The item parameters for the 1-parameter model using 3-parameter responses came from a simulated item pool of 181 items. Item difficulty was uniformly distributed from -3 to + 3, discrimination was fixed at .588 and pseudo-guessing fixed at .12.

True θ parameters were generated from a uniform distribution ranging from - 3 to +3 at .25 intervals. Responses were generated using monte carlo simulation based on the item parameters and the θ parameters for the 1- and 3 -parameter models and replicated 25 times using different seed values in order to generate independent sets. For the third model, the probability of a correct response used in the ratio calculations (see Equation 33) were based on the 1-parameter model, but the probability of a correct response generated from the monte carlo simulation was based on the 3-parameter model (i.e., b_i was uniformly distributed from -3 to + 3, a_i was fixed at .588 and c_i fixed at .12). This misspecification was created in order to examine the effects of guessing on classification using the 1-parameter model (Reckase, 1979).

Items were initially selected using CAT with fixed sized branching until at least one correct and one incorrect response had been given. Then item selection was based on maximum item information and θ estimates were updated using maximum likelihood estimation.

Instead of using the IRT definition of mastery, however, SPRT was employed instead. The criterion for mastery was set at $\theta_c = 0$ for all conditions. The SPRT ratio was computed after each item administration using $\alpha = .02$ and $\beta = .10$. The rationale for using these values was not reported. The maximum test length was 20 items. If a decision could not be reached by then, then a truncated SPRT model was used where θ values above 1.0 were classified as mastery and below 1.0 as nonmastery.

Results. Reckase's data were evaluated by two SPRT functions: the operating characteristic function (OC) and the average sample number function (ASN). The OC function is defined as the probability of accepting H_o as a function of the true proportion of the item pool known by the examinee (Reckase, 1979). It gives an indication of the accuracy of decisions made. The steeper the function the more accurate the SPRT decision. The ASN function is defined as the expected number of items needed to make a decision as a function of the true proportion of the item pool known by the examinee (Reckase, 1979). It gives an indication of the efficiency of the decision (number of items needed). The lower the curve, the more efficient the decision. Both functions were approximated for this study (see Reckase, 1979 for a more detailed description of the approximation derivation).

The OC functions for the 1-parameter logistic model were similar across the three regions of indifference (boundaries for continuation of testing decision), indicating similar levels of decision accuracy regardless of the width of the indifference region. For the ± .3 region, the actual values were different than the theoretical values, which Reckase (1979) assumed was due to the effects of restrictive stopping rules. The ASN functions showed narrower regions of indifference, required more items to make a decision. For all three regions, the largest expected number was near the criterion cutoff.

Like the 1-parameter logistic model, the OC curves for the three regions were similar for the 3-parameter logistic model, except the ± 1 region at 1 and - 1 showed a decline in precision. The ASN showed similar results to the 1-parameter model. The narrower the range, the more items required and all three regions required the most items near the cutoff. Compared to the 1-parameter model, however, the 3-parameter model required fewer items to make a decision. Reckase (1979) noted that the 3-parameter ASN was more asymmetrical due to the guessing component.

For the 1-parameter model using 3-parameter responses the three OC curves were similar but shifted to the left as compared to the other two models, yielding an effective criterion of -1.5 rather than 0. More examinees were classified as masters than with the 1-parameter model without the guessing effect. For the ASN functions, the same relationship was preserved: the narrower the region, the more items required with the most being required near the criterion. The difference, however, was that the effective criterion had shifted left and more items were required than the 1-parameter model without guessing.

Conclusion. The 3-parameter SPRT model performed the best using the fewest items and narrower ranges of indifference required more items and wider regions required fewer items to make a decision. Reckase did not vary α and β. It would have been interesting to see how varying this condition might influence results. It would also have been appropriate to compare the performance of the IRT-based SPRT model in terms of accuracy and efficiency of decision-making with a pure SPRT model (random selection) and a pure IRT model for the data sets. This would have given a better idea of how well each model worked. Similarly, different distributions of θ (normal vs. negatively skewed) should have been compared with these three models (pure SPRT, pure IRT, and IRT based SPRT).

MODEL COMPARISON STUDIES

Several researchers (Frick, 1990; Kingsbury & Weiss, 1980, 1983; Reckase, 1979; and Spray & Reckase; 1996) have compared: adaptive mastery testing, SPRT, and/or conventional testing in terms of test length reduction and accuracy of classification.

Adaptive Mastery Testing vs. SPRT and Conventional Testing

Kingsbury and Weiss (1980, 1983) were interested in comparing the performance of the three models under varying data structures using monte carlo simulation.

Method/Design

Four item pools of 100 items each were generated: uniform, *b*-variable, *a*- and *b*-variable, and *a*-, *b*-, and *c*-variable pools to represent different types of data. The uniform pool consisted of items with fixed values for *a, b,* and *c* (1.00, 0.00, and 0.20 respectively) in accordance with the SPRT assumption of equality of items. The other three item pools varied item parameter values in accordance with the 1-, 2-, and 3-parameter IRT models respectively.   The item parameters were evenly distributed across selected values. For the a_i parameter, the values were .5, 1.0, 1.5 and 2.0. For the b_i parameter, the values were -2.5, -2.0, -1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5, 2.0, and 2.5. For the c_i parameter, the values were .1, .2, and .3. 500 simulee θ parameters were drawn from a normal distribution with mean 0 and standard deviation of 1. Item responses for each of the items in each of the pools were generated using monte carlo simulation based on the item parameters and the simulee θ parameters. Conventional tests of three lengths (10, 25, and 50 items) were created by randomly drawing items from each of the item pools with the condition that the 25-item test contained the first 10 items from the 10-item test and the 50-item test contained the first 25 items from the 25-item test.   These tests lengths were also used as a maximum test lengths for the AMT and SPRT procedures.

The conventional test was of fixed length and scored two different ways: (1) proportion correct (CT/PC) and (2) Bayesian (CT/B). For CT/PC scoring, mastery was defined as answering at least 60% of the items correctly.   For CT/B scoring, mastery was defined as scoring above θ_m, where θ_m is on the converted θ metric from 60% on the TRF. After administration of the fixed-length test, the simulee's score, $\hat{\theta}$, was obtained from the response pattern and item parameters using Bayesian scoring assuming a normal (0, 1) prior distribution.

SPRT used an area of indifference of .5 and .7 with values of α and β set at .10. Decisions were made using SPRT criteria unless the item pool was exhausted. If a decision could not be made before exhausting the item pool, a proportion correct of .60 was used to define mastery.

AMT, like CT/B, used θ_m, a converted value from the TRF equivalent to 60% correct for each of the item pools. Decisions were made using 95% Bayesian confidence intervals as defined by Kingsbury and Weiss (1979). If the item pool was exhausted before the AMT criteria could be met then a decision was based on $\hat{\theta}$ being above or below θ_m.

Results

Models were evaluated according to four criteria: (1) test length, (2) correspondence with true mastery status, (3) correspondence with true mastery status as a function of test length, and (4) frequencies and types of decision errors.

Test length. Both SPRT and AMT reduced the length of the test, and the percent reduction increased with increased maximum test length. For the uniform pool, SPRT reached a decision faster and therefore had a greater reduction in maximum test length (MTL) than AMT (12% vs. 9.7% for the 10-item test, 48% vs. 36% for the 25-item test and 69% vs. 54% for the 50-item test). For the *b*-variable pool, MTL reduction was 6%, 28%, and 46% for the 10-item, 25-item, and 50-item tests respectively whereas the reduction for SPRT was 4%, 33% and 57%. For the *a*-, *b*-, and *c*-variable pool, the AMT reduction in MTL was 13%, 35%, and 53% and the MTL reduction for SPRT was 14%, 46%, and 69%.

Correspondence with true mastery status. A phi correlation was computed between observed mastery status determined by the various models and true mastery status. This correspondence was more influenced more by test length than by model type. All models showed increased correspondence with increasing maximum test length. Within a test length, the IRT-based decision models (AMT and CT/B) performed better than SPRT across the four types of item pools except for the 50-item uniform pool where CT/PC performed the best.

Correspondence with true mastery status as a function of test length. This criterion was used to evaluate efficiency, which is defined as the highest correspondence coefficient combined with the shortest test length (Kingsbury & Weiss, 1980). For the uniform pool, SPRT proved to be more efficient, whereas for the other pools AMT was more efficient.

Frequencies and types of decision errors. Results were not as clear cut for this criterion. The 1- and 2-parameter pools tended to elicit more false mastery errors across all methods whereas the uniform and 3-parameter pools tended to elicit more false nonmastery errors. For total errors, AMT or CT/B had the lowest percent across all item pools and test lengths except for the condition uniform pool and 50 items. AMT also had a more equal dispersion of error types across most maximum test lengths and item pools.

Conclusions

Kingsbury and Weiss (1980) concluded that both AMT and SPRT reduced test length while at the same time maintained the same or higher phi correlation than conventional tests for most conditions. The three exceptions were (1) uniform pool and 50-item test, (2) *a*-, *b*-, and *c*-variable pool and 25-item test and (3) *a*-, *b*-, and *c*-variable pool and 50-item item test. SPRT was more efficient when using a uniform item pool whereas AMT was more efficient when using nonuniform item pools. Unfortunately, they did not vary decision criteria, such as mastery level cutoff, like they did in the Kingsbury and Weiss 1979 study, nor the θ distribution. It would have been interesting to see if either of these conditions would have had an influence on performance.

SPRT vs. AMT and the Beta Model

Frick (1990) was interested in replicating Kingsbury and Weiss (1979) with the intention of demonstrating that SPRT is an adequate model for small-scale testing when the model is used conservatively (i.e., α and β are small and the area of indifference is sufficiently wide; Frick, 1990).   Frick further argued that as long as the item sample chosen is representative of the entire pool and α and β are sufficiently small to control decision error, SPRT can be used despite its equality of items assumption.

Many of Kingsbury and Weiss' (1979) conditions were repeated.   The item pool generated was 100 items in length with the same distribution of parameters. The same three maximum test lengths: 10, 25, and 50 items were retained.   Also, the .60 mastery level was used. Frick (1990) differed from the Kingsbury and Weiss study in three aspects: (1) a beta model was used in lieu of the conventional model; (2) $\alpha = \beta = .025$ rather than .01, which is comparable to a 95% confidence interval used in AMT; and (3) a second, wider area of indifference $P_m = .8$ and $P_n = .4$ was used as well as the original $P_m = .7$ and $P_n = .5$

Method

True θs were generated for 500 simulees based on a normal distribution with mean = 0 and variance = 1. For each simulee, a subset of items of size 10, 25, or 50 was chosen at random. Test items were selected from the subset, not the entire item pool. Responses were generated using monte carlo simulation.   A second distribution of true θs was also generated from a normal distribution of mean 0 but the variance was 10. Frick was interested in seeing if there were differences using a flatter θ distribution.

Once responses were simulated, each model selected items and made decisions whether to terminate or continue testing according to its respective rules. If the subset of items was exhausted before a decision could be made, then a truncated decision was made based on the current estimate.

Accuracy of decisions was made by the following criteria:

1) True $\theta > \theta_c = 0$ and a mastery decision was made, classify as a hit
2) True $\theta < \theta_c = 0$ and a nonmastery decision was made, classify as a hit

3) Otherwise, classify as a miss

Results

For all three test length conditions, SPRT (.8 vs. .4) had the shortest tests whereas SPRT (.7 vs. .5) had the longest tests. This implies that a wider interval tends to reduce test length with the SPRT model.

AMT was more accurate across all conditions, but all conditions had greater than the 5% expected error rate. Frick (1990) believed this was due to two reasons: (1) a large proportion of the simulees true θs were near the criterion ($\theta_c = 0$) and (2) many truncated decisions were made, especially in the 10-item maximum test length conditions. For the distribution with a larger variance, there were shorter tests and increased accuracy across all conditions.

Frick (1990) concluded that AMT was the most accurate and efficient when maximum test length constraints were not made and if testing was large-scale (e.g. in school district where there are hundreds if not thousands of examinees). He argued that SPRT is more practical and adequate to use in small-scale testing (less than 200 examinees) provided that the items are well piloted (items are of appropriate difficulty for the examinees with empirically established mastery and nonmastery levels). He also suggested that the beta model is adequate for mastery testing decisions but advised requiring a minimum of 5 to 10 items to overcome a tendency to prematurely terminate test if the examinee experiences any start-up problems.

Conclusions

Frick's (1990) results tend to agree with Reckase (1979) that a wider interval of indifference leads to shorter tests. Unfortunately, like Reckase, he did not vary conditions such as the size of α and β with a comparable width of confidence interval for the AMT condition. Varying this condition could check on accuracy of the decision as well. In addition, the rationale for comparing SPRT and AMT to the beta model was not clear. The beta model requires more work initially (posterior probabilities are calculated in advance for a given cutoff with all possible combinations of successes and failures). Frick also stated that it is also prone to error if a student gets off to a poor start. It probably would have been more informative to compare SPRT and AMT with Reckase's modified SPRT model rather than the beta model. A final concern was the use of subsets rather than the entire item pool for item selection. For the random models, the representativeness of the subsets to the item pool is in question. Results may capitalize on idiosyncrasies of a particular subset chosen. For AMT, this is even more serious since item selection is not random but based on the item characteristics. Urry (1977) recommended having an item pool of at least 100 items. Using Frick's method, the maximum item pool size during the simulation was 50 items, and that might have affected AMT performance. The one condition that Frick did vary that the others did not is the θ distribution. In addition to increasing the variance, it would be interesting to compare results with a negatively skewed distribution, since in a mastery situation this type of distribution is very possible.

EXSPRT vs. AMT

Luk (1991) compared EXSPRT [intelligent (EXSPRT-I) and random (EXSPRT-R)] with AMT. Unlike Frick's (1992) study with relatively small samples. Luk (1991) compared the two procedures when (1) there was a sufficient number of examinees to make calibrations using a 3-parameter model and (2) when the item parameter estimation sample was a different set of examinees from the validation sample.

Method

The test used was for this study was a Spanish placement test used by the Spanish department at the University of Indiana. The test was a paper-and-pencil test consisting of a 40-item listening subtest and a 60-item reading subtest.   All items were multiple-choice with four alternatives. Based on a total of 1,672 examinees, Cronbach's α was .81 and .87 for the listening and reading subtests, respectively. From this group of examinees, 1,000 were randomly selected to be part of the item parameter estimation group. Estimation of IRT item parameters and EXSPRT rules were based on responses from this group. The remaining 672 examinees were used in the validation group. Their responses were used in a real-data simulation for three adaptive algorithms: AMT, EXSPRT-I, and EXSPRT-R. A real-data simulation is the same as Frick's (1989) retroactive simulation (i.e., select items in the same order they were presented originally, examine the response pattern, compute the probability ratio, and make a decision of mastery, nonmastery, or continuation of testing).   Item parameter estimation and simulations were done separately for the two subtests: listening and reading.

Mastery cutoffs were set at .60 for the listening subtest and .50 for the reading subtest, based on the point of inflection of the test response function.   Examinees were categorized as masters and nonmasters for each subtest based on these cutoffs.

For the AMT post-hoc simulation, the conditions of Weiss and Kingsbury (1984) were closely followed. The initial prior for θ was set at 0 with an initial prior variance of 1. Items were selected based on maximum information, and θ was updated after each response based on Owen's (1969) algorithm.   Status was determined using a 95% confidence interval. If all the items were exhausted before a decision could be reached, then a decision was made based on the most recent θ estimate. For each examinee, data concerning the number of items and agreement with the classical decision (based on a conventional, full-length test) were collected.

For the EXSPRT-I post-hoc simulation, P_{om} (prior probability of the examinee being a master) and P_{on} (prior probability of the examinee being a nonmaster) were set to .5, indicating that there was an equal chance of being a master or nonmaster at the beginning of the test. $\alpha = \beta = .025$, which is comparable to a confidence interval of .95 in the AMT procedure. Items were then selected based on their utility indices (see Equation 23) with data collected on the number of items and agreement with the classical decision. For the EXSPRT-R simulation, the conditions were the same as the EXSPRT-I except that items were chosen randomly rather than based on utility indices.

Results

Means of the parameter estimation group and validation group for each of the subtests were compared using *t*-tests and found to be nonsignificant, indicating that the groups were similar. Two variables were used to compare the various selection procedures: (1) decision accuracy (i.e., the percentage of agreement with classical decisions) and (2) efficiency (i.e., the number of items needed to reach a decision). For each subtest, AMT was the most accurate overall for both types of decisions. When this was broken down into mastery and nonmastery decisions, AMT was the least accurate for the mastery decisions and the most accurate for the nonmastery decisions. The reverse occurred for EXSPRT. Both EXSPRT-I and EXSPRT-R were more accurate for mastery decisions and less accurate for nonmastery decisions, with EXSPRT-I doing slightly worse than EXSPRT-R.

In regard to efficiency, for nonmasters EXSPRT-I required the least number of items to make a decision, followed by AMT and then EXSPRT-R. For masters, the pattern was EXSPRT-I, then EXSPRT-R with AMT requiring the most items.

Conclusions

Luk (1991) concludes that EXSPRT-I is a "strong, viable alternative to the 3-parameter AMT model" (Luk, 1991, p. 20) and is more accurate and more efficient than AMT for mastery decisions. There were some problems, however, with the compatibility of the sample and the models used. First, the θ distributions were negatively skewed. While this is not unusual for achievement tests, misspecifying the prior can have an effect on the accuracy of decisions as well as the number of items required to reach that decision (Weiss & McBride, 1984). This might have been a reason why AMT did so well classifying nonmasters but had trouble classifying masters. Skewness is often a condition that has been ignored in mastery testing and its effects should be further investigated.

Secondly, the proportion of masters and nonmasters was quite different. In a group of 672 examinees, only 62 were classified as masters in the listening subtest and 64 in the reading subtest. For the parameter estimation group only 89 out of 1,000 were masters for the listening subtest and 109 out of 1000 for the reading test. This violated the assumption in the EXSPRT-I procedure that the likelihood of being a master or nonmaster is assumed to be 50%. It also seems unusual that there were so few masters with negatively skewed data, especially given the low cutoffs. This is an indication that perhaps the test was not at an appropriate level for the group (i.e., too difficult) or the cutoffs were too stringent. Unfortunately, the distribution of item parameter estimates was not reported.

Despite these problems, this is the only study that compared EXSPRT-I with AMT using comparable sample sizes. The fact that AMT did well with nonmastery decisions and EXSPRT-I did well with mastery decisions requires further investigation. This type of investigation should involve varying conditions such as the θ distribution (normal vs. skewed), item parameter distributions (uniform vs. normal), percentages of masters and nonmasters, and cutoffs.

SPRT vs. Sequential Bayesian (AMT)

Spray and Reckase (1996) did a simulation study that compared SPRT and AMT, which they referred to as "Sequential Bayes". In their study they matched the two procedures on expected error rates and then compared expected test length to determine which procedure was more efficient.

Method

Simulees were generated from a normal distribution with mean 0 and standard deviation of 1. Items administered to the simulees were from a pre-calibrated 200-item pool that fit the 3-parameter logistic model. Decision points, δ, ranged in value from -3.0 to +3.0 in increments of .25. Items were selected based on maximum information at the decision point, δ. Although not specifically stated, it appears that this selection algorithm was used for both procedures, which is not truly AMT. For the SPRT procedure, the probability of mastery, $p(\delta)$, at each level was a fixed value based on the average of probabilities across the item pool. The termination criterion was based on the log of the likelihood ratio (log of Equation 20) and the decision boundaries based on α and β discussed earlier in the section on SPRT. If a decision could not be made before 50 items were administered, then a truncated decision was made based on whichever decision boundary was closest to the log of the likelihood ratio. For the Sequential Bayesian procedure, probability was allowed to vary from item to item. The termination criterion was based on whether or not the Bayesian confidence interval around the θ estimate included the criterion. If a decision could not be made before 50 items were administered, then a truncated decision was made based on whether the most recent point estimate was above or below the criterion.

Two steps were involved for matching the two procedures. First, the Sequential Bayes simulations were conducted for each of the 25 examinee levels using four confidence levels (η) and 25 different decision points (δ). Each test was replicated 1,000 times for each of the 25 decision point values. Next the replications were analyzed and error rates at two points symmetric to each δ were then used as the α and β values for the SPRT simulations (Spray & Reckase, 1996). The expected classification error was computed at each decision point. If this expected classification error was approximately the same for the two procedures across values of δ, then error rates were considered to be matched. Once error rates were matched, expected test length was computed for each procedure and compared.

Results

Spray and Reckase (1996) found that when error rates were matched, SPRT yielded shorter tests. This difference was more pronounced as the confidence interval became narrower. They concluded that when SPRT and Sequential Bayes are matched, SPRT is a more efficient model.

Conclusions

In this comparison study using a 3-parameter model, SPRT was more efficient than Sequential Bayes. This evidence, however, is not conclusive and some factors need to be considered. First, the item selection rule for both SPRT and Sequential Bayes was different than prior studies. SPRT usually employed random item selection (Frick, 1990; Kingsbury & Weiss 1979, 1980, 1983). For AMT, Kingsbury and Weiss (1979, 1980, 1983) utilized maximum information of the most updated θ estimate.

Spray and Reckase (1996), on the other hand, based item selection on maximum information at the decision point. Depending on θ's distance from the decision point, information may have been lower than when using Sequential Bayes, resulting in selection of a different set of items. This in turn, might have affected efficiency. Second, the method for determining α and β is not clear and seems to be confounded with the data. It is not clear why Spray and Reckase (1996) did not halve (1-η), the area of indifference, like Frick (1990), for each level, so α = β = .005, .025, .05, and .125. Symmetric points around each value of δ could have been chosen and the SPRT ratios computed in this manner. It would be interesting to see what proportion of expected classification error would match the Sequential Bayes and if SPRT would still be more efficient. Despite these shortcomings, this study varied a condition that had not been varied in the past (confidence interval) and varied a condition to a much greater extent (decision points). These conditions should be considered in future studies along with the original form of AMT.

TESTLETS

Lee, Brennan, and Frisbie (2000) define testlets (a term first used by Wainer and Kiely, 1987) quite broadly as "a subset of the items in a test form that is treated as a measurement unit in test construction, administration, and/or scoring." Most of the literature, however, limits the measurement unit to test construction of what Lee et al (2000) refer to as "stimulus-based" testlets. A smaller portion of the literature incorporates testlets in selection or branching, which is related to administration. In stimulus-based testlets, testlets are a subgroup or "bundle" of items shorter in length than a test, that are interconnected through some common stimulus beyond the trait being measured. Three examples where these kinds of testlets might be useful are: reading comprehension, interpretation of tables and graphs, or analytical type story problems. A reading comprehension test is generally composed of several passages. Each passage is followed by items related to that passage and the subset of items can be thought of as a testlet, since in addition to the trait, the content and information in the passage ties the items together more closely than items across passages. In the table interpretation type example, the table is the common stimulus, and the items related to the table make up the testlet. In the analytical example, there is a story problem, followed by items related to the text.

Criticisms of Classical Adaptive Tests.

Research on testlets (Lewis & Sheehan, 1990; Luecht (2003), Sheehan & Lewis, 1992; Wainer, Kaplan, & Lewis, 1992; Wainer & Kiely, 1987; and Wainer & Lewis, 1990) came about due to dissatisfaction with classical single-item branching adaptive tests. This dissatisfaction was due to three main reasons: (1) inappropriateness of individual items for measuring certain traits, (2) contextual effects and (3) test security.

Inappropriateness was alluded to earlier with the example of the reading comprehension test. From a practical standpoint it is time-consuming and inefficient for an examinee to read an entire passage and answer only one item per passage. It can also be questioned whether it is really sound or fair practice to ask only one item from such a stimulus. A solution to that problem is to shorten the passage. This increases efficiency but most likely changes the trait being measured (Wainer & Lewis, 1990). From a measurement standpoint, multiple items per passage might violate the IRT assumption of local independence (Thissen, Steinberg & Mooney, 1989). It is still possible, however, to use a unidimensional IRT model, if there is unidimensionality across testlets even if there is not within testlets (Lee, 2000; Rosenbaum, 1988).

Contextual effects are the influence of previously administered items on the performance of later items due to content and/or location. Because adaptive testing constructs the test "on line" certain contextual problems can occur that might not arise in a conventionally constructed test where there is a chance to examine the test as a whole "off-line" prior to administration (Wainer, Kaplan & Lewis, 1992; Wainer & Kiely, 1987; and Wainer & Lewis, 1990) . Wainer and Kiely (1987) and Wainer and Lewis (1990) provide an excellent example in the following items:

1. Carbon dioxide (CO_2) is added to water to make:
   a) Carbolic acid
   b) Seltzer
   c) Sloe gin fizz
   d) Hydrogen carbonate

2. CO_2 is known as:
   a) Carbon dioxide
   b) Carbon monoxide
   c) Carbonated water
   d) Ozone (Wainer & Kiely, 1987).

The answer to Item 2 is embedded in the stem of Item 1. In a conventionally constructed test, this context effect would have been corrected by rewording one item [probably deleting “(CO_2)” from Item 1] and possibly change the proximity of the two items prior to administration. With adaptive testing, given the enormous number of possible combinations of items, it is more difficult to foresee all the possible contextual problems. With testlets, a certain degree of control is retained, since the testlets, themselves can be examined prior to on-line testing (Luecht, 2003; Mills & Stocking, 1996). While it only solves the problem of within testlet contextual effects, contextual effects across testlets can be greatly reduced due to the fact that different testlets usually contain different content and similar items will no longer be located adjacent to one another (Wainer & Kiely; 1987; Wainer & Lewis; 1990).

A second concern of contextual effects is content balancing. In a test designed to cover more than one content area, it is quite feasible in a classical adaptive test to have a disproportionate number of items selected from a single content area. Testlets can control for this problem by having content balanced within testlets. It is not necessary, however, to use testlets in order to have a content-balanced test. Kingsbury and Zara (1989, 1991) suggested using a constrained item selection procedure called constrained-CAT, where in addition to item information, the percentage of items answered in each content area is determined. When an area is below the prespecified percentage, an item with maximum information is chosen from the deficient content area.   This does, however, increase the minimum number of items needed for the item pool (Kingsbury & Zara, 1989). Items have to be written with varying parameters across each content area, but it still allows for content balancing in situations where it is more appropriate to measure examinees using individually administered items.

Test security, in theory, should be enhanced with adaptive testing. Each examinee gets a potentially different subset of items, so cheating or discussion of items after the test is largely irrelevant. In actuality, it can, in some ways be more of a security problem.   The main problem with adaptive testing is increased exposure rate.   Unlike paper-and-pencil tests, which are usually administered to large groups at infrequent intervals, adaptive tests are often administered to small groups at frequent intervals. Even though each examinee is administered a different set of items, there is some overlap of items, especially well-written items and/or items of certain difficulty levels. As a result, a large portion of the items may be exposed in a short amount of time (Mills & Stocking, 1996).   A second problem arises with repeat administrations.   If the items or testlets are well calibrated and the examinee answers in accordance with the model on each administration, it is quite feasible that a large number of the same items will be selected again. It then becomes necessary to keep track of which items were previously administered or create enough of a time gap to allow for forgetting (Kingsbury & Zara, 1989; Luecht, 2003; Wainer & Kiely, 1987; and Wainer & Lewis, 1990). A third security problem, especially with high-stakes tests is break-ins at testing sites. If all the items are in a single pool and the item pool is stolen, then all the items are compromised (Mills & Stocking, 1996). Mills and Stocking (1996) suggest subdividing a single large pool into smaller subsets.

Lewis and Sheehan (1990) argue, however, that testlets increase test security since they potentially enhance the chances of each examinee receiving a different subset of testlets, and it is less burdensome to keep track of previously administered testlets rather than items for tests that allow repeat administrations. While it is true that it is easier to keep track of a half a dozen previously administered testlets than 30 items, the same advantages and disadvantages of adaptive testing are true for testlets as they are for items. If anything, a testlet bank is probably smaller than a bank of single items and likely would incur more security problems, not fewer. Kingsbury and Zara (1989) make another suggestion called randomesque item selection. Instead of choosing the item with maximum information, randomly select an item from a group of 2-10 (or some other range of) items with similar item information levels. This can work with either items or testlets. Once again, the disadvantage of any additional constraints in item selection is that it will increase the minimum size of the item or testlet bank and the length of the tests administered from them.

Non-IRT Testlet Models.

While this paper emphasizes IRT-based adaptive mastery testing or studies leading up to IRT-based adaptive mastery testing, these studies are mentioned here because either (1) the concepts could be used with IRT-based selection or (2) the testlets themselves were constructed using IRT, even though the selection procedure did not use IRT. In the case of the fixed- branching models, IRT is not used for either construction of the items or selection of the items, so the discussion will be limited to the structure of the models mentioned.

Fixed branching. In Wainer and Kiely's (1987) hierarchical models, item selection is based on fixed branching rather than maximum information. In the full-size hierarchical model, a testlet is administered. Based on the responses to the items, either a more difficult or any easier testlet is administered. There is a unique path to any particular outcome

Although Wainer and Kiely's (1987) approach to testlets uses fixed branching and validity-based scoring (Lewis, 1989), the structure could easily be modified for use with IRT-based selection and scoring. Adaptive selection could occur either between and/or within testlets and be based on maximum information. For certain content areas (e.g., algebra), it could follow Wainer and Lewis' (1990) suggestion that testlets themselves be administered linearly (all examinees are administered all testlets) to ensure content balancing, but within a testlet the items should be hierarchically organized (i.e., administered adaptively).

Comparison of hierarchical and linear testlets. Wainer, Kaplan, and Lewis (1992) compared hierarchical and linear testlets with varied item pool size and testlet length. Items were generated based on a 3-parameter logistic model with a_i fixed at 1.0, c_i fixed at .20 and b_i uniformly distributed from -3 to +3.

Wainer, Kaplan, and Lewis (1992) evaluated the accuracy of decisions through η^2, a measure that relates status or decision dictated by the branching procedure with true (θs) of the simulees

$$\eta^2 = 1 - (SSE/SST) \tag{37}$$

where SSE is the pooled within cell sums of squares of θ
SST is the total sums of squares of θ

(Wainer, Kaplan, & Lewis, 1992).

Therefore, the best testlet was chosen that maximized η^2 (i.e., minimized SSE).

Wainer, Kaplan, and Lewis (1992) concluded that linearly based testlets, when calibrated from larger item pools could compete with adaptive testlets. They also concluded that because the linear test was peaked at the middle of the distribution where the majority of the simulees were located, it did well but adaptive tests would be better for the tails of the distribution.

Bayesian decision theory selection. The following studies used non-IRT based scoring but testlet selection was based on Bayesian decision theory instead of fixed branching.

Lewis and Sheehan (1990), like Wainer, Kaplan, and Lewis (1992), used IRT for testlet construction and non-IRT based scoring. Their study used data from the National Council of Architectural Registration Boards (NCARB) licensure test of seismic knowledge. The original pencil-and-paper test was comprised of 60 items across two content areas in a 60/40 ratio. Scoring was on a scale from 1 to 100 and the cut score was 75 (Lewis & Sheehan; 1990).

Lewis and Sheehan (1990) were interested in constructing a testlet-based adaptive test that took the 60/40 content ratio into account. The testlets were constructed based on IRT. The entire item pool consisted of 110 items and calibrations indicated that items fit a 3-parameter logistic model. The item pool was divided into 10-item testlets. Testlets were constructed to be as parallel as possible in terms of content balance, difficulty, and discrimination. The reason for parallel testlets was for simplicity. Parallel testlets allowed for random testlet selection. This, in turn, meant that θ did not have to be updated after every item, and it was easier to construct parallel testlet pools near the cutoff rather than testlets across the entire θ scale (Lewis & Sheehan, 1990). The six most parallel testlets were selected and two additional checks were made. Each testlet was examined for context effects and interchangeability. If context effects were found, then items were edited. Interchangeability was checked by determining if there was a difference in average likelihood of a number-correct score at two cut points: θ_n (highest score of a nonmastery) and θ_m (lowest score of a master). Average number-correct score at those two cut points was demonstrated to be approximately the same for all six testlets. A minimum of two testlets were administered with a maximum of six testlets.

As mentioned previously, testlet selection was random but the adaptive stopping rule was based on Bayesian decision theory. If the expected loss plus the cost of another testlet was less than the expected loss prior to administration of the testlet, testing continued; if the expected posterior loss was greater, or the testlet bank was exhausted, then testing was discontinued. The loss function parameters *A* (false positive), *B* (false negative), and *C* (cost of an additional testlet) were determined through a combination of theory (seriousness of misclassification based on theory) and empirical evidence (simulation studies). A was considered twice as serious as *B* so *A* = 2*B*. *B* was determined through a separate simulation study to be 20. *C* was set equal to 1 (Lewis & Sheehan, 1990).   Lewis and Sheehan (1990) chose to score using total number-correct score because it required fewer calculations.

Their results were in accordance with their model. When probability of mastery was plotted against expected loss, expected loss for a pass decision decreased linearly as the posterior probability of a master increased, expected loss of a fail decision increased linearly as the probability of master increased, and the probability of continuation of testing shrunk with increased test length.

In addition to real data, simulated data were also used to determine classification error against true θ. Data were simulated for 100 simulees for 41 levels of θ for a total of 4,100 simulees. Many loss functions (*A*/*B*) were used but only three were reported: 20/20, 40/20, and 100/20. The loss functions were evaluated in terms of fixed- versus variable-length tests (Lewis & Sheehan, 1990). For each loss function, average test length, pass rate, false positives and false negatives were compared across both types of tests. Average test length for the variable-length test ranged from 25 to 30 items whereas the fixed-length test used 60 items. Pass rates were comparable, ranging from 60% to 69%. For each respective loss function there were similar error rates (especially for the false positives) yet the variable-length tests were much shorter, providing support for use of an adaptive test. The loss function chosen also influenced results. More severe losses (i.e.,100/20) led to longer tests but with fewer classification errors. Also in accordance with a variable-length test, as true θ approached the cutoff it required more items (50 - 60) to make a decision whereas θ on either extreme required fewer items (20 - 30) to make a decision.

Lewis and Sheehan (1990) did succeed in creating a working model of adaptive testing using content-balanced testlets. There are, however, some problems with the model. First and foremost is the use of number-correct score. The items were calibrated using a 3- parameter model but θ was estimated using number-correct score, which is not a sufficient statistic for this model. The use of parallel testlets is also somewhat of an inflexible constraint as well. As mentioned in their study, these conditions were used to simplify calculations, which in 1990, might have been an issue with on-line time lag for administration of items/testlets. Given the speed and capability of computers over a decade later, it is probably not necessary to take “short-cuts” to reduce on-line time lag. Furthermore, selection was at the testlet level, but scoring was at the item level. If some sort of IRT-based testlet scoring was available, then it would not be necessary to construct parallel testlets. Selection of testlets could be adaptive, as well. Moreover, if there was IRT-based scoring at the item level as well, it would be possible to have adaptive testing within a testlet as well as across testlets. Lastly, IRT-based scoring would allow for a less subjective selection procedure, such as maximum information, rather than cost which is quite subjective.

Comparison of equivalent and non-equivalent testlets. Sheehan and Lewis' (1992) simulation study was a continuation of the Lewis and Sheehan (1990) study. In the later study, nonequivalent testlets were considered as well and compared with equivalent testlets and paper-and-pencil tests in terms of average test length, accuracy of classification, and overall pass rate. Like the 1990 study, all testlets had the same number of items and covered the same content areas. What varied was the cut scores for the nonequivalent testlets.

For both equivalent and nonequivalent testlets, the posterior probability ($P_{m|i}$) of being a master, given the i^{th} testlet was administered is:

$$P_{m|i} = P(\theta = \theta_m \mid X_1, X_2, ..., X_i) = \frac{P(X_i \mid \theta = \theta_m) P_{m|i-1}}{P(X_i \mid \theta = \theta_m) P_{m|i-1} + P(X_i \mid \theta = \theta_n) P_{n|i-1}} \qquad (38)$$

where $P(X_i|\theta = \theta_m)$ conditional probability of observing number-correct score X for testlet i given proficiency level θ_m.
$P(X_i|\theta = \theta_n)$ conditional probability of observing number-correct score X for testlet i given proficiency level θ_n (Sheehan & Lewis, 1992).

In other words, the probability of being a master given the i^{th} testlet was administered is dependent on performance from the previous testlet and mastery status. What differed between these models is how $P(X_i|\theta = \theta_m)$ was determined for each mode. For the equivalent mode (parallel testlets)

$$P(X_i = s \mid \theta = \theta_m) = \exp\left\{1/T \sum_{t=1}^{T} \ln\left[P(X_i = s \mid \theta = \theta_m, t)\right]\right\} \qquad (39)$$

where s = the number-correct score at the i^{th} stage of testing
t = testlet administered at stage i.

In other words, the probability of obtaining a particular number-correct score is averaged across all testlets.

For the nonequivalent mode (nonparallel testlets)

$$P(X_i = s|\theta = \theta_m, t) = \sum \prod_{j=1}^{n} P_{jt}(\theta_m)^{x_j} \left[1 - P_{jt}(\theta_m)\right]^{1-x_j} \tag{40}$$

where summation is taken over all response patterns such that the number-correct (NC) score is *s*, x_j is 1 or 0 for the *j*th item, and $P_{jt}(\theta_m)$ is the conditional probability of a correct response to the j^{th} item on testlet *t* given a proficiency level θ_m.

Instead of taking the average of the testlet likelihood functions, score probabilities were conditional on testlet administration (see Equation 39). Expected loss for nonequivalent testlets was based on a weighted average of all possible outcomes where weights corresponded to predictive probabilities (Sheehan & Lewis, 1992).

The number of computations needed to determine the number and patterns of future testlet administrations tends to escalate quite rapidly. For example, for a testlet pool of 10 testlets consisting of four stages, there are 5,040 (10 × 9 × 8 × 7) possible permutations of testlet administrations. For each permutation there are eight possible outcomes:

1. P_i
2. F_i
3. $C_i\ P_{i+1}$
4. $C_i\ F_{i+1}$
5. $C_i\ C_{i+1}\ P_{i+2}$
6. $C_i\ C_{i+1}\ F_{i+2}$
7. $C_i\ C_{i+1}\ C_{i+2}\ P_{i+3}$
8. $C_i\ C_{i+1}\ C_{i+2}\ F_{i+3}$

where P_i = pass immediately
F_i = fail immediately
C_i = continue testing (Sheehan & Lewis, 1992).

Because of the large number of computations needed, testlets in the nonequivalent mode were selected "off-line". For each examinee, a permutation was randomly selected prior to administration of the first testlet and all computations were made in accordance with the particular permutation selected.

The data used were from the Architect Registration Examination (ARE), a professional certification examination covering eight areas. Two areas were chosen: Division E and Division D/F. Both had been previously administered as paper-and-pencil tests (P&P). For Division E there were six testlets composed of 10 items each. A minimum of two testlets were administered (20 items) with a maximum of six testlets (60 items). Division D/F consisted of ten 25-item testlets. A minimum of two testlets were administered (50 items) with a maximum of five testlets (125 items). Items from both Divisions fit the 3-parameter logistic model. For the equivalent testlets, the mean difficulty and discrimination were similar across testlets with no large differences.

Data were simulated for each Division based on distributions from previously administered P&P tests. Simulees were simulated at 40 different θ levels (± 20 NC score points above and below the cut score). The results were weighted to match observed distributions of a recently administered P&P test. For each simulee, responses were generated according to a 3-parameter logistic model using actual ARE parameters (Sheehan & Lewis, 1992).

For each Division, θ_n was set at 1.5 standard errors below the cut score and θ_m was set at 1.5 standard errors above the cut score. The rationale for these values was not reported. The loss function parameters *A*, *B*, and *C* were set at 40, 20, and 1 respectively. The loss function parameter values were based on the simulation results from the 1990 study.

For the P&P simulated data, the entire test was considered a single testlet. Posterior expected loss was minimized by deciding nonmastery status if NC score was below the cut score and mastery status if the NC score was above the cut score.

For simulated computerized mastery testing (CMT) data using equivalent testlets, pool-wide LFs were averaged for masters and nonmasters separately and used in computing expected loss at each stage. Based on these averages, there was a maximum fail score and a minimum pass score for each stage.

For simulated CMT data using nonequivalent testlets, cut scores for nonequivalent testlets were based on testlet-specific LFs. Division E used all 6! or 720 permutations and computed testlet cut scores. In general, only 2 or 3 unique pass or fail cut scores existed for each stage of testing. The highest frequency cut score was the same as the equivalent mode. For Division D/F, it would have involved 30,240 permutations, so 1,000 were randomly selected for calibration of cut scores. In this case, there were five to six unique cut scores for each stage of testing.

Results from the simulation studies indicate that CMT with equivalent testlets and CMT with nonequivalent testlets performed quite similarly -- pass rates were almost the same. The unweighted classification of status as a function of θ was almost identical. The number of classification errors for the two CMT methods were almost the same. The mean test length for both methods were very close.

Based on their results, Sheehan and Lewis (1992) concluded that using nonequivalent testlets had a negligible impact on classification and it is valid to use an equivalent testlet design for the current ARE testlet pools. New testlets would have to be validated for equivalence by the same procedure.

One question remains, though -- what was the degree of nonequivalency of their nonequivalent testlets? First, item parameter statistics for the nonequivalent testlets were not reported. It was not known how much testlets varied in average item difficulty and discrimination. Secondly, Sheehan and Lewis (1992) said that variation in cut scores was an indication of variation in testlet difficulty. For Division E, the unique cut scores never varied more than ±1 point from the NC score on equivalent testlets. For example, the cut score from Stage 2 of the equivalent testlets was 14 and for the nonequivalent testlets was 15. For Division D/F, the unique cut scores never varied more than ±3 from the NC score on equivalent testlets (62 for the nonequivalent and 65 for the equivalent at Stage 4). Moreover, the mode was equal to the equivalent testlet cut score. It appears from these results that the nonequivalent testlets, based on cut scores, in fact were quite similar to one another, especially in Division E. Division D/F showed slightly more variation. Likewise, it can be asked if the unweighted classifications were identical, because the nonequivalent testlets were not so nonequivalent or because both modes truly had the same degree of accuracy. The item parameters of the items comprising the nonequivalent testlets need to be varied in a more systematic manner before accepting their conclusions.

A second question, which is a carry-over from the 1990 study is: Why not use IRT scoring and select testlets based on performance rather than randomly? It would be interesting to compare results of truly nonequivalent testlets using random selection versus selection based on performance.

Lastly, a condition from the 1990 study, distribution of item parameters, particularly item difficulty, near the cutoff vs. along the entire range of the continuum, would be interesting to consider for nonequivalent testlets. Since only a dichotomous decision is being made in mastery testing, it should be investigated whether or not its precision is compromised or improved by having items at levels near the cutoff rather than across the entire ability continuum. If precision is not compromised, it can save substantially on time and expense in item writing, particularly with testlets.

IRT-Based Testlets.

Maximum information testlet selection. Kingsbury and Zara (1991) were interested in measuring the degree to which content control influenced efficiency and accuracy of adaptive testing. It was expected that content control would reduce measurement efficiency and their study was designed to quantify that loss. While they did not examine mastery testing directly, it is included here because one of the conditions used testlets and another condition simulated an ability distribution that represented a mastery type testing situation.

Three types of adaptive testing were compared: traditional computerized adaptive testing (CAT), constrained computerized adaptive testing (CCAT), and testlet-based adaptive testing. CAT uses maximum information to select the next item and the content balancing issue is ignored. CCAT uses maximum information but only after a specific content area is chosen. Content area is chosen by comparing the content percentages of items already administered with the prespecified percentages. The content area with the largest discrepancy from its specified percentage is chosen and then maximum information selection is applied only to that content area. Testlet-based adaptive testing involves structuring each testlet to be content-balanced. Testlets, rather than items, are selected based on maximum information of the testlet, where testlet information is essentially an average of the item information functions comprising the testlet.

Design. Two groups of 10,000 simulees each were simulated to represent two types of θ distributions. Group 1 simulees were randomly drawn from a normal distribution with a mean of 0 and standard deviation of 1. Group 2 simulees were randomly drawn from a normal distribution with a mean of 1.5 and standard deviation of 1. Group 1 represented a "survey group" and Group 2 a mastery group (Kingsbury & Zara, 1991).

In addition to θ distributions, item pool size was also varied. There were three item pools of size 100, 300, and 500, comprised of four content areas of equal sizes. For each content area A, B, C, and D, b_i was normally distributed but varied in terms of mean and standard deviation across content areas [A: N ~ (0, 1), B: N ~ (-1, 1), C: N ~ (1, 1), and D: N ~ (0, .0625)]. This was done to represent different difficulty levels for different content areas in real data situations (Kingsbury & Zara, 1991). Discrimination, however, did not vary across content areas, and *a* was normally distributed with a mean of 1.0 and standard deviation of .25. Pseudo-guessing, c_i, was fixed at .20.

Each simulated group (μ = 0 and μ = 1.5) was administered nine tests (3 adaptive types × 3 item pool lengths). Item selection was based on true item parameters. For all three adaptive methods, $\hat{\theta}$ was updated after each item administered.

Tests were a fixed length of 48 items. For CCAT, the content-balance specifications were: 33.3% of items from content A, 33.3% from content B, 16.7% from content C, and 16.7% from content D. This ratio was preserved in the testlets as well. Testlets consisted of six items, 2 from A, 2 from B, 1 from C and 1 from D given in the order ABCDAB.

The three adaptive procedures were compared in terms of mean absolute error, information, and bias of maximum likelihood trait estimates.

Results. For test lengths greater than 10 items, the same pattern held true for all the conditions: the mean absolute error of $\hat{\theta}$ was smallest for the CAT, followed by CCAT, with testlets having the largest mean absolute error. Holding mean absolute error constant, CCAT required 10% to 15% more items while testlets required 60% to 105% more items depending on the size of the item pools. Also the mean difference between CAT and CCAT was consistently smaller than the mean difference between testlets and CAT or testlets and CCAT (Kingsbury & Zara, 1991).

A similar situation occurred when examining the information function of the item pool. For test lengths greater than 5, CAT provided the most information followed by CCAT and then testlets. When holding information constant, CAT was 5% to 15% shorter than CCAT and 30% to 50% shorter than testlets.

Kingsbury and Zara (1991) used bias to evaluate the various selection procedures. Bias is the average signed difference between the true and estimated parameters.

$$\text{Bias} = \frac{\sum_{j=1}^{n}(\theta_j - \hat{\theta})}{n} \qquad (41)$$

where θ_i is the parameter for person *j* (Cassuto, 1996; Yoes, 1993). It is an indication of whether a parameter has been over- or underestimated.

For θ, bias was substantial and variable for test lengths less than 15 items. For the μ = 0 distribution, bias was positive (indicating underestimation of θ) for CAT and CCAT but negative (overestimating θ) for testlets. For the μ = 1.5 distribution, bias was negative for all three types and 3 to 5 times larger. For longer tests (15 to 48 items) bias was slightly positive and converged toward 0 rather quickly. (Kingsbury & Zara, 1991).

Kingsbury and Zara (1991) concluded that for content control, CCAT is a better selection method with a relatively small price tag (5% to 10% increase in number of items and a 4% to 14% decrease in information as compared to CAT). There were some issues, however, that were not completely addressed. First, is the scrambled content order. For a simulation study, mixed content order does not influence performance, but this might or might not be the case with live examinees. Would examinees be able to perform as well when content is scrambled as opposed to taking each content area one at a time? This needs to be investigated empirically.

Another issue not well addressed is the two different ability distributions. Unlike previously mentioned studies, Kingsbury and Zara (1991) recognized that θ is not always normally distributed with a mean of 0 and standard deviation of 1. Unfortunately, they gloss over the differences between these two distributions and concentrate mainly on adaptive test type and item pool size. Overall, the μ = 1.5 distribution did not perform as accurately as the μ = 0 distribution. The mean absolute error was larger, the information level was lower, and bias took longer to stabilize. It is important to keep in mind that given the difficulty distributions of the content areas, it was an “easy” test for the $\mu = 1.5$ distribution. To determine which adaptive procedure is better it is necessary to vary both item and θ distributions, including skewed distributions and not use only normal distributions.

Another point that was not made by the authors is that for all conditions, as pool size increased, mean absolute error decreased and the number of items to achieve a specified mean absolute error decreased. This indicates that larger pool sizes increase accuracy and efficiency of adaptive testing.

Lastly, content control was the only impetus for creating the testlet. Often, this is not the reason why testlets were created. If a stimulus is time consuming such as a reading passage, it is more efficient and potentially less fatiguing for the examinee to be administered more than one item per passage. It is obvious that a testlet introduces a constraint and is not as efficient as individually administered items, but in some circumstances there may not be a viable choice. This study specified how to quantify that constraint. The next step is to compare different testlet selection procedures. For example, which is more efficient? Selection based on maximum item information with the rest of the items from the testlet administered along with it vs. maximum testlet information, which is the average item information across a testlet, could be compared against unconstrained CAT as a baseline

Comparison of adaptive selection procedures. Schnipke and Reese (1997) were interested in investigating various adaptive testlet selection procedures with Law School Admission Test (LSAT) items. Unlike Kingsbury and Zara (1991), who were interested in content-balancing, Schnipke and Reese wanted to use testlets in the context of a common stimulus such as a reading comprehension text. Schnipke and Reese were also interested in providing an opportunity for review without upsetting the precision and efficiency of adaptive testing. This is possible when the items are fixed within a testlet but adaptation occurs across testlets.

Schnipke and Reese (1997) simulated examinees and responses to compare θ estimates from six different test designs: a two-stage testlet design, a two-stage testlet design with changing levels (rerouting allowed for misclassification), a multi-stage testlet design, a standard maximum-information item-level design, a maximum-information testlet-based design, and a paper-and-pencil conventional test design.

Two groups of simulees were simulated. 50,000 simulees were randomly selected from a standard normal distribution for the purpose of establishing cutoffs. A second group of 25,000 simulees was simulated from a uniform (flat) distribution with 1,000 θs at 25 equal increments ranging from -3 to +3. This smaller group of simulees was used for the various test designs.

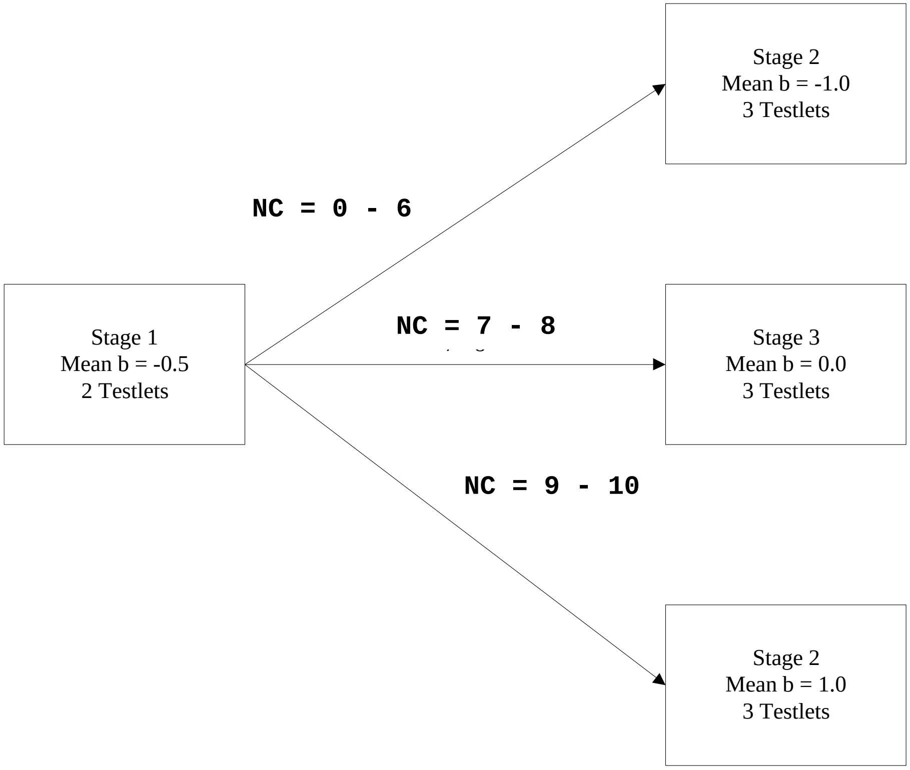
Testlets were created for the two-stage and multi-stage designs. The testlets from the multi-stage design were also used for the maximum-information testlet-based designs. Testlets of five items in length were simulated in the following manner. For Stage 1, b_i values were randomly selected from a normal distribution with a mean of -0.5 and a standard deviation of 0.8. For Stage 2, three levels of testlets were created by centering b_i at -1.0 (low), 0.0 (medium) or 1.0 (high) with a standard deviation of 0.8. The rationale for using this value for the standard deviation was not reported. Stage 3 contained 4 levels: b_i = -1.25, -.75, .75, or 1.25. Stage 4 contained 5 levels: b = -1.5, -1.0, 0.0, 1.0 or 1.5. All levels and stages had a standard deviation of 0.8. If the difference between the lowest and highest b_i within a level was between 1.5 and 2.0 and the testlet mean was within .3 of the specified mean then the testlet was retained (Schnipke & Reese 1997).

For the a_i parameter, Stage 1 a_is were selected from a normal distribution of mean 0.8 and standard deviation of 0.22. For the other stages, a_i was drawn from a normal distribution with a mean of .9 and standard deviation of 0.22, with the reasoning of saving more discriminating items for later.   The c_is for all stages/levels were drawn from a uniform distribution ranging from 0.15 to 0.25 to simulate a 5-alternative multiple-choice test.

Once the testlets were created, cutoff criteria were established. It is not clear why, but the cutoff criterion was based on NC score rather than an IRT-based score. Using the 50,000 normally distributed simulees, each simulee was administered two testlets from Stage 1 and a NC score was determined.   All simulees were administered three randomly selected testlets from the low level of Stage 2, θ was estimated, and mean squared error (MSE) determined, where MSE is the average squared discrepancy between true θ and estimated θ. Stage 2 was repeated for the medium and high level testlets with θ being re-estimated and a new MSE calculated for each level. The routing pattern was based on the pattern that yielded the lowest MSE. Once these criteria were established, the two-stage test design was implemented using the 25,000 simulees from the uniform distribution.   Simulees were administered two testlets (a total of 10 items) in Stage 1. Based on their NC score, they were routed to one of three levels: low (0 - 6 items correct), medium (7 - 8 items correct), or high (9 - 10 items correct: see Figure 1).

Figure 1

Two-Stage Testlet Routing



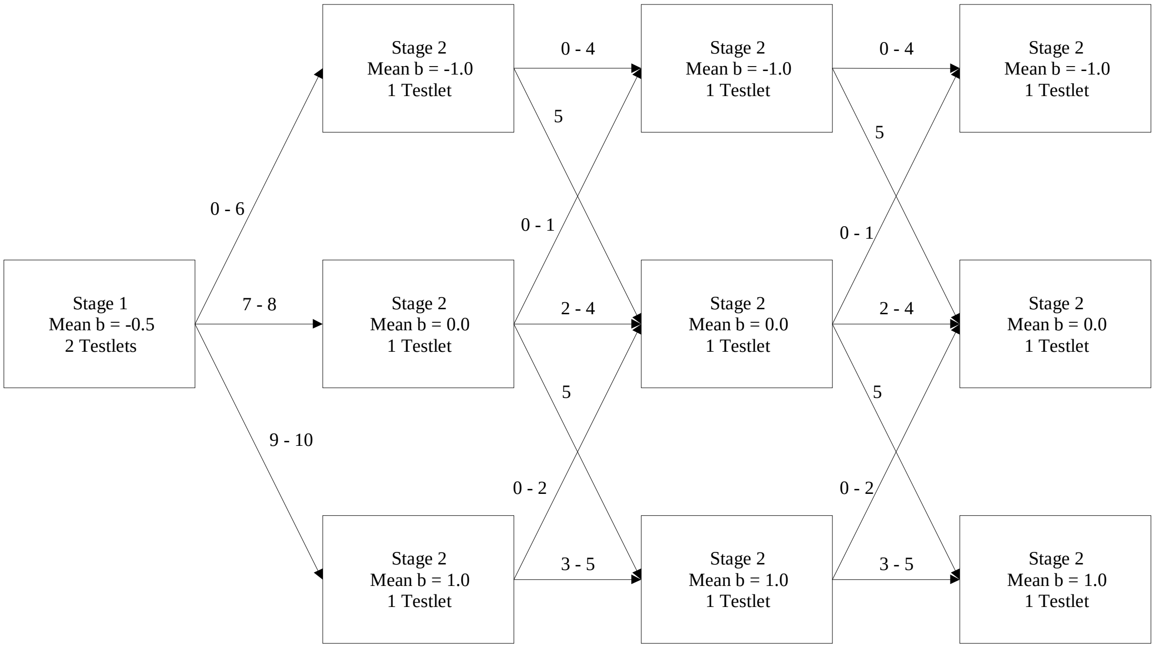
(Schnipke & Reese, 1997).

At Stage 2, three testlets were randomly selected from the appropriate level and administered. After all 25 items were administered, θ was estimated using Bayesian modal scoring with a standard normal prior.

A second two-stage test design was also implemented. It was designed to simulate rerouting in Stage 2 when simulees were misclassified. Simulees were administered two testlets from Stage 1, routed to one of three levels in Stage 2, and administered one testlet. Each simulee was administered one testlet from each of the other levels (see Figure 2).

Figure 2

Two-Stage Routing with Reroute at Stage 2 with NC listed at Branch



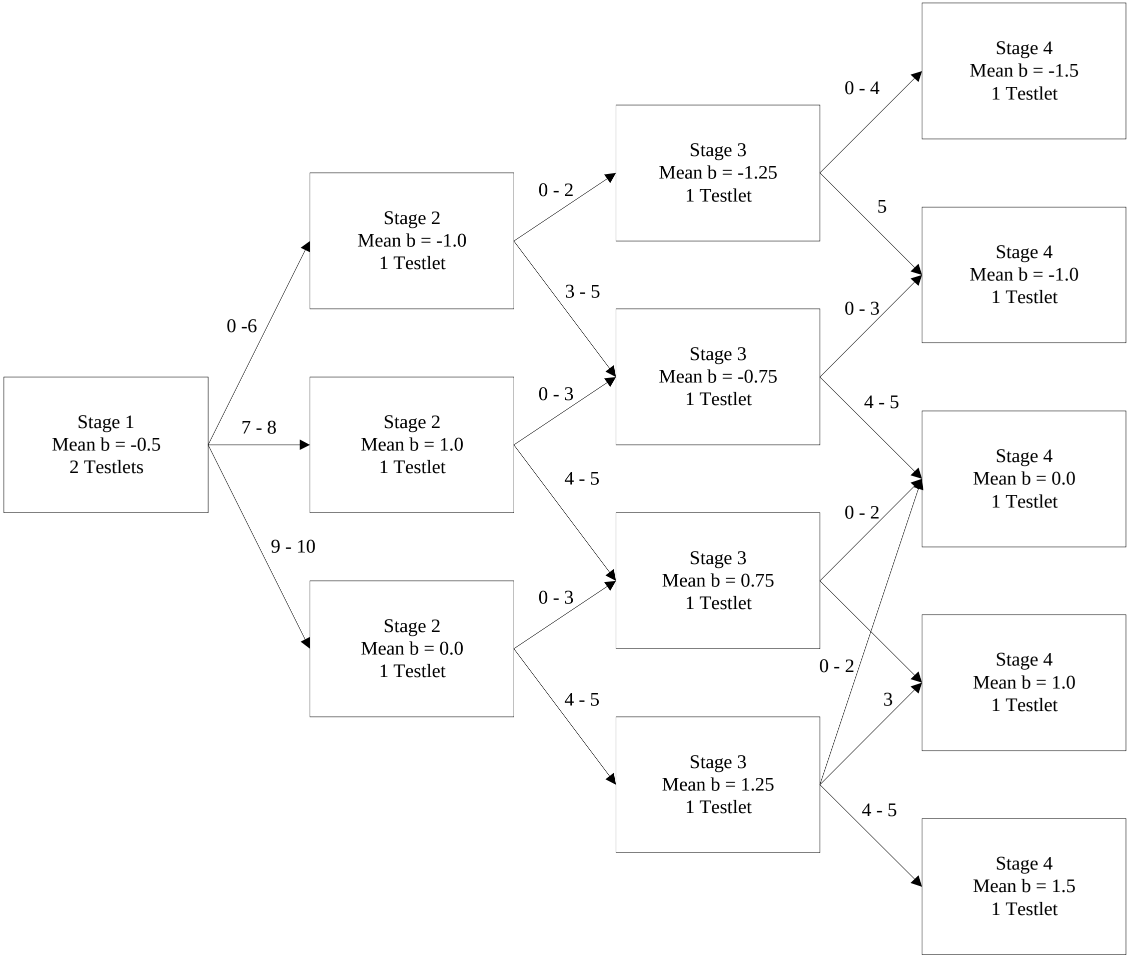
(Schnipke & Reese, 1997).

Like the previous test design, cutoffs were determined in the same manner from the normally distributed simulees based on θ estimates and MSE. The uniform group was essentially routed twice, once after Stage 1 and once after the first testlet in Stage 2. Final θ estimates were established in the same manner based on all 25 items using Bayesian modal scoring.

In the multi-stage testlet design (Figure 3), simulees took two testlets from Stage 1, one testlet from Stage 2 (3 levels), one testlet from Stage 3 (4 levels) and one testlet from Stage 4 (5 levels). Cutoffs, like in the previous designs, were based on the lowest MSE of simulees from the normal distribution, in a given level with a given number-correct score.

Figure 3

Multi-Stage Design with NC Listed at Branch



(Schnipke & Reese, 1997).

The fourth test design, standard maximum information item-level design, simulated a traditional CAT where testlet structure is ignored and items are selected based on maximum information for a given level of θ. Item information was calculated for 37 levels of θ ranging from -2.25 to +2.25 in increments of 0.125. There was one constraint, however, referred to as an exposure-control method (Kingsbury & Zara, 1991; Schnipke & Reese, 1997). The first item was selected randomly from the 10 items with the highest information at the current θ estimate. θ was re-estimated and the next item was randomly selected from the 9 most informative items not yet selected. The third item was randomly selected from the 8 most informative items not yet selected. This pattern continued until the 10^{th} item was based on maximum information. After all items had been administered, a Bayesian modal score was calculated and used as the final $\hat{\theta}$.

The fifth test design, maximum information testlet-based design, selected a testlet based on the summated information of the individual items comprising the testlet, or testlet information function. A total of five testlets (25 items) were chosen using the same exposure control method, except information was testlet-based rather than item-based.

The sixth and final test design was the traditional paper-and-pencil test. Items were taken from two intact sections of the LSAT that were considered to best measure middle ability (Schnipke & Reese, 1997). The length of each section was 25 and 26 items respectively. Responses were simulated for the 25-item length and the combined sections or 51-item test.

Since all the test designs (with the exception of the paper-and pencil) were of the same length (25 items), only accuracy of estimation of θ was considered. Two measures were used: root mean square error (RMSE) and Bias (Equation 41). RMSE is a measure of the discrepancy between true and estimated parameter (θ).

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n}\left(\theta_j - \hat{\theta}_j\right)}{n}} \qquad (42)$$

(Cassuto, 1996; Yoes, 1993).

Results. Several trends were observed their study. First, the standard maximum-information item-level design (the traditional CAT) performed the best, with the least amount of error and the least amount of bias. Second, the 25-item paper-and-pencil had the most error and bias. Third, the two-stage, multi-stage, and standard maximum-information item-level performed similarly in terms of RMSE and bias to the longer 51-item paper-and-pencil test for θs less than 1.5. For θs greater than 1.5, the 51-item paper-and-pencil test performed better than the two-stage and multi-stage test. Lastly, the maximum information testlet based design did slightly better than the longer paper-and-pencil test, especially in the tails of the θ distribution.

Schnipke and Reese (1997) concluded that the testlet-based designs improved precision over the same length pencil-and-paper and were almost as precise as the longer paper-and-pencil test. This is true, but it is not clear what, if any, differences there were among the various testlet-based designs. No tests of significance were performed. Also, because all the designs were of a fixed length (25 items), it is not clear which designs were more efficient. In the Kingsbury and Zara (1991) study, the testlet-based adaptive testing was the least efficient and was a major concern.

In addition, there are some questions about the design of the study. It is not clear why the b_i parameters were simulated from a standard normal distribution. Usually b_i parameters are selected from a uniform distribution to ensure a sufficient number of items for each level of θ. It is also unknown why the θs for the test simulation were from a uniform distribution yet the θs from the cutoff group were from a standard normal distribution. Furthermore, IRT-based adaptive testing did not occur until Stage 3. The first two stages for all the designs were based on fixed branching using number-correct scores. It is not clear how changing testlet selection strategies mid-stream affected performance. The next logical step is to compare various testlet-based adaptive test designs in terms of both precision and efficiency using an IRT-based selection procedure throughout the entire test.

CONCLUSIONS

Adaptive mastery testing began its development over 30 years ago with Ferguson's (1969) computerized arithmetic test and has continued to develop ever since. With the lightning speed development of computers and software, there is an increase in the utilization of computerized adaptive exams ranging from licensure/certification type exams (Nursing Boards and Architecture certification) to aptitude tests (SAT, GRE, TOEFL) to name a few. As the demand for different types of domains and testing environments sprouted, so did different models to meet those demands. Several adaptive mastery testing models (AMT, SPRT, EXSPRT, IRT-based SPRT, beta binomial, and testlets) have been proposed and studied.

SPRT and EXSPRT models (Frick, 1990; Frick 1992; and Luk, 1991) were investigated as an alternative to AMT for small-scale testing where large samples were not available for item calibration. Conclusions about which models performed the best are still in debate for some conditions and seems to be definitive for others. For example, all the studies that compared SPRT with other models (Frick, 1990; Kingsbury & Weiss, 1980, 1983; Luk, 1991; Spray & Reckase, 1996, and Weiss & Kingsbury, 1984) concluded that SPRT models required fewer items than AMT, but there was disagreement as to which model is more accurate concerning mastery decisions. Both Frick (1990) and Kingsbury and Weiss, (1980, 1983) found AMT to be more accurate, but Spray and Reckase (1996) found just the opposite although they did not implement AMT as originally proposed. Luk (1991) found that accuracy level depended on the type of errors. In Luk's (1991) study, AMT was more accurate for nonmastery decisions, but EXSPRT was more accurate for mastery decisions. This differential accuracy has been investigated only for these two models and should be investigated for other models and designs.

SPRT also seems sensitive to varying conditions. Kingsbury and Weiss (1980, 1983) found that SPRT performed well with items from a uniform distribution, and both Frick (1990) and Reckase (1979) found that the area of indifference influenced test length. A wider area of indifference required fewer items and a narrower area of indifference required more items, but the narrower indifference area was also more accurate in terms of mastery decisions.

Testlets were originally investigated as a means of controlling contextual effects (Wainer, Kaplan, & Lewis, 1992; Wainer & Kiely, 1987; and Wainer & Lewis, 1990). The Wainer et. al studies were concerned with the structure of the testlets and did not incorporate testlets with IRT. The types of issues examined (testlet length, item pool length, adaptive within vs. across testlets), though, are all issues that could and should be examined in the context of IRT. Another implementation of testlets was in the context of content balancing (Kingsbury & Zara, 1989, 1991; Lewis & Sheehan, 1990; and Sheehan & Lewis, 1992). All the testlet studies are in agreement that testlets allow a greater degree of control in terms of avoiding contextual problems in adaptive testing prior to administration. Whether or not testlets are an efficient means for content balancing is in dispute. Lewis and Sheehan (1990) and Sheehan and Lewis (1992) found that testlets did shorten testing, but this was in comparison to a full-length fixed test. Kingsbury and Zara (1991), on the other hand, found that testlets did not shorten tests, but their baseline was a traditional CAT rather than a paper-and-pencil test.

Recently, testlets have been investigated for the purpose of domains that require a common stimulus and to allow review of items without loss of efficiency (Schnipke & Reese, 1997). This study was an example of a real type of testing situation where both the domain and test administration constraints must be considered. Both Kingsbury and Zara (1991) and Schnipke and Reese (1997) demonstrated that a pure CAT with no constraints is the most efficient method of IRT-based adaptive mastery testing. Unfortunately, not all domains fit within that format. As adaptive mastery testing continues to become increasingly more popular, constraints such as domain and environment, and how to implement them in that context have created a need to develop more sophisticated test construction techniques. As a result, it is necessary to empirically investigate real data conditions and constraints in an effort to determine which method(s) provide the most accurate decisions with the least loss in efficiency. Some of these conditions have begun to be investigated (testlet length, item pool size) and other conditions have not (skewed θ distributions, varying item parameter distributions, IRT-based testlet selection procedures). Of particular interest, is the latter. Different methods of testlet selection where testlets are constructed, selected, and scored based on IRT, and varying conditions such as testlet length, ability distribution, and item parameter distributions need to be investigated in a systematic manner. To date, no research has directly compared these two selection procedures: (1) maximum testlet information and (2) maximum item information that preserves testlet structure (i.e., all the items connected to the testlet are administered along with the selected item, not just the item with maximum information) within the context of the 3-parameter dichotomous model.

CHAPTER 2: METHOD

Purpose

The use of AMT based on the selection of single items in the context of a 3-parameter IRT model, selecting items by maximum information, is well documented (Kingsbury & Weiss,1983; Kingsbury & Weiss, 1981; Kingsbury, 1980; Kingsbury & Weiss, and Weiss & Kingsbury, 1984).   Use of AMT to administer testlets in the context of a 3-parameter IRT model using maximum information is still not well developed. Lewis and Sheehan (1990) and Sheehan and Lewis (1992) used testlets in their AMT, but testlet selection was based on Bayesian decision theory rather than maximum information. Thissen, Steinberg, and Mooney (1989) also employed AMT with testlets, but in the context of a polytomous IRT model. Only two studies (Kingsbury & Zara, 1991; Schnipke & Reese, 1997) have used a dichotomous, 3-parameter model with testlet selection based at least partially on maximum information. Kingsbury and Zara (1991) were interested in comparing different selection techniques in an effort to control content balance on the Nursing Boards examination. Content-balanced testlets were one of the selection techniques. Schnipke and Reese (1997) used testlets in the context of a reading comprehension test, but it was a mixed model using both IRT and non-IRT selection methods. To date, no research has compared *different* IRT-based testlet selection techniques based on maximum information while systematically controlling for variables such as θ level and number of items per testlet.

The present study was designed to determine which conditions or combination of conditions rendered the “best” estimate of θ and the most accurate mastery decision to assess the accuracy and usefulness of different IRT-based testlet selection procedures.

Requirements and Criteria for AMT

Weiss (1985) and Weiss and Yoes (1991) stated that there are five necessary requirements for adaptive testing: (1) an item bank, (2) a starting point which can be fixed or variable, (3) an item selection algorithm/procedure, (4) a scoring procedure and (5) a termination rule. In adaptive mastery testing, an additional criterion is needed: mastery/nonmastery status. Because true mastery/nonmastery status of real examinees can never be known, in the present study person data were simulated with mastery status known a priori.

Item bank. In the context of testlets, creation of an item bank can be the most difficult requirement to meet. Because items within a testlet are administered as a "package," two distributions emerge: testlet parameters and item parameters. In order to make meaningful generalizations about performance of a particular testlet selection model under different conditions, it is best if the testlet parameter distribution parallels the item parameter distribution from the item pool from which it came.

In addition to this constraint, the number of items in the item bank itself will have to be large. Urry (1977) had suggested a minimum of at least 100 items. Kingsbury and Zara (1989, 1991) stated that when constraints are added to adaptive testing, a larger item bank is needed, so that there will be sufficient items for all desired conditions. Administration of testlets is a type of constraint. In this study, two item bank sizes were selected that allowed for at least 100 testlets of different lengths; one had 600 items and the other 900 items.

Variable starting point. A variable starting point can be advantageous if there is prior information about θ. When the initial item administered is close in difficulty to a person's θ level, less time and fewer items are needed to make an accurate decision. In actuality, accurate information about θ level prior to testing is rarely known. A best guess might be to administer an item of average difficulty ($b_i = 0$), which was done in this study. For this study, the initial estimate of θ before any items/testlets were selected was 0.

Item selection procedure. Comparison of item selection procedures was the main thrust of this study. Three item selection procedures were examined: (1) maximum testlet information (MTI), (2) maximum item information which preserves testlet structure (MII-T), and (3) maximum item information that ignores testlet structure, i.e., traditional CAT (TCAT). The first two selection procedures (MTI and MII-T) implement the adaptive testing procedure using testlets. Testlet information is similar to test information in that it sums information across the items comprising the testlet. MTI uses testlet information and selects the next, not previously administered testlet that maximizes testlet information nearest an examinee's most updated $\hat{\theta}$. MII-T searches for the item with maximum information at the current $\hat{\theta}$ but, having identified that item in the bank, administers all items within the testlet in which that item resides. The last selection method, TCAT, selects the item with the maximum information at the current $\hat{\theta}$, but does not necessarily administer the other items in the testlet. TCAT is AMT without the constraint of testlets and served as a baseline or reference for the other two selection procedures.

Scoring and termination criterion. Scoring was based on maximum likelihood estimation (Hambleton & Swaminathan, 1985). The criterion for termination was if the 95% confidence interval around $\hat{\theta}$ was completely above or completely below the criterion cutoff (Kingsbury & Weiss, 1979). If the 95% confidence interval was neither completely above or below the cutoff, then termination occurred when all items in the bank had been administered. The criterion cut-off was based on the 77% cut-off used in the foreign language data described in detail below.

Data Generation

Item parameters. Item parameters for the 600- and 900-item banks were simulated independently using PARDSIM (Yoes, 1998) in such a way that testlet distributions paralleled the item difficulty distribution of the bank. Item difficulty, b_i, was uniformly distributed in each bank from –3 to +3 to ensure that there were an adequate number of items available for use at every level of θ. Testlets were created so that the average difficulty of the testlets was uniformly distributed from -3 to +3, as well. This is described in more detail below.

Whenever possible, it is best to have simulated data resemble real data. The simulated data were modeled after real data acquired from the French, German, and Spanish Reading Comprehension sections of the Entrance Proficiency Tests (EPT) developed by the Center for Advanced Research on Language Acquisition (CARLA) and administered at the University of Minnesota. Each test was 35 items in length and comprised of several testlets. Each testlet consisted of a passage preceded by a sentence or two describing the setting, a relevant picture or graphic, and then a set of two to five questions (items) related to the passage. All items were multiple-choice with four alternatives. Item parameters were calibrated using XCALIBRE (Assessment Systems Corporation, 1995). The 3-parameter logistic model was used with item parameters estimated by marginal maximum likelihood estimation. Tables A-1, A-2, and A-3 in Appendix A give the item parameter estimates as well as the number of examinees and testlets from the EPT administered in Spring, 1997.

Since achievement tests tend to have more moderate discriminations (Yoes, 1993), the distribution of item discrimination, a_i, mimicked the discriminations of items from the CARLA data. Item discrimination from the German dataset was approximately normally distributed with μ_a = .73 and σ_a = .10 and was used for the simulation since it was slightly higher than the other two languages (μ_a = .63 for French and σ_a = .65 for

Spanish). Pseudo-guessing, c_i, was fixed at .25 to reflect a multiple-choice test with four alternatives. All three data sets had mean c_i parameters that were either at or quite close to .25 with a standard deviation of .01, indicating very little variance (see Table A-4 in Appendix A). Since the variance was so small, c_i was fixed at .25. To ensure enough items at all levels of difficulty, b_i was uniformly distributed from -3 to +3. A more detailed explanation is discussed in under the creation of testlets section.

θ *parameter.* In order to determine if there were differences among the CAT procedures as a function of θ, seven discrete levels of θ were selected at equal intervals across the range of the continuum from -3 to +3. One thousand simulees were generated at θ = -3, -2, -1, 0, +1, +2, and +3 for a total of 7000 simulees. Having 1,000 simulees at each level allowed for any idiosyncratic simulee data that might arise to be overwhelmed and factored out by the vast majority of "usual" simulee data. A relatively moderate number of levels was chosen to keep the number of conditions down to a reasonable level.

Item responses. Once the item and θ parameters were determined, item responses were generated using Monte Carlo simulation from PARDSIM (Yoes, 1998). Although the vast majority of simulees only "took" a subset of the item bank in the adaptive test, responses to all the items in the pool were generated since it was unknown a priori which items would be selected. Also, if a decision could not be made based on a 95% confidence interval, exhausting the item pool was a secondary way to terminate the adaptive testing procedure. Therefore, it was necessary for each simulee to have responses to all the items in the pool. In the Monte Carlo simulation, a random number from a uniform distribution ranging from 0 to 1 was generated. This number was compared to the probability of a correct response based on the 3-parameter logistic model (Equation 1) for that item and person. If the random number was less than the probability of a correct response, then a 1 or correct response was assigned to the simulee. If the random number was greater than the probability of a correct response, a 0 or incorrect answer was assigned to the simulee (Yoes, 1993; Cassuto, 1996).   This procedure was repeated across all items and across all simulees, so that each of the 7,000 simulees had a response vector for all the items in each of the two pools.

Creation of Testlets

Testlet width. As mentioned previously, to make meaningful comparisons across different selection methods, it was crucial that the b_i distribution of the testlets paralleled the b_i distribution of the item bank. Furthermore, the distance between the highest b_i and the lowest b_i within a testlet, or testlet width, was fixed across all testlets.   Having testlets of different widths might confound selection procedures. It is similar to the logic behind the homogeneity of variance assumption in ANOVA.   In ANOVA, variance is assumed to be fixed, so that any differences that are found across groups can be attributable to differences in means. Likewise, if testlets are of fixed widths, any differences found can then be attributable to conditions manipulated in the study.   For example, suppose two testlets have testlet information functions with the same peak position but one testlet has a wider testlet information function than another.   If MTI is the selection procedure, since they have the same peak, the selection outcome would be the same. If, however, MII-T was used, because the information functions are of different widths, the selection outcome might be different. It would, therefore, be difficult to determine whether the different outcomes were due to different selection procedures or due to varying testlet widths.   Hence, it was necessary to fix testlet width. Therefore, to obtain parallel distributions of testlets and items and to fix the width of the testlets, generation of the b_is was not purely random.

An appropriate fixed width was determined based on the widths in the CARLA data. When item difficulty was calibrated, it was determined that the average testlet difficulty width ranged from 1.18 to 1.46 (see Table A-5 in Appendix A).   For convenience, testlet difficulty width was fixed at 1.2 which allowed for varying the number of items within a testlet. The logic will become evident shortly.

Testlet length and bank. There were three different length testlets: 2-item, 3-item, and 5-item. Based on the CARLA data, the average length was 3 items but ranged from 2 to 6 item testlets. The 2-item testlets represented a "short" testlet while the 5-item testlet represented a "long" testlet. In order to have at least 100 testlets for every testlet length, there needed to be at least 500 items in the item bank. Unfortunately, this item pool size would not work for the 3-item testlets. Therefore, an item pool size of 600 was selected. This allowed for 300 two-item testlets, 200 three-item testlets, and 120 five-item testlets. Since it was not known whether 600 items was sufficient with these constraints, an item pool size of 900 was used as well. All banks were generated independently for a completely crossed design.

Creating testlet uniformity. To create a uniform distribution of the testlets that maintained equal testlet widths, the b_i of the initial item in the testlet was selected at discrete points along the b_i/θ continuum ranging from –3 to +3. For the 600-item pool, the interval was every 0.02 b_i units for the 2-item testlet, 0.03 b_i units for the 3-item testlet and 0.05 b_i units for the 5-item testlet. For the 900-item pool, the interval was 0.0133 b_i units for the 2-item testlets, 0.02 b_i units for the 3-item testlets and 0.0333 b_i for the 5-item testlets. For negative values of b_i, this initial point served as the lower endpoint of the testlet and represented the easiest item in the testlet. For positive values of b_i, this initial point served as the upper endpoint of the testlet and represented the most difficult item in the testlet. If the initial b_i value was negative, 1.2 was added to the value to create the highest b_i value in the testlet. If the initial b_i value was positive, 1.2 was subtracted from that value to create the lowest b_i value in the testlet. Selection in this manner created 2-item testlets that all had a fixed width while at the same time maintaining a uniform distribution of b for both the items and testlets. To create 3-item testlets, the midpoint of the 2-item testlet was the b_i of the third item. To create 5-item testlets, the ¼ and ¾ points became the b values for the 4th and 5th items respectively. A hypothetical example is illustrated in Figures 4 - 9.

To create testlets that were equally distributed along the b_i continuum, the initial values were selected at equal increments. For the 300 2-item testlets, .02 intervals were used. This meant that the easiest item of the first testlet was –3, the easiest item of the second testlet was –2.98, the easiest item of the third testlet was –2.96 and so forth. The value of $b_i = 0$ was skipped since it is neither positive nor negative. Then increments began again with $b_i = .02, .04, .06$, etc. For the 200 3-item testlets the increment was .03 (-3, -2.97, -2.94…). For the 120 5-item testlets, the increment was 0.05 (-3, -2.95, -2.90…). A similar pattern was used for the 900-item bank. For the 450 2-item testlets the increment was .0133 (-2.9858, -2.9725, -2.9592…), for the 300 3-item testlets, the increment was .02 (-3.00, -2.98, -2.96…), and for the 180 5-item testlets, the increment was .0333 (-2.9803. -2.9470, -2.9137…). To better understand how the creation of the item difficulty of the items comprising a testlet works, a hypothetical example is given.

Hypothetical example. Suppose the value of b_i is -2.5. Since this value is negative, this will be the easiest item in the testlet. The most difficult item in the testlet will then have a value of -1.3 since -2.5 + 1.2 = -1.3 (Figure 4). For the 3-item testlet, the first two items have the same b_i values of -2.5 and -1.3. The third item has a difficulty midway between the first two items $b_i = -1.9$ (Figure 5). For the 5-item testlet, the 4th and 5th items will have difficulty values at the ¼ and ¾ location in the range of difficulty $b_i = -2.2$ and $b_i = -1.6$ (Figure 6).

Figure 4

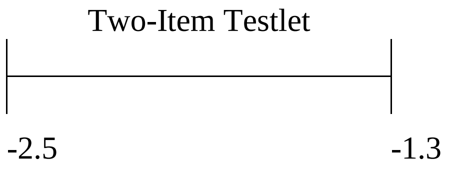


Figure 5

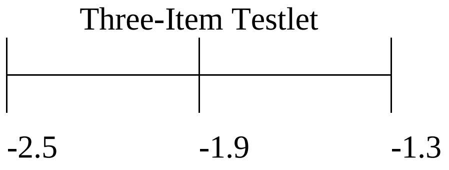
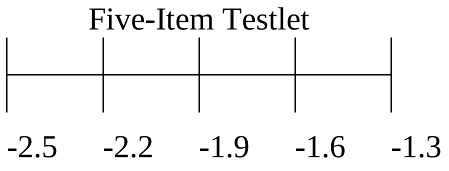


Figure 6



If the initial value of b_i was positive, say +2.5, the entire process was reversed. This value is the most difficult item in the testlet. The easiest item will have a value of +1.3. If it is a 3-item testlet, there will be an item with difficulty b = +1.9. If it is a 5-item testlet, there will be items of difficulty values b_i = +1.6 and +2.2, (Figures 7, 8, and 9).

Figure 7

Two-Item Testlet

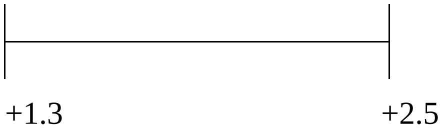


Figure 8

Three-Item Testlet

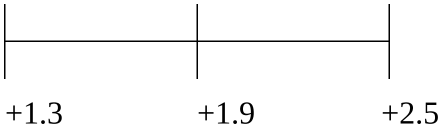
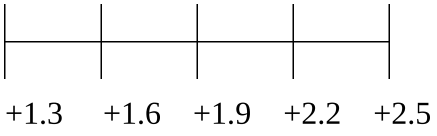


Figure 9

Five-Item Testlet



Testlet Selection Procedure

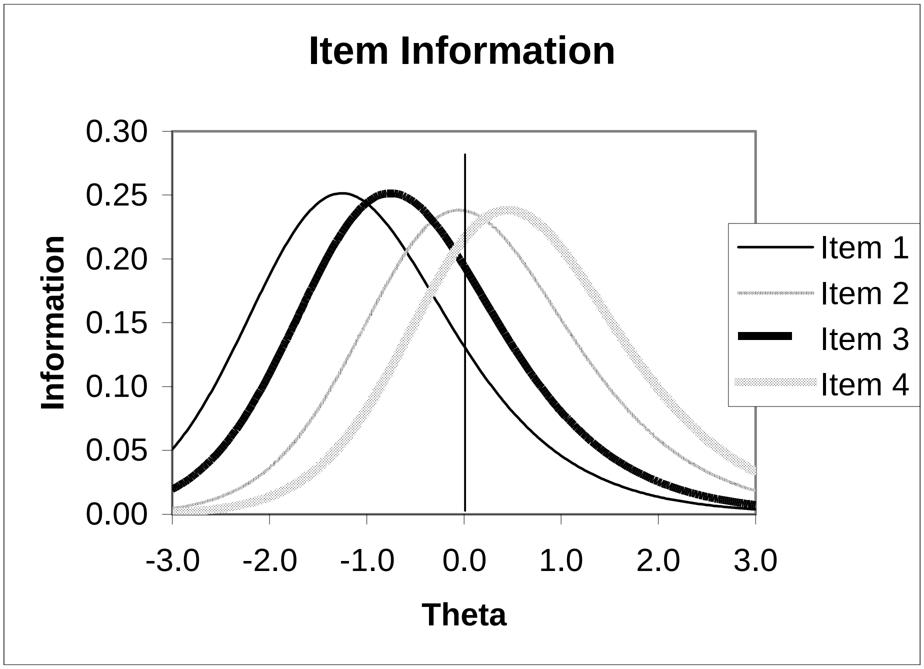
A simplified example of how the two testlet procedures operated is given using the item parameters from four items listed in Table 5. Items 1 and 2 belong to Testlet 1, and Items 3 and 4 belong to Testlet 2.

Table 5
Sample Item Parameters for Two 2-Item Testlets

	Item Parameters			
	Item	*a*	*b*	*c*
Testlet				
	1	.75	-1.5	.25
1	2	.73	-1.3	.25
	3	.75	-1.0	.25
2	4	.73	0.2	.25

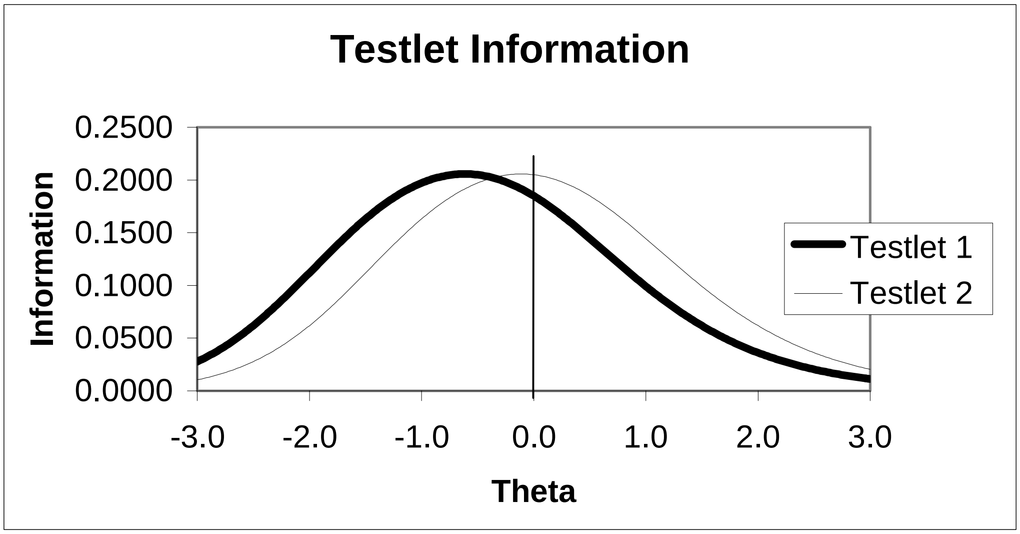
Figure 10a illustrates the item information functions for the four items. Suppose an examinee's most updated estimation of θ was 0. Item 2 has the most information at this point on the θ continuum. If using MII-T, Item 2 will be administered along with the rest of the items in Testlet 1 (in this case Item 1).

Figure 10a



If, however, testlets are chosen using MTI (the sum of the individual item information within a testlet), then Testlet 2 maximizes information at this θ level and would be administered next (Figure 10b).

Figure 10b



Using the baseline condition of MII, Item 2 would be chosen and administered. Based on the most updated estimate of θ, another item would be selected that maximizes information. Since testlet structure is not a constraint in this condition, the item could be any item from the pool not yet chosen.

Software

For the actual simulation itself, a modified version of POSTSIM (Weiss, 2005) was used. This program is designed to perform post-hoc simulation, i.e., to simulate an adaptive test when responses are known a priori; item response vectors were generated by Monte Carlo simulation. Items or testlets are based on a selection rule (e.g. maximum information) with $\hat{\theta}$ being updated dependent on the pattern of responses and the item parameters. An adaptive termination criterion is used to end the "testing session."

In the modified version of POSTSIM, the three types of selection rules employed were: TCAT, MII-T, and MTI. A minimum of 20 items were administered in all conditions. While it is possible to make a decision with far fewer items (3 or 4, for some examinees), it is not advisable, particularly when simulating a real data application. In a real testing situation where a mastery decision may be high-stakes (e.g., graduation requirement), there would be much concern if such an important decision were made after administration of only a few items, particularly if the decision was nonmaster. Twenty items is an acceptable minimum. Also, if an item bank is used for multiple purposes, such as decision and ability estimation, a longer test might be needed to increase the precision of $\hat{\theta}$.   A more precise $\hat{\theta}$, in turn, will reduce misclassification. Lastly, in a Monte Carlo simulation, as well as with real data, there may be a small percentage of aberrant response vectors. Having a minimum number of items allows for "correction" of $\hat{\theta}$ and reduction of misclassification, particularly, if the aberrancy occurs in the first few items.

In this simulation, termination occurred when the 95% confidence interval around $\hat{\theta}$ was completely above or completely below the mastery criterion. The 77% proportion correct criterion from the CARLA data set was used as mastery status. This converted to a mastery cutoff of θ = 1 for all the TRFs from each of the testlet length/item bank combinations (see Appendix B, Figures B-1 through B-6).

Design

Figure 11 illustrates the general design of the study with all the conditions. There were 3 testlet lengths (TL) × 3 selection procedures (SM) × 7 levels of θ × 2 item pool sizes (PS) for a total of 126 conditions. Each cell was generated by an independent dataset for a completely crossed design.

Figure 11
General Design of the Study
(With Item Pool Sizes of 600 and 900)

		Item/Testlet Selection Procedure								
Pool Size	θ Level	MTI			MII-T			TCAT		
	(N= 1000 per θ level)	Testlet Length			Testlet Length			Testlet Length		
		2	3	5	2	3	5	2	3	5
PS = 600	θ = -3									
	θ = -2									
	θ = -1									
	θ = 0									
	θ = +1									
	θ = +2									
	θ = +3									
PS = 900	θ = -3									
	θ = -2									
	θ = -1									
	θ = 0									
	θ = +1									
	θ = +2									
	θ = +3									

Dependent Variables

Three criteria were relevant in evaluating this study: (1) estimation of θ, (2) quality of decisions, and (3) efficiency of the selection procedures.

Estimation of θ. To evaluate how well each selection procedure estimated θ, three evaluative criteria were used: Pearson product-moment correlation (*r*), root mean square error (RMSE) (Baker, 1994; Cassuto, 1996; Yoes, 1993), and bias (Cassuto, 1996; Yoes, 1993).

The Pearson product-moment correlation is the correlation between true θ and $\hat{\theta}$. This dependent variable indicates how well rank order was maintained across the true and estimated θ parameters.

RMSE is a measure of average discrepancy between the true parameter and the estimated parameter. In this case, the parameter is θ. It is the square root of the average squared discrepancy (see Equation 42). A small value indicates little estimation error, whereas a large value indicates more estimation error.

In addition to the magnitude of estimation error, direction of the error needed to be examined as well. Bias indicates whether $\hat{\theta}$ is over-estimating or under-estimating θ (see Equation 41). If the bias value is positive, this indicates that the estimation procedure has a tendency to under-estimate θ: if the value is negative, this indicates that the estimation procedure has a tendency to over-estimate θ.

Quality of classification decisions. To evaluate how well each selection procedure classified simulees, two evaluative criteria were used: (1) classification percentages and (2) average decision error.

Five types of classification percentages were examined: (1) percentage of correctly identified masters, (2) percentage of correctly identified nonmasters, (3) percentage of misidentified masters, (4) percentage of misidentified nonmasters and (5) percentage of no decision. The first four percentages were obtained from a 2 × 2 decision table comparing actual mastery/nonmastery status and master/nonmastery decision (Table 6).

Table 6
True Status vs. Decision Made

	True Status	
Decision	Master	Nonmaster
Master	Percentage of Correctly Identified Masters	Percentage of Misidentified Masters
Nonmaster	Percentage of Misidentified Nonmasters	Percentage of Correctly Identified Nonmasters

The fifth percentage was based on simulees where the entire item pool was administered but a 95% confidence interval decision could not be made. It is important to quantify these simulees, as well. A larger percentage is indicative of a less economical decision procedure whereas a smaller percentage is indicative of a more economical procedure.

In addition to the rate or percentage of decision errors made, magnitude was also measured. When an incorrect decision was made, there was a measurable distance of $\hat{\theta}$ from the cutoff of $\theta = 1$. A large value indicates that $\hat{\theta}$ was far from the mastery cutoff of $\theta = 1$ and a worse decision error was made than if the value was small indicating closer proximity to the cutoff. Since incorrect decision were made in both directions, this value was squared and then the square root taken:

$$\text{Average decision error (ADE)} = \sqrt{\sum_{k=1}^{m} \frac{\left(\hat{\theta}_k - \theta_c\right)^2}{m}} \qquad (43)$$

where $\hat{\theta}_k$ is the estimate for an incorrect decision of person *k*
θ_c is the cutoff
m is the total number of misclassifications

Efficiency of the selection procedure. There are three ways of examining efficiency of a selection procedure: (1) decision accuracy, (2) test length, and (3) a combination of the two. Decision accuracy can be measured through the use of phi correlations. The two variables of interest were true status (true master and true nonmaster) and decision (classified master and classified nonmaster). The higher the correlation, the higher the agreement between true and classified status. Test length is simply the number of items administered to a simulee when a 95% confidence decision was reached. Shorter tests are indicative of a more efficient procedure. A better way to look at efficiency, however, is to combine accuracy and test length. A particular selection procedure can be accurate in terms of decisions but administer long tests. Likewise, a selection procedure can shorten the adaptive process but the decisions are less accurate. An ideal procedure will have shorter tests that maintain a high level of accuracy.

Efficiency was defined as the proportion of correct decisions made divided by average test length:

$$\text{Efficiency} = \frac{\frac{1}{\mathrm{N}}\sum_{j=1}^{N} d_j}{\frac{1}{\mathrm{N}}\sum_{j=1}^{\mathrm{N}} l_j} \tag{44}$$

Where d_j = 1 if a correct decision was made for person j
l_j = the number of items administered (test length) for person j

Since the number of simulees or sample size is in both the numerator and denominator, this equation can be simplified to Equation 45.

$$\text{Efficiency} = \frac{\sum_{j=1}^{N} d_j}{\sum_{j=1}^{\mathrm{N}} l_j} \qquad (45)$$

If long tests are administered, the denominator becomes large and efficiency is reduced. As the number of correct decisions increases (i.e., precision), the numerator becomes large and efficiency is increased.

Table 7 summarizes the 12 dependent variables used in the analyses.

Table 7
Dependent Variables

Dependent Variable
θ Estimation
Correlation
RMSE
Bias
Decision
Percentage of True Masters
Percentage of False Masters
Percentage of True Nonmasters
Percentage of False NonMasters
Percentage of No Decision
Decision Error
Efficiency
Phi Correlation
Test Length
Accuracy/Test Length
Bivariate Plot of Correct Decision and Test Length

Analysis

Data transformation. To analyze the data by means of ANOVA, certain transformations were needed. The first transformation was for the correlation between θ and $\hat{\theta}$. When the population correlation between two variables ρ_{XY} is other than zero, the distribution of the sample correlation r_{XY} tends to be skewed (Hayes, 1988, p. 589). In this study, r_θ was negatively skewed. This skewness can be corrected through the use of a function known as the Fisher *r*-to-z transformation.

$$Z = \frac{1}{2}\ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right) \qquad (46)$$

For positive skewness related to RMSE and average bias, logarithmic transformations were made (Cassuto, 1996; Howell, 1992, p. 311-314; VanLoy, 1996; Yoes, 1993). The RMSE was transformed to a logarithmic mean square error (LMSE) defined as:

$$\text{LMSE} = \log_{10}(\text{RMSE}^2) \qquad (47)$$

To avoid problems related to bias values either close to zero or at zero, a constant of 1 was incorporated into the logarithmic transformation of bias (Cassuto, 1996; Howell, 1992, p. 314). The LBIAS transformation was defined as follows:

$$\text{LBIAS} = \log_{10}(\text{bias} + 1) \qquad (48)$$

Decision error was found to be positively skewed, as well, and required a transformation. It is similar to RMSE in that it is measuring distance from true θ, but it would be problematic to use a logarithmic function when values could be 0 (no error). Therefore a square root transformation was used instead:

$$\text{Sqrt(Error)} = \sqrt{ADE} \qquad (49)$$

where ADE is average decision error

The last transformation dealt with proportions. The five classification proportions: no decision, correctly identified masters, correctly identified nonmasters, incorrectly identified masters, and incorrectly identified nonmasters along with converted proportions from the percentage values in the 2 × 2 decision table (Table 6) were transformed as follows:

$$Y = 2 \text{ arcsine } \sqrt{p} \qquad (50)$$

where Y = transformed value
p = proportion

(Howell, 1992).

Once these transformations were made, analysis of variance (ANOVA) was conducted for the various dependent variables.

Analysis of variance. To interpret the results in a cogent manner, a completely crossed ANOVA design was conducted for each of twelve dependent variables listed in Table 7. Table 8 lists the independent variables: item/testlet selection method, testlet length, θ level, and item pool size. Because the sample size was so large (N = 7,000), it was not practical to do tests of significance (because power is related to sample size, small differences will be significant). Instead, the purpose of conducting these ANOVA was to examine the strength of association or effect size. Effect size “assesses the amount of total variance in the dependent variable that is predictable from knowledge of the levels of the independent variable” (Tabachnick, 1996, p. 53). It can be estimated as:

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \qquad (51)$$

As suggested by Yoes (1993), an $\eta^2 \geq .05$ was considered “sizable”.

Because there was only one observation per cell (dependent variable) it was necessary to pool the 3-way and 4-way interactions to create a residual error term. For each of the main effects -- selection method, testlet length, θ level, and pool size and their 2-way interactions -- η^2 was computed for each of the evaluative criteria (dependent variables). No *F* and *p* values were computed, since the main purpose was to summarize results and identify the independent variables influencing θ recovery, decision, and test efficiency. Table 8 shows the ANOVA table for the main effects, 2-way interactions and degrees of freedom (*df*).

Table 8
Sources of Variation and Degrees of Freedom (*df*)
for ANOVAs of the Dependent Variables

Source of Variation	*df*
Main Effect	
Item/Testlet Selection Method (SM)	2
Testlet Length (TL)	2
θ Level (θ)	6
Pool Size (PS)	1
2-way Interactions	
SM × TL	4
SM × θ	12
SM × PS	2
TL × θ	12
TL × PS	2
θ × PS	6
Residual	76
Total	125

For those main effects that had η^2 values ≥ .05, secondary analyses were conducted to determine levels of the main effect in question contributed to the large effect size. To be consistent with the concept of η^2, a proportion of effect size (PES) was calculated by taking the ratio of $SS_{level}/SS_{between}$ for each of its components or levels of the main effect. By calculating PES, it is possible to examine the influence of each level on effect size in terms of a proportion for a given dependent variable. A large PES value is generated when a particular level is quite different from the other levels since it is based on deviations from the mean. Smaller values occur when several levels have similar means. PES quantifies the influence of a particular level and is a way of examining level differences when significance testing is not appropriate.

To determine if a condition with a large PES is a favorable condition, the magnitude of PES was considered in conjunction with the magnitude of the mean for a particular dependent variable. For some dependent variables, such as Fisher *r* to *z* and ϕ correlations, where a large mean indicates either more accurate estimation or more precise decisions, a favorable condition would have the highest mean and the largest PES. For those dependent variables involving error such as RMSE, bias or ACE, a favorable condition would have the smallest mean and the largest PES. Lastly, for a favorable condition, a large PES was defined as having a magnitude of 50% or larger. In other words, it accounted for at least 50% of the effect size.

For 2-way interactions, the same principles were applied. For those with η^2 values $\geq .05$, PES was calculated. Each level of the interaction was a particular permutation. For example, in a SM × PS interaction there were 6 possible combinations: (TCAT and 600-item pool, TCAT and 900-item pool, MII-T and 600-item pool, MII-T and 900-item pool, MTI and 600-item pool, and MTI and 900-item pool). For other interactions there could be as many as 21 combinations. Due to the large number of permutations created from interactions, only the value of PES showing the largest magnitude was reported. Because of the way that sums of squares (SS) are calculated for interactions ($SS_{A \times B} - SS_A - SS_B$), at the individual component level, it is possible to elicit both negative PES values and positive PES values greater than 1 for the various components of a particular interaction. When this occurs, they do sum to 1 but only PES values between 0 and 1 (legitimate proportions) were considered for interpretation purposes.

CHAPTER 3: RESULTS

PARAMETER RECOVERY

θ Estimation

Correlations

Table 9 shows θ recovery in terms of the product-moment correlation (r_θ) and mean test length (MTL). Correlation demonstrates the similarity of the ordering of true ability (θ) and, ordering of estimated ability ($\hat{\theta}$). For this particular evaluative criterion there were only 18 conditions rather than 126, because true θ level was fixed for each set of conditions. The seven θ levels were discrete points rather than distributions, leaving no variation with which to compute a correlation.

Table 9
Correlation of θ and $\hat{\theta}$ (r_θ) Estimated by TCAT, MII-T and MTI
by Testlet Length (TL), Item Pool size (PS), and Mean Test Length (MTL)

	PS= 600			PS= 900		
TL	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
r	.957	.925	.921	.962	.929	.944
MTL	69.31	72.54	72.67	82.42	97.09	92.59
3 items						
r	.951	.885	.899	.958	.886	.923
MTL	66.18	73.03	71.89	88.81	92.77	95.46
5 items						
r	.957	.854	.904	.952	.813	.827
MTL	68.92	74.08	72.43	94.89	99.13	95.50

Correlations ranged from .813 to .962. TCAT had mean correlations higher than the other two selection procedures. The 2-item testlets had higher correlations than the longer testlets, particularly in the larger item pool. MII-T had lower correlations than MTI in all but one data set.

Comparing item pools, MTL was longer across all three selection methods and across all three testlet lengths for the larger pool (PS = 900). Within each item pool, the MTL was shortest for TCAT across all three testlet lengths and both pool sizes.

Effect sizes (η^2). Table 10 shows the η^2 values from the factorial ANOVA on the Fisher's *r* to z transformation of the r_θ correlations between "true" θ and $\hat{\theta}$.

Table 10

η^2 from the ANOVA on Fisher *r* to *z* values of r_θ

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	.79	2	.658
Testlet Length (TL)	.24	2	.198
Pool Size (PS)	.00	1	.000
2-way Interactions			
SM × TL	.08	4	.067
SM × PS	.00	2	.002
TL × PS	.06	2	.054
Residual	.03	4	.021
Total	1.19	17	

There were two main effects (SM and TL) and two interactions (SM × TL and TL × PS) that were considered sizable (i.e., $\eta^2 > 0.05$). Selection method was by far the largest effect, accounting for almost 66% of the variance. Testlet length was next in magnitude ($\eta^2 = .198$). The two interactions involving testlet length combined accounted for approximately 12% of the variance.

The means and standard deviations of r_θ for each level of the two main effects with $\eta^2 > .05$ are shown in Table 11. TCAT and shorter testlets had the highest means whereas MII-T and longer testlets had the lowest means.

Table 11

Means and SDs of r_θ for the Main Effects of Selection Method and Testlet Length

Variable	Mean	SD
Selection Method		
TCAT	0.9561	0.0042
MII-T	0.8820	0.0439
MTI	0.9031	0.0405
Testlet Length		
2 items	0.9397	0.0175
3 items	0.9169	0.0320
5 items	0.8846	0.0625

PES analyses. Since selection method had a sizable η^2, sums of squares and (PES) were calculated by taking the ratio of $SS_{level}/SS_{between}$ for each of its components or levels of the selection method main effect. Table 12 lists the lists the means, *n*s, SS, and PES of Fisher *r* to *z* of r_θ for the three selection methods.

Table 12

Means, *n*s, SS, and PES

of Fisher *r* to *z* of r_θ for the Three Levels of the SM Main Effect

Variable	Mean	*n*	SS	PES
Selection Method				
TCAT	1.9005	6	0.5015	0.6379
MII-T	1.4137	6	0.2345	0.2983
MTI	1.5200	6	0.0501	0.0638
Main Effect	1.6114	18	0.7861	

TCAT had by far the largest contribution to the SS, since its PES (64%) reflected almost two-thirds of the main effect. MII-T had the second largest contribution and MTI had very little contribution. When interpreting which levels were different from one another, it is important to combine the magnitude of PES with the magnitude of the means as well. In the case of Fisher *r* to *z*, a larger mean indicates a higher correlation between θ and $\hat{\theta}$, which is more desirable. Since TCAT had both the highest mean and the largest PES, this means that TCAT was very different from MII-T and MTI in terms of correlations between θ and $\hat{\theta}$. MII-T had some influence, with a PES of 30%, but its mean was quite similar to the MTI mean. In this case, TCAT was a more favorable method.

Testlet length also had a sizable η^2 although much smaller in magnitude than selection method. Table 13 lists the means, *n*s, SS, and PES of Fisher *r* to *z* of r_θ for the three testlet lengths.

Table 13

Means, *n*s, SS, and PES

of Fisher *r* to *z* of r_θ for the Three Levels of TL

Variable	Mean	*n*	SS	PES
Testlet Length				
2-item	1.7544	6	0.1227	0.5202
3-item	1.6056	6	0.0002	0.0008
5-item	1.4741	6	0.1130	0.4790
Main Effect	1.6114	18	0.2360	

Both the shortest and longest testlets had the most contribution to PES and they were distributed somewhat evenly (52% and 48% respectively). The 2-item testlet, however, had the largest mean and therefore was the most favorable condition of the testlet lengths.

The means and standard deviations of Fisher *r* to *z* of r_θ for the 2-way interaction of SM × TL (selection method and testlet length) are shown in Table 14.

Table 14

Means and SDs of Fisher *r* to *z* of r_θ for the SM × TL Conditions

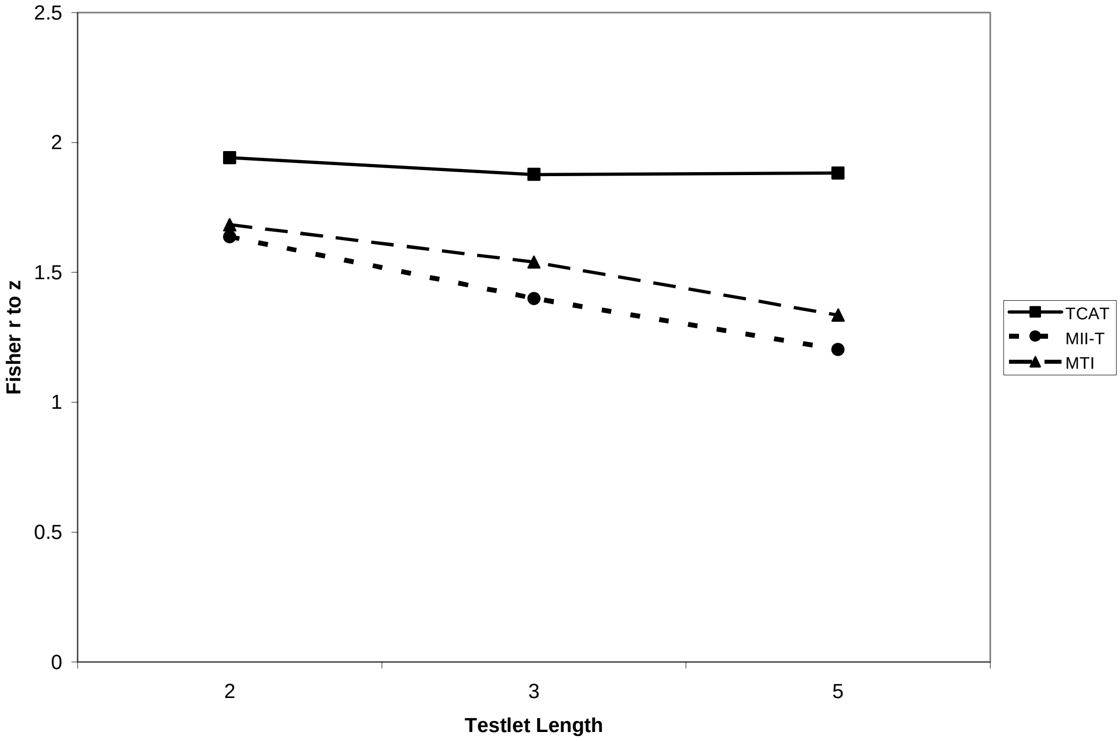
	TL = 2		TL = 3		TL = 5	
Selection Method	Mean	SD	Mean	SD	Mean	SD
TCAT	1.9421	0.0467	1.8770	0.0561	1.8824	0.0385
MII-T	1.6372	0.0241	1.3999	0.0045	1.2040	0.0968
MTI	1.6840	0.1287	1.5399	0.1006	1.3361	0.2222

The combination of TCAT and TL = 2 had the highest mean (1.9421) and MII-T and TL = 5 had the lowest mean (1.2040). TCAT had the highest mean for all three testlet lengths. TL = 2 had the highest mean and TL = 5 had the lowest mean.

Due to the large number of permutations created from interactions, only the value(s) of PES showing the most different and/or favorable condition(s) are reported and figures are used to illustrate any trends.

Figure 12 shows the two-way interaction of Fisher *r* to *z* between SM and TL.

Figure 12
2-Way Interaction of Fisher *r* to *z* between SM and TL



TCAT maintained the highest correlation across all three testlet conditions but performed marginally best with the shortest length testlets. MII-T performed slightly better than MTI but with a similar trend. As testlet length increased, performance decreased for all three methods but had a sharper drop for the two testlet-based methods. Only two permutations had PES values between 0 and 1: TCAT using TL = 2 and MII-T using TL = 3. The PES for both permutations were quite similar (13% and 14% respectively), but TCAT using TL = 2 had the higher mean (1.942). Therefore, while both of these permutations were different from all the other permutations, TCAT using TL = 2 was the most favorable condition.

The means and standard deviations of Fisher *r* to *z* of r_θ for the 2-way interaction of PS × TL are shown in Table 15. Means range from 1.3899 (900-item pool and 5-item testlet) to 1.8014 (900-item pool and 2-item testlets). Shorter testlets had higher means and larger means were observed for the larger pool size, except for TL = 5.

Table 15

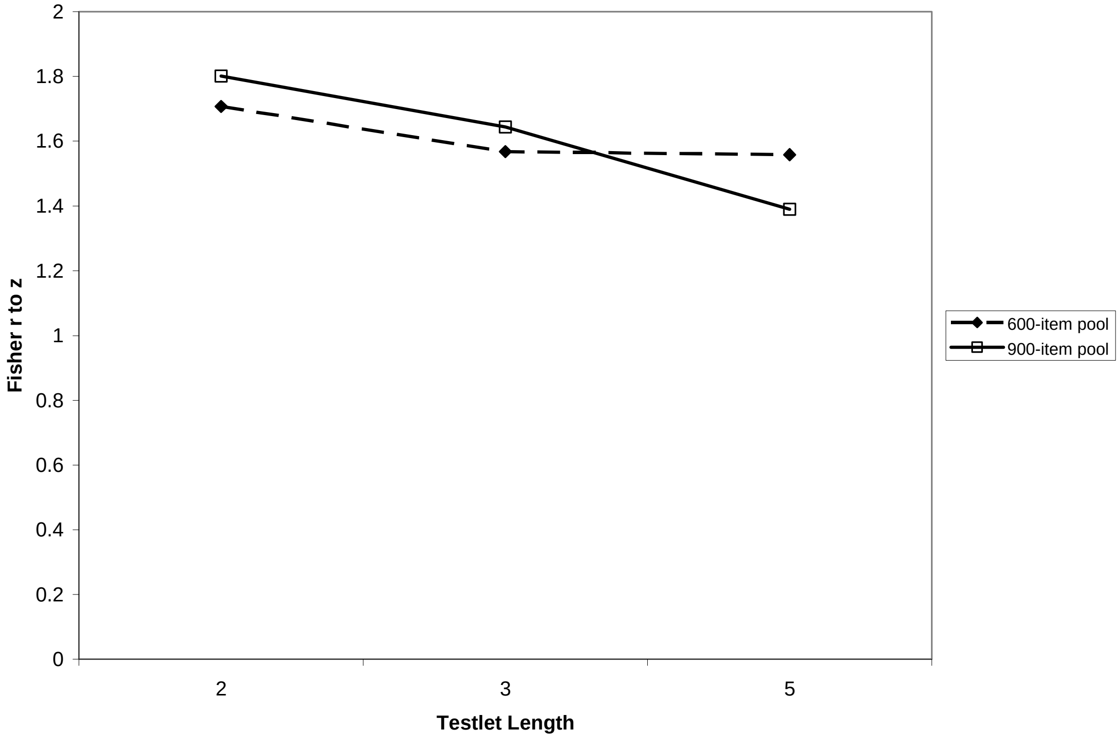
Means and SDs of Fisher *r* to *z* of r_θ for the PS × TL Conditions

	TL = 2		TL = 3		TL = 5	
Pool Size	Mean	SD	Mean	SD	Mean	SD
600 items	1.7074	0.1752	1.5676	0.2364	1.5584	0.3235
900 items	1.8014	0.1621	1.6436	0.2583	1.3899	0.4035

Figure 13 shows the two-way interaction of Fisher *r* to *z* between PS and TL.

Figure 13

2-Way Interaction of Fisher *r* to *z* Between PS and TL



For both pool sizes, shorter testlet lengths yielded higher correlations and TL = 5 yielded the lowest correlation. The 900-item pool had a slightly higher mean correlation than the 600-item pool for TL = 2 and TL = 5 but had a lower mean correlation for TL = 5. The highest mean correlation was for the TL = 2 with the 900-item pool, but unfortunately, the PES was negative. Only one condition yielded a PES between 0 and 1: PS = 600 and TL = 3 (PES = 13%). Since this mean was lower than several other means, there was no favored condition. In other words, no combination of conditions performed better than the others, but this condition performed worse.

RMSE

Table 16 shows the recovery of θ in terms of RMSE.

Table 16

RMSE of θ and $\hat{\theta}$ Estimated by TCAT, MII-T and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = -3	1.1418	1.9599	2.0676	1.0411	1.8290	1.4864
θ = -2	0.5666	0.5466	0.5403	0.5336	0.5799	0.6150
θ = -1	0.5249	0.4901	0.5016	0.4719	0.4893	0.4900
θ = 0	0.3876	0.3727	0.3820	0.3741	0.4183	0.4053
θ = 1	0.4329	0.4409	0.4367	0.4094	0.4394	0.4385
θ = 2	0.3756	0.4201	0.4347	0.3982	0.4279	0.4128
θ = 3	0.4922	0.5362	0.5342	0.4745	0.5754	0.5537
3 items						
θ = -3	1.3470	2.7359	2.5156	1.1685	2.6221	2.0064
θ = -2	0.5584	0.6872	0.5909	0.5866	1.0041	0.6368
θ = -1	0.5406	0.5026	0.4963	0.4791	0.5144	0.4818
θ = 0	0.3969	0.4035	0.4010	0.3914	0.3925	0.3907
θ = 1	0.4263	0.4496	0.4585	0.4169	0.4766	0.4290
θ = 2	0.3895	0.4266	0.4330	0.3762	0.4014	0.3922
θ = 3	0.5187	0.5587	0.5676	0.4921	0.5733	0.5552
5 items						
θ = -3	1.1874	3.1116	2.2559	1.3140	4.0052	3.7319
θ = -2	0.5470	1.5202	1.0040	0.5688	1.3220	1.5533
θ = -1	0.4650	0.5007	0.5028	0.5204	0.6493	0.7058
θ = 0	0.4030	0.4324	0.3928	0.3718	0.4290	0.4099
θ = 1	0.4183	0.4579	0.4724	0.3848	0.4607	0.4694
θ = 2	0.3798	0.4808	0.4578	0.3633	0.4387	0.4527
θ = 3	0.5169	0.6255	0.6141	0.4830	0.5959	0.5737

The minimum RMSE (.3633) (i.e., the set of conditions with the most accurate recovery of θ) was with PS = 900, TL = 5, and θ = 2 using TCAT. The maximum RMSE (4.0052) (i.e., the least accurate recovery of θ, was with PS = 900, TL = 5, and θ = -3 using MII-T.

Effect Sizes (η^2*).* Table 17 shows the η^2 values from the factorial ANOVA of the LMSE transformation of RMSE. Only the η^2 from the main effect θ level was considered sizable and explained almost 85% of the variance.

Table 17
η^2 from the ANOVA on LMSE

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	1.03	2	0.035
Testlet Length (TL)	0.53	2	0.018
Theta Level (θ)	24.73	6	0.847
Pool Size (PS)	0.00	1	0.000
2-way Interactions			
SM × TL	0.27	4	0.009
SM × θ	1.14	12	0.039
SM × *n*	0.02	2	0.001
TL × θ	0.71	12	0.024
TL × *n*	0.04	2	0.001
θ × *n*	0.05	6	0.002
Residual	0.66	76	0.024
Total	29.19	125	

Table 18 shows the means and standard deviations of RMSE for the main effect of θ level. Means were smallest at and near the mastery cutoff of θ = 1.0, indicating greater accuracy in θ recovery.

Table 18

Means and SDs of RMSE for the Main Effect of θ Level

Variable	Mean	SD
θ Level		
θ = -3	2.0849	0.8941
θ = -2	0.7756	0.3494
θ = -1	0.5181	0.0617
θ = 0	0.3975	0.0175
θ = 1	0.4399	0.0240
θ = 2	0.4145	0.0323
θ = 3	0.5467	0.0443

PES analyses. Table 19 lists the means, *n*s, SS, and PES of LMSE for the seven θ levels. θ = -3 accounted for the largest portion of the main effect (74%) followed by θ = 0 (9%) and θ = 2 (8%). The interpretation of which levels were different from the others identified levels with absolute mean value closest to 0 and a large PES. The level with the most influence or highest PES was θ = -3. It had a medium amount of relative estimation error, but the highest PES (74%). The second most influential level was θ = 0. It had the most amount of estimation error. The least influential level was θ = 3. Since the level with the smallest error (θ = -2) had a very low PES (2%), there was no most favored condition for LMSE. The least favored condition was θ = 0 -- it had the highest estimation error and the second highest PES (9%).

Table 19

Means, *n*s, SS, and PES of LMSE for the Seven Levels of True θ

Variable	Mean	*n*	SS	PES
True θ				
-3	0.5652	18	18.3371	0.7415
-2	-0.2865	18	0.4472	0.0181
-1	-0.5761	18	0.3139	0.0127
0	-0.8021	18	2.3076	0.0933
1	-0.7145	18	1.3165	0.0532
2	-0.7674	18	1.8816	0.0761
3	-0.5272	18	0.1243	0.0050
Main Effect	-0.4441	126	24.7281	

Bias

Table 20 shows the recovery of θ in terms of bias$_\theta$. The two best values of bias, (closest to zero) were TCAT with TL = 5, θ = -1 and PS = 600 (bias = -0.0009) and TCAT with TL = 3 and θ = 0 and PS = 900 (bias = 0.0010). The two worst values of bias (i.e., furthest from zero) were MII-T with TL = 5, θ = -3 and PS = 900 (bias = 1.125) and MTI with TL = 5, θ = -3 and PS = 900 (bias = 1.0846). As positive values they were both underestimating θ.

Table 20

Bias of θ and $\hat{\theta}$ Estimated by TCAT, MII-T and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = -3	0.0490	0.2948	0.4108	0.0547	0.2758	0.2381
θ = -2	-0.0721	-0.0542	-0.0727	-0.0982	-0.0310	-0.0541
θ = -1	-0.0557	-0.0218	-0.0358	-0.0395	-0.0382	-0.0434
θ = 0	0.0042	0.0580	0.0457	0.0220	0.0575	0.0684
θ = 1	-0.0488	-0.0501	-0.0362	-0.0098	-0.0316	-0.0356
θ = 2	-0.0784	-0.1145	-0.1054	-0.0797	-0.0959	-0.0770
θ = 3	-0.0220	-0.0249	-0.0348	-0.0392	-0.0429	-0.0474
3 items						
θ = -3	0.1576	0.5500	0.5557	0.1041	0.5220	0.3893
θ = -2	-0.0309	-0.0401	-0.0395	-0.0189	0.0438	-0.0142
θ = -1	-0.0538	-0.0064	-0.0432	-0.0297	0.0143	-0.0383
θ = 0	0.0191	0.0515	0.0341	0.0010	0.0374	0.0331
θ = 1	-0.0567	-0.0572	-0.0750	-0.0643	-0.0638	-0.0624
θ = 2	-0.0546	-0.1085	-0.1187	-0.0526	-0.0776	-0.0698
θ = 3	-0.0817	-0.0451	-0.0761	-0.0525	-0.0600	-0.0661
5 items						
θ = -3	0.0982	0.7345	0.4388	0.0985	1.0846	1.1225
θ = -2	-0.0167	0.1385	0.0278	-0.0311	0.0195	0.1310
θ = -1	-0.0009	-0.0115	0.0085	-0.0375	-0.0166	-0.0067
θ = 0	0.0359	0.0405	0.0603	0.0092	0.0698	0.0491
θ = 1	-0.0422	-0.0576	-0.0686	-0.0334	-0.0509	-0.0667
θ = 2	-0.0559	-0.1477	-0.1419	-0.0377	-0.1241	-0.1328
θ = 3	-0.0334	-0.1255	-0.0702	-0.0360	-0.0710	-0.0519

Effect sizes (η^2). Table 21 shows the η^2 values from the factorial ANOVA of the LBIAS. Like RMSE, these values are related to the distance between true and estimated θ, but take into account the direction of that difference.

One main effect, θ level, and two interactions: SM × θ and TL× θ had sizeable η^2. θ level explained 69% of the variance, the SM × θ interaction explained about 14% of the variance and the TL × θ interaction only 6 percent.

Table 21

η^2 from the ANOVA on LBIAS

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.010	2	0.024
Testlet Length (TL)	0.010	2	0.014
Theta Level (θ)	0.390	6	0.690
Pool Size (PS)	0.000	1	0.001
2-way Interactions			
SM × TL	0.000	4	0.002
SM × θ	0.080	12	0.138
SM × PS	0.000	2	0.001
TL × θ	0.030	12	0.059
TL × PS	0.000	2	0.002
θ × PS	0.000	6	0.001
Residual	0.040	76	0.068
Total	0.560	125	

The means and standard deviations of bias for each level of the θ main effect with $\eta^2 > .05$ are shown in Table 22.

Table 22

Means and SDs of Bias for the Main Effect of θ Level

Variable	Mean	SD
θ Level		
θ = -3	0.39883	0.32442
θ = -2	-0.01185	0.06406
θ = -1	-0.02534	0.02098
θ = 0	0.03870	0.02111
θ = 1	-0.05060	0.01653
θ = 2	-0.09294	0.03291
θ = 3	-0.05448	0.02493

Positive bias occurred only at θ = -3 and θ = 0. The remainder of the θ levels had negative bias. The smallest means, or least bias, were at θ levels of 0, -1 and -2 whereas the largest or most bias was at θ = -3, which is furthest from the mastery cutoff of 1.

PES analyses. For LBIAS, like LMSE, the most sizable η^2 was for the main effect θ level. Table 23 lists the *n*s, SS, and PES of LBIAS for the seven θ levels.

Like LMSE, θ = -3 was the most influential level (PES = 78%) and had the most bias. θ = 2 was the second most influential level (PES = 11%), but had much less bias but was still the second highest in magnitude. The level with the least amount of bias, θ = -2, also had the least influence (PES < 1%). As a result, there was no most favored condition, but θ = -3 and θ = 2 were different from the other levels.

Table 23

Means, *n*s, SS, and PES of LBIAS for the Seven Levels of True θ

Variable	Mean	*n*	SS	PES
True θ				
-3	0.1357	18	0.3007	0.7776
-2	-0.0060	18	0.0028	0.0072
-1	-0.0112	18	0.0056	0.0146
0	0.0164	18	0.0018	0.0046
1	-0.0226	18	0.0152	0.0393
2	-0.0426	18	0.0434	0.1121
3	-0.0245	18	0.0172	0.0445
Main Effect	0.0064	126	0.3867	

The means and standard deviations of LBIAS for the SM × θ interaction are shown in Table 24.

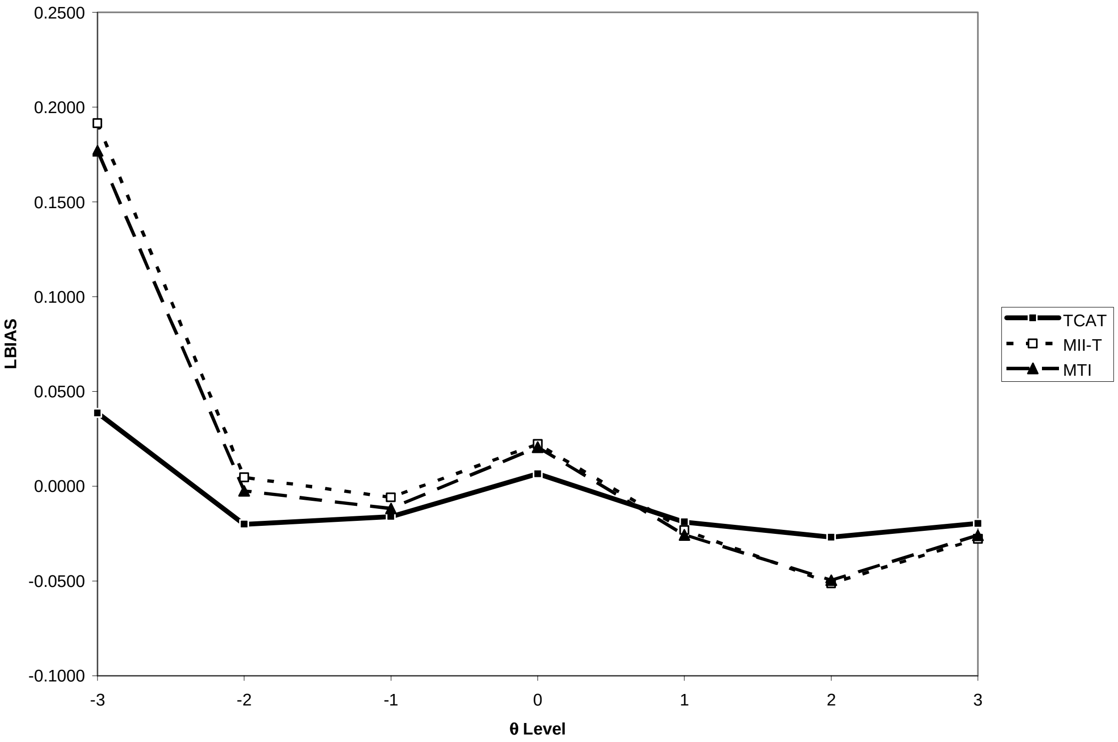
Table 24

Means and SDs of LBIAS for the θ × SM Conditions

	TCAT		MII-T		MTI	
θ Level	Mean	SD	Mean	SD	Mean	SD
-3	0.0387	0.0155	0.1915	0.0803	0.1770	0.0801
-2	-0.0201	0.0152	0.0046	0.0302	-0.0026	0.0315
-1	-0.0161	0.0089	-0.0059	0.0076	-0.0117	0.0097
0	0.0065	0.0056	0.0222	0.0050	0.0205	0.0058
1	-0.0189	0.0087	-0.0231	0.0051	-0.0257	0.0079
2	-0.0268	0.0075	-0.0514	0.0117	-0.0496	0.0142
3	-0.0197	0.0096	-0.0279	0.0165	-0.0259	0.0072

Figure 14 shows the 2-way interaction between selection method and θ level for LBIAS.

Figure 14
2-Way Interaction of LBIAS between SM and θ



TCAT had the least amount of LBIAS especially at θ = -3. As all three selection procedures approach θ = 1, the LBIAS became negative in value but TCAT still had the least amount of LBIAS except for values θ = -1 and θ = -2. MII-T and MTI had similar amounts across all θ levels. MTI with θ = -3 had a PES of 95% accounting for almost all the influence of this interaction with quite a large amount of LBIAS (.1770). MII-T has an even higher LBIAS (.1915) but a PES larger than 1. At every level of θ, the selection method with the smallest absolute mean value had a negative PES, leaving no favorable condition.

The means and standard deviations of LBIAS for the 2-way interaction of TL × θ are shown in Table 25. LBIAS was largest for θ = -3 regardless of testlet length. LBIAS was positive for θ = 0 and θ = -3. For negative values, LBIAS was largest for θ level = 2 regardless of testlet length.

Table 25

Means and SDs of LBIAS for θ × TL Conditions

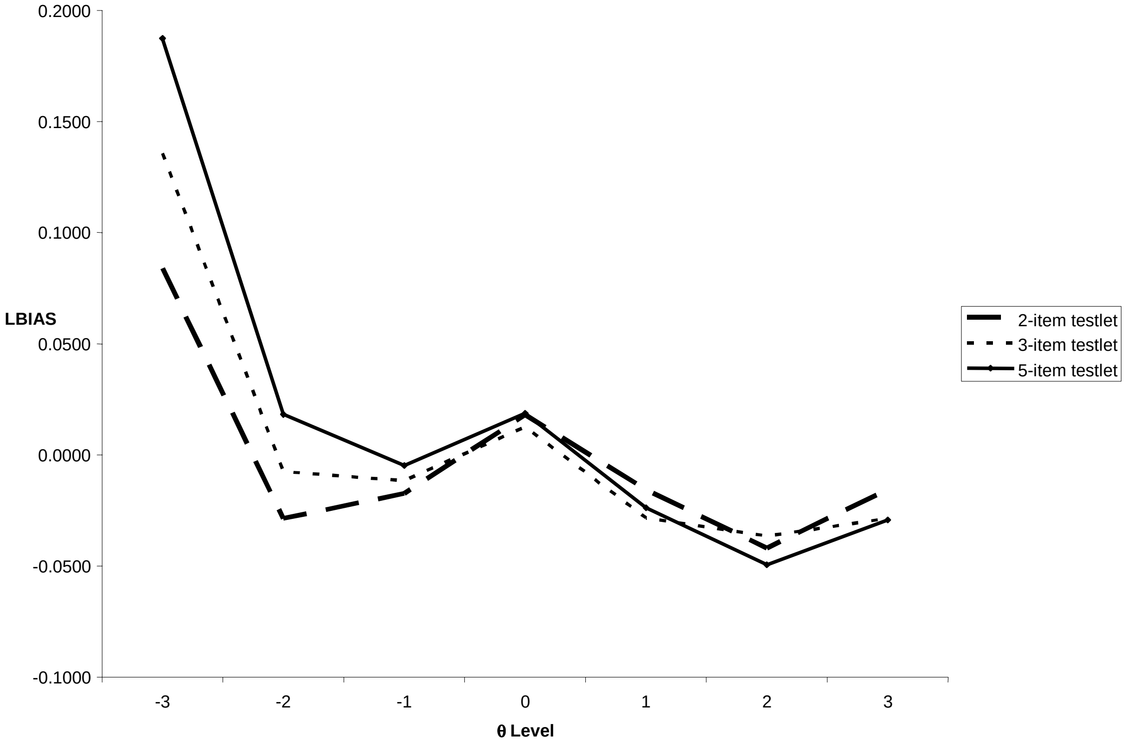
	TL = 2		TL = 3		TL = 5	
θ Level	Mean	SD	Mean	SD	Mean	SD
-3	0.0840	0.0516	0.1357	0.0666	0.1874	0.1291
-2	-0.0287	0.0106	-0.0075	0.0136	0.0182	0.0300
-1	-0.0173	0.0050	-0.0116	0.0113	-0.0048	0.0069
0	0.0180	0.0103	0.0125	0.0073	0.0187	0.0089
1	-0.0157	0.0065	-0.0284	0.0031	-0.0238	0.0063
2	-0.0419	0.0076	-0.0365	0.0131	-0.0495	0.0227
3	-0.0156	0.0045	-0.0286	0.0065	-0.0293	0.0160

Figure 15 shows the interaction between TL and θ level. LBIAS was quite high for all three testlet lengths at θ = -3 and dropped considerabally for θ between -2 and 1. Between θ = 1 and 3, there are only slight fluctuations in magnitude. TL = 2 exhibited the least bias for the two extreme values of θ = -3 and 3 as well as the cutoff of θ = 1, whereas TL = 5 showed the least amount of LBIAS only at θ = -1. LBIAS was lowest for the remainder of the θ levels for TL = 3. Only 7 of the 21 interaction conditions had

PES values between 0 and 1. Of those 7, the two largest in magnitude were: TL = 2 at θ = -2 (PES = 18%) and TL = 5 at θ = 2 (PES = 11%). Since neither of these had the least amount of LBIAS at that θ level, there was no most favorable condition.

Figure 15

2-Way Interaction of LBIAS between TL and θ Level



Summary

In terms of correlations, recovery of θ was best with TCAT, the selection method with the least amount of constraints. Likewise, shorter testlets (a condition most similar to TCAT, where branching occurs more frequently), appeared to recover θ better than longer testlets. In the PES analyses for correlations, TCAT and 2-item testlets were most different from the other levels both for the main effects and the interactions. It was also the most favored condition, with the highest mean correlation and the highest PES value.

For RMSE and bias, θ level was the main effect with the most influence. For RMSE, θ = -3, 0, and 2 had the largest values of PES indicating the largest differences from the other values. Bias had similar results with θ = -3 and 2 having large PES values. Neither dependent variable had a most favorable condition. The two bias interactions, SM ×θ and TL × θ were more difficult to interpret, since there were several interactions with PES values outside the 0-1 range defined for proportions in this study. Conditions that appeared to have the most influence were: MTI and θ = -3, 2-item testlet at θ = -2 and 5-item testlet at θ = 2.

QUALITY OF CLASSIFICATION DECISIONS

Proportions

Tables 26, 30, 34, 38, and 43 show the percent of no-decision, correctly identified masters, correctly identified nonmasters, incorrectly identified masters and incorrectly identified nonmasters, respectively for each of the conditions.

Table 26 shows that there was a proportion of no-decision greater than zero only at the θ = 1 level for all sets of conditions. This indicated that only at the mastery cut-off was there an issue of whether or not a decision could be made.

Table 26

Proportion of No Mastery Decision for TCAT, MII-T and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = -3	0.000	0.000	0.000	0.000	0.000	0.000
θ = -2	0.000	0.000	0.000	0.000	0.000	0.000
θ = -1	0.000	0.000	0.000	0.000	0.000	0.000
θ = 0	0.000	0.000	0.000	0.000	0.000	0.000
θ = 1	0.552	0.573	0.579	0.457	0.561	0.529
θ = 2	0.000	0.000	0.000	0.000	0.000	0.000
θ = 3	0.000	0.000	0.000	0.000	0.000	0.000
3 items						
θ = -3	0.000	0.000	0.000	0.000	0.000	0.000
θ = -2	0.000	0.000	0.000	0.000	0.000	0.000
θ = -1	0.000	0.000	0.000	0.000	0.000	0.000
θ = 0	0.000	0.000	0.000	0.000	0.000	0.000
θ = 1	0.501	0.559	0.552	0.505	0.523	0.540
θ = 2	0.000	0.000	0.000	0.000	0.000	0.000
θ = 3	0.000	0.000	0.000	0.000	0.000	0.000
5 items						
θ = -3	0.000	0.000	0.000	0.000	0.000	0.000
θ = -2	0.000	0.000	0.000	0.000	0.000	0.000
θ = -1	0.000	0.000	0.000	0.000	0.000	0.000
θ = 0	0.000	0.000	0.000	0.000	0.000	0.000
θ = 1	0.542	0.578	0.566	0.554	0.563	0.543
θ = 2	0.000	0.000	0.000	0.000	0.000	0.000
θ = 3	0.000	0.000	0.000	0.000	0.000	0.000

These proportions ranged from .457 to .579. In every case except one (5 item-testlet and 900-item pool), TCAT had the smallest proportions, indicating that decisions were made more often than the other selection methods. For the majority of cases, MII-T had a slightly higher proportion of no decision than did MTI. It appeared that MTI more closely resembled TCAT.

Effect sizes (η^2). Table 27 shows the η^2 values from the factorial ANOVA of the transformed proportion of no decision (NODEC). θ level accounted for almost 100% of the variance (η^2 = 99.8%). There were no other sizable main effects or interactions for NODEC.

Table 27

η^2 from the ANOVA on NODEC

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.00	2	0.000
Testlet Length (TL)	0.00	2	0.000
Theta Level (θ)	42.38	6	0.998
Pool Size (PS)	0.00	1	0.000
2-way Interactions			
SM × TL	0.00	4	0.000
SM × θ	0.02	12	0.000
SM × PS	0.00	2	0.000
TL × θ	0.01	12	0.000
TL × PS	0.00	2	0.000
θ × PS	0.01	6	0.000
Residual	0.02	76	0.000
Total	42.45	125	

Table 28 lists the means and standard deviations for the proportion of no mastery decision for the seven levels of θ. For θ = 1, the mastery cutoff, a little more than half the time (54%), a decision could not be made. For all other levels of θ, a decision was made prior to administering the entire item bank.

Table 28

Means and SDs of Proportion of No Mastery Decision for the Main Effect of θ Level

Variable	Mean	SD
θ Level	0.0000	0.0000
θ = -3	0.0000	0.0000
θ = -2	0.0000	0.0000
θ = -1	0.0000	0.0000
θ = 0	0.0000	0.0000
θ = 1	0.5432	0.0311
θ = 2	0.0000	0.0000
θ = 3	0.0000	0.0000

PES analyses. Table 29 lists the *n*s, SS, and PES of NODEC for the seven θ levels. θ = 1 was the most influential level with a PES of 86%. It was also the only level where a mastery decision could not be made, thereby making it the least favorable condition.

Table 29

Means, *n*s, SS, and PES of NODEC for the Seven Levels of True θ

Variable	Mean	*n*	SS	PES
True θ				
-3	0.0000	18	1.0090	0.0238
-2	0.0000	18	1.0090	0.0238
-1	0.0000	18	1.0090	0.0238
0	0.0000	18	1.0090	0.0238
1	1.6573	18	36.3251	0.8571
2	0.0000	18	1.0090	0.0238
3	0.0000	18	1.0090	0.0238
Main Effect	0.2368	126	42.379	

Table 30 lists the proportion of correctly identified masters, where “master” is defined as having a true θ ≥ 1. This proportion included simulees for whom their true θ level was ≥ 1 and a mastery decision was made, out of all simulees with a true θ ≥ 1. Since only θ levels ≥ 1 are relevant, only those values are reported. In Table 30, there was a trend of low proportions at the θ = 1 level (range of correctly identified masters was from .245 to .335), very high proportions at the θ = 2 level (range of correctly identified masters was between .989 to 1.000), and almost certainty at the θ = 3 level (all but one proportion was 1.000). For the most part, TCAT had higher proportions than MII-T or MTI. In general, MII-T tended to have proportions slightly lower than MTI.

Table 30

Proportion of Correctly Identified Masters (θ ≥ 1) by TCAT, MII-T, and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = 1	0.290	0.257	0.248	0.322	0.263	0.288
θ = 2	0.999	0.993	0.996	0.989	0.996	0.998
θ = 3	1.000	1.000	1.000	1.000	1.000	1.000
3 items						
θ = 1	0.335	0.270	0.281	0.286	0.289	0.296
θ = 2	0.999	0.997	0.996	0.989	0.997	0.998
θ = 3	1.000	1.000	1.000	0.998	0.999	1.000
5 items						
θ = 1	0.284	0.245	0.250	0.286	0.263	0.267
θ = 2	0.998	0.998	0.999	0.999	1.000	0.998
θ = 3	1.000	1.000	0.999	1.000	1.000	1.000

Effect sizes (η^2). Table 31 shows the η^2 values from the factorial ANOVA of the transformed proportion of correctly identified masters (TrueM).

Table 31

η^2 from the ANOVA on TrueM

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.00	2	0.000
Testlet Length (TL)	0.00	2	0.000
Theta Level (θ)	228.33	6	1.000
Pool Size (PS)	0.00	1	0.000
2-way Interactions			
SM × TL	0.00	4	0.000
SM × θ	0.02	12	0.000
SM × PS	0.01	2	0.000
TL × θ	0.03	12	0.000
TL × PS	0.00	2	0.000
θ × PS	0.00	6	0.000
Residual	0.04	76	0.000
Total	228.44	125	

Only the main effect, θ level, had a sizable η^2 and accounted for 100% of the variance. Table 32 lists the means and standard deviations of the proportion of correctly identified masters for the three mastery levels of θ. Only levels of $\theta \geq 1$ were relevant and are included in the table. At the mastery cutoff, $\theta = 1$, the proportion was quite low (.28). As θ increased in magnitude, the proportion of correctly identified masters increased significantly (> .99).

Table 32

Means and SDs of Proportion of Correctly Identified Masters for the Main Effect of θ Level

Variable	Mean	SD
θ Level		
θ = 1	0.2789	0.0241
θ = 2	0.9966	0.0032
θ = 3	0.9998	0.0005

PES analyses. Table 33 lists the *n*s, SS, and PES of TrueM for the three mastery levels of θ.

Table 33

Means, *n*s, SS, and PES of TrueM for the Three Mastery Levels of True θ

Variable	Mean	*n*	SS	PES
True θ				
θ = 1	1.1121	18	0.0941	0.0006
θ = 2	3.0368	18	71.7826	0.4770
θ = 3	3.1296	18	78.6122	0.5224
Main Effect	1.0398	126	150.4889	

The PES values for θ = 2 and θ = 3 were quite similar in magnitude (48% and 52% respectively). In addition to a slightly larger PES value, θ = 3 had a slightly higher mean making it a more favorable condition. Since the differences in magnitude were relatively small, θ = 2 was a very close second. The least favorable condition was θ = 1, at the cutoff, with the lowest mean and very small PES.

Table 34 lists the proportion of correctly identified Non-Masters where nonmastery is defined as having a true θ < 1.

Table 34

Proportion of Correctly Identified Non-Masters (θ < 1) by TCAT, MII-T and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = -3	0.991	0.997	1.000	0.997	1.000	0.999
θ = -2	0.994	1.000	0.998	0.996	0.999	0.998
θ = -1	0.990	0.997	0.995	0.989	0.995	0.996
θ = 0	0.954	0.991	0.975	0.946	0.970	0.980
3 items						
θ = -3	0.995	0.999	1.000	0.999	0.999	1.000
θ = -2	0.995	1.000	1.000	0.991	0.998	0.997
θ = -1	0.993	0.998	0.999	0.995	0.998	0.995
θ = 0	0.955	0.980	0.975	0.978	0.984	0.980
5 items						
θ = -3	0.997	1.000	1.000	0.994	1.000	1.000
θ = -2	1.000	1.000	0.999	0.997	0.999	0.999
θ = -1	0.996	1.000	0.998	0.991	1.000	0.996
θ = 0	0.957	0.979	0.990	0.969	0.988	0.985

This is the proportion of simulees for whom their true θ level was < 1 and a nonmastery decision was made out of all simulees with a true θ < 1. Since only θ levels < 1 are relevant, only those values were reported in this table.

For all θ levels below the mastery cutoff, the proportions were quite high, ranging from .946 to 1.000. While the range of proportions was quite narrow, TCAT tended to have slightly lower proportions for correctly identifying non-masters over the two testlet selection methods.

Effect sizes (*η^2*). Table 35 shows the η^2 values from the factorial ANOVA of the transformed proportion of correctly identified masters (TrueNM).

Table 35

η^2 from the ANOVA on TrueNM

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.08	2	0.000
Testlet Length (TL)	0.02	2	0.000
Theta Level (θ)	278.41	6	0.999
Pool Size (PS)	0.00	1	0.000
2-way Interactions			
SM × TL	0.00	4	0.000
SM × θ	0.08	12	0.000
SM × PS	0.00	2	0.000
TL × θ	0.02	12	0.000
TL × PS	0.00	2	0.000
θ × PS	0.01	6	0.000
Residual	0.08	76	0.000
Total	278.71	125	

Only the main effects, θ level, had sizable η^2 and accounted for almost 100% of the variance.

Table 36 lists the means and standard deviations of the main effect, θ level, for correctly identified nonmasters. The only relevant levels of θ were those < 1. The negative levels performed equally well but there was a slightly lower mean and larger standard deviation at θ = 0.

Table 36

Means and SDs of Proportion of Correctly Identified NonMasters for the Main Effect of θ Level

Variable	Mean	SD
θ Level		
θ = -3	0.9982	0.0026
θ = -2	0.9978	0.0025
θ = -1	0.9956	0.0032
θ = 0	0.9742	0.0133

PES analyses. Table 37 lists the *n*s, SS, and PES of TrueM for the four nonmastery levels of θ. The three negative values of θ had similar values for PES, indicating that they contributed to the main effect θ to a similar degree. θ = 0 was moderately less influential. In terms of means, as θ became more distant from the cutoff, the mean increased in magnitude, indicating more accurate decisions. Since θ = -3 had the highest mean and the largest PES, it was the most favorable condition.

Table 37

Means, *n*s, SS, and PES of TrueNM for θ Level

Variable	Mean	*n*	SS	PES
θ Level				
θ = -3	3.0822	18	33.6893	0.2814
θ = -2	3.0664	18	32.9129	0.2749
θ = -1	3.0219	18	30.7814	0.2571
θ = 0	2.8286	18	22.3575	0.1867
Main Effect	1.7142	126	119.7411	

Table 38 lists the proportion of incorrectly identified masters which is one type of decision error. This means that true θ was < 1 but the decision made was "master." Only cells with θ < 1 are relevant and therefore are included in the table. Proportions of misidentified masters were quite small ranging from 0.000 to 0.0540. Proportions were smaller at θ levels further from 1 and became larger as θ approached the mastery cutoff. TCAT tended to have higher proportions than MII-T and MTI, indicating that it had a higher error rate. There were no 0 error rates for TCAT as there were for MII-T and MTI. MII-T tended to have a lower error rate than MTI. Longer testlets (particularly TL = 5) tended to have lower error rates than shorter testlets.

Table 38

Proportion of Incorrectly Identified Masters by TCAT, MII-T and MTI
by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = -3	0.009	0.003	0.000	0.003	0.000	0.001
θ = -2	0.006	0.000	0.002	0.004	0.001	0.002
θ = -1	0.010	0.003	0.005	0.011	0.005	0.004
θ = 0	0.046	0.009	0.025	0.054	0.030	0.020
3 items						
θ = -3	0.005	0.001	0.000	0.001	0.001	0.000
θ = -2	0.005	0.000	0.000	0.009	0.002	0.003
θ = -1	0.007	0.002	0.001	0.005	0.002	0.005
θ = 0	0.045	0.020	0.025	0.022	0.016	0.020
5 items						
θ = -3	0.003	0.000	0.000	0.006	0.000	0.000
θ = -2	0.000	0.000	0.001	0.003	0.001	0.001
θ = -1	0.004	0.000	0.002	0.009	0.000	0.004
θ = 0	0.043	0.021	0.010	0.031	0.012	0.015

Effect sizes (η^2). Table 39 shows the η^2 values from the factorial ANOVA of the transformed proportion of incorrectly identified masters (FalseM). There were two sizable η^2, SM and θ level. θ level accounted for 82% of the variance whereas SM accounted for only 5%.

Table 39

η^2 from the ANOVA on FalseM

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.08	2	0.051
Testlet Length (TL)	0.02	2	0.009
Theta Level (θ)	1.36	6	0.819
Pool Size (PS)	0.00	1	0.001
2-way Interactions			
SM × TL	0.00	4	0.001
SM × θ	0.08	12	0.049
SM × PS	0.00	2	0.001
TL × θ	0.02	12	0.009
TL × PS	0.00	2	0.000
θ × PS	0.01	6	0.009
Residual	0.08	76	0.048
Total	1.66	125	

Table 40 lists the means and standard deviations of the two main effects, SM and θ level, for incorrectly identified masters. The only relevant levels of θ were those < 1.

Table 40

Means and SDs of Proportion of Incorrectly Identified Masters for the Main Effects

SM and θ Level

Variable	Mean	SD
SM		
TCAT	0.0081	0.0142
MII-T	0.0031	0.0068
MTI	0.0035	0.0069
θ Level		
θ = -3	0.0018	0.0026
θ = -2	0.0022	0.0025
θ = -1	0.0044	0.0032
θ = 0	0.0258	0.0133

Since this was a type of error, a smaller mean was more desirable. MII-T and MTI, the two testlet methods, did better at avoiding this type of error than did TCAT. With regard to θ level, as θ moved away from the cutoff, there was a reduction in this type of error.

PES analyses. Table 41 lists the *n*s, SS, and PES of FalseM for the three selection methods. TCAT was the most influential selection method with a PES of 66%. MII-T was the next most influential selection method with MTI having the least amount of influence on effect size. MII-T performed the best, since its mean for this type of error was lowest. Because it had the lowest mean and the second highest PES it would be considered the more favored condition.

Table 41

Means, *n*s, SS, and PES of FalseM for SM

Variable	Mean	*n*	SS	PES
SM				
TCAT	0.1173	42	0.0552	0.6553
MII-T	0.0588	42	0.0208	0.2472
MTI	0.0670	42	0.0082	0.0976
Main Effect	0.0810	126	0.0843	

Table 42 lists the *n*s, SS, and PES of FalseM for the four relevant θ levels.

Table 42

Means, *n*s, SS, and PES of FalseM for θ Level

Variable	Mean	*n*	SS	PES
θ Level				
θ = -3	0.0594	18	0.0085	0.0084
θ = -2	0.0752	18	0.0006	0.0006
θ = -1	0.1197	18	0.0270	0.0268
θ = 0	0.3129	18	0.9681	0.9641
Main Effect	0.0810	126	1.0041	

θ = 0 was quite influential with a PES value of 96%. Unfortunately, it was also the worst performing level, since it had the highest mean, indicating the highest rate of this type of error. θ = -3 was the best performing level but had almost no influence (< 1%) on effect size.

Table 43 lists the proportions for the second type of error decision: incorrectly identified nonmasters. This means that true θ was ≥ 1 but the decision made was "nonmaster." Only cells with θ ≥ 1 are relevant and therefore included in the table.

Table 43

Proportion of Incorrectly Identified Non-Masters by TCAT, MII-T and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = 1	0.158	0.170	0.173	0.221	0.176	0.183
θ = 2	0.001	0.007	0.004	0.011	0.004	0.002
θ = 3	0.000	0.000	0.000	0.000	0.000	0.000
3 items						
θ = 1	0.164	0.171	0.167	0.209	0.188	0.164
θ = 2	0.001	0.003	0.004	0.011	0.003	0.002
θ = 3	0.000	0.000	0.000	0.002	0.001	0.000
5 items						
θ = 1	0.174	0.177	0.184	0.160	0.174	0.190
θ = 2	0.002	0.002	0.001	0.001	0.000	0.002
θ = 3	0.000	0.000	0.001	0.000	0.000	0.000

Proportions were relatively high at the θ = 1, level which was the mastery cutoff, ranging from 0.160 to 0.209. At θ = 2, proportions were quite small except for TL = 3 for TCAT with PS = 900. At θ = 3, the proportions approached zero indicating very little occurrence of this type of error. Across all three levels of θ, TCAT did slightly better than the other selection methods for PS = 600 and for PS = 900 with TL = 5.

Effect sizes (η^2). Table 44 shows the η^2 values from the factorial ANOVA of the transformed proportion of incorrectly identified nonmasters (FalseNM).

Table 44

η^2 from the ANOVA on FalseNM

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.00	2	0.000
Testlet Length (TL)	0.00	2	0.000
Theta Level (θ)	11.33	6	0.992
Pool Size (PS)	0.00	1	0.000
2-way Interactions			
SM × TL	0.00	4	0.000
SM × θ	0.00	12	0.000
SM × PS	0.01	2	0.001
TL × θ	0.01	12	0.001
TL × PS	0.00	2	0.000
θ × PS	0.00	6	0.000
Residual	0.05	76	0.004
Total	11.42	125	

The main effect, θ level, was the only sizable η^2 and it accounted for 99% of the variance.

Table 45 lists the means and standard deviations of the main effect, θ level, for incorrectly identified nonmasters. The only relevant levels of θ were those ≥ 1. This type of error was relatively high at the cutoff, θ = 1, and decreased in magnitude as θ increased in distance from that cutoff.

Table 45

Means and SDs of Proportion of Incorrectly Identified NonMasters

for the Main Effect θ Level

Variable	Mean	SD
θ Level		
θ = 1	0.1779	0.0163
θ = 2	0.0034	0.0032
θ = 3	0.0002	0.0005

PES analyses. Table 46 lists the *n*s, SS, and PES of FalseNM for the relevant levels of θ.

Table 46

Means, *n*s, SS, and PES of FalseNM for θ Level

Variable	Mean	*n*	SS	PES
θ Level				
θ = 1	0.8703	18	9.5722	0.9673
θ = 2	0.1048	18	0.0236	0.0024
θ = 3	0.0120	18	0.2996	0.0303
Main Effect	0.1410	126	9.8954	

The mastery cutoff, θ =1, had the largest PES value (97%) and influenced effect size the most. It was also the worst performing level, since it had the highest mean (.87). The best performing level (mean of .14), θ = 3 had very little influence, with a PES value of 3%. Therefore, there was no favored condition and θ = 1 was the least favored condition.

Decision Error

Decision error is similar to RMSE except instead of quantifying the distance between θ and $\hat{\theta}$, the distance between $\hat{\theta}$ and the criterion cutoff θ_c was computed. Unlike RMSE, which is calculated for all conditions, decision error was calculated only for incorrect decisions. Like RMSE, the distribution of decision errors was positively skewed and had to be transformed. Because there were conditions in which no errors were made, a transformation using a function other than log or ln was needed. In this case, the square root of ADE ($\sqrt{ADE}$) was used since it produced the least amount of skewness without creating any missing data.

Table 47 lists the ADE for each of the 126 conditions.

Table 47

ADE for TCAT, MII-T and MTI by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
$\theta = -3$	3.6246	4.8994	0.0000	0.9257	0.0000	0.8908
$\theta = -2$	0.7309	0.0000	0.6406	0.9592	0.7760	1.0991
$\theta = -1$	1.2526	0.8676	0.7525	1.3694	0.9207	1.8639
$\theta = 0$	0.5240	0.7154	0.6723	0.5737	0.6029	0.5091
$\theta = 1$	0.6345	0.6388	0.6469	0.5509	0.6791	0.6484
$\theta = 2$	0.2643	0.3964	0.4790	0.6096	0.2163	0.4851
$\theta = 3$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3 items						
$\theta = -3$	4.6645	1.1205	0.0000	0.5536	0.1298	0.0000
$\theta = -2$	2.8541	0.0000	0.0000	0.9885	0.4350	3.8154
$\theta = -1$	2.3320	0.6566	0.7069	0.3928	1.0425	1.1725
$\theta = 0$	0.5471	0.4234	0.6193	0.6060	0.5574	0.6818
$\theta = 1$	0.6015	0.6428	0.6238	0.5089	0.6618	0.6222
$\theta = 2$	0.3484	0.5013	0.2104	0.5676	0.1620	0.3769
$\theta = 3$	0.0000	0.0000	0.0000	0.4337	0.6386	0.0000
5 items						
$\theta = -3$	4.1851	0.0000	0.0000	2.5367	0.0000	0.0000
$\theta = -2$	0.0000	0.0000	0.4955	1.6226	0.0831	0.4537
$\theta = -1$	0.5437	0.0000	0.7814	2.2640	0.0000	0.5729
$\theta = 0$	0.6856	0.7561	0.5755	0.5625	0.5138	0.7053
$\theta = 1$	0.6184	0.6449	0.6409	0.5917	0.6696	0.6119
$\theta = 2$	0.3770	0.6581	0.1220	0.3285	0.0000	0.7158
$\theta = 3$	0.0000	0.0000	0.2213	0.0000	0.0000	0.0000

Magnitude of the ADE ranged from 0.000 to 4.8994.

Effect sizes (η^2*).* Table 48 shows the η^2 values from the factorial ANOVA on $\sqrt{ADE}$. Selection method and θ level were the two sizable main effects and SM × θ and θ × PS were the two sizable interactions. The two sizable main effects explained about 31% of the variance whereas the two interactions explained about 29% of the variance.

Table 48

η^2 from the ANOVA of $\sqrt{ADE}$

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	2.22	2	0.072
Testlet Length (TL)	0.63	2	0.021
Theta Level (θ)	7.47	6	0.243
Pool Size (PS)	0.01	1	0.000
2-way Interactions			
SM × TL	0.67	4	0.022
SM × θ	6.38	12	0.208
SM × PS	0.44	2	0.014
TL × θ	1.17	12	0.038
TL × PS	0.03	2	0.001
θ × PS	2.58	6	0.084
Residual	9.15	76	0.298
Total	30.74	125	

The means and standard deviations for the two main effects are listed in Table 49. TCAT had the highest $\sqrt{ADE}$ among the selection methods. Error rates were smallest for θ levels above the mastery cutoff and of moderate magnitude for θ = 0 and 1.

Table 49

Means and SDs of ADE for the

Main Effects of SM and θ Level

Variable	Mean	SD
Selection Method		
TCAT	0.9937	1.1223
MII-T	0.5002	0.7776
MTI	0.5575	0.6484
θ Level		
θ = -3	1.3073	1.8039
θ = -2	0.8308	1.0351
θ = -1	0.9718	0.6612
θ = 0	0.6017	0.0869
θ = 1	0.6243	0.0416
θ = 2	0.3788	0.1948
θ = 3	0.0719	0.1802

PES analyses. Table 50 lists the means, *n*s, SS, and PES of $\sqrt{ADE}$ for the three selection methods. TCAT accounted for the majority of the effect size (63%) followed by MII-T (31%) and lastly MTI (6%). TCAT was different from the two testlet methods. Ideal conditions would have minimal error and large PES. However, TCAT had the largest $\sqrt{ADE}$. MII-T was the selection method with the least amount of error, but it was not as influential in the SS as TCAT.

Table 50

Means, *n*s, SS, and PES of $\sqrt{ADE}$ for SM

Variable	Mean	*n*	SS	PES
Selection Method				
TCAT	0.8460	42	1.4034	0.6323
MII-T	0.5349	42	0.6912	0.3114
MTI	0.6086	42	0.1248	0.0562
Main Effect	0.6632	126	2.2194	

Table 51 lists the means, *n*s, SS, and PES of $\sqrt{ADE}$ for the seven levels of θ.

Table 51

Means, *n*s, SS, and PES of $\sqrt{ADE}$ for the Seven Levels of the True θ

Variable	Mean	*n*	SS	PES
True θ				
-3	0.7769	18	0.2329	0.0312
-2	0.7111	18	0.0414	0.0055
-1	0.9003	18	1.0120	0.1354
0	0.7738	18	0.2202	0.0295
1	0.7897	18	0.2882	0.0385
2	0.5833	18	0.1149	0.0154
3	0.1071	18	5.5653	0.7445
Main Effect	0.6632	126	7.4749	

θ = 3 was by far the influential level accounting for 74% of the main effect. The second most influential level was θ = -1 accounting for 13%. θ = 3 was the most favorable condition since it had the smallest $\sqrt{ADE}$ with the largest PES (74%). θ = -1 was the least favorable condition with the highest $\sqrt{ADE}$ and the second highest PES (13%).

Table 52 lists the means and standard deviations for the interaction of θ level by selection method for the $\sqrt{ADE}$.

Table 52

Means and SDs of $\sqrt{ADE}$ for the θ × SM Conditions

	TCAT		MII-T		MTI	
Theta Level (θ)	Mean	SD	Mean	SD	Mean	SD
-3	1.5680	0.5895	0.6054	0.8889	0.1573	0.3853
-2	0.9653	0.5594	0.3048	0.3839	0.8633	0.6377
-1	1.1142	0.3757	0.6204	0.4854	0.9662	0.2231
0	0.7629	0.0367	0.7676	0.0816	0.7908	0.0479
1	0.7639	0.0309	0.8100	0.0101	0.7952	0.0095
2	0.6376	0.1059	0.5027	0.2890	0.6094	0.1794
3	0.1098	0.2688	0.1332	0.3262	0.0784	0.1921

Figure 16 shows the 2-way interaction. MII-T had the smallest $\sqrt{ADE}$ at θ = -2, -1 and 2. For the two extremes (θ = -3 and 3), MTI had the smallest $\sqrt{ADE}$. For negative values of θ, $\sqrt{ADE}$ was quite different across all three selection methods. The pattern was quite similar, however, for θ levels of 0 or greater. At these θ levels, as θ increased in magnitude, $\sqrt{ADE}$ was reduced, indicating better performance.

Eighteen of the 21 interaction conditions had PES values between 0 and 1. Of those 18, the two conditions with the highest PES values were: TCAT at θ = -3 (PES = 23%) and TCAT at θ = -1 (PES = 11%). The $\sqrt{ADE}$ rates were quite high for these conditions, so there were no favorable conditions for this particular interaction.

Figure 16

2-Way Interaction of $\sqrt{ADE}$ between Selection Method (SM) and Theta Level (θ)

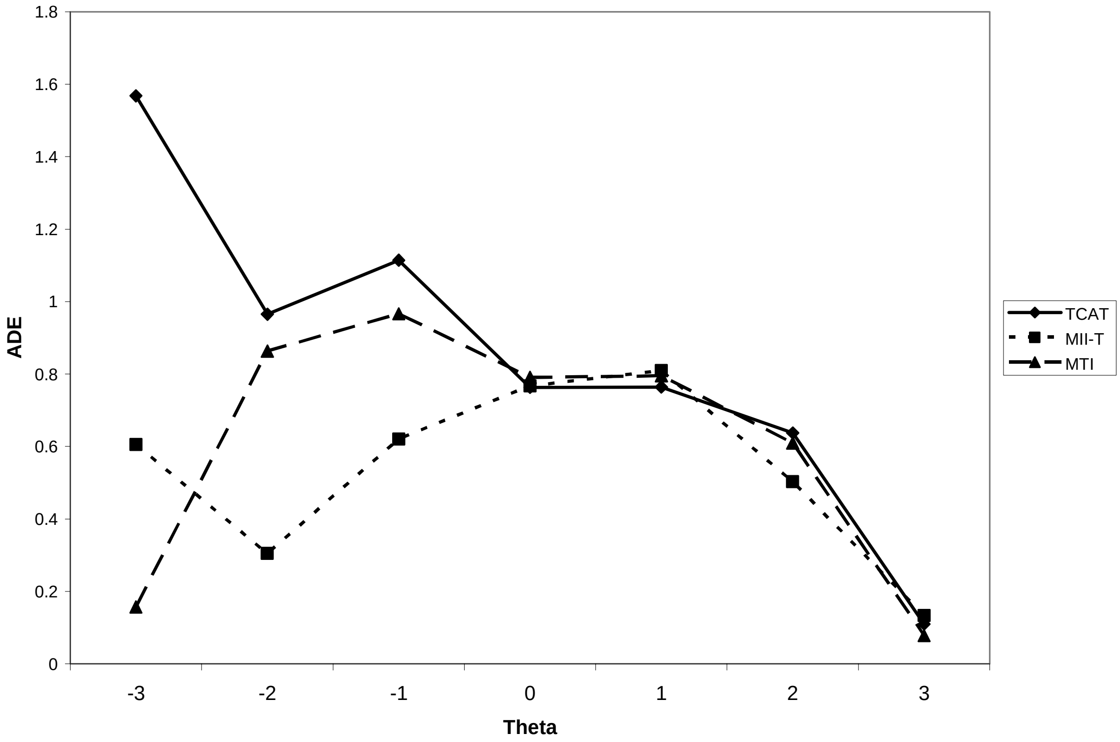


Table 53 lists the means and standard deviations for the interaction θ level by PS by for the $\sqrt{ADE}$.

The smallest $\sqrt{ADE}$ for PS = 600 were for θ levels -2, 2, and 3. The smallest $\sqrt{ADE}$ for PS = 900 were for θ levels -3, 2, and 3. Both pool sizes had similar magnitudes of average decision error for middle ranges of θ.

Table 53

Means and SDs of $\sqrt{ADE}$ for the θ level × PS Interaction

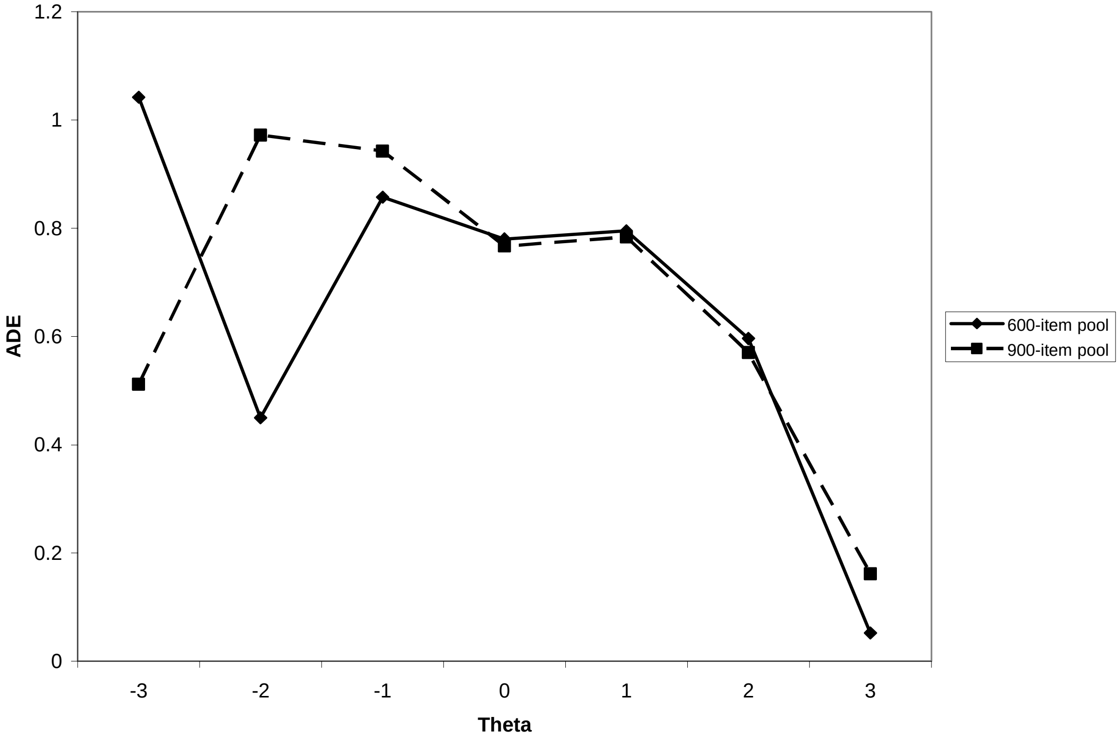
	PS = 600		PS = 900	
θ Level	Mean	SD	Mean	SD
-3	1.0424	1.0437	0.5114	0.5791
-2	0.4498	0.6021	0.9724	0.4641
-1	0.8575	0.3993	0.9430	0.4465
0	0.7803	0.0692	0.7672	0.0435
1	0.7953	0.0095	0.7841	0.0372
2	0.5964	0.1397	0.5702	0.2589
3	0.0523	0.1568	0.1620	0.3233

Figure 17 illustrates the interaction between PS and θ level for $\sqrt{ADE}$. For θ = -3, the $\sqrt{ADE}$ was the higher for the smaller pool size; from θ = 0, pool sizes performed similarly in terms of $\sqrt{ADE}$. Like the previous interaction, the magnitude of the $\sqrt{ADE}$ dropped rapidly as θ approached 3.

The two interaction conditions with the most influence were: PS = 600 items and θ = -3 (PES = 46%) and PS = 900 items and θ = -2 (PES = 33%). In both cases, $\sqrt{ADE}$ was quite large, rendering no most favored condition. The condition with the least amount of $\sqrt{ADE}$, pool size 600 and θ = 3, has a PES value of 22%.

Figure 17

2-Way Interaction of $\sqrt{ADE}$ between PS and θ Level



Summary.

For all six dependent variables related to quality of decisions (NODEC, TrueM, FalseM, TrueNM, FalseNM, and $\sqrt{ADE}$), θ level had the largest, and often the only sizable, η^2. Selection method had a sizable η^2 for FalseM and $\sqrt{ADE}$. Only $\sqrt{ADE}$ had any interactions with sizable η^2s.

Performance was worst at the cutoff, θ = 1, for the three proportion criteria that included it: NODEC, TrueM and FalseNM. Performance improved as θ moved away from the cutoff for all five proportion criteria. In terms of PES for θ level, it was largest at θ = 1 for NODEC and FalseNM; largest at θ = 3 for TrueM and $\sqrt{ADE}$; largest at θ = 0 for FalseM; and approximately equally distributed for TrueNM.

With regard to selection method, TCAT was the most influential but worst performing selection method for both FalseM and $\sqrt{ADE}$. In both cases, MII-T slightly outperformed MTI. MII-T had the least amount of $\sqrt{ADE}$ and some influence (29%), even though TCAT was most influential on effect size.

The pattern of $\sqrt{ADE}$ for θ level differed from the proportion criteria with values of θ ≥ 0 eliciting smaller $\sqrt{ADE}$ than negative values of θ. θ = 3 was the most favorable condition since it had the lowest $\sqrt{ADE}$ and the highest PES.

For the SM × θ interaction, MII-T had the lowest $\sqrt{ADE}$, except for the two extremes and the cutoff. All three selection methods had similar patterns for θ ≥ 0. Interaction conditions that were different from the others were TCAT with θ = -3 and TCAT with θ = -1. No interaction condition emerged as favorable. For the TL × θ interaction, both pool sizes exhibited a similar pattern for θ ≥ 0 and both performed best for θ = 2 and θ = 3. Conditions with high PES values indicating differences were PS = 600-item with θ = -3 and PS = 900 with θ = -2.

EFFICIENCY

ϕ Correlations

ϕ correlations are related to accuracy of classification decisions. Like the correlation between θ and $\hat{\theta}$, because true θ is a discrete point at each level with no variation with which to perform a correlation, there were only 18 conditions (3 methods × 2 item pools × 3 testlet lengths).

Table 54 shows the ϕ correlations of true vs. decision mastery status. The range of correlations was quite narrow (.9033 to .9388). This indicated that all conditions performed similarly in terms of classifications made relative to true status.

Table 54

ϕ Correlation of True and Adaptive Mastery/NonMastery Status by SM ,TL, and PS

	PS = 600			PS = 900		
TL	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items	0.9241	0.9370	0.9310	0.9023	0.9290	0.9310
3 items	0.9262	0.9354	0.9355	0.9162	0.9309	0.9366
5 items	0.9258	0.9342	0.9349	0.9308	0.9388	0.9309

Effect sizes (η^2). Like the Pearson product-moment correlation, ϕ was negatively skewed and was transformed using the function arcsin of the square root of ϕ (Equation 66). Table 55 shows the η^2 values from the factorial ANOVA on the transformed correlations.

Table 55

η^2 from the ANOVA on arcsin(sqrt) of ϕ

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.01	2	0.527
Testlet Length (TL)	0.00	2	0.111
Pool Size (PS)	0.00	1	0.058
2-way Interactions			
SM × TL	0.00	4	0.093
SM × PS	0.00	2	0.035
TL × PS	0.00	2	0.081
Residual	0.00	4	0.095
Total	0.02	17	

All three main effects (SM, TL, and PS), as well as two of the 2-way interactions (SM × TL and TL × PS), had sizable η^2. Selection method was by far the largest in magnitude (η^2 = .527) accounting for over half (53%) of the variance.

The means and standard deviations of arcsin(sqrt) of ϕ for each level of the three main effects with $\eta^2 > .05$ are shown in Table 56. MII-T, 5-item testlets, and the 600-item pool had slightly higher means, indicating more accurate classifications. TCAT, TL = 2, and PS = 900-items had slightly lower transformed correlations.

Table 56

Means and SDs of arcsin(sqrt) of ϕ for the

Main Effects of SM, TL, and PS

Variable	Mean	SD
SM		
TCAT	2.5724	0.0372
MII-T	2.6230	0.0149
MTI	2.6193	0.0105
TL		
2-items	2.5909	0.0447
3-items	2.6073	0.0302
5-items	2.6165	0.0176
PS		
600-items	2.6125	0.0194
900-items	2.5973	0.0419

PES analyses. Table 57 lists the means, *n*s, SS, and PES of arcsin(sqrt) of ϕ for the three selection methods. Like Fisher *r* to *z*, a desirable transformed ϕ has a level mean above the main effect mean and a relatively large PES. In Table 57, the results are mixed. TCAT had the highest PES accounting for 66% of the effect size, but it had the lowest mean. Both testlet methods had means above the main effect mean but even combining their PES values still accounted for only 33% of effect size. MII-T did seem slightly more favorable than MTI with a slightly higher mean and the second highest PES (21%).

Table 57

Means, *n*s, SS, and PES of arcsin(sqrt) of ϕ for SM

Variable	Mean	*n*	SS	PES
Selection Method				
TCAT	2.5724	6	0.0063	0.6638
MII-T	2.6230	6	0.0020	0.2057
MTI	2.6193	6	0.0012	0.1304
Main Effect	2.6049	18	0.0095	

Testlet length had the next largest η^2 among the main effects. Table 58 lists the means, *n*s, SS, and PES of arcsin(sqrt) of ϕ for the three testlet lengths.

Table 58

Means, *n*s, SS, and PES of arcsin(sqrt) of ϕ for TL

Variable	Mean	*n*	SS	PES
TL				
2-item	2.5909	6	0.0012	0.5818
3-item	2.6073	6	0.0000	0.0167
5-item	2.6165	6	0.0008	0.4015
Main Effect	2.6049	18	0.0020	

As testlet length increased, so did the mean of transformed ϕ. The pattern for PES did not quite follow in that manner. The 2-item testlet had the lowest mean and the highest PES (58%). The 5-item testlet, however, had the highest mean and the second highest PES accounting for 40% of the effect size, making the 5-item testlet the more favorable condition.

Table 59 lists the means, *n*s, SS, and PES of arcsin(sqrt) of ϕ for the two pool sizes.

Table 59

Means, *n*s, SS, and PES of arcsin(sqrt) of ϕ for PS

Variable	Mean	*n*	SS	PES
PS				
600 item	2.6125	9	0.0005	0.5000
900 item	2.5973	9	0.0005	0.5000
Main Effect	2.6049	18	0.0010	

For pool size the PES was equally distributed. The 600-item pool mean was slightly above the main effect mean, indicating a more favorable condition although there was very little difference between the means.

Table 60 lists the means and standard deviations for the first sizable interaction, selection method by testlet length for arcsin(sqrt) of ϕ.

Table 60

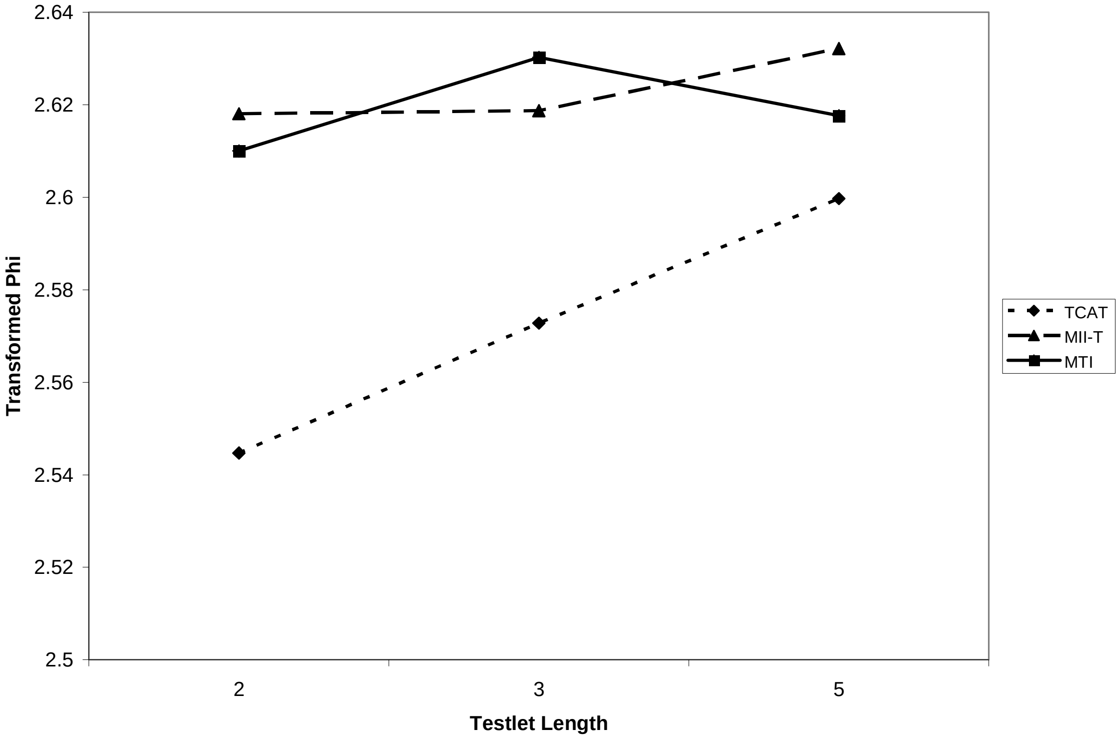
Means and SDs of arcsin(sqrt) of ϕ for the Interaction SM × TL

	TL = 2		TL = 3		TL = 5	
SM	Mean	SD	Mean	SD	Mean	SD
TCAT	2.5447	0.0549	2.5728	0.0260	2.5997	0.0137
MII-T	2.6181	0.0227	2.6187	0.0127	2.6321	0.0135
MTI	2.6100	0.0001	2.6302	0.0033	2.6176	0.0112

Figure 18 illustrates the interaction between pool size and θ level for transformed ϕ.

Figure 18

2-Way Interaction of arcsin(sqrt) of ϕ between SM and TL



TCAT clearly performed poorest across all three testlet lengths. MII-T performed best for TL = 2 and TL = 5, but MTI performed best for TL = 3. For PES, only two conditions had values between 0 and 1: (1) MII-T and TL = 5, with a PES of 33%; and (2) MTI and TL = 3, with a PES of 51%. These two conditions were the most favorable since they had the highest means and higher PES values.

Table 61 lists the means and standard deviations for the second sizable interaction, PS × TL for arcsin(sqrt) of ϕ.

Table 61

Means and SDs of arcsin(sqrt) of ϕ for the Interaction PS × TL

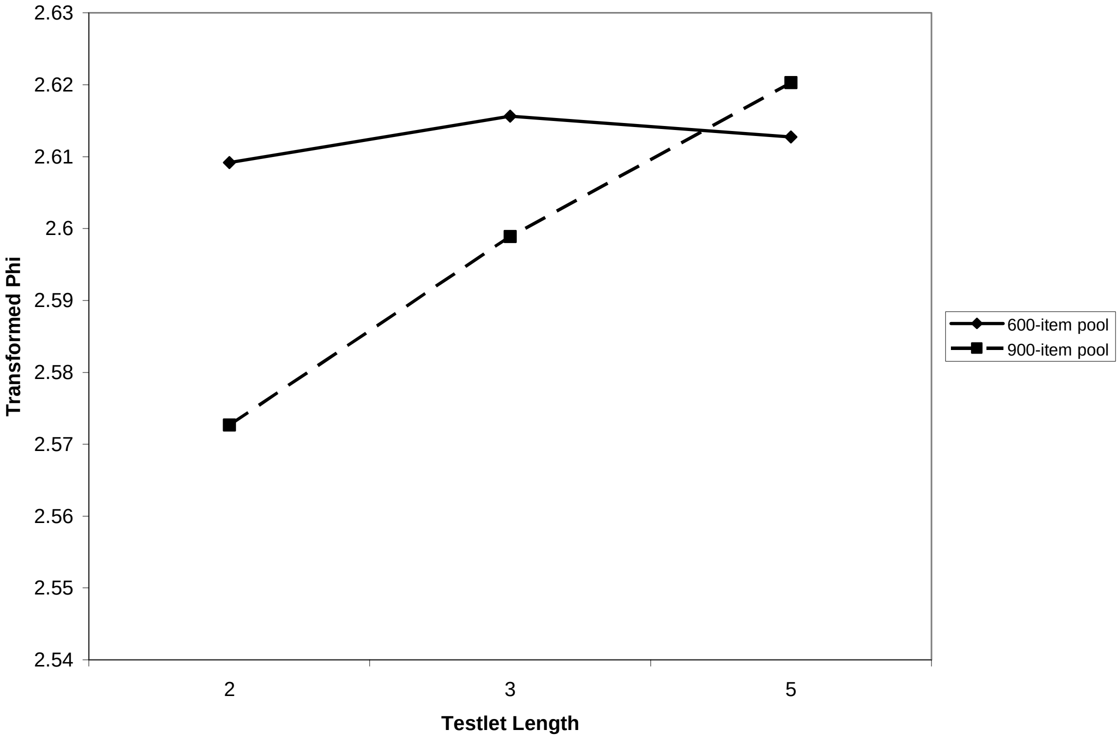
	TL = 2		TL = 3		TL = 5	
PS	Mean	SD	Mean	SD	Mean	SD
600 items	2.6092	0.0253	2.6156	0.0211	2.6127	0.0197
900 items	2.5727	0.0580	2.5989	0.0402	2.6203	0.0185

The means ranged 2.5725 to 2.6203, indicating very little variation among the means of the pool size × testlet length interaction.

Figure 19 illustrates the interaction between PS and θ level for transformed ϕ. The 600-item pool performed fairly steadily across all thee testlet lengths with a slight increase at TL = 3. The 900-item pool increased in performance as testlet length increased. The 600-item pool performed better than the 900-item-pool for TL = 2 and TL = 3, but slightly worse for TL = 5. In terms of PES, there were only two conditions with a value between 0 and 1. The highest PES was 11% for TL = 3 and 600-item pool . Since this was also the second highest mean, it was the most favorable condition for transformed ϕ.

Figure 19

2-Way Interaction of arcsin(sqrt) of ϕ between PS and TL



Test Length

Test length is an indication of how long it takes for a particular set of conditions to make a status classification**.**

Table 62 shows the means and standard deviations of mean test length (MTL). At the criterion cutoff of θ = 1, where it was most difficult to make classifications, MTL was quite long, ranging from 337.891 to 557.1350. For all other levels of θ, MTL ranged 20.0040 to 27.5450 items.

Table 62

MTL by SM by TL, θ Level, and PS

	PS = 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
$\theta = -3$	20.041	20.226	20.164	20.019	20.326	20.144
$\theta = -2$	20.004	20.010	20.006	20.011	20.006	20.014
$\theta = -1$	20.013	20.082	20.078	20.016	20.132	20.098
$\theta = 0$	22.796	25.196	25.098	22.060	24.570	24.418
$\theta = 1$	360.007	378.104	379.624	453.209	549.592	519.030
$\theta = 2$	22.303	24.068	23.678	21.591	24.810	24.312
$\theta = 3$	20.026	20.116	20.046	20.009	20.164	20.082
3 items						
$\theta = -3$	20.052	21.570	21.255	20.055	21.222	21.087
$\theta = -2$	20.009	21.048	21.009	20.015	21.003	21.006
$\theta = -1$	20.033	21.129	21.171	20.061	21.087	21.090
$\theta = 0$	23.089	26.682	26.106	22.925	26.781	26.280
$\theta = 1$	337.891	373.668	367.638	496.669	511.446	532.671
$\theta = 2$	22.146	25.893	24.888	21.911	26.505	24.891
$\theta = 3$	20.027	21.198	21.150	20.019	21.315	21.186
5 items						
$\theta = -3$	20.080	20.855	20.620	20.043	21.060	20.595
$\theta = -2$	20.035	20.170	20.055	20.005	20.435	20.105
$\theta = -1$	20.049	20.220	20.150	20.027	20.270	20.205
$\theta = 0$	22.686	26.480	25.710	23.021	27.545	25.415
$\theta = 1$	357.279	383.860	374.085	539.086	557.135	535.695
$\theta = 2$	22.320	26.490	25.945	22.021	26.980	26.140
$\theta = 3$	20.019	20.475	20.415	20.024	20.485	20.340

Effect sizes (η^2*).* Table 63 shows the η^2 values from the factorial ANOVA of MTL. Only θ level had a sizable η^2 and explained 96% of the variance.

Table 63

η^2 from the ANOVA of MTL

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	940.05	2	0.00
Testlet Length (TL)	241.64	2	0.00
Theta Level (θ)	2761645.64	6	0.96
Pool Size (PS)	15183.01	1	0.01
2-way Interactions			
SM × TL	142.10	4	0.00
SM × θ	3211.28	12	0.00
SM × PS	34.84	2	0.00
TL × θ	1340.31	12	0.00
TL × PS	159.39	2	0.00
θ × PS	90981.94	6	0.03
Residual	3776.92	76	0.00
Total	2877657.14	125	

Table 64 shows the means and standard deviations for MTL for the main effect of θ level. As previously noted, means were largest at and near the mastery cutoff indicating that a lengthier test was required in order to make a classification.

Table 64

Means and SDs of MTL for the Main Effect of θ Level

Variable	Mean	SD
θ Level		
θ = -3	20.5230	0.5205
θ = -2	20.2748	0.4208
θ = -1	20.3284	0.4412
θ = 0	24.8254	1.6970
θ = 1	444.8161	82.5655
θ = 2	24.2718	1.8413
θ = 3	20.3942	0.4788

PES analyses. Table 65 lists the means, *n*s, SS, and PES of MTL for the seven levels of θ. θ = 1, the cutoff, was by far the most influential level of effect size accounting for 86% of the effect size. It was, however, the worst performing condition, averaging 445 items. A shorter test that does not compromise accuracy of estimation or classification is one of the main advantages of adaptive mastery testing. Therefore, conditions with smaller means for MTL but a large PES value were the more favorable conditions. All values of θ other than the cutoff had extremely small PES values, probably due to the fact that they had similar MTLs.

Table 65

Means, *n*s, SS, and PES of MTL for the Seven Levels of the True θ

Variable	Mean	*n*	SS	PES
True θ				
-3	20.5230	18	68483.6213	0.0248
-2	20.2748	18	69035.9190	0.0250
-1	20.3284	18	68916.4458	0.0250
0	24.8254	18	59263.0477	0.0215
1	444.8161	18	2366764.4835	0.8570
2	24.2718	18	60412.2510	0.0219
3	20.3942	18	68769.8766	0.0249
Main Effect	82.2048	126	2761645.6448	

Precision and Test Length

In addition to accuracy of decisions and test length, the most efficient test will be a combination of the two. Efficiency here was defined as the proportion of correct decisions divided by average number of items administered (Equation 61).

Table 66 shows the means and standard deviations of efficiency. Values ranged from .0005 to .0500. Efficiency was quite low at the mastery cutoff of θ = 1 and quite high for values of θ far from the cutoff.

Table 66

Average Efficiency by SM, TL, θ Level, and PS

	PS= 600			PS = 900		
TL and θ level	TCAT	MII-T	MTI	TCAT	MII-T	MTI
2 items						
θ = -3	0.0494	0.0493	0.0496	0.0498	0.0492	0.0496
θ = -2	0.0497	0.0500	0.0499	0.0498	0.0499	0.0499
θ = -1	0.0495	0.0496	0.0496	0.0494	0.0494	0.0496
θ = 0	0.0418	0.0393	0.0388	0.0429	0.0395	0.0401
θ = 1	0.0008	0.0007	0.0007	0.0007	0.0005	0.0006
θ = 2	0.0448	0.0413	0.0421	0.0458	0.0401	0.0410
θ = 3	0.0499	0.0497	0.0499	0.0500	0.0496	0.0498
3 items						
θ = -3	0.0496	0.0463	0.0470	0.0498	0.0471	0.0474
θ = -2	0.0497	0.0475	0.0476	0.0495	0.0475	0.0475
θ = -1	0.0496	0.0472	0.0472	0.0496	0.0473	0.0472
θ = 0	0.0414	0.0367	0.0373	0.0427	0.0367	0.0373
θ = 1	0.0010	0.0007	0.0008	0.0006	0.0006	0.0006
θ = 2	0.0451	0.0385	0.0400	0.0451	0.0376	0.0401
θ = 3	0.0499	0.0472	0.0473	0.0499	0.0469	0.0472
5 items						
θ = -3	0.0497	0.0480	0.0485	0.0496	0.0475	0.0486
θ = -2	0.0499	0.0496	0.0498	0.0498	0.0489	0.0497
θ = -1	0.0497	0.0495	0.0495	0.0495	0.0493	0.0493
θ = 0	0.0422	0.0370	0.0385	0.0421	0.0359	0.0388
θ = 1	0.0008	0.0006	0.0007	0.0005	0.0005	0.0005
θ = 2	0.0447	0.0377	0.0385	0.0454	0.0371	0.0382
θ = 3	0.0500	0.0488	0.0489	0.0499	0.0488	0.0492

Effect sizes (η^2). Table 67 shows the η^2 values from the factorial ANOVA of the efficiency. Only θ level had a sizable η^2 and explained 99% of the variance.

Table 67

η^2 from the ANOVA of Average Efficiency

Source of Variation	SS	*df*	η^2
Main Effect			
Item/Testlet Selection Method (SM)	0.00	2	0.00
Testlet Length (TL)	0.00	2	0.00
Theta Level (θ)	0.03	6	0.99
Pool Size (PS)	0.00	1	0.00
2-way Interactions			
SM × TL	0.00	4	0.00
SM × θ	0.00	12	0.00
SM × PS	0.00	2	0.00
TL × θ	0.00	12	0.00
TL × PS	0.00	2	0.00
θ × PS	0.00	6	0.00
Residual	0.00	76	0.00
Total	0.03	125	

Table 68 shows the means and standard deviations of the θ level main effect. At θ = 1, the cutoff, efficiency was close to 0 indicating that the accuracy of correct classification (numerator) was quite small. The two adjacent levels of θ (0 and 2) had smaller magnitudes of efficiency relative to the more extreme values of θ.

Table 68

Means and SDs of Efficiency for the Main Effects of θ Level

Variable	Mean	SD
θ Level		
θ = -3	0.0487	0.0012
θ = -2	0.0492	0.0010
θ = -1	0.0490	0.0010
θ = 0	0.0394	0.0023
θ = 1	0.0006	0.0001
θ = 2	0.0413	0.0031
θ = 3	0.0490	0.0011

PES analyses. Table 69 lists the means, *n*s, SS, and PES of efficiency for the seven levels of θ. Ideal conditions for efficiency are a relatively large mean and large PES. There was no clear-cut favorable condition. θ = 1, the cutoff, had the highest PES (81%) indicating that it was most influential, but had the lowest mean efficiency. The condition θ = -2, had the highest mean efficiency and the second highest PES but accounted for only 5% of the effect size and, therefore, was not very influential. The two values adjacent to the cutoff, θ = 0 and θ = 2, had the lowest PES values and the lowest means after θ = 1, indicating they were also not favorable conditions in terms of efficiency.

Table 69

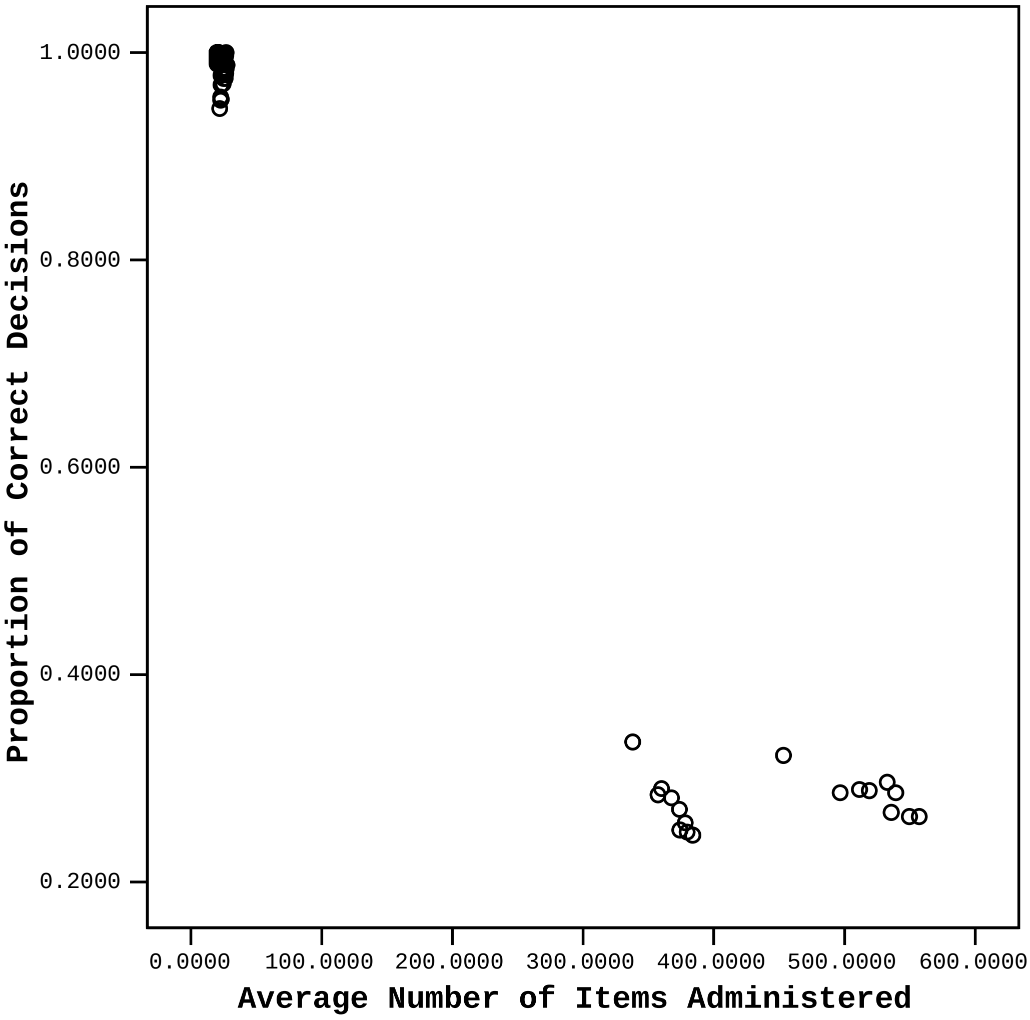
Means, *n*s, SS, and PES of Efficiency for the Seven Levels of True θ

Variable	Mean	*n*	SS	PES
True θ				
-3	0.0487	18	0.0015	0.0438
-2	0.0492	18	0.0017	0.0494
-1	0.0490	18	0.0016	0.0471
0	0.0394	18	0.0000	0.0000
1	0.0006	18	0.0273	0.8106
2	0.0413	18	0.0001	0.0015
3	0.0490	18	0.0016	0.0476
Main Effect	0.0396	126	0.0337	

Another way to examine the relationship between accuracy of decision and number of items administered is to look at its bivariate plot. Figure 20 shows this bivariate plot.

Figure 20

Proportion of Correct Classification by MTL



In this figure there are three distinct groupings. The most efficient group (high proportion of correct decisions relative to a small number of items administered) contained 108 cases. These cases consisted of all levels of θ, except the cutoff θ = 1. Proportion of correct decisions ranged from .95 to 1.00 with a mean of .99 and average number of items administered ranged from 20 to 28 items with a mean of 22 items. The middle group and group farthest to the right contained all those cases where $\theta = 1$. The middle group was all from the item pool of 600 items, whereas the group on the right all contained all those cases where $\theta = 1$ from an item pool of 900 items. These data show that as long as the θ level was not exactly at the cutoff, efficiency was quite similar in magnitude across different testlet lengths, selection methods, and item pool size. Efficiency dropped drastically when θ was at the cutoff, but was influenced by the item pool size.

Summary

Because θ was fixed at discrete points, leaving no variation with which to compute a ϕ correlation, θ level dropped out as a condition making it difficult to compare across all three evaluative criteria.

For ϕ correlations, three main effects: SM, TL, and PS and two interactions: SM × TL and PS × TL had sizable η^2s. Both MII-T and MTI performed similarly, although MII-T was the most accurate in making classification decisions and TCAT was the least accurate. TCAT, however, was the most influential for selection method. For testlet length, TL = 5 was the most accurate but TL = 2 had the most influence and was the least accurate. For pool size, PS = 600-item was the more accurate although both pool sizes were equally influential. In terms of the SM × TL interaction, MII-T was slightly more accurate for TL = 2 and TL = 5, whereas MTI was more accurate for TL = 3. MTI with TL = 3 was also the most influential condition. For the PS × TL interaction, PS = 900 and TL = 5 was most accurate and most influential. In several cases, MII-T and longer testlets were more accurate in making classifications but TCAT and shorter testlets were more influential.

With regard to MTL and efficiency, θ level was a very powerful main effect accounting for the vast majority of variance (96% and 99% respectively). For both MTL and efficiency, $\theta = -2$ was the best performing level and $\theta = 1$, the cutoff, was the most influential. In general, performance was quite poor at the mastery cutoff of $\theta = 1$, which was expected, and increased as θ moved away from the cutoff.

CHAPTER 4: DISCUSSION AND CONCLUSION

The purpose of this study was to examine two different adaptive testlet selection procedures under varying conditions, in comparison to traditional adaptive testing. These conditions, while simulated, were based on real data. The width of difficulty in each testlest and the length of the testlets, as well as the mastery proportion correct, were based on CARLA data. Even the different levels of θ were realistic. Reading comprehension of a foreign language is an unusual trait in the sense that an examinee could have a very high θ level on reading comprehension in his/her own native language but be very low on reading comprehension of a foreign language. Therefore it was necessary to look at the entire θ continuum. The only condition that was theoretical was item pool size.

There were three main areas of interest: parameter recovery of θ, quality of classifications decisions made, and efficiency of those classifications. In any type of mastery testing, adaptive or not, estimation of ability near or at the cutoff is critical. If estimation is not precise, then the accuracy of classification decisions made is compromised. Much of the study therefore hinged on the quality of person parameter (θ) estimation. Even if estimation and accuracy of decisions are good, however, an adaptive test should be efficient as well. One of the major advantages of an adaptive test is that it shortens the length of each test while at the same time maintaining or increasing accuracy of estimation. A shorter test tends to avoid the problem of fatigue or lagging motivation, two undesirable factors that can affect performance. An extremely long test takes away this major advantage of adaptive testing. Therefore, the most ideal test is one which is precise yet relatively short.

Influential Factors in θ Estimation

Correlation of θ and $\hat{\theta}$. Recovery of θ was good. Correlations between θ and $\hat{\theta}$ for all conditions were generally between .80 and .90. Correlations tended to be highest for TCAT and shorter testlets. Shorter testlets generate more frequent branching, which is more similar to a traditional CAT. Because θ was fixed at each level, θ level could not be examined as a main effect with correlations. With θ level collapsed across conditions, selection method emerged as the main effect with the most influence. TCAT had the highest mean correlation (most accurate) and PES (most influential) relative to the other two methods making it the most favorable condition or best performing selection method with respect to the correlation between θ and $\hat{\theta}$. Testlet length was the other main effect with a sizable η^2. The shortest testlets (2-item) had the highest mean correlation and PES relative to the other two testlet lengths, indicating it was the best performing length. The same sort of pattern emerged for the interaction of selection method and testlet length. The SM × TL interaction with a PES between 0 and 1 and the highest mean correlation was TCAT with TL = 2. The PS × TL interaction indicated better performance for both pool sizes with shorter testlets.

RMSE. RMSE was quite low at or near the mastery cutoff and increased in magnitude as true θ reached the more extreme ends of the θ continuum. Test length tended to be longer near the cutoff, which in turn, tended to increase accuracy of estimation. RMSE also tended to be lower for shorter testlets and for TCAT, where more frequent branching occurred. While in an adaptive mastery environment, accurate estimation is critical at and near the mastery cutoff but less important further away from the mastery cutoff, accuracy of estimation can be a concern if the item bank is used for multiple purposes where it might be necessary to estimate accurately across all levels of θ. For example, the Minnesota Basic Skills Test (MBST), which is a mastery test administered in 8th grade, is used as part of the graduation requirement for students throughout the state. Many schools, however, use it for placement in programs or to predict scores on future state tests. Using a mastery test for multiple purposes can be a problem since extreme values, which are furthest from the mastery cutoff, allow for a quicker decision, but also might result in less accurate estimates.

Another possible problem with estimation at the extreme ends of the continuum is the range of b in the item bank. In the design of this study, the range of b exactly matched the range of θ. For $\theta = -3$ and $\theta = 3$, there were not the same number of items/testlets available above and below their respective levels to choose from, as there were for middle values of θ. For example, at $\theta = 3$, the most difficult items were exactly at $b = 3$. There were no more difficult items to choose from and that value was the endpoint of the testlet, whereas at $\theta = 0$, there were many items just above and just below $b = 0$. The reverse was true for $\theta = -3$. There were many items that were more difficult, but none available that had difficulties below $b = -3$. In essence, for extreme levels of θ there was essentially a more limited item bank. Thus, while there was reduced quality of estimation at the extreme ends of the θ continuum, further research is needed on the effect of extending the range of *b*s when examining θ recovery.

In terms of main effects for RMSE, θ level was the only influential effect. RMSE tended to decrease in magnitude at θ levels further from the cutoff. $\theta = 0$ was the level with the smallest RMSE, indicating the most accuracy. The estimation program was set up using $\theta = 0$ as an initial estimate and then, depending on the response, branching up to $\theta = 3$, or down to $\theta = -3$. Since the initial estimate of θ was equal to true θ, accuracy of estimation was increased, which was evidenced in the smallest mean and standard deviation of RMSE across θ levels. The rationale for choosing $\theta = 0$ is that there was no prior information about the examinees and a middle of the continuum estimate was a “best guess.” Varying the initial θ for adaptive mastery testing is a possibility for further research and is discussed further below.

In the PES analyses, the condition with the largest PES (i.e., the most influential level) was $\theta = -3$. It also had the largest RMSE and was therefore, the worst performing level of θ. This influence, however, was not symmetrical on the θ continuum. At $\theta = 3$, the other extreme of the continuum, the PES was lowest in value, indicating that it was least influential, yet the RMSE was somewhat moderate in magnitude. While both $\theta = -3$ and $\theta = 3$ had more limited item banks to choose from, as discussed previously, apparently better θ estimates could be obtained with items that were relatively easy side rather than items that were above θ level in difficulty. This is probably due to pseudo-guessing at the $\theta = -3$ level. This is particularly important to keep in mind with any future research that uses individuals rather than simulees, where psychological factors may play a part. No favored condition emerged for RMSE.

Bias. Like RMSE, bias was smallest at or near the mastery cutoff and increased in absolute value as true θ reached the more extreme ends of the θ continuum. Bias was positive (underestimation) for θ values of 0 and -3 across all main effects and mainly negative (overestimation) for all other conditions. Like RMSE, bias was smallest for TCAT and TL = 2, conditions in which more frequent branching occurred. Negative bias tended to be largest for MTI and θ = 2. Positive bias tended to be largest for TL = 5 and θ values of 0 and -3, but was evenly distributed among selection methods MII-T and MTI.

Since, bias, like RMSE is based on deviations between θ and $\hat{\theta}$, some of the PES analyses yielded the same results. Like RMSE, θ level was the sole influential main effect for bias. θ = -3 had the largest PES and the largest absolute mean value, indicating that it was the most influential level with the poorest estimation (pseudo-guessing may have played a part). For both RMSE and bias, θ = -2 was the level with the smallest absolute mean value, indicating most accurate estimation but the least influential level.

Bias, unlike RMSE, where direction of the estimation error was taken into account, had two interactions with sizable η^2: θ × SM and θ × TL. For the interaction SM × θ, the two testlet methods (MII-T and MTI) had the smallest absolute mean values at θ = -2 but very small PES values, so they were not very influential. MTI at θ = -3, on the other hand, had the largest PES and the second largest absolute mean value for bias. This was also one of the two θ values with positive bias. While all three selection methods had positive bias at θ = -3, it was quite large for the two testlet methods, yet fairly close to 0 for TCAT. Apparently, testlet methods elicited a stronger positive bias than TCAT for this extreme level of θ. There was no favorable condition for this particular interaction.

The other interaction, TL × θ, also had large positive bias at θ = -3 for all three testlet lengths. TL = 2 with θ = -2 had the largest PES between 0 and 1, but in absolute terms was not of major influence at 18%. All three testlet lengths performed similarly for values of θ ≥ 0. Differences across testlet lengths occurred only for negative values of θ. There was no favorable condition. Like correlations, pool size did not have much influence.

The interesting aspect of the results was that the majority of bias values were negative. It is not clear from the design of this study whether this result was due to the selection method, data, or options selected in the software. For θ = -3, the two testlet methods had large, positive bias. At θ = 0, while the positive bias was small, it was larger for the two testlet methods as compared to TCAT. This positive bias only occurred, on the other hand, for two of the seven levels of θ, which may be due to fluctuations in data rather than option selection in the software. Further studies are needed to determine whether this small amount of overestimation was due to the selection method, anomalous data, or particular set of options used with the software.

In sum, θ estimation was most precise where it was required -- at and near the mastery cutoff. θ level was the most influential condition, but when it was collapsed across groups, as it was for correlations, then selection method and testlet length became influential factors which were usually favored by TCAT and shorter testlets. The endpoint θ = -3 and the initial estimate θ = 0, seemed to be influential; the reasons for this finding should be investigated in further studies.

Influential Factors on Quality of Classification Decisions

*No decision***.** The sole condition in which a classification could not be made was when θ level was exactly at the cutoff of 1. It was also the most influential level. Of all the no-decision proportions at θ = 1, TCAT had the smallest proportion, indicating a lower rate of indecision. Apparently there is an inverse relationship between more frequent branching and rate of indecision.

Correctly identified masters. With regard to proportion of correctly identified masters, for all testlet lengths and methods θ = 1 had the lowest proportion followed by θ = 2 with θ = 3 having the highest proportion. PES values also increased in magnitude as θ became more distant from the cutoff. At θ = 1, TCAT tended to have a higher proportion than the two testlet selection procedures; this was probably a result of having better estimation which increased accuracy of decision. Otherwise, there was no definite pattern regarding pool size or testlet length.

Correctly identified nonmasters. Proportion of correctly identified nonmasters was quite high (90% to 100 %). For the 600-item pool, in general, MII-T had slightly higher proportions. In the 900-item pool, MTI and MII-T had similar percentages. It is interesting to note that the two testlet selection methods did slightly better than TCAT. Apparently TCAT had more of a propensity to identify simulees as masters than the two testlet procedures, which lowered the proportion of this type of correct decision. Positive bias, or underestimation, occurred to a greater degree for the two testlet methods, which may have contributed somewhat to more identification of nonmasters. Also, having less frequent branching with “extra” items in the testlets seemed to allow for better decisions in the nonmastery direction, even if not all the items provided maximum information at the most updated θ estimate.

Misidentified masters. The proportion of misidentified masters was low and tended to be lower as θ level was further from the mastery cutoff. MII-T and MTI were the two selection procedures for which there were no occurrences of this type of error under certain conditions. TCAT tended to have higher proportions of this type than MII-T and MTI, indicating that it had a higher error rate. This may be due to the direction of bias across the three selection methods or, perhaps, when branching occurs more frequently like it does with TCAT, although it elicits a tighter distribution around $\hat{\theta}$ it may simply be in the wrong direction leading to more misclassifications. Since there was a prevalence of negative bias, this would lead to overestimation of θ which, in turn, would lead to misclassification of masters, particularly for TCAT.

MII-T tended to have a lower error rate than MTI. Longer testlets tended to have lower error rates than shorter testlets. These conditions also tended to have less accurate estimation, probably due to less frequent branching and “extra or unneeded” items in the testlets. For example, in the MII-T selection procedure, if the next testlet selected was chosen because the maximum information was in the most difficult or easiest item in the testlet, the other items administered through MII-T were somewhat less useful since information would not be maximized. This was especially true for longer testlets. These extra items, however, seemed to have the opposite effect for decisions and tended to reduce this type of misclassification.   In terms of influence, TCAT was the most influential selection method and θ = 0 was the most influential θ level.

Misidentified nonmasters. Proportion of misidentified nonmasters was almost nonexistent at θ = 3 and increased (although remained small) as θ approached 1. Because the cutoff was included in the definition of masters, there were higher proportions of misidentified nonmasters than in the misidentified masters group.   At θ = 2, TCAT tended to have the smallest proportions (least amount of this type of error) with the 600-item pool but was more evenly split between MII-T and MTI for the 900-item pool. This tended to follow TCAT's propensity to identify simulees as masters. Since fewer were identified as nonmasters, there were also less misclassifications in this direction, as well. In terms of influence, θ = 1 was the most influential level.

Thus, all three methods were fairly accurate in correctly identifying masters and nonmasters, but did demonstrate some differences in performance depending on the hit-miss condition. TCAT fared best when a decision could not be made. Because of its propensity to identify simulees as masters, it had more accurate identification of true masters (particularly at the mastery cutoff) and a slightly higher rate of misclassification of masters. MII-T and MTI performed better in nonmaster identification -- they were evenly matched on identifying true nonmasters and lower rates of error on misidentifying nonmasters. MII-T also had the lowest proportion of misidentified masters, especially with longer testlets.

Decision error. MII-T more frequently had no error, particularly with a longer testlet. The worst conditions were with TCAT and θ = -3. In terms of differences across the three selection methods, the largest discrepancy was between TCAT and MII-T, with

TCAT having a much larger mean. The denominator of the decision error criterion was the number of errors for a particular condition. Having a smaller denominator will increase the magnitude of $\sqrt{ADE}$ regardless of the direction of the error. Since TCAT had a slight propensity to overestimate θ, this could increase the error distance in the numerator for misclassifications of masters. Increasing the numerator while at the same time decreasing the denominator probably accounted for the difference in between TCAT and MII-T.

In terms of PES analyses, TCAT had the largest PES value, but did not have the smallest $\sqrt{ADE}$. MII-T had the smallest $\sqrt{ADE}$ but its PES accounted for only about 31% of the effect. Therefore, there was no favorable condition for selection method. With regard to θ level, θ = 3 was the most influential level, with the largest PES value and the smallest $\sqrt{ADE}$, making it the most favored condition as well.

In the SM × θ interaction, there were definite differences across the three selection methods for negative levels of θ, particularly at θ = -3 where pseudo-guessing was a factor, but all three methods performed similarly for levels of θ ≥ 0. TCAT at θ = -3 was the most influential condition but had the highest $\sqrt{ADE}$, making it an unfavorable condition. TCAT at θ = -3 had the highest PES value (23%) but also had the highest $\sqrt{ADE}$, making it an unfavorable condition. No favorable condition could be determined.

A similar pattern occurred in the PS × θ interaction. There were large differences across the two pool sizes for negative values of θ, but both performed similarly for levels of θ ≥ 0. The 600-item pool and θ = -3 condition had the largest PES, making it the most influential condition. It also had the highest $\sqrt{ADE}$, making it an unfavorable condition. The smallest $\sqrt{ADE}$ was with pool size 600 and θ = 3. Unfortunately the PES was relatively modest (22%), making it not very influential. Like the previous interaction, no favorable condition could be determined.

There was no symmetry in terms of performance for levels near the cutoff. For example, θ = -1, which was the same distance from the cutoff as θ = 3, had the largest $\sqrt{ADE}$ and showed differences across pool size and across selection methods, yet this did not occur at θ = 3. Whether this was directly caused by the prevalence of negative bias or limitations on the item bank at the extreme of θ, or a combination of both, would have to be examined in further studies.

Influential Factors on Efficiency

ϕ correlations . ϕ correlations, like Fisher *r* to *z* correlations, were quite high and brought out the effect of selection method once θ levels were collapsed. TCAT had the highest PES (66%) but the lowest mean correlation of the three selection methods. MII-T had the highest mean correlation but was much less influential than TCAT, with a PES of 21%. There was no favored selection method.

There was minimal variation across testlet length means. Influence was somewhat more evenly distributed across the 2- and 5-item testlets, with PES values of 58% and 40%, respectively. Since the mean differences between these two testlet lengths was minimal, it could be argued that TL = 2 was the favored condition. There was no favored pool size. Mean correlations were quite similar and PES was split exactly at 50% for each pool size.

With regard to the SM × TL interaction, both testlet methods performed better than TCAT. The most influential condition was MTI and the 3-item testlet, with a PES of 51% and one of the highest mean correlations making it the most favored condition.

In the PS × TL interaction, the smaller 600-item pool performed better for shorter testlets but the 900-item pool did better with the 5-item testlets. The 900-item pool had a fairly linear increase in performance as testlet length increased. The largest PES value was 11%, indicating that there was no primary influential condition. This was probably due to the very small variation in means across the various interaction permutations.

Test length. MTL was solely influenced by θ level and only at θ = 1, the mastery cutoff. All other sets of conditions produced similar test lengths, although there was a slight increase in MTL for θ values adjacent to the cutoff. The PES analyses confirmed that θ = 1 was quite influential, with a PES of 86%. This overpowering main effect was probably due to the huge difference in length at the mastery cutoff, where it was difficult to make a decision.

Efficiency. When precision and test length were combined, again θ level had an overpowering influence, with a PES value of 81%. Efficiency was worst at the cutoff. This is expected because of the difficulty of making a correct decision (numerator) and very long test length (denominator). Efficiency increased as θ levels deviated from the mastery cutoff, but individual levels of θ had very little influence, with PES values ≤ 5%.

Bivariate plot. The bivariate plot between proportion of correct decisions and test length tended to congregate into three groups: (a) All levels of θ except θ = 1 (108 conditions), (b) θ = 1 and the item pool of 600 items (9 conditions), and (c) θ = 1 and the item pool of 900 items (9 conditions). The first group had the highest proportion of correct classifications with relatively few items, indicating a high level of efficiency. Of the two θ = 1 groups, the smaller item pool had a shorter average test length than the larger pool. When test length was involved, pool size did have some influence. Because there were more items available in the bank, the test length tended to be longer with the larger pool size.

Conclusions

Table 70 summarizes the PES analyses in terms of conditions with the maximum PES for both the main effects and the 2-way interactions of the 12 dependent variables. In this chapter, the maximum PES value was discussed, but it was always in relation to the other PES values for a particular criterion. It is also useful to look at PES values in absolute terms. In Table 70, cells with PES values ≥ 50% are in bold. Having at least 50% is defined here as indicating a strong, sizable influence beyond simply the highest value relative to other levels of each criterion discussed previously.

Table 70
Conditions Eliciting the Best Performance, Most Influence, and Maximum PES for the 12 Criteria

	Source of Variance								
Criterion	SM	TL	θ	PS	SM × TL	SM × θ	TL × θ	TL × PS	θ × PS
Fisher *r* to *z*									
Best Mean	TCAT	2-item			TCAT 2-item			900 2-item	
Most Influential	**TCAT**	**2-item**			MII-T 3- item			600 3-item	
PES Value	**64%**	**52%**			14%			13%	
LMSE									
Best Mean			θ = 0						
Most Influential			**θ = -3**						
PES Value			**74%**						
LBIAS									
Best Mean			θ = -2			MTI θ = -2	3-item θ= -2		
Most Influential			**θ = -3**			MTI θ= -3	2-item θ= -2		
PES Value			**78%**			95%	18%		
NODEC									
Best Mean			θ ≠ 1						
Most Influential			**θ = 1**						
PES Value			**86%**						
TrueM									
Best Mean			θ = 3						
Most Influential			**θ = 3**						
PES Value			**52%**						

-continued on the next page-

Table 70 Cont.

Conditions Eliciting the Best Performance, Most Influence, and Maximum PES for the 12 Criteria

	Source of Variance								
Criterion	SM	TL	θ	PS	SM × TL	SM × θ	TL × θ	TL × PS	θ × PS
TrueNM									
Best Mean			θ = -3						
Most Influential			θ = -3						
PES Value			28%						
FalseM									
Best Mean	MII-T		θ = -3						
Most Influential	**TCAT**		**θ = 0**						
PES Value	**66%**		**96%**						
FalseNM									
Best Mean			θ = 3						
Most Influential			**θ = 1**						
PES Value			**97%**						
$\sqrt{ADE}$									
Best Mean	MII-T		θ = 3			MTI θ = 3			600 θ = 3
Most Influential	**TCAT**		**θ = 3**			TCAT θ = -3			600 θ = -3
PES Value	**63%**		**74%**			23%			46%
Transformed ϕ									
Best Mean	MII-T	5-items		600	MII-T 5-item			900 5-item	
Most Influential	**TCAT**	**2-item**		**600** **900**	**MTI 3-item**			600 3-item	
PES Value	**66%**	**58%**		**50%**	**51%**			11%	
MTL									
Best Mean			θ = -2						
Most Influential			**θ = 1**						
PES Value			**86%**						
Efficiency									
Best Mean			θ = -2						
Most Influential			**θ = 1**						
PES Value			**86%**						

For the majority of the evaluative criteria, θ level was by far the most influential factor across all conditions except for the correlations where it was not used. Because the mastery cutoff, where it was very difficult for an accurate decision to be made, was included as one of the θ levels, it tended to overpower the other main effects. Table 70 indicates that the most influential levels were either at the extreme ends of the continuum ($\theta = -3$ and $\theta = 3$) or the mastery cutoff of $\theta = 1$. For both TrueM and $\sqrt{ADE}$, $\theta = 3$ was the best performing, the most influential and had an absolute PES magnitude greater than 50%. For the two types of correlations, however, where θ level could not be included, other effects emerged. For the most part they were related to selection method, with a more moderate influence of testlet length. Pool size was the least influential effect and its influence extended mainly to Fisher *r* to *z*, transformed ϕ and $\sqrt{ADE}$.

TCAT and shorter testlets seemed to do better than the two testlet selection methods for estimation. This was also confirmed in the PES analyses in which TCAT and 2-item testlets were the most influential and favored conditions. The opposite was true, however, for decision making in the nonmastery direction. TCAT seemed to have a slight propensity to overidentify masters, which affected the quality of its mastery/nonmastery decisions; this result, however, might have been a function of the tendency of TCAT to underestimate θ. Because of this, the two testlet selection methods, as well as longer testlets, had fewer classification errors and did a better job identifying true nonmasters. This was seen, as well, in the PES analyses where the MTI was most influential in decision error.

In sum, whether or not it is advisable to use testlets as a selection method really depends on the purpose of the test. If the main purpose is for accurate estimation, TCAT or shorter testlets yield more accurate results. If the main purpose is decision classification and little tolerance for error is allowed, particularly for a false master, then it is preferable to use MII-T or MTI and longer testlets. If the test is to be used for multiple purposes, which is often the case, then the stakes of the test, the tolerance for error, and loss of accuracy in estimation by using testlets must all be taken into account.

Further Research and Design

The design of this study was based on real data, but no overall condition or set of conditions emerged that was superior across all three types of evaluation: estimation, quality of classification decisions, and efficiency. There were patterns for a particular evaluative criterion but not for all three. The main reason is probably due to the overpowering effect of θ level. To avoid that problem, one extension of this study should be to examine the effects of θ levels near, but not including, the mastery cutoff. The question to be investigated is at what distance from the mastery cutoff does θ level tend to overpower or to stop overpowering other main effects in terms of estimation, which in turn affects the quality of the decisions. The design should involve θ levels that approach but do not include the mastery cutoff in smaller increments. Another condition might be to do similar analyses but collapse across θ levels (as was done for correlations), or simply not to include the cutoff but all other levels of θ in the analyses.

A related issue is the actual mastery cutoff itself. $\theta = 1$ (77%) was selected because it matched real data. Other reasonable cutoffs could vary between 70% and

85%. Would changing the cutoff value to the more positive or negative end of the ability continuum make a difference? Varying the mastery cutoff should be part of future research.

Another condition that was not varied in this study was the minimum number of items, which was set at 20 items. While it is certainly possible to investigate the same sets of conditions with no minimum number of items, it would not be tolerated very well in the public eye. A mastery test is usually associated with "high-stakes testing," such as a requirement for high school graduation or to receive a driving license. Depending on the degree of importance associated with the outcome of a mastery test, most people would object to having an important decision made based on four or five items, even if it is psychometrically possible to make a decision after so few items. Both a smaller and a larger minimum should be investigated. A reasonable small number might be 10 or 15 items, whereas a larger minimum might extend up to 30 items. It would be interesting to see if differences would emerge between the two testlet selection procedures and/or testlet lengths with lower or even higher minimum number items than this study's 20 items.

The other conditions not varied are related to the construction of the testlets themselves. The testlet lengths were all quite similar (2, 3, and 5 items), which may be a reason it had so little influence. Again the rationale was intended to replicate real data. The item writers were interviewed and stated that for such short passages it is extremely difficult to write more than five items per passage without having overlap. It was, therefore, not realistic to simulate an increase in the length of the testlets as a condition for this content area. It might be practical for another content area, such as law. In any case, it might be appropriate in future research to simulate testlets in such a way that the difficulty of the items comprising the testlets were not evenly distributed even if the width remained constant. This would also probably be more realistic, too, in terms of how items are written. Different combinations of easier vs. harder items within a testlet length could be investigated to see how well the two different testlet selection methods coped with it.

Another more realistic design would be to have an adaptive mastery test using testlets of different lengths, keeping the other conditions the same. Instead of summing testlet information, it would then have to be averaged across the number of items within its respective testlet. Variation of testlet length is how most testlets were constructed in the CARLA data set. Perhaps by collapsing this condition, other main effects might emerge.

Related to construction of the testlets, could be the range of difficulty in the item bank. If the range of simulees is –3 to 3 then the range of item difficulty could be from –3.5 to 3.5 or –4.0 to 4.0 and provide less limited choice of items for θ levels at the extreme end of the θ continuum.

There are some estimation issues as well. The ability to make correct classifications is only as good as the accuracy of estimation. RMSE was not as good for more extreme levels of θ and bias tended to be negative. Perhaps this was due to the limited item bank. By increasing the range of item difficulty to extend beyond the range of θ, RMSE and bias would improve, which in turn would increase the accuracy of decisions made. Another cause of these idiosyncrasies may be the type of estimation chosen. It might be worthwhile to compare slightly different estimation procedures such as maximum likelihood vs Bayesian estimation of θ. Certain estimation procedures may work better than others with different selection methods and testlet lengths. Another related issue is the starting value of θ. Because it is unusual to have prior information about θ, especially if it is the first attempt at a mastery test, initial θ was always set to 0. The level of initial θ could be varied to see if that has any influence on estimation. Using prior information about an examinee would not have to be limited to an estimate from a previous test. For example, if it were known that an examinee was receiving special services such as English as a Second Language (ESL) or special education, a lower initial θ could be used. If the examinee was in a Gifted and Talented program, a higher initial level of θ could be used.

Because the same initial θ was used, the same initial items tended to be used as well. While this is less of a problem with simulated data, it could be a real issue with real data in terms of item overexposure. In order to avoid item overexposure, using randomesque item selection (Kingsbury & Zara, 1989) might be more appropriate in future studies. When randomesque item selection is used, instead of selecting the item or testlet with the absolute highest maximum information, an item or testlet is selected randomly from a group of, say, 10 or 15 items with similar amounts of information. It would be interesting to see if a similarly designed study using randomesque item/testlet selection yielded results that were similar or different from this study. This would be valuable information for any test using real examinees, the ultimate goal of any simulation study.

Two other issues related to future research are content balancing and stakes of the test. Kingsbury and Zara (1991) demonstrated that constrained CAT was a more efficient selection method than testlets when content balancing was a concern. In that study, however, the testlets did not have a common stimulus. It is certainly possible to write hybrid testlets that have both a common stimulus and incorporate content balancing. For example, suppose it was desired to have reading passages that had a vocabulary item, literal comprehension item, and inferential item. Like any other trait, certain content areas lend themselves more easily to be written in this manner than others.

The stakes of the tests are mainly dependent on the error tolerated by the declared master for future performance. If it is a mastery test for some skill which involves life and death decisions, little or no error is tolerated in future performance and the stakes are quite high. If there is a higher tolerance for error (e.g., typing test) then the stakes are much lower. The stakes, however, can be manipulated two ways: changing the cutoff criterion and allowing or not allowing retakes. Raising the bar for the cutoff criterion raises the stakes of the test and lowering the bar lowers the stakes. Likewise, not allowing repeat administrations raises the stakes and allowing repeat administrations lowers the stakes. Allowing for repeat administrations has an additional benefit: a prior $\hat{\theta}$ is available. Prior $\hat{\theta}$ and/or correct and incorrect mastery classifications could be generated based on proportions from this study to simulate “retakes” and to examine conditions that influence them. Further, when real data are available, growth across time could be examined for these “retakes.”

REFERENCES

Anastasi, A. (1988). *Psychological Testing* (6th ed.) New York: Macmillan Publishing.

Assessment Systems Corporation (1995). XCALIBRE: *Marginal maximum-likelihood IRT parameter estimation program*. St. Paul, MN: Author

Baker, F. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Baker, F. (1987). Methodology review: IRT parameter estimation. *Applied Psychological Measurement, 11(2),* 111-141.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6,* 258-276.

Cassuto, N. (1996). *The performance of the linear logistic test model under different testing conditions.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. (Eds.) Orlando, FL: Harcourt, Brace, Janovich College Publishers.

Ferguson, R. (1969). *Computer-assisted criterion-referencedmeasurement*, University of Pittsburgh Learning Research and Development Center, Pittsburgh.

Fischer, G. (1973). The linear logistic test mode as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Frick, T. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research, 5,* 89-114.

Frick, T. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research, 6,* 479-513

Frick, T. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8,* 187-213.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. *American Psychologist, 18,* 519-521.

Hambleton, R. & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159-170.

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

Jaeger, R. (1987). Two decades of revolution in educational measurement!? *Educational Measurement: Issues and Practice, Winter*, 6-14.

Kaplan, R. & Succuzo, D. (1997). *Psychological testing: Principles, applications, and issues.* (4th ed.) Pacific Grove, CA: Brooks/Cole Publishing.

Kingsbury, G. & Weiss, D. (1979*). An adaptive testing strategy for mastery decisions.* (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Kingsbury, G. & Weiss, D. (1980). *A comparison of adaptive, sequential, and conventional testing strategies for mastery decisions* (Research Report 80-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program

Kingsbury, G. & Weiss, D. (1981). *A validity comparison of adaptive and conventional strategies for mastery testing* (Research Report 81-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Kingsbury, G. & Weiss, D. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing*, 257-283, New York: Academic Press.

Kingsbury, G. & Zara, A. (1989). Procedures for selection items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

Kingsbury, G. & Zara, A. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

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

Lee, G., Brennan, R., & Frisbie, D. (2000). Incorporating the testlet concept in test score Analyses. *Educational Measurement: Issues and Practice*, *19*, 9-15.

Lewis, C. (1989). *Validity-based scoring*. Unpublished Manuscript.

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

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

Lord, F. (1984). *Maximum likelihood and Bayesian parameter estimation in item response theory* (Research Report RR-84-30-ONR). Princeton, NJ: ETS

Luecht, R. (2003). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the National Council on Measurement in education. Chicago, IL

Luk, H. (1991). *An empirical comparison of an expert systems approach and an IRT approach to computer-based adaptive mastery testing*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago,IL

Medina-Diaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement 17,* 117-130

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

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). Applied linear statistical models (4th ed). Chicago: Irwin.

Novick, M. & Lewis, C. (1974). *Prescribing test length for criterion-referenced measurement,* American College Testing Program, Technical Bulletin No. 18, Iowa City, Iowa

Owen, R. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Reckase, M. (1979). Some decision procedures for use with tailored tests. In D. J. Weiss (Ed), *Proceedings of the 1979 Computerized Adaptive Testing Conference,* Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 79-100.

Reckase, M. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing,* 238-256, New York: Academic Press.

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

Ross, S. (1984). *A first course in probability* (2nd edition). New York: Macmillan Publishing.

Samejima, F. (1980*). Is Bayesian estimation proper for estimating the individual's ability?* (Research Report No. 80-3). Knoxville: University of Tennessee. Department of Psychology.

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

Sheehan, K. & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16,* 65-76.

Spray, J. & Reckase, M., 1996. Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405-414.

Tennyson, R. Christensen, D. & Park, S. (1984). The Minnesota adaptive instructional system: An intelligent CBI system, *Journal of Computer-Based Instruction, 11,* 2-13.

Urry, V. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Wald, *A. Sequential analysis*. New York: Wiley, 1947.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of hierarchical and linear testlets. *Journal of Educational Measurement, 29(3),* 243-251.

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

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

Weiss, D. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

Weiss, D. J. (2005) Manual for *POSTSIM: Post-hoc (real-data) Simulation of computerized adaptive testing*. St. Paul, MN: Assessment Systems Corporation.

Weiss, D. & McBride, J. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.

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

Weiss, D. & Yoes, M. (1991). Item Response Theory. In R. K. Hambleton & J. Zall (Eds.), *Advances in educational and psychological testing*. Boston: Kluwer-Nijhoff.

Yoes, M. (1993). *A comparison of the effectiveness of item parameter estimation techniques used with the 3-parameter logistic item response theory model.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Yoes, M (1998). *PARDSIM: Parameter and response data simulation.* St. Paul, MN: Assessment Systems Corporation.

APPENDIX A

Historical Development of the Entrance Proficiency Test (EPT)

In 1986, due to dissatisfaction with second language (L2) functional proficiency of the students, the College of Liberal Arts (CLA) changed the criteria for fulfilling the foreign language requirement from course work to demonstration of proficiency (i.e., mastery) on a proficiency-based test called the EPT. The Center for Advanced Research on Language Acquisition (CARLA) took over the test development and construction for the three most commonly taught languages: French, German, and Spanish. Four modalities were tested: reading comprehension, writing, listening, and speaking. Reading and listening were in multiple-choice format, and writing and speaking used an open-ended format. IRT calibrations for this dissertation were based on the multiple-choice reading comprehension section.

The Entrance Proficiency Test (EPT)

The reading test was 35 items in length. Each testlet consisted of a passage which was preceded by a sentence or two describe the setting, some sort of picture or graphic, and then a set of two to five items related to the passage. All items were multiple-choice with four alternatives. Instructions, setting, and the items were written in English, the first language of most examinees, and only the passages were in the target language. The rationale for constructing the test in this manner was two-fold: (1) the test was intended to measure reading comprehension of the passages, not comprehension of the items; and (2) having the items in English allowed for much more flexibility and ease in construction of test items

Embedded within the middle of each reading comprehension test were five pilot items, resulting in a 40-item test. Examinees were not informed which were the pilot items and were encouraged to answer all items on the test. Each language had several versions depending on which sets of pilot items were embedded within the test. French had five versions for a total of 25 pilot items, German had four versions for a total of 20 pilot items, and Spanish had six versions for a total of 30 pilot items. Having these embedded pilot items allowed the item bank to be increased from 35 standard items to 60 items for the French test, to 55 items for the German, and to 65 items for the Spanish students without requiring everyone to take all the items.

Tables A-1, A-2, and A-3 list the item parameter calibrations for the three reading comprehension tests (French, German, and Spanish) as well as the number of examinees answering each question (N), the testlet to which each item belonged, and the status of the item: standard or pilot.

Table A-1

Item Parameter Estimates, N, Testlet Number, and Status for the French EPT

Item	a	b	c	N	Testlet	Status
1	0.49	-1.72	0.26	844	1	Standard
2	0.49	-1.95	0.26	844	1	Standard
3	0.48	-1.37	0.26	844	1	Standard
4	0.77	-2.26	0.25	844	2	Standard
5	0.82	-0.18	0.24	844	2	Standard
6	0.61	-0.99	0.25	844	3	Standard
7	0.76	-1.56	0.25	844	3	Standard
8	0.41	-2.48	0.26	844	3	Standard
9	0.42	-2.19	0.26	844	3	Standard
10	0.69	0.22	0.25	844	4	Standard
11	0.49	-0.36	0.25	844	4	Standard
12	0.75	-1.04	0.25	844	5	Standard
13	0.58	-1.95	0.25	844	5	Standard
14	0.59	-0.21	0.25	844	5	Standard
15	0.67	-0.76	0.25	844	5	Standard
16	0.77	-1.12	0.25	844	6	Standard
17	0.64	-1.07	0.25	844	7	Standard
18	0.43	-2.64	0.26	844	7	Standard
19	0.61	0.47	0.25	171	8	Pilot
20	0.72	0.54	0.24	171	8	Pilot
21	0.61	-0.56	0.25	171	8	Pilot
22	0.58	-1.91	0.25	171	9	Pilot
23	0.60	-0.56	0.25	171	9	Pilot
24	0.63	0.52	0.25	162	10	Pilot
25	0.57	-1.82	0.25	162	10	Pilot
26	0.74	-0.81	0.25	162	10	Pilot
27	0.63	-0.05	0.25	162	10	Pilot
28	0.76	0.11	0.24	162	11	Pilot
29	0.66	-0.25	0.25	161	12	Pilot
30	0.69	-1.37	0.25	161	13	Pilot
31	0.74	-1.62	0.25	161	13	Pilot
32	0.72	-1.73	0.25	161	13	Pilot
33	0.68	-0.31	0.25	161	13	Pilot
34	0.61	0.51	0.25	173	14	Pilot
35	0.72	0.27	0.24	172	14	Pilot
36	0.64	-0.75	0.25	172	14	Pilot
37	0.59	0.03	0.26	172	14	Pilot
38	0.70	1.18	0.23	172	14	Pilot
39	0.62	0.05	0.25	176	15	Pilot

-continued on the next page-

Table A-1 Cont.

Item Parameter Estimates, N, Testlet Number, and Status for the French EPT

Item	a	b	c	N	Testlet	Status
40	0.65	-2.22	0.25	176	15	Pilot
41	0.56	-0.26	0.26	176	15	Pilot
42	0.56	-0.26	0.25	176	16	Pilot
43	0.71	0.02	0.24	176	16	Pilot
44	0.65	-2.2	0.25	842	17	Standard
45	0.45	-0.75	0.26	842	17	Standard
46	0.71	-2.02	0.25	842	17	Standard
47	0.52	-1.43	0.26	840	18	Standard
48	0.73	-1.29	0.25	840	18	Standard
49	0.40	-0.52	0.26	836	18	Standard
50	0.49	-0.43	0.26	834	18	Standard
51	0.60	-2.29	0.25	827	19	Standard
52	0.64	-0.74	0.25	825	19	Standard
53	0.66	0.29	0.24	821	19	Standard
54	0.75	-1.91	0.25	818	19	Standard
55	0.77	-0.98	0.24	812	19	Standard
56	0.67	-0.39	0.24	807	20	Standard
57	0.72	0.02	0.24	806	20	Standard
58	0.58	0.61	0.26	805	20	Standard
59	0.66	-0.04	0.25	799	20	Standard
60	0.52	-0.41	0.26	799	20	Standard

Table A-2

Item Parameter Estimates, N, Testlet Number, and Status for the German EPT

Item	a	b	c	N	Testlet	Status
1	0.86	-3.00	0.25	754	1	Standard
2	0.52	-0.82	0.25	754	1	Standard
3	0.64	-1.44	0.25	754	1	Standard
4	0.62	-2.39	0.25	754	1	Standard
5	0.64	-0.91	0.25	754	1	Standard
6	0.74	-2.62	0.25	754	2	Standard
7	0.56	-0.86	0.25	754	2	Standard
8	0.82	-0.60	0.25	754	3	Standard
9	0.85	0.41	0.25	754	3	Standard
10	1.01	0.30	0.22	754	4	Standard
11	0.59	-0.59	0.25	754	4	Standard
12	0.82	-2.26	0.25	754	4	Standard
13	0.66	-0.37	0.25	753	5	Standard
14	0.93	3.00	0.19	753	5	Standard
15	0.53	0.52	0.26	753	5	Standard
16	0.68	-1.09	0.25	753	6	Standard
17	0.85	-1.29	0.25	753	6	Standard
18	0.74	-1.54	0.25	196	7	Pilot
19	0.75	-1.10	0.25	196	7	Pilot
20	0.73	-2.61	0.25	196	8	Pilot
21	0.66	-1.78	0.25	196	8	Pilot
22	0.79	-1.73	0.25	196	8	Pilot
23	0.63	0.57	0.26	188	9	Pilot
24	0.76	1.60	0.24	188	9	Pilot
25	0.81	1.55	0.23	188	9	Pilot
26	0.62	-1.55	0.25	188	10	Pilot
27	0.83	1.40	0.22	188	10	Pilot
28	0.71	-1.14	0.25	176	11	Pilot
29	0.79	-0.61	0.24	176	11	Pilot
30	0.71	-0.65	0.24	176	12	Pilot
31	0.71	-0.19	0.25	176	12	Pilot
32	0.64	-0.13	0.25	176	12	Pilot
33	0.68	-0.18	0.25	193	13	Pilot
34	0.71	-0.82	0.25	193	13	Pilot
35	0.58	0.83	0.26	193	14	Pilot
36	0.87	0.91	0.24	193	14	Pilot
37	0.73	0.90	0.24	193	14	Pilot
38	0.87	-1.77	0.24	753	15	Standard
39	0.69	-0.78	0.25	753	15	Standard

-continued on the next page-

Table A-2 Cont.

Item Parameter Estimates, N, Testlet Number, and Status for the German EPT

Item	a	b	c	N	Testlet	Status
40	0.76	-0.75	0.25	753	15	Standard
41	0.86	0.22	0.23	752	16	Standard
42	0.74	-1.17	0.25	752	17	Standard
43	0.74	-1.44	0.25	752	18	Standard
44	0.64	-1.45	0.25	752	19	Standard
45	0.64	0.32	0.25	746	20	Standard
46	0.87	0.55	0.22	746	20	Standard
47	0.68	-0.99	0.25	742	21	Standard
48	0.80	-1.79	0.24	741	21	Standard
49	0.79	-1.92	0.25	740	22	Standard
50	0.64	-1.27	0.25	739	22	Standard
51	0.78	0.15	0.26	731	23	Standard
52	0.68	-1.10	0.25	728	23	Standard
53	0.56	0.49	0.25	726	24	Standard
54	0.77	0.44	0.24	721	25	Standard
55	0.76	-0.99	0.25	721	25	Standard

Table A-3

Item Parameter Estimates, N, Testlet Number, and Status for the Spanish EPT

Item	a	b	c	N	Testlet	Status
1	0.37	-1.14	0.26	2138	1	Standard
2	0.43	-1.39	0.25	2138	2	Standard
3	0.65	0.60	0.23	2138	3	Standard
4	0.48	-0.98	0.25	2138	4	Standard
5	0.84	2.92	0.15	2138	5	Standard
6	0.82	0.31	0.24	2138	5	Standard
7	0.62	0.28	0.23	2138	6	Standard
8	0.55	-1.71	0.25	2138	6	Standard
9	0.73	-1.03	0.25	2138	7	Standard
10	0.77	-1.29	0.24	2138	7	Standard
11	0.57	-1.43	0.25	2138	7	Standard
12	0.58	-1.91	0.25	2136	8	Standard
13	0.75	-1.20	0.24	2136	8	Standard
14	0.35	-1.59	0.25	2136	8	Standard
15	0.77	-2.03	0.24	2136	8	Standard
16	0.70	-1.34	0.24	2136	8	Standard
17	0.70	-0.03	0.24	348	9	Pilot
18	0.79	-0.83	0.24	348	9	Pilot
19	0.56	0.83	0.26	348	9	Pilot
20	0.73	-1.10	0.24	348	9	Pilot
21	0.80	1.90	0.20	348	10	Pilot
22	0.54	-1.84	0.25	360	11	Pilot
23	0.61	0.23	0.25	360	11	Pilot
24	0.59	-0.69	0.24	360	11	Pilot
25	0.69	-1.33	0.25	360	12	Pilot
26	0.57	-0.01	0.25	359	12	Pilot
27	0.67	2.04	0.25	377	13	Pilot
28	0.60	2.02	0.28	376	13	Pilot
29	0.52	0.61	0.25	376	14	Pilot
30	0.63	-0.72	0.24	376	14	Pilot
31	0.62	0.04	0.25	376	14	Pilot
32	0.61	-1.68	0.25	342	15	Pilot
33	0.66	0.65	0.25	342	15	Pilot
34	0.50	-0.19	0.25	342	16	Pilot
35	0.68	0.24	0.24	342	16	Pilot
36	0.74	0.03	0.24	342	16	Pilot
37	0.66	1.60	0.24	368	17	Pilot
38	0.65	1.54	0.23	368	18	Pilot
39	0.75	1.76	0.21	368	18	Pilot

-continued on the next page-

Table A-3 Cont.

Item Parameter Estimates, N, Testlet Number, and Status for the Spanish EPT

Item	a	b	c	N	Testlet	Status
40	0.80	3.00	0.18	368	18	Pilot
41	0.74	1.99	0.23	368	18	Pilot
42	0.67	0.00	0.25	340	19	Pilot
43	0.58	1.32	0.25	340	19	Pilot
44	0.66	-0.80	0.24	340	20	Pilot
45	0.76	-1.71	0.24	340	20	Pilot
46	0.71	-0.72	0.24	339	20	Pilot
47	0.69	1.22	0.23	2130	21	Standard
48	0.59	-0.55	0.25	2129	21	Standard
49	0.69	-0.97	0.24	2128	21	Standard
50	0.68	-0.52	0.24	2126	22	Standard
51	0.94	-1.38	0.25	2124	22	Standard
52	0.71	-0.89	0.24	2122	22	Standard
53	0.52	-0.79	0.24	2117	23	Standard
54	0.54	-0.83	0.24	2116	23	Standard
55	0.58	-0.35	0.24	2106	24	Standard
56	0.40	-1.06	0.26	2105	24	Standard
57	0.83	0.20	0.25	2103	24	Standard
58	0.71	-0.38	0.24	2101	24	Standard
59	0.51	1.03	0.25	2087	25	Standard
60	0.30	-0.05	0.27	2081	25	Standard
61	0.65	-0.44	0.27	2075	25	Standard
62	0.91	0.70	0.24	2065	25	Standard
63	0.86	-0.28	0.23	2063	25	Standard
64	0.77	-0.88	0.24	2055	26	Standard
65	0.85	-0.69	0.25	2055	26	Standard

Table A-4 shows the means and standard deviations of the final item parameter estimates for each of the languages:

Table A-4

Means and Standard Deviations of the Item Parameter Estimates for French, German, and Spanish EPT

	a		*b*		*c*	
Language	Mean	SD	Mean	SD	Mean	SD
French	0.63	0.10	-0.81	.93	0.25	0.01
German	0.73	0.10	-0.61	1.19	0.24	0.01
Spanish	0.65	0.13	-0.18	1.21	0.24	0.01

Table A-5 shows the average width of the item difficulty of the testlets for each of the three languages. Some passages had only one item associated with it. Only passages for which there were two or more items associated with it were included in the average.

Table A-5

Width of Testlet Difficulty and Mean Width Difficulty for French, German, and Spanish EPT

Testlet	French	German	Spanish
1	0.58	2.18	
2	2.08	1.76	
3	1.49	1.01	
4	0.58	2.56	
5	1.74	3.37	2.61
6		0.20	1.99
7	1.57	0.44	0.40
8	1.1	0.88	0.83
9	1.35	1.03	1.93
10	2.34	2.95	
11		0.53	2.07
12		0.52	1.32
13	1.42	0.64	0.02
14	1.93	0.08	1.33
15	2.27	1.02	2.33
16	0.28		0.43
17	1.45		
18	1.00		1.46
19	2.58		1.32
20	1.02	0.23	0.99
21		0.80	2.19
22		0.65	0.86
23		1.25	0.04
24			1.26
25		1.43	1.47
26			0.19
Mean Width	1.46	1.18	1.25

APPENDIX B

Appendix B contains the test response functions for each combination of testlet length/item bank size: (Testlet length = 2, 3 and 5 for 600 items and testlet length = 2, 3 and 5 for 900 items.

Figure B-1

Test Response Function for 2-Item Testlets in 600-Item Bank

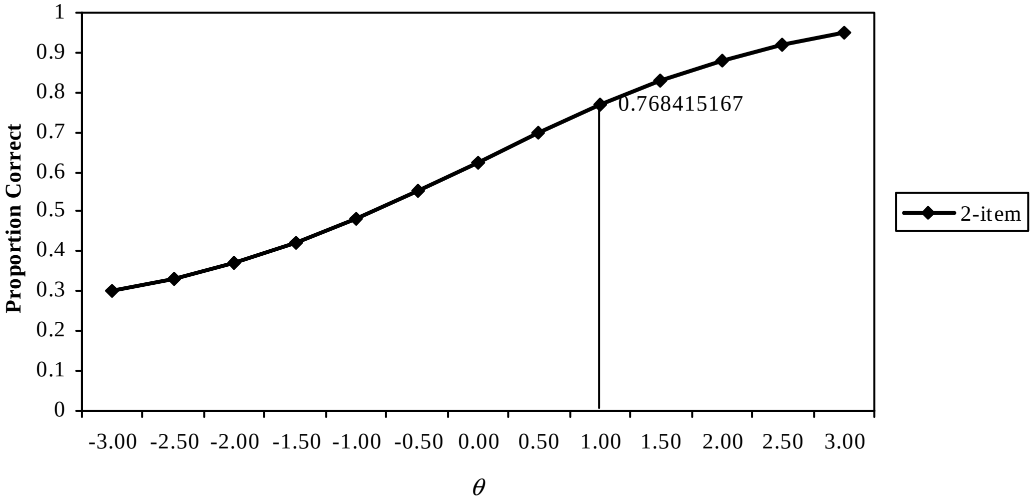


Figure B-2
Test Response Function for 3-Item Testlets in 600-Item Bank

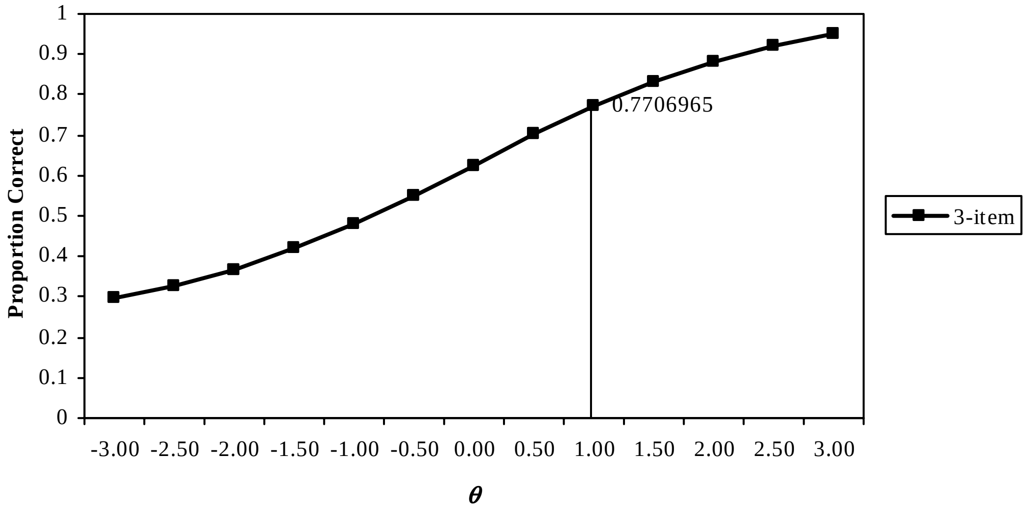


Figure B-3
Test Response Function for 5-Item Testlets in 600-Item Bank

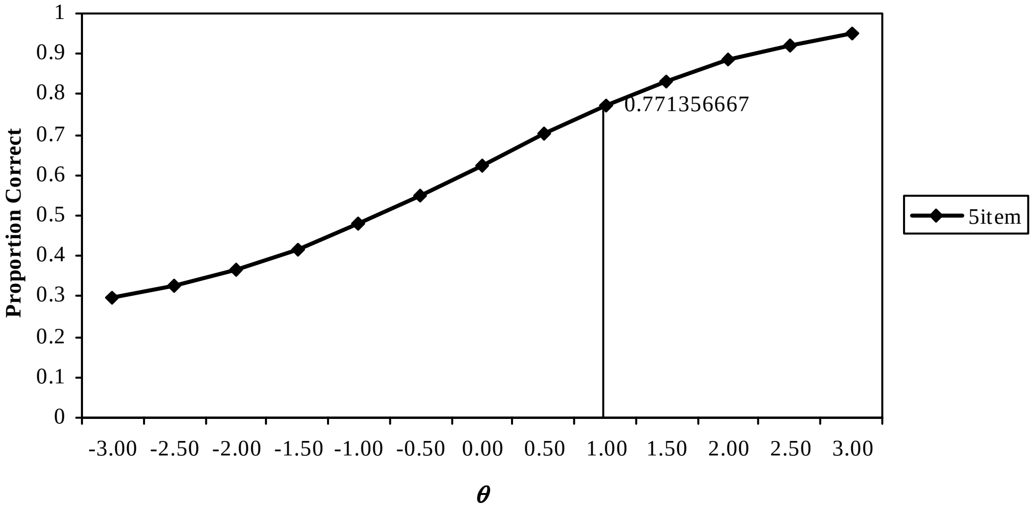
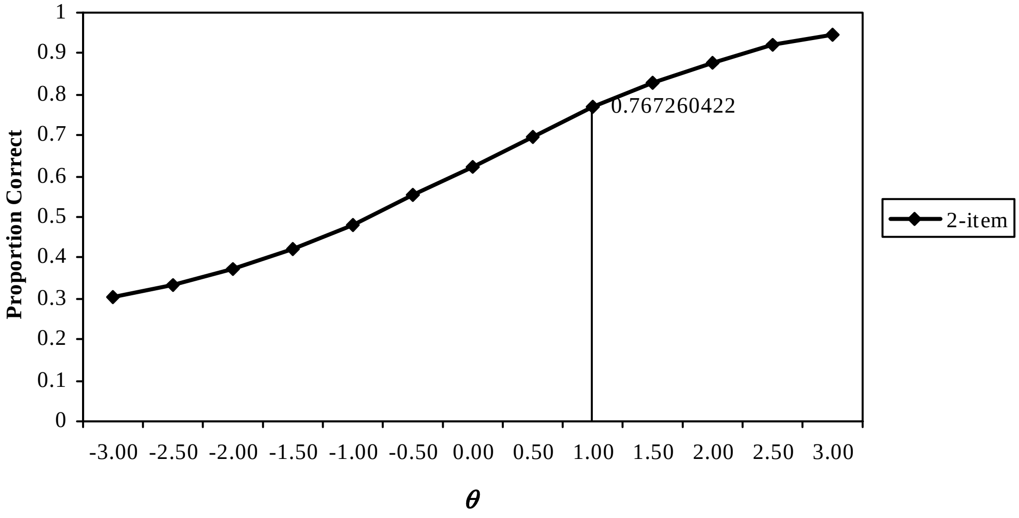


Figure B-4

Test Response Function for 2-Item Testlets in 900-Item Bank



Test Response Function for 3-Item Testlets in 900-Item Bank

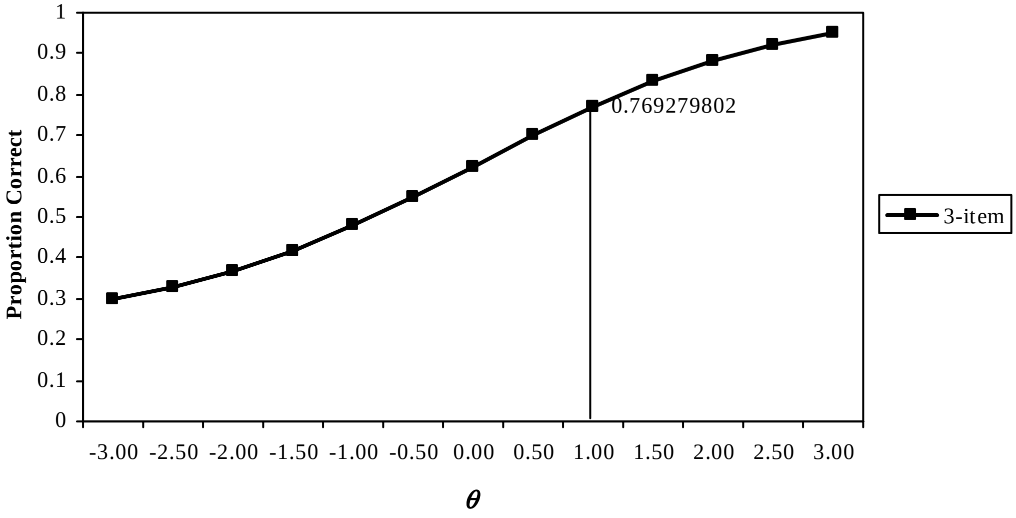
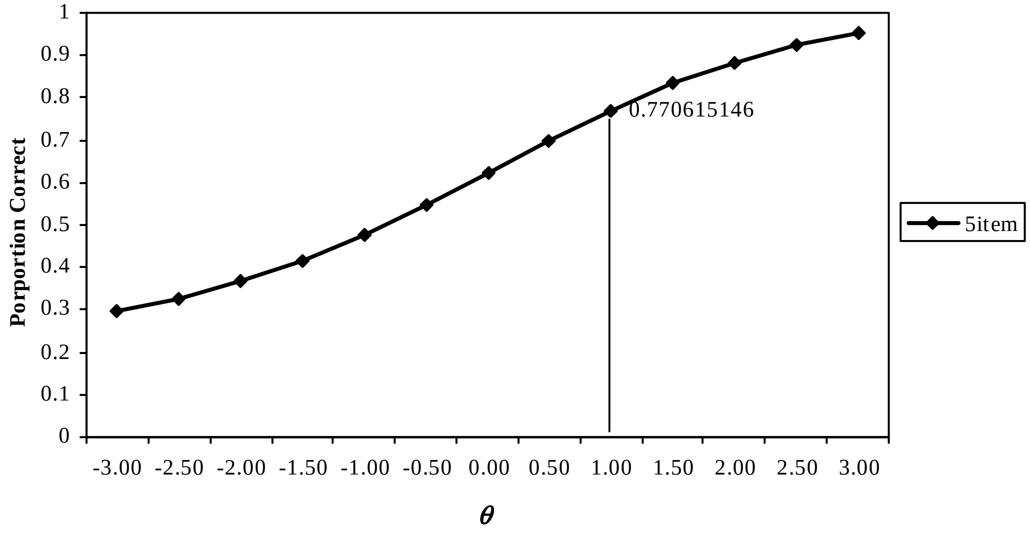


Figure B-6
Test Response Function for 5-Item Testlets in 900-Item Bank


Porportion Correct
5item
θ