2007

An investigation of a Bayesian decision-theoretic procedure in the context of mastery tests

Ming-Chuan Hsieh
University of Iowa

AN INVESTIGATION OF A BAYESIAN DECISION-THEORETIC PROCEDURE IN THE CONTEXT OF MASTERY TESTS

by

Ming-Chuan Hsieh

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

December 2007

Thesis Supervisor: Associate Professor Timothy N. Ansley

ABSTRACT

The purpose of this study was to extend Glas and Vos's (1998) Bayesian procedure to the 3PL IRT model by using the MCMC method. In the context of fixed-length mastery tests, the Bayesian decision-theoretic procedure was compared with two conventional procedures (conventional- Proportion Correct and conventional- EAP) across different simulation conditions. Several simulation conditions were investigated, including two loss functions (linear and threshold loss function), three item pools (high discrimination, moderate discrimination and real item pool) and three test lengths (20, 40 and 60). Different loss parameters were manipulated in the Bayesian decision-theoretic procedure to examine the effectiveness of controlling false positive and false negative errors. The degree of decision accuracy for the Bayesian decision-theoretic procedure using both the 3PL and 1PL models was also compared.

Four criteria, including the percentages of correct classifications, false positive error rates, false negative error rates, and phi correlations between the true and observed classification status, were used to evaluate the results of this study. According to these criteria, the Bayesian decision-theoretic procedure appeared to effectively control false negative and false positive error rates. The differences in the percentages of correct classifications and phi correlations between true and predicted status for the Bayesian decision-theoretic procedures and conventional procedures were quite small. The results also showed that there was no consistent advantage for either the linear or threshold loss function. In relation to the four criteria used in this study, the values produced by these two loss functions were very similar.

One of the purposes of this study was to extend the Bayesian procedure from the 1PL to the 3PL model. The results showed that when the datasets were simulated to fit the 3PL model, using the 1PL model in the Bayesian procedure yielded less accurate results. However, when the datasets were simulated to fit the 1PL model, using the 3PL model in the Bayesian procedure yielded reasonable classification accuracies in most cases. Thus, the use of the Bayesian decision-theoretic procedure with the 3PL model seemed quite promising in the context of fixed-length mastery tests.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

AN INVESTIGATION OF A BAYESIAN DECISION-THEORETIC PROCEDURE IN THE CONTEXT OF MASTERY TESTS

by

Ming-Chuan Hsieh

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
(Educational Measurement and Statistics)
in the Graduate College of
The University of Iowa

December 2007

Thesis Supervisor: Associate Professor Timothy N. Ansley

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Ming-Chuan Hsieh

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations (Educational
Measurement and Statistics) at the December 2007 graduation.

Thesis Committee: _____
Timothy N. Ansley, Thesis Supervisor

_____
Stephen B. Dunbar

_____
Michael J. Kolen

_____
Mary Kathryn Cowles

_____
Andrew D. Ho

To my parents and Te-Min

ACKNOWLEDGMENTS

I would like to show my deeply appreciation for my advisor, Professor Timothy Ansley, for his advise, encouragement and guidance. I am truly grateful to Professor Kate Cowles, whom is always helpful and provides me so many valuable suggestions whenever I met Bayesian problems. I am also thankful to my other committee members, professors Michael Kolen, Stephan Dunbar and Andrew Ho, their suggestions make this dissertation valuable.

I would like to thank the Iowa Testing Program, Pearson Education, and ACT for providing me the financial support and graduate assistantship experience for my PhD study. To my supervisors in American Institutes for Research, Drs. Stephan Ahadi, Gary Phillips and Tao Jiang, thanks for giving me the flexible working schedule, so I can finish my dissertation. Thanks to my best friends in Iowa City- Xia Mao, David Shin, Dongmei Li, Jianlin Hou, and Jin Zhang -their friendship constitutes the sweetest memory in my PhD study.

My parents and three elder brothers deserve the greatest appreciation for their long-term support and faith in me. Their love and encouragement is the best support for me. Finally, I am thankful to my husband, Te-min, for his love and understanding. It is impossible to finish my study without his support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Many test applications involve sorting examinees into two categories, often called masters and non-masters, based on whether or not the examinees have sufficient knowledge and understanding of a particular content domain. Today, mastery tests are applied widely in many fields. Examples include: minimum competency exams for school graduation, the selection of personnel, the assignment of clients to therapeutic treatments, pass-fail decisions in instructional units, certification and licensure exams, the choice of careers in vocational situations, and the evaluation of training programs. These tests are designed as gate keepers for different content areas or professions as well as diagnostic tools for various clinical and counseling situations.

Mastery testing can be regarded as one of the most basic forms of criterion-referenced testing (Kingsbury & Weiss, 1983). The main purpose of mastery testing is to compare an examinee's ability estimate with one or more cut scores and categorize the examinee into an appropriate mastery group. Thus, the focus of mastery testing is correctly classifying people into ability "blocks" instead of estimating individual proficiency levels on a continuous scale.

When mastery tests are used to make high stakes decisions, the correct classification of examinees is of utmost importance. For example, under the No Child Left Behind Act, each student is classified as either proficient or non-proficient. For most states, if the proportion of non-proficient students in a school is less than the Adequate Yearly Progress Index for two successive years, the school can be labeled as needing

improvement. Since any inaccurate classification of student proficiencies can have a serious impact on a school's evaluation, accurate categorization is critically important. Traditionally, mastery tests have been administered in a paper-pencil format. With advances in computer technology and in item response theory, however, assessing mastery using a computerized format is now feasible. This specific type of mastery testing is referred to as computerized classification testing (CCT).

CCT has many potential advantages over traditional paper and pencil testing. In CCT, the test is tailored to fit either an examinee's ability level or to maximize information at the cut point. The selection of items for an examinee is based either on the examinee's previous responses or on the amount of information at the cut point (Bleiler, 1998). With the use of appropriate mastery decision algorithms, administering a CCT can significantly reduce test length and improve measurement precision compared to conventional paper-pencil tests (Frick, 1990; Kingsbury & Weiss, 1983; Reckase, 1983). Other advantages include flexibility in scheduling the test, improvement in test security, ease in administering new types of tests, and ease in collecting data (Wainer, 2000). However, some potential disadvantages of CCT should also be considered in practice. For example, test security may be a problem if the item exposure rates are not controlled in the CCT algorithm. Examinees could memorize the items and share them with other examinees. Eventually, the entire item pool could be exposed and the credibility of the test to classify examinees would be questioned. In addition, the content in CCT must be balanced across whatever areas of content are represented in the test. It is challenging to ensure every examinee is administered the same content domains since every one could get different items in a CCT adaptive format (Wang, 1995).

Regardless of format, a conventional procedure to determine an examinee's mastery status is to compare the examinee's performance with the cut point. An examinee is declared a master if his/her score on an achievement test is as high or higher than the pre-specified cut point, or an examinee is declared a non-master if his/her score on the test is lower than the cut point. This simple procedure is widely used for many mastery tests. With the demands for more efficient and accurate decision making, new testing procedures have been developed.

Generally, most commonly used testing procedures fall into three categories: (1) Sequential Probability Ratio Testing (SPRT; Ferguson, 1969; Reckase, 1983; Spray & Reckase, 1987, 1996; Wald, 1947) (2) IRT-based Adaptive Mastery Testing (AMT; Kingsbury & Weiss, 1983); and (3) various Bayesian models for Mastery Testing (Lewis & Sheehan, 1990; Glas & Vos, 1998; Vos, 2000, 2002; Smith & Lewis, 1995). The first two procedures are mainly used for the CCT format and the third testing procedure can be applied to either the paper-pencil or the CCT format.

SPRT, which was proposed by Wald (1947), was developed as a method to determine if a batch of manufactured goods was acceptable or not based on a very small sample. The SPRT procedure is implemented sequentially to test if the product has met the specified quality. Two hypotheses (pass and fail) are compared with each other to decide which is more likely for the selected sample. The likelihood ratio under these two hypotheses is calculated. If this ratio is greater than a pre-specified value, the lot is passed; otherwise, the lot is failed. This procedure was first employed for mastery testing by Ferguson (1969) who used the binomial model to represent the likelihood of an examinee's response. Reckase (1983) extended the SPRT further to a computerized

adaptive testing situation using item response theory. SPRT is widely used in quality control applications in fields such as agriculture (Binns, 1994) and health care (Spiegelhalter et al., 2003). Many testing programs (e.g. ACT) have also implemented this procedure to make mastery decisions for licensure or certification tests.

The AMT model is IRT based and uses Bayesian confidence intervals around the ability estimate to classify examinees. In this procedure, the confidence interval is constructed around the examinee's provisional estimates of ability (posterior mean). If the confidence interval excludes the cut score, the examinee is passed or failed. If the cut score is still contained in the confidence interval, testing continues (Kingsbury & Weiss, 1983).

A more recent approach for mastery testing was developed by Lewis and Sheehan (1990). Their testing process selects testlets randomly from a pool of parallel testlets for administration. The decision is based on minimizing the posterior expected loss at each stage of testing. Vos (1997, 2000, 2002) has applied a similar procedure at the item level instead of the testlet level by using a binomial probability model. His procedure was named the Bayesian Sequential Rule (Bayes).

Several studies have been done to compare these three procedures (Kingsbury & Weiss, 1983; Spray & Reckase, 1996 ; Vos, 2000 ; Yi et al., 2001). There is no consensus regarding which procedure is preferable since each procedure requires some different test characteristics or parameters to be established prior to testing (Bleiler, 1998). However, the Bayes procedure has drawn some attention recently for its ability to take the relative costs of erroneous testing outcomes into consideration explicitly. That is, the relative seriousness of false negative errors, false positive errors or administering additional items

can be considered in advance in the Bayes procedure by specifying the appropriate values, which are called loss parameters (Vos, 1997, 2000, 2002).

In some cases, the test users may desire to minimize a specific kind of classification error. For example, for a board exam to select medical specialists, it is possible that the test user intends to avoid the risk of selecting examinees whose actual ability levels are below the pass level who might be declared masters. On the other hand, eliminating some truly qualified examinees in the exam may be relatively less serious. In this specific condition, the test user can set a higher weight for the loss parameter for false positive errors and a smaller loss weight for false negative errors to decrease the chance of making false positive errors in the Bayes procedure.

In the Bayes procedure, prior knowledge about an examinee's true ability is assumed to be available. The prior information is used in conjunction with the likelihood function to estimate the posterior distribution parameters. After the posterior distribution is determined, a posterior expected loss associated with mastery, non-mastery, and continuing sampling can be calculated. If the expected loss plus the cost of another item is less than the expected loss prior to administration of the item, testing is continued; on the other hand, if the expected posterior loss is greater or the maximum test length is attained, the testing is stopped. An optimal sequential decision rule is chosen which can minimize the posterior expected losses associated with all possible outcomes (mastery, non-mastery, or continued sampling) at each stage of sampling (Vos,1997).

Vos used an expected loss function as the criterion to make mastery decisions. He also employed a beta prior for the ability of the examinees and a binomial likelihood for the item responses. This binomial model assumes items are sampled at random and the

probability of answering each item correctly is equal. While this was mathematically convenient, the assumptions of the binomial model ignore that items may vary in difficulty level, discrimination power or easiness of guessing. In Vos's (2000) approach, each item has the same item parameters, which is not reasonable in most real testing situations (Fischer, 1973; Medina-Diaz, 1992).

Glas and Vos (1998) proposed a Bayesian procedure (named Adaptive Sequential Mastery Testing) which employed the Rasch model to make mastery decisions. The procedure selects a number of testlets which contain maximum information from an examinee's previous responses. A mastery decision is made at each stage when a complete testlet is given to the examinee. Their procedure assumed local independence holds within and between testlets; that is, all responses are independent given the theta level (Glas & Vos, 1998).

Based on Glas and Vos's (1998) discussion, the most challenging part in implementing the Bayesian procedure for mastery testing is calculating the expected loss function from the posterior distribution. As mentioned before, the posterior distribution involves a prior multiplied by a likelihood function. The prior is usually assumed to follow the standard normal distribution for a Bayesian model in achievement tests. If the likelihood is an item response model (e.g. Rasch model), which is an exponential function, then integrating the posterior distribution over a certain interval to obtain the expected loss function can be challenging.

When mastery testing involves many examinees and items, a larger number of possible response patterns could make the posterior distribution function quite complicated. In the Rasch model, the expected loss function of the posterior distribution

can be explicitly estimated since the number correct score is the minimum sufficient statistic for theta. Glas and Vos (1998) derived a general form to estimate the probability of response patterns by incorporating the sufficient statistics into the exponential function of the Rasch model. The general form they derived could possibly extend to the 2PL IRT model; however, their framework has some difficulties in applications of the 3PL IRT model because of the lack of sufficient statistics.

Different testing procedures certainly have some impact on the accuracy of decision making; however, the quality of the mastery test also depends in large part on the quality of the item pool. According to Lord (1980), constructing an item pool is one of the major tasks in the development of a mastery test. The item pool is constructed based on a test blueprint and normally contains a large number of calibrated items. Calibrating an item involves administering the item to a large representative sample of examinees under a pre-testing design, then employing a statistical analysis to determine the item characteristics such as discrimination, difficulty, and pseudo-guessing parameters (Wainer, 2000). Item pool quality plays an important role not only for the efficiency of selecting items but also for the accuracy of classifying examinees into different categories (Bleiler, 1998).

The Purposes

Mastery tests play an important role in a variety of professional areas in our society. Many companies and academic associations use such tests to maintain the quality of the people who practice a profession. For the conventional mastery tests, an examinee can be defined as a master when he/she is able to answer the minimum percentage of items from a set of items. Most mastery testing situations require all examinees to answer

the same set of test questions in order to make the test scores comparable. Although this conventional procedure is straightforward and simple, some studies have shown that it may not be very efficient and accurate under some conditions (Kingsbury & Weiss, 1983). With the demand of more accurate decisions in the mastery tests, new testing procedures have emerged.

The Bayesian method is a relatively new approach, which was developed by Lewis and Sheehan (1990) and expanded by Glas and Vos (1998). There are two basic elements in the application of Bayesian methods to decision making: the psychometric model used to relate the examinee's observed test scores and true ability level, and the loss structure employed to evaluate the costs of all possible decision outcomes (Vos, 1997). Most classification tests use an IRT model to connect the examinee's responses and his/her true level of ability. However, employing an IRT model to estimate the posterior expected loss structure under the Bayesian method can be quite challenging due to the complexity of the posterior distribution. Glas and Vos (1998) proposed Bayesian sequential decision theory with the Rasch model. As stated above, although their procedure can be extended to the 2-PL IRT model, it is not applicable to the 3PL IRT model. In this study, an alternative procedure to estimate the expected loss function for the 3PL IRT model is proposed. This modified Bayesian decision-theoretic procedure adopts the logic from Glas and Vos's (1998) work. However, a Markov Chain Monte Carlo method is employed to solve the expected loss function problem.

There are three purposes of this study. The first purpose is to extend Vos's Bayesian procedure from the 1PL to the 3 PL IRT model. To demonstrate the efficacy of the 3PL procedure compared to that of the 1PL procedure, the degree of accuracy of the

Bayesian decision-theoretic procedure using both 3PL IRT and 1PL IRT model is compared. In addition, the effectiveness of controlling the false positive errors and false negative errors by setting the loss parameters in the Bayesian decision-theoretic procedure is investigated. Conventional methods are used to serve as the baseline to evaluate the performance of Bayesian decision-theoretic procedure.

The second purpose is to examine the effects of the item pool on the accuracy of the pass/fail classifications. Two simulated item pools (a high discrimination item pool and a moderate discrimination item pool) and one real item pool are considered in this study. These three pools have different statistical characteristics, and they will be used to explore the impact of the item pool on the classification decisions under different conditions.

The third purpose is to investigate the impact of test length on classification decisions. In a mastery test, the test length depends on the purpose and the stakes of the test. For example, for teacher certification exams, the test length is typically around 120 items (Florida Department of Education, 2006). But for a math placement exam in a university, the examinees may only need to take 40 items. Although most people would be concerned about making an important decision based on only a few questions, it is psychometrically possible in some situations to make an accurate decision after administering just a few items. It would be of interest to investigate the impact of test length on classification decisions.

The Importance

Mastery tests have drawn considerable interest in the psychological, educational and professional fields. In these fields, mastery tests are designed to evaluate the examinees' knowledge, skill and competence to determine if they are qualified to practice their specialty. Making mastery decisions involves a variety of challenging measurement problems. As implementing mastery tests on computers is becoming increasingly common, relevant research is needed to ensure that the best methods are applied for important testing situations. Thus, by carefully evaluating new procedures, more efficient, accurate, and flexible mastery tests can be developed and put into practice.

This study is organized as follows. The next chapter provides the theoretical background for this study. Previous studies related to mastery tests, the testing procedures, and MCMC analyses are reviewed. Chapter III outlines the specifications of components in the mastery tests used for this study, data generation methods, implementation of the testing procedures, research design and the evaluation criteria. In Chapter IV, the results are presented and illustrated. Conclusions, limitations, and future directions of work are provided in Chapter V.

CHAPTER II

LITERATURE REVIEW

This chapter serves to provide a general background and theoretical framework for this study. There are three sections in this chapter. Section I describes four important components in mastery tests, including the item pool, the item selection method, the cut score and the testing procedure for making mastery decisions. Four testing procedures which are commonly applied in the mastery tests: Sequential Probability Ratio Test (SPRT), Adaptive Mastery Testing (AMT), Conventional Procedures and Bayesian Procedures are reviewed. Studies comparing these testing procedures are discussed in Section II. In Section III, some topics related to Markov Chain Monte Carlo (MCMC) methods are described. A summary is given at the end of this chapter.

Components in Mastery Testing

A typical mastery testing program consists of the following four components: an item pool, a method for selecting items, a procedure for setting the cut score(s) and a rule for making mastery decisions. This section provides a description of each of these components and the relevant issues associated with them.

Item Pool

The development of an item pool usually requires several steps. First, content specialists are employed to write items based on the test specifications. After a comprehensive review, the items are administered to large, representative samples of examinees under non-operational conditions; then a statistical analysis is performed to determine the discrimination value, difficulty level, and the guessing probability for each

item. Following calibration, items are stored in the item pool along with their keyed answers, parameter values and other relevant attributes such as item content, cognitive level, and item format (van der Linden, 2005). There are usually a large number of calibrated items in the item pool; it may include a few hundred to several thousand items.

Some requirements for the item parameters have been suggested for adaptive tests when the purpose is to locate examinees over the entire range of the ability scale. According to Urry (1977), a satisfactory item pool for such tests is characterized by items with a-parameters greater than 0.8, with b-parameters evenly and widely distributed over at least the range (-2.0, 2.0) and c-parameters smaller than 0.3. However, for mastery testing, the nature of a "high quality" item pool may depend on which testing procedures are used. For example, Bleiler (1998) showed that a high quality pool for the SPRT procedure means many highly discriminating items concentrated near the cut theta values. However, for the AMT procedure, it means the item pool contains a sufficient number of highly discriminating items over the whole range of ability.

Item Selection

The efficient selection of items from an item bank is a major issue in mastery testing, especially when the tests are administered using computers. In the unconstrained computerized classification test (CCT), which neglects non-statistical issues such as content balancing, the standard approach is to select an item that maximizes the Fisher information function at the examinee's current estimated trait level or at the cut theta. The item information at a given ability level is determined by the item parameters. To facilitate the item selection process, it is possible to create an "information table" before the administration of the CCT. The construction of the information table depends on how

θ is set in the item information curve. If theta is set as the examinee's current ability estimate, then items in the table are rank ordered by the amount of information they can provide for each ability level. To select the item each time, the computer program can search the information table, locate the column that contains the provisional ability estimate and find the top ranked item not yet administered. On the other hand, if theta is set as the cut point, the items are rank ordered based on the amount of information at the cut point. All examinees receive items in the same order and the test can be viewed as a conventional fixed-item test. However, in this case, the validity of the test may be challenged since there might exist some motivation problems for examinees whose ability level is substantially below or above the cut score. In practice, selecting the most informative items at an examinee's given trait level seems preferable.

Bleiler (1998) compared the results of selecting items based on the maximum information at the examinee's trait level and at the cut point for both the Adaptive Mastery Testing (AMT) and Sequential Probability Ratio Test (SPRT) procedures. The AMT procedure performed more efficiently when items were selected to be maximally informative at the examinee's current theta estimate. However, the SPRT procedure performed more efficiently when items were selected to be maximally informative at the cut score.

Information-based item selection methods need to be used with caution since they can give rise to extremely skewed item exposure rates. For example, some items could be used frequently, whereas, others might never be used. This would decrease test security and potentially increase test cost. However, some randomization schemes can be

employed in practice to control item exposure rates (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995; Sympson & Hetter, 1985).

Cut Scores

Cut scores are used in mastery testing as standards to define the mutually exclusive categories along a score scale. The cut score can be regarded as a standard which is a dividing line between two categories. Examinees with scores equal to or greater than the standard are considered to have mastered the content; they are regarded as passing the test which measured the relevant skills or content domain.

There are five methods commonly adopted to establish the cut score: Nedelsky's method (Nedelsky, 1954), Angoff's method (Angoff, 1971), the borderline-group method (Livingston & Zieky, 1982), the bookmark procedure (Mitzel et al, 2001) and the contrasting groups method (Crocker & Algina, 1986). These methods are briefly described below.

Nedelsky's method can only be used with multiple-choice tests. In this method, the judges are asked to decide the number of distractors that the "minimally competent" or "borderline" examinees could identify to be incorrect choices for each item. These judgments are used to estimate, for each item, the probability that a borderline examinee would choose the correct answer. Then the summed value of these probabilities is set as the cut score for the test.

Angoff's method is basically the same as Nedelsky's method except that the judges are asked to specify the probabilities that borderline students would correctly respond to the item directly. In addition, the use of Angoff's method is not limited to multiple-choice tests.

The borderline-group method is similar to Nedelsky's and Angoff's methods. This method is based on the premise that the passing score is equal to the score obtained by a typical borderline examinee. However, instead of making judgments for each item, the judges select a specific group of examinees as having a borderline level of the knowledge required by the test. The median of those examinees' scores is set as the cut score.

For the bookmark procedure, panelists are asked to place "bookmarks" in an ordered item booklet at the point where they feel the difficulty of the items exceed the level of ability of a minimally qualified student. With the information of how students actually performed and panelists' discussion for several rounds, the cut score is set at the median of all scores which panelists provided. With the bookmark procedure, the constructed-response and multiple-choice items can be considered together by the panelists.

As for the contrasting-group method, the judges classify individual examinees as masters or nonmasters. The passing score is usually chosen to minimize false positive error rates and false negative error rates. The passing scores can also be chosen to minimize a weighted sum of these two types of wrong decisions.

In general, the selection of a cut score is subjective. It has been found that different methods will produce different passing scores. Therefore, it is argued that the validity of a test depends in part on the method used to set the cut score (Norcini et. al, 1997). Although there is no "gold standard" for establishing a passing score, Downing et. al. (2006) indicated that the key to determining a defensible passing score depends on the choice of a credible panel of judges and the use of a systematic approach of collecting

their judgments. All standards are ultimately policy decisions which reflect the collective, subjective opinions of a panel of experts.

Testing Procedures

The implementation of testing procedures in mastery tests is usually closely related to the format of the tests. Mastery tests can be either fixed-length format or variable-length format. In a fixed-length mastery test, examinees are classified as masters or non-masters after a fixed number of items are administered. This test format can be implemented in either paper-pencil or computerized test administrations. As for the variable-length format, each examinee is administered a different number of items and the classification decisions would not be made until an acceptable level of measurement precision has been attained. This mastery test format can only implemented with computerized administrations. The main advantage of variable-length tests is that it is possible to shorten the test lengths for examinees whose mastery levels are clearly above or below the standard and increase the test length for those whose mastery levels are still uncertain (Glas & Vos, 1998).

Theoretically, adaptive tests can provide equal precision at all points on the ability scale through the use of variable-length tests; however, Stocking (1987) showed that variable-length testing can cause biased proficiency estimates, especially when the test is short. This may not be a problem if a large number of items are administered, or if the purpose of the test is to make a classification decision. However, if the examinee's test score is reported on a continuous scale and the number of items in the test is small, fixed-length tests may be preferable.

Different test formats are usually associated with different test procedures. For example, for the variable-length mastery tests, the SPRT and AMT procedures are frequently used. As for the fixed-length mastery tests, more conventional procedures can be used. Some testing procedures, such as Bayesian methods, can be used in both fixed and variable-length formats. These test procedures are described in detail next.

1. Procedures Used in Variable-length Format Tests

(1) Sequential Probability Ratio Test (SPRT)

The SPRT procedure was developed by Wald (1947) to control the quality of light bulbs using small samples. Wald demonstrated that when the sampling was done sequentially and the SPRT rule was applied after each observation, conditional on the same degree of decision accuracy, the total number of samples needed can be reduced by approximately half, compared to that for conventional methods. The SPRT procedure has been used often as a vehicle for quality control in manufacturing industries.

The SPRT was employed in educational testing by Ferguson (1969) for individually prescribed instruction and then extended by other researchers for mastery testing (Reckase, 1983; Spray & Reckase, 1987). In mastery testing, the observation of light bulbs becomes the observation of the examinees' correct or incorrect answers to the items in the test. Two hypotheses are defined:

H_0 : examinee is a non-master, and

H_1 : examinee is a master.

After the student's correct or incorrect responses to items are observed, a probability ratio is computed sequentially to choose between these two hypotheses.

Before implementing the SPRT procedure, several components must be defined by the test user:

(1) Two probabilities around the cut off score: i)the probability of any given item being answered correctly by an examinee in the mastery status, and ii) the probability of any given item being answered correctly by an examinee in the non-mastery status. These two probabilities are obtained based on a representative sample of examinees. The region between these two probabilities is known as the indifference region (Ferguson, 1967; Reckase, 1983).

(2) Type I error probability (α) and Type II error probability (β), where Type I error is passing a non-master and Type II error is failing a master. Type I and Type II error probabilities can be set at different values to match the purpose of the test.

In the early stages of applying SPRT to mastery tests, the examinees' responses to items were assumed to follow a Bernoulli distribution. That is, item responses of a given examinee are independent of each other and items administered are randomly selected from the item pool without replacement. For each item that has been administered to the examinee, a probability ratio is computed (Frick, 1990):

$$\text{PR}= \frac{p_1^m (1-p_1)^{n-m}}{p_0^m (1-p_0)^{n-m}}$$

where n represents the total number of items;

m represents the number of correct responses;

p_0 represents the probability of a correct response to the item by a nonmaster;

p_1 represents the probability of a correct response to the item by a master.

This probability ratio is evaluated after every observation based on the pre-specified Type I error probability (α) and Type II error probability (β). These error probabilities are then used to construct the boundaries for the probability ratio. The boundaries are A and B where

Lower boundary =B $\geq \beta/(1-\alpha)$

Upper boundary =A $\leq (1-\beta)/\alpha$

The mastery decision is made by comparing the likelihood ratio with the boundaries,

if PR $\geq$ A, then the examinee is a master;

PR $\leq$ B, then the examinee is a non-master;

otherwise, another item is administered.

Although employing the binomial model in the SPRT is simple and straightforward, some criticisms have been made because of the nature of this model. In the binomial model, each item has equal difficulty and discrimination levels, which is not generally true in real educational testing. Fischer (1973) and Medina-Diaz (1992) noted that the binomial assumption is appropriate for relatively simple tasks such as pure math computation, but when the tasks become more complicated, such as in advanced algebra or calculus, the binomial model is not appropriate. In addition, in the binomial model, items are assumed to be randomly selected from the pool, without replacement. If a number of extremely easy (or hard) items are selected by chance at the early stage of test administration, a premature and incorrect mastery decision could be made for examinees (Frick, 1990).

Reckase (1983) extended the SPRT from the binomial model to an IRT model. Within an educational situation, using the IRT model, the decisions must be made based on the relative likelihood of two hypotheses,

$H_0 : \theta_i = \theta_0$ and $H_1 : \theta_i = \theta_1$,

where θ_i is the examinee's true theta level after each item has been answered;

θ_0 represents the theta that corresponds to the lower limit;

θ_1 represents the theta that corresponds to the upper limit.

When the use of SPRT is based on an IRT model, each item is regarded as different and has unique item parameters such as discrimination, difficulty and level of guessing to capture the item characteristics, depending on the model. Items are selected that either contain the most information at the cut point or at the examinee's current ability estimate. In addition, a cut score is identified on the theta scale instead of the proportion correct scale.

In order to control the testing time, some constraints are typically used for the maximum test length. A forced classification is usually made when the maximum test length is reached and an undecided classification status has occurred. Basically, when the maximum test length is reached, the classification will be made by comparing the likelihood ratio with the two boundaries. That is, accept H_1 when the $L(x_1, x_2 ..., x_n)$ is greater than the midpoint of the interval $[\beta/(1-\alpha), (1-\beta)/\alpha]$; otherwise, accept H_0.

(2) Adaptive Mastery Testing (AMT)

In this procedure, before administering the first item, a prior distribution for theta needs to be specified, and it is typically assumed to be normal. Each time an item is administered, the examinee's theta is estimated by Owen's method based on the

parameters of the items administered and the cumulative pattern of responses. The item that provides the maximum information at the current theta estimate is selected next (Kingsbury & Weiss, 1983). In AMT, if the cut score is within the confidence interval of the examinee's current theta estimate, the test continues and another item is administered; otherwise, the test is terminated and the examinee's mastery status is determined. The decision rule can be summarized as:

If $[E(\theta) - \Phi^{-1}(Z_{1-\gamma})V(\theta)^{1/2}] \geq \theta_c$, then the examinee is a master;

$[E(\theta) + \Phi^{-1}(Z_{1-\gamma})V(\theta)^{1/2}] < \theta_c$, then the examinee is a non-master;

otherwise, another item is administered;

where $E(\theta)$ is the posterior mean based on Owen's theta estimation;

$V(\theta)$ is the posterior variance based on Owen's theta estimation;

θ_c is the cut theta;

$1-\gamma$ is the density region in the posterior distribution of θ_i;

Φ^{-1} is the quantile function for the normal distribution.

The choice of γ is arbitrary and there is no direct way to control the relative weights of false negative error rates and false positive error rates in this procedure. The AMT procedure uses a symmetrical confidence interval, implying equal costs for the two error types.

2. Procedures Used in Fixed-length Format Tests

(1) Conventional Procedure- Proportion Correct

In this context of mastery tests, it is common to define the criterion for mastery using the percentage of the items on the test correctly answered. After all items in the conventional test are administered, if the examinee's score is equal to or exceeds the proportion correct cut score, the examinee is declared a master. Otherwise, the non-master decision is declared.

In some cases, if IRT is used, it may be necessary to convert the latent trait score metric to the proportion correct score scale. This can be done by using the Test Characteristic Curve (TCC; Lord, 1977). The TCC for the 3PL IRT model can be expressed in the following form:

$$E(P \mid \theta) = \sum_{i=1}^{n} \left[c_i + (1 - c_i) \frac{1}{1 + \exp[-1.7 a_i (\theta - b_i)]} \right] \Big/ n$$

where $E(P|\theta)$ is the expected value of the proportion correct score on the test given theta;

n is total number of items on the test;

a_i, b_i, c_i are the discrimination, difficulty and pseudo-guessing parameters, respectively for item i;

θ is the given latent trait ability level.

(2) Conventional-EAP

Some testing programs use the examinee's ability estimate to determine if the examinee passes or fails the test. There are many ability estimation methods available, such as maximum likelihood estimation, maximum a posteriori estimation, Owen's

Bayesian estimation and expected a posteriori estimation (EAP). The estimators from any of these methods could be used to determine an examinee's mastery status. However, EAP has been shown to have significantly less bias than the other methods (Wang, 1995). Thus, it is common to use EAP to estimate examinees' ability levels.

For the expected a posteriori (EAP) method, the examinee's ability estimate is the mean of the posterior distribution of theta. It can be approximated by a rectangular quadrature procedure (Thissen & Wainer, 2001):

$$\overline{\theta_i} = \frac{\sum_{k=1}^{q} X_k L_i(X_k) W(X_k)}{\sum_{k=1}^{q} L_i(X_k) W(X_k)}$$

where $\overline{\theta_i}$ is the mean of the posterior distribution of theta;

L_i is the likelihood function after i items;

X_k is one of q quadrature points;

$W(X_k)$ is the weight associated with the quadrature points. The weights are normed so that the sum of these weights equals 1.

In EAP estimation, the weights are the probabilities at the corresponding points of a discrete prior distribution. In some contexts of education, normal prior distributions for the points and weights are usually assumed to improve the accuracy of the numerical approximation of the integral (Bock & Mislevy, 1982). In this Conventional-EAP procedure, the examinee's ability level is estimated by the EAP method after all items in the conventional test are administered. The EAP estimator is compared with the cut theta. If the EAP estimator is equal to or greater than the cut theta, then the examinee is

classified as a master; on the other hand, if the EAP estimator is smaller than the cut theta, the examinee is classified as a non-master.

3. Procedures Used in Both Variable-length and Fixed-Length Format Tests

Bayesian Procedures

Computerized Mastery Testing (CMT) and Adaptive Sequential Mastery Testing (ASMT) are Bayesian procedures that are commonly used with mastery tests, and they are very similar to each other. Both of these procedures use Bayesian decision theory to make classification decisions. Some details of these two procedures are described next. Lewis and Sheehan (1990) developed a Computerized Mastery Testing procedure (CMT) which is a sequential testlet-based method to make mastery decisions. In this procedure, the testlets must be constructed to be as parallel as possible. That is, the testlets in the pool are composed of the same number of items, are equivalent with respect to content coverage, and are equivalent with respect to the likelihood of particular number correct scores at proficiency levels near the cut score. This procedure randomly selects each testlet from the pool, seeking to minimize both test length and classification errors.

There are two main components in the CMT procedure. The first component is to construct the loss structure. The loss parameters for false positive errors and false negative errors need to be specified in advance in this procedure. The flexibility to specify the loss parameters has been considered a great advantage of Bayesian procedures since the test users can control the undesired misclassification errors (e.g. false positive errors). The second component of CMT is the decision rule. A detailed description of these two components is given next.

Loss Functions

The CMT procedure uses a decision theory approach to control classification error rates. A loss function is specified in association with each decision made. As shown in Table 1, there are four possibilities for making decisions in a fixed-length mastery test (Lewis & Sheehan, 1990). For situation (1), the true theta level is less than the cut score and the examinee is declared a non-master. This correct mastery decision incurs no loss. For situation (2), the true theta is less than the cut score, but the examinee is declared a master. This incorrect decision incurs a loss for making a false positive error which is represented by a constant *L*. For situation (3), the true theta is greater than or equal to the cut score, but the examinee is classified as a non-master. In this case, the loss is a false negative error which is represented by a constant *M*. For situation (4), the true theta is greater than or equal to the cut score, and the examinee is classified as a master. In this case, no loss occurs.

Table 1. Threshold Loss Function Defined for a Fixed-length Test

True Theta Level Decision Made	Theta < Cut	Theta ≥ Cut
Non-master	*0* (1)	*M* (3)
Master	*L* (2)	*0* (4)

Note: $-\infty < \theta < \infty$. Theta: examinee's true theta. *L*: the loss associated with a false positive error. *M*: the loss associated with a false negative error

For a variable-length mastery test, the loss of delivering one testlet is specified as a constant C. For the sake of simplicity, following Lewis and Sheehan (1990), this kind

of loss is assumed to be equal for each decision outcome and for each sampling of a testlet. Suppose the variable-length mastery test given to an examinee consists of *s* stages, and one of the available testlets is administered at each stage. The total losses of delivering *s* testlets can be represented as *sC.* As indicated in Table 2, this type of loss exists even when a correct mastery decision is made. Basically, loss functions that are associated with higher values of *L* and M, relative to *C*, will result in longer tests. *L* and *M* are usually expressed in the same units as those for *C*.

Table 2. Threshold Loss Function Defined at Stage *s* for a Variable-length Test

True Theta Level Decision Made	Theta < Cut	Theta ≥ Cut
Non-master	sC (1)	$M + sC$ (3)
Master	$L + sC$ (2)	sC (4)

Note: $-\infty < \theta < \infty$. Theta: examinee's true theta. *L*: the loss associated with a false positive error. *M*: the loss associated with a false negative error. *C*: the loss associated with administering one testlet

Decision Rule

The decision rule is designed to minimize posterior expected loss associated with the possible outcomes at each stage of testing. For computational simplicity, the threshold loss is used. That is, the same loss is incurred whether an examinee's score is close to the cut point or far from it.

Two levels of mastery are specified in CMT: θ_m is the lowest boundary at which an examinee is determined to be a master, and θ_n is the highest boundary at which an

examinee is determined to be a non-master. The posterior probabilities are calculated conditional on the observed number correct score in a testlet. The posterior probability distribution of mastery at the *i*th stage of testing is given as (Lewis & Sheehan, 1990),

$$P[\Theta = \theta_m \mid X_i] = \frac{P(X_i \mid \Theta = \theta_m)P_{mli-1}}{P(X_i \mid \Theta = \theta_m)P_{mli-1} + P(X_i \mid \Theta = \theta_n)P_{nli-1}}$$
$$\equiv P_{mli}$$

where X_i is the number correct score on the *i*th testlet;

P_{mli-1} is the prior probability based on the first *i* -1 stages given proficiency level θ_m;

P_{nli-1} is the prior probability based on the first *i* -1 stages given proficiency level θ_n;

$P(X_i \mid \Theta = \theta_m)$ is the conditional probability of observing a number correct score of X_i, given the proficiency level θ_m;

$P(X_i \mid \Theta = \theta_n)$ is the conditional probability of observing a number correct score of X_i, given the proficiency level θ_n.

For a fixed-length test, the posterior expected loss for passing at stage i is:

$$E[l(pass \mid \Theta)] = M * (1 - P_{mli})$$

And the posterior expected loss for failing is

$$E[l(fail \mid \Theta)] = L * P_{mli}$$

If $E[l(pass \mid \Theta)] < E[l(fail \mid \Theta)]$, the examinee is regarded as a master. Otherwise, the examinee is classified as a non-master.

The decision rule for variable-length tests is more complicated than that for fixed-length tests. In the variable-length tests, the losses associated with pass, fail and continue testing need to be calculated at each stage of testing. If the expected posterior loss plus the cost of another testlet is less than the expected loss prior to administration of the testlet, testing is continued; if the expected posterior loss is greater, or the testlet pool is exhausted, then testing is discontinued.

In the Lewis and Sheehan (1990) model, the prior and posterior distributions are defined only on two points, θ_m and θ_n. Glas and Vos (1998) extended the CMT to an IRT Rasch model that accommodates a continuous latent trait ability scale and posterior distributions. The procedure was renamed Adaptive Sequential Mastery Testing (ASMT). In ASMT, the mastery decision is made based on the entire vector of observed item responses, and the mode of the posterior distribution for the ability level is estimated by the Rasch model. The decision rule is the same as in the CMT procedure.

Glas and Vos's (1998) procedure follows the format of a sequential Bayes procedure where, at each stage, the previous posterior distribution of the unknown parameter serves as the new prior. After an item is administered, the likelihood is updated and combined with the previous posterior distribution to calculate a new posterior distribution. The expected loss function is computed from the posterior distribution at each stage. The procedure is repeated until a mastery decision is made.

In Glas and Vos's study, the prior was based on a normal distribution and the likelihood function followed a 1PL IRT model. Since the class of normal priors is not a conjugate form for the IRT model, there is no simple way to calculate the true posterior distribution. However, Glas and Vos derived a closed-form approximation for the true

posterior distribution by using the existence of sufficient statistics in the 1PL IRT model. Their procedure can be extended to the 2PL IRT model by either of two methods. First, estimate the expected loss function by computing the quadrature points of the grid of the posterior distribution integrals (Glas & Beguin, 1996). Second, use an exponential family model and elementary symmetric functions to approximate the expected loss (Verhelst & Glas, 1995; Verstralen, 1996). However, for the 3 PL IRT model, the two methods for the 2PL model cannot be applied. Monte Carlo simulation might be a promising method for the computation of the expected loss in this case (Glas & Vos, 1998).

Although the Bayesian procedure has not been expanded into the 3PL IRT model, it is the only procedure to provide direct control over Type I and Type II error rates. It would be interesting to evaluate the performance of this procedure under the 3PL model.

Model Comparison Studies

Many testing procedures could be used in mastery tests. However, which procedure is preferable in a certain situation is open to question. This section provides some comparison studies of these procedures.

SPRT and Adaptive Mastery Testing

Kingsbury and Weiss (1983) compared the SPRT and AMT procedures in terms of test length and accuracy of classification. In their study, four item pools were generated: uniform, b-variable, a- and b-variable, and a-, b-, and c- variable pools to represent different types of data. Each pool consisted of 100 items. For the uniform pool, a-, b-, and c- parameters were fixed at constants for all items (a=1, b=0, c=0.2). The b-variable pool varied from the uniform pool only in that the items had different b-parameters, ranging from [-2.5, 2.5] with 0.5 increments. The a- and b- variable pool

differed from the b-variable pool in that the a-parameters were allowed to vary in a range from .5 to 2. As for the a-, b- and c-variable pool, in addition to the a- and b- parameters varying, the c-parameters were set as .1, .2, or .3. Five hundred simulees were generated from a standard normal distribution. The cut score was defined as answering at least 60% of the total items correctly.

There were several conclusions in Kingsbury and Weiss's study. First, SPRT was more efficient and accurate when using a uniform item pool. For other types of item pools, SPRT yielded the shorter test length, but using the AMT procedure typically resulted in more accurate mastery decisions. Second, the AMT procedures resulted in higher correlations between estimated mastery status and true mastery status. However, the correlation was more directly associated with test length and type of item pool rather than testing procedure.

Kingsbury and Weiss's (1983) study has been criticized since they did not match the two procedures in a statistical sense. For SPRT, it is required to set four known parameters, α, β, θ_o and θ_1 before actual testing. The AMT procedure, on the other hand, needs to set the parameter, γ. It is not easy to match the SPRT and AMT procedures on the test parameters directly. However, Spray and Reckase (1996) adopted a sequence of "matching steps" to make the comparison more reasonable. First, they performed a simulation of the AMT procedure. The estimates of Type I and Type II errors from the AMT procedure became the input parameters for the SPRT simulation. After ensuring the expected Type I and Type II error rates for these two procedures were almost exactly the same, the average test lengths for these two procedures were compared.

Spray and Reckase used 200 items from a real pool, which were calibrated with the 3PL IRT model. Items were selected based on the maximum information at the decision point. Twenty five thousand examinees' ability parameters were generated from a rectangular distribution on the range from -3 to 3 in increments of .25.

The results showed that, in most cases, the SPRT procedure yielded shorter test lengths than the AMT procedure under the same degree of accuracy. It could also be inferred that, with similar test lengths, the SPRT procedure produced smaller classification errors than the AMT procedure.

There were some limitations in Spray and Reckase's study. The item selection rule for the AMT procedure was not the same as in previous studies. Kingsbury and Weiss's original idea for the AMT procedure was to select items at the maximum information of the most current theta estimate instead of selecting items based on the maximum information at the cut point. The different item selection method may have affected the efficiency of the AMT procedure, especially for those examinees whose thetas were far away from the decision point.

SPRT and Bayesian Sequential Testing

Yi et. al. (2001) conducted a series of simulation studies to compare the performance of the SPRT and CMT procedures. A total of 240 items from ACT Mathematics Tests were used as the item pool. The a-parameters in the pool had a mean of 1.08 and a standard deviation of 0.36. The average b-parameter was -0.06 with a standard deviation of 1.09. And the c-parameters had a mean of 0.17 and a standard deviation of 0.07. Twenty five hundred examinees were simulated from a standard normal distribution for each condition. Five cut points (-0.61, 0, 0.61, 1.05, 1.61) and two

widths of the indifference region (0.5, 1.0) were considered. For their study, two sets of loss parameters were investigated: L=20, M=40, C=1 and L=40, M=20, C=1, which indicated that the cost of making a false negative error was twice the cost of making a false positive error and the cost of making a false positive error was twice that of a false negative error, respectively. The parameter C was fixed to be 1 which implied that the costs of administering an additional testlet were rather small relative to the costs associated with incorrect decisions.

In the SPRT procedure, the items administered to an examinee were selected based on the maximum information at the cut point. For the CMT procedure, a testlet was randomly selected from a pool of parallel testlets to administer to an examinee. Each testlet had approximately equivalent content specifications and classical statistics. The results of their study showed that the error rates were functions of the cut points, widths of the indifference regions, and loss parameters. When the error rates were closely matched between the two procedures, the SPRT procedure had a shorter average test length than the CMT procedure, especially at the lower cut points. Generally, the SPRT procedure had a small advantage over CMT in terms of classification accuracy and efficiency (Yi. et al., 2001).

A similar comparison had been done by Vos (2000); however, he compared the SPRT and Bayesian Sequential Testing (BAYES). The BAYES procedure was essentially the same as the ASMT procedure except it assumed a beta prior for the ability of the examinees and a binomial likelihood for the item responses. This binomial model assumed that items were sampled at random and the probability of correctly responding was constant across items. To reflect the choice of the binomial model, one hundred

items with *b*-parameters all equal to 0 constituted the item pool. One thousand simulees were generated from a standard normal distribution. Vos's study focused on a variable-length mastery test. Three different maximum test lengths, 10, 25, and 50 items were randomly selected from the item pool. A value of 0.6 on the proportion correct metric was used as the cut score in this study. For the SPRT procedure, the indifference region was set at 0.2 and the values of Type I and Type II error rates were set equal to 0.1. For the BAYES procedure, the loss parameters for *L* and *M* were both set to be 100, and *C* was set as equal to 1. This implies that the costs of making false positive error and false negative error were assumed to be equal and the cost for administering another random item was assumed to be very small.

Simulations were conducted to compare these two procedures in terms of average test length, correspondence with true mastery status and classification accuracy. The results showed that the BAYES procedure produced lower classification errors and shorter average test lengths than the SPRT procedure for all test conditions.

In sum, many attempts have been made to compare testing procedures in the context of CCT; however, no consensus has been reached. Although these testing procedures yield similar outcome measures, the different requirements for certain test characteristics or parameters prior to testing in these procedures make meaningful comparisons difficult. For example, before administering the items, the AMT procedure requires the specification of confidence intervals, the SPRT procedure requires an indifference region, and the Bayesian procedure needs loss functions. These parameters are not at all similar to each other. This makes it very difficult to compare these procedures in an attempt to find the best one in a particular mastery testing situation (Bleiler, 1998).

Among these procedures, the Bayesian procedure has drawn increased attention for its flexibility to control the costs of each type of error by setting the appropriate loss parameters. In the SPRT procedure, although the error rates, α and β, can be set independently of one another to reflect different tolerance for each type of error, the actual error rates are not efficiently controlled (Wald, 1947). As for the AMT procedure, the error rates are mainly controlled by setting the confidence interval. However, the AMT procedure uses a symmetrical confidence interval and assumes equal costs for the two error types. There is no direct way to control each type of error when the relative costs of erroneous testing outcomes are different.

Although the Bayesian procedure seems to be a promising way of controlling different types of errors, little research has been done to investigate how effectively the actual error rates can be controlled. In addition, as mentioned in the last section, Bayesian procedures are limited to the 1PL and 2PL models currently. In order to extend the Bayesian procedure to the 3PL model, the Markov Chain Monte Carlo (MCMC) technique may be required. Some topics related to MCMC analyses are described in the next section.

MCMC Related Topics

Markov Chain Monte Carlo (MCMC) methods are used to fit complex Bayesian statistical models. Analytic methods for Bayesian inferences often require the evaluation of complex high dimensional integrals to obtain posterior distributions for the unobserved quantities of interest in the model. Simulation based methods avoid this problem by instead drawing samples from the posterior distribution and basing inferences on them.

MCMC is extremely useful when it is impossible or not computationally efficient to draw independent samples from the posterior distribution.

WinBUGS

WinBUGS (Spiegelhalter, Thomas, Best & Lunn, 2003) is a computer program which employs MCMC methods to facilitate Bayesian analysis in many application areas. It is written in Component Pascal running in Oberon Microsystems's Black Box environment. The user specifies a model either by drawing a directed graph (Lauritzen & Spiegelhalter, 1988) or by using an S-like language. The software then constructs the transition kernel for a Markov Chain to generate samples from the joint distribution of the unknowns in the model (Cowles, 2004). The user provides the data and initial values, the number of parallel MCMC chains to be run, the number of iterations, the unknown model quantities to monitor for analysis and the types of convergence evaluation and summaries. The final output provides numerical and graphical summaries of the requested model quantities. WinBUGS can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.

WinBUGS uses different sampling methods for different types of models. For the simplest case, a conjugate prior distribution is used with a standard likelihood to yield a posterior distribution from which the parameters in the model can be directly sampled. For more complicated cases in which it is not possible to get the samples directly from the posterior distribution, some forms of Gibbs sampling and Metropolis-Hastings sampling are used to sample from the posterior distribution.

Gibbs sampling algorithms are used by WinBUGS to construct the transition kernels for its Markov Chain samplers. At each iteration, a value of each model unknown

is drawn from its full conditional distribution. For nonstandard but log-concave full conditional distributions, derivative-free adaptive rejection sampling is used (Gilks, 1996).

The slice sampling algorithms (Neal, 1997) are used by WinBUGS for non log-concave densities on a restricted range. This has a tuning phase of 500 iterations. For nonconjugate continuous full conditional distributions with an unrestricted range, the random walk Metropolis algorithm (Metropolis et al, 1953) is used. The use of Metropolis algorithms in WinBUGS is based on a symmetric normal proposal distribution, whose standard deviation is tuned over the first 4000 iterations to obtain an acceptance rate of between 20% and 40%. More information about WinBUGS sampling methods can be found in the WinBUGS user manual (Spiegelhalter, Thomas, Best & Lunn, 2003).

Assessing Convergence of MCMC

An MCMC sampler can be regarded as converged when the samples obtained truly represent the stationary distribution of the Markov Chain. However, how to determine if it is appropriate to terminate sampling and use the samples to estimate the characteristics of the distribution of interest is not easy. This is because what has been produced by the MCMC procedures at convergence is not a single number or a distribution, but rather a group of correlated samples from a distribution. Correlations between the samples will slow the algorithm convergence and make the calculation of appropriate variances of model characteristics complicated (Cowles & Carlin, 1996).

Many researchers have tried to find a solution to the problem of determining MCMC algorithm convergence (Gelman & Rubin, 1992; Raftery & Lewis, 1992; Garren

& Smith, 1993; Geweke, 1992; Johnson, 1994; Liu, Liu, & Rubin, 1992; Mykland, et.al, 1995; Ritter & Tanner, 1992; Roberts, 1992; Yu, 1994; Yu & Mykland, 1994; Zellner & Min, 1995). Their approaches have generally applied some diagnostic tools to the output produced by the algorithms (e.g. comparing the empirical distributions of output produced at consecutive iterations). Because the stationary distribution is unknown in practice, it is difficult to conclude whether any samplers are converged or not. Still, some statisticians depend on such diagnostics since "a weak diagnostic is better than no diagnostic at all" (Cowles & Carlin, 1996). Cowles and Carlin (1996) and Mengersen et. al. (1999) provided a complete review of these convergence diagnostic approaches.

Although many diagnostic approaches have shown promise in detecting a lack of convergence, it is still possible that those methods fail in some cases. Cowles and Carlin (1996) suggested that the convergence conclusion should be based on a variety of diagnostic tools instead of just a single plot or statistic. For example, three to five parallel chains, with different starting values, should be investigated by overlaying their sampled values for each parameter on a common graph. Several diagnostics, such as Gelman and Rubin statistics and lag 1 autocorrelations, should also be examined before any convergence conclusion is made.

Gelman et. al. (2004) argued that even if an iterative process does appear to converge and has passed all tests of convergence, it is still possible that the drawn samples are far from the target distribution. Although it is never possible to conclude that the drawn samples are fully representative of an underlying target distribution, the convergence diagnostics still provide some evaluation of the algorithm's performance.

Although MCMC is a relatively new estimation method in item response theory, it has been employed widely in many measurement applications.(Albert, 1992; Patz & Janker, 1999a, 1999b; Bradlow, Wainer & Wang, 1999; Wainer , Bradlow, & Du, 2000) The previous studies indicated that MCMC has a great potential to be an efficient and versatile estimation procedure in item response theory. However, most of these studies were focused on using MCMC to estimate the item or ability parameters. It would be interesting to evaluate how much benefit can be obtained by employing the MCMC procedure in the context of CCT.

Summary

This chapter presented a review of literature related to mastery tests. The following topics were reviewed. In the first section, four components of mastery testing were described: item pool, item selection method, cut score and testing procedure. In mastery testing, the optimal type of item pool and the best item selection strategy depend on which testing procedures are used for mastery testing. The selection of the cut score involves a subjective judgment and different methods commonly result in different cut scores. Even so, it is still possible to determine a defensible cut score by organizing a credible panel of experts and using a systematic approach to collect their judgments. Different formats of mastery tests require different testing procedures. For the variable-length format, the SPRT and AMT procedures are frequently used. For the fixed-length format, the examinee's performance score (number correct score or EAP estimator) is directly compared with a cut score. If the examinee's score is equal to or exceeds the proportion correct cut score, the examinee is declared a master. Otherwise, the non-master decision is declared.

A Bayesian procedure can be used in both fixed-length and variable-length format mastery tests. This method was originally proposed by Lewis and Sheehan (1990) and modified by Glas and Vos (1998). In Glas and Vos's procedure, the standard normal distribution was used as a prior and the Rasch model was used for the likelihood function. Although their procedure can be extended to the 2-PL IRT model, their logic is not applicable to the 3PL IRT model. The problem occurs in the 3PL model because of the difficulty evaluating the high-dimensional integrals. However, the MCMC algorithm might be helpful in solving this problem.

The second section examined studies comparing the testing procedures. Conclusions about which procedures performed best are still inconclusive. For example, Kingsbury and Weiss (1983) found AMT to be more accurate than SPRT, but Spray and Reckase (1996) found just the opposite conclusion. Yi et al. (2001) found SPRT required fewer items than the Bayesian method, however, Vos (2000) showed that the Bayesian procedure produced lower classification errors and shorter average test lengths than the SPRT procedure when a uniform item pool was used.

The third section reviewed some topics related to MCMC. MCMC has the ability to reduce complex multidimensional problems to a sequence of much simpler problems, (Cowles & Carlin, 1996). Recently, MCMC has been used widely in many educational areas. However, not many studies have applied the MCMC within the context of mastery tests.

The current study aims to extend Glas and Vos's (1998) procedure to the 3PL IRT model using the MCMC method. Two conventional methods (Conventional-Proportion Correct and Conventional-EAP) are used as the baseline to evaluate the performance of

this new procedure under conditions where test lengths and item pool quality are systematically manipulated. The next chapter contains a complete description of the methods to be used in this investigation to address the issues raised above.

CHAPTER III

DESIGN OF STUDY

The purpose of this study was to extend Glas and Vos's (1998) ASMT procedure to the 3PL IRT model by using the MCMC method in the context of mastery tests. This extension, referred to here as the Bayesian decision-theoretic procedure, was examined to see if it produced acceptable accuracy in terms of Type I and Type II errors under a variety of testing conditions. In order to evaluate the performance of this new procedure, two conventional methods, namely Conventional-Proportion Correct and Conventional-EAP, were used as the baselines for comparison. Specially, the following research questions were addressed.

(1) To what extent are the classification errors controlled by setting constraints for false positive and false negative errors in the Bayesian decision-theoretic procedure for the 3PL?

(2) Does this Bayesian decision-theoretic procedure perform better than the conventional procedures in terms of the degree of accuracy for the 3PL?

(3) How does the item pool quality affect the performance of the testing procedures for making mastery decisions?

(4) Is the accuracy of the testing procedures affected by the length of the test?

(5) Is the accuracy of the Bayesian decision-theoretic procedure affected by the use of different IRT models such as the three-parameter model or the one-parameter model?

There are four sections in this chapter. In the first section, the specification of components in mastery tests and the data generation methods are described. The mastery

procedures investigated in this study are illustrated in the second section. The research design is summarized in the third section. In the last section, the criteria to evaluate the mastery procedures in this study are described.

Specification of Components in Mastery Tests and Data Generation

Specification of Components in Mastery Tests

Item Response Model

The item responses were generated either based on the three-parameter logistic model (3PL) or one-parameter logistic (1PL) model. The probability function for the 3PL model can be expressed as follows:

$$P_{ij}(X_{ij} \mid a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i) * \frac{1}{1 + \exp(-1.7 * a_i * (\theta_j - b_i))}$$

where P_{ij} is the probability of a correct response on item *i* for a person j;

a_i, b_i, c_i are the discrimination, difficulty and pseudo-guessing parameters, respectively for item i;

θ_j is an ability parameter for person j.

The 1PL model defines the probability of a correct response as:

$$P_{ij}(X_{ij} \mid b_i, \theta_j) = \frac{1}{1 + \exp(-1.7 * (\theta_j - b_i))}$$

Where P_{ij}, b_i and θ_j are defined the same as in the 3PL model.

Item Pools

Three item pools were used: two simulated item pools and one real item pool. Each pool contained 60 items. Two simulated pools, a high discrimination pool and moderate discrimination pool, were considered. For each item, three parameters were generated: discrimination (*a*-parameter), difficulty (*b*-parameter) and pseudo-guessing (*c*-parameter). These two simulated pools differed only in the values of their *a*-parameters, and they were generated based on the common characteristics of a real-world computerized test item pool discussed by Wang (1995).

For the high discrimination pool, *a*-parameters were generated from the normal distribution with mean equal to 1.9. The value 1.9 was chosen because some item parameter calibration programs set 2 as the upper bound for *a*-parameters (Wang, 1995). The *a*-parameters for the moderate discrimination item pool were also generated from a normal distribution, but with mean equal to 1. The value 1 was chosen because it was very close to discrimination parameters found in real item pools (Wang, 1995). For both item pools, the standard deviation of the a-parameters was set as 0.1. This small value was chosen in order to constrain the range of the a-parameters so that the quality of simulated item pool can be controlled. The *b*-parameters were generated from a normal distribution with mean equal to 0 and standard deviation equal to 2. This range was chosen since it provided enough coverage for the ability levels generated in this study. The *c*-parameters were set to be 0.15 because the value is close to the mean of c-parameters for real item pools. The resulting parameter values for these two simulated item pools were consistent with Urry's (1977) recommendations for the range of item parameters in the context of a computerized adaptive test.

The item parameters for the real item pool were calibrated using items from the ACT mathematics test. The ACT mathematics test consists of 60 multiple choice items. This test is designed to assess the mathematics skills that students have typically obtained from high school courses. Six main areas are included in this test: Pre-Algebra, Elementary Algebra, Intermediate Algebra, Coordinate Geometry, Plane Geometry and Trigonometry. A total of 60 discrete multiple-choice items from the ACT mathematics test were used. The item parameter estimates were calibrated from real data using the 3PL IRT model with the computer program BILOG (Mislevy & Bock, 1990).

Cut Point

Although the ACT mathematics test was not designed as a classification test, ACT test scores are frequently used as a standard to place students into different levels of mathematics courses or as a criterion to award scholarships. Some colleges use an ACT scale score equal to or greater than 22 as the minimum score for admission. For this study, a scale score equal to 22 (approximately equal to the 63rd percentile) was used as the cut score. In a normal distribution, the corresponding cut score on the theta scale was equal to 0.4.

Test Length

A fixed-length format was adopted to terminate the tests. Three test lengths, 20, 40, and 60 items were considered. They represented short-, medium- and long-length tests, respectively. For the short- and medium-length tests, the items were randomly drawn from the item pool with the stipulation that the items in the short-length test serve as the first portion of the medium-length test. As for the long-length test, the first 40 items were from the medium-length test.

Data Generation

Sets of ability parameters for five thousand examinees were generated to fit a standard normal distribution (mean = 0 and standard deviation =1.0). A normal distribution is commonly used for ability parameters in simulation studies. The item parameters along with the ability parameters were used in either the 1PL or 3PL logistic model to obtain the probability of a correct response for each examinee. The probability of a correct response (P_{ij}) was compared with a random deviate (d_{ij}) which was drawn from a uniform distribution in the range [0,1]. If $P_{ij} > d_{ij}$, the item was scored as correct (1); otherwise, the item was scored as incorrect (0). For the 3PL dataset, the responses were generated based on the a-, b- and c-parameters in each item pool. For the 1PL dataset, the responses were generated based on the b-parameters in the high discrimination item pool. The responses of examinees to 20, 40, and 60 items were simulated.

The simulation and data generation were conducted using the computer software program R (Downloaded from CRAN, version 2.2.1). R is public domain, open source software available from the website http://www.r-project.org/.

Mastery Procedures

Conventional Procedure- Proportion Correct

For this conventional procedure, if the examinee's score was larger than or equal to the proportion correct cut score, the examinee was declared a master. Otherwise, the non-master decision was declared. As mentioned previously, the cut score on the theta scale was set as 0.4. The corresponding cut score on the proportion correct scale was estimated through the test characteristic curve.

Conventional Procedure- EAP

In this Conventional-EAP procedure, the examinee's ability level was estimated by the EAP method after all items in the conventional test were administered. The EAP estimator was compared with the cut theta. If the EAP estimator was equal to or greater than the cut theta, the examinee was classified as a master; on the other hand, if the EAP estimator was smaller than the cut theta, the examinee was classified as a non-master.

Bayesian Decision-theoretic Procedure

There are two main components in the Bayesian decision-theoretic procedure. The first component is to construct the loss structure. The second component is the decision rule. The description of these two components and an application of a Markov Chain Monte Carlo (MCMC) procedure using WinBUGS for making optimal mastery decisions will be illustrated next.

Loss function

Generally speaking, a loss function specifies the total costs for each possible decision outcome. These costs usually incorporate all relevant psychological and social consequences associated with decisions. There are two kinds of loss functions used in this study: threshold loss function and linear loss function.

The specification of threshold loss functions is the same as Lewis and Sheehan (1990) described for the CMT procedure. As mentioned in Chapter II, for the CMT procedure, the expected losses associated with making a false positive error and a false negative error are specified by constants L and M, respectively. Although the threshold loss function is simple and has been frequently used in the literature, it may be unrealistic in some situations (Glas and Vos, 2006). A major criticism of threshold loss is that no

matter how far the examinee's ability level is from the cut score, the loss is assumed to be equal (See Figures 1 and 2).

Figure 1. Threshold Loss Function for Declaring an Examinee as a Master

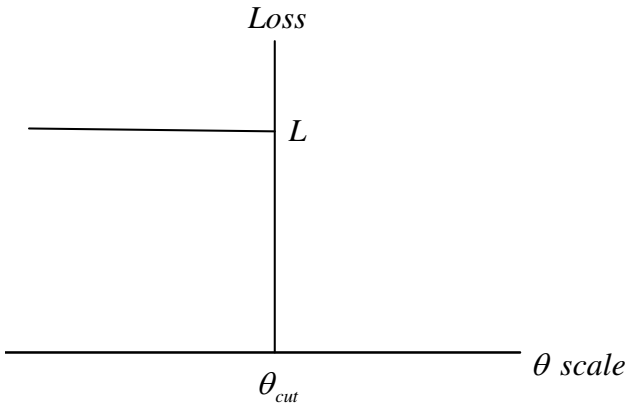
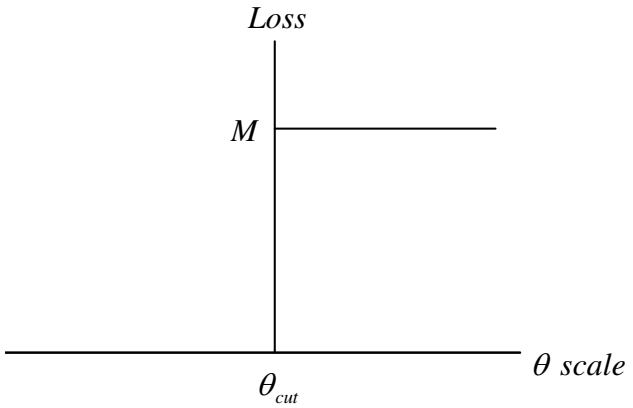


Figure 2. Threshold Loss Function for Declaring an Examinee as a Nonmaster



To overcome the limitation of threshold loss, van der Linden and Mellenbergh (1997) proposed a linear loss function for fixed-length mastery tests, which assumes the loss to be a continuous function of the examinee's theta level. For a linear loss function, examinees with different theta levels have different loss functions. If an examinee is declared a master but his/her true theta level is below the cut score, then the linear loss function is a decreasing function of theta (Figure 3). On the other hand, if an examinee is declared a nonmaster but his/her true theta level is above the cut score, then the linear loss function is an increasing function of theta (Figure 4). The expected losses associated with making a false positive error and a false negative error are specified as L(θ_{cut} - θ) and M(θ - θ_{cut}). Table 3 provides the linear loss functions for four possible outcomes.

Figure 3. Linear Loss Function for Declaring an Examinee as a Master

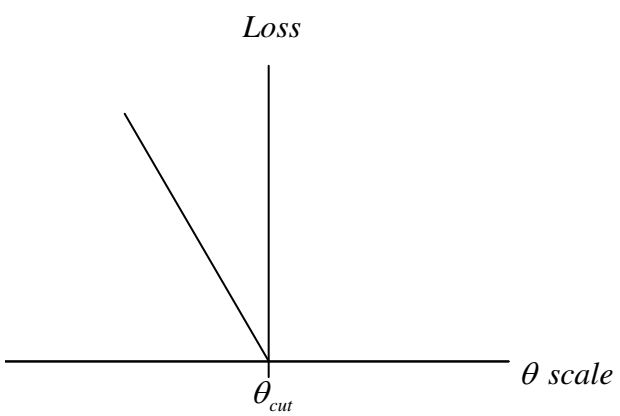


Figure 4. Linear Loss Function for Declaring an Examinee as a Nonmaster

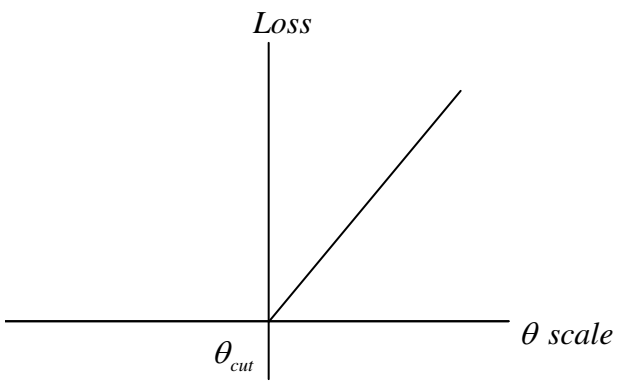


Table 3. Linear Loss Function Defined for a Fixed-length Test

True θ Level Decision Made	$\theta < \theta_{cut}$	$\theta \geq \theta_{cut}$
Non-master	*0*	M(θ - θ_{cut})
Master	L(θ_{cut} - θ)	*0*

Note: $-\infty < \theta < \infty$. θ_{cut} : cut score. *L* and M are specified the same as in threshold loss functions.

Decision Rule for a Fixed-length Mastery Test

There are two kinds of errors that can occur when making mastery decisions based on test scores: passing examinees who are true non-masters and failing examinees who are true masters (Hambleton et.al., 1978). The probabilities of these two classification errors are controlled through the use of loss functions. After administering a

set of items, a Bayesian decision rule for a fixed-length mastery test is used to minimize the posterior expected loss associated with the two classification decisions. Suppose $P(\theta \mid X_n)$ represents the posterior distribution of theta after *n* items are administered; E[Loss(master, $P(\theta \mid X_n)$)] and E[Loss(non-master, $P(\theta \mid X_n)$)] represent the posterior expected loss for making the mastery and non-mastery decisions, respectively. Based on Vos (2000), the examinee is classified as a master when E[Loss(master, $P(\theta \mid X_n)$)] < E[Loss(non-master, $P(\theta \mid X_n)$)]. Otherwise, the examinee is classified as a non-master.

When making the mastery decision, the only possible loss is to make a false positive error. In this case, the examinee's true ability is below the cutoff score, but the mastery decision was given. The posterior expected loss, E[Loss(master, $P(\theta \mid X_n)$)], can be estimated by integrating the loss function over the posterior distribution of theta for the range [$-\infty$, cutoff]. Note that for threshold loss, the loss function is specified as L; for linear loss, the loss function is specified as L(θ_{cut} - θ). Similarly, the loss for giving a non-mastery decision is to make a false negative error. In this case, the examinee's true ability is above the cut score, but the nonmastery decision was given. Thus, the posterior expected loss, E[Loss(non-master, $P(\theta \mid X_n)$)], can be estimated by integrating the loss function over the posterior distribution of theta for the range [cutoff, ∞]. Note that for the threshold loss, the loss function is specified as M; for linear loss, the loss function is specified as M(θ - θ_{cut}). Replacing these terms in the previous equations, under the threshold loss function, the examinee is declared a master when

$$L * \int_{-\infty}^{\theta_{cut}} P(\theta \mid X_n) d\theta < M * \int_{\theta_{cut}}^{\infty} P(\theta \mid X_n)\, d\theta$$

And under the linear loss function, the examinee is declared a master when

$$L * \int_{-\infty}^{\theta_{cut}} (\theta_{cut} - \theta) P(\theta \mid X_n) d\theta < M * \int_{\theta_{cut}}^{\infty} (\theta - \theta_{cut}) P(\theta \mid X_n)\, d\theta$$

The posterior distribution, $P(\theta \mid X_n)$, is the product of the prior and the likelihood. For this study, a vague prior for θ was used because it was desired that the prior distribution play a minimal role in the posterior distribution and inferences. A vague prior expresses vague or general information about a variable. The main reason to use this kind of prior distribution is to "let the data speak for themselves" so that the inferences are not affected by any information besides the current data (Gelman et.al, 2004). For this study, a noninformative prior was set for theta, and this prior followed a normal distribution with mean equal to μ and the precision equal to τ (the precision is the inverse of the variance). The hyperprior on μ was a normal distribution with mean equal to 0 and the precision equal to 0.01. In order to constitute a conjugate prior distribution, the hyperprior on τ was a gamma distribution with parameters (0.01, 0.01). In sum, the following priors were used:

$$P(\theta) = \text{normal}\ (\mu,\ \tau)$$

$$P(\mu) = \text{normal}\ (0,\ 0.01)$$

$$P(\tau) = \text{gamma}\ (0.01,\ 0.01)$$

The likelihood is the product of the probabilities associated with each examinee's item responses. Suppose there are a total of *N* examinees responding to n items. The likelihood function can be expressed as:

$$Likelihood = \prod_{j=1}^{N} \prod_{i=1}^{n} P_{ij}^{x_{ij}} (1 - P_{ij}^{1-x_{ij}})$$

where x_{ij} is the item response on item i for a person j , x_{ij} =1 or 0 ;

P_{ij} is the probability of a correct response on item i for a person j

based on either the 1PL or 3PL IRT model ;

There are many ways to estimate the integrals of the expected loss function for the 1PL model. However, there is no simple way for the 3PL model because the class of normal priors is not a conjugate form for the 3PL model. Some numerical integration methods can provide the solution to this kind of complicated integration; however, it can be solved easily by employing an MCMC procedure. For this study, a computer program, WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), was used to implement the MCMC procedure to estimate the expected loss functions E[Loss(master, $P(\theta \mid X_n)$)] and E[Loss(non-master, $P(\theta \mid X_n)$)] for both the 1PL and 3PL Bayesian models.

The following steps illustrate the procedures to implement the MCMC procedure using WinBUGS.

1. Establish the Bayesian model. The vague prior was set for the ability parameter and the likelihood function was set as either the 1PL or 3PL model.
2. Create a Boolean variable (b_j^*) at each MCMC iteration; if the value drawn for θ was less than the cut score, b_j^* equals 1; otherwise, b_j^* equals 0.
3. Create two variables, Lm[j] and Ln[j], which were used to estimate the posterior expected loss for making a mastery decision and a nonmastery decision, respectively.

Under a threshold loss function,

Lm[j] <- b_j^* * L

Ln[j] <- (1-b_j^*) * M

Under a linear loss function,

Lm[j] <- b_j^* * L* ($\theta_{cut} - \theta_j$)

Ln[j] <- (1-b_j^*) * M* ($\theta_j - \theta_{cut}$)

4. Run three parallel Markov Chains with different starting values for μ and τ.
5. Evaluate convergence for each parameter in the WinBUGS model. The following outputs were used to evaluate sampler performance:
    a. history plots
    b. autocorrelation plots
    c. Brooks, Gelman, and Rubin diagnostic plots
    d. Monte Carlo errors.
6. After convergence is assessed for each parameter, record the mean of Lm[j] and Ln[j] for each examinee. These two means were used to estimate the posterior expected loss.
7. The output from WinBUGS was then read into a program for analyzing the mastery/nonmastery decisions for each examinee.

Loss function in the Bayesian Decision-theoretic Procedure

For both the threshold loss function and the linear loss function, three conditions were considered in this study: 2L=M, L=M, L=2M, which represent: the cost of making a false negative error was twice as serious as making a false positive error; the costs of

making a false positive error and a false negative error were equal, and the cost of making a false positive error was twice as serious as making a false negative error, respectively. These ratios have been used in previous studies (Lewis & Sheehan, 1990; Yi et. al, 2001; Vos, 2000)

Research Design

A series of Monte Carlo simulations were conducted for this study. The Bayesian decision-theoretic procedure with different loss functions was compared with two conventional procedures (Conventional- Proportion Correct and Conventional- EAP) with regard to test length and item pool quality. In addition, the efficacy of employing the 3 PL procedure compared to that of the 1PL procedure using the Bayesian decision-theoretic procedures was also evaluated. Most of these experimental conditions have been described in the previous sections; however, this section summarizes all of the conditions manipulated in this study.

Seventy-two different combinations of conditions were investigated (3 item pools $\times$3 test lengths $\times$ 8 testing procedures) to answer research questions (1) to (4). These simulations were run using simulated responses generated from the unidimensional 3PL model for 5000 simulated examinees. For each of the testing procedures, a simulation was carried out using three different test lengths for each of the three item pools. Table 4 summarizes the experimental conditions employed to examine the performance of Bayesian decision-theoretic procedure using the 3PL model.

In order to answer research question (5), which required a comparison of the efficacy of employing the 3PL model or the 1PL model using the Bayesian decision-theoretic procedures, another 72 (3 test lengths $\times$ 6 Bayesian procedures $\times$ 2 IRT models

× 2 datasets) simulations were run. For each Bayesian procedure, the likelihood function was applied to either the 1PL or the 3PL IRT model. Two types of datasets were employed to facilitate the comparison. The first dataset was generated to fit the 3PL model using item parameters from the high discrimination item pool. The second dataset was generated to fit the 1PL model using the b-parameters from the high discrimination item pool. Table 5 summarizes the experimental conditions for comparing the Bayesian decision-theoretic procedure using either the 1PL or 3PL model.

Table 4. Experimental Conditions to Evaluate the Bayesian Decision-theoretic Model

	Item Pool								
	High Discrimination Pool			Moderate Discrimination Pool			Real Pool		
Testing Procedure	Test Length			Test Length			Test Length		
	N=20	N=40	N=60	N=20	N=40	N=60	N=20	N=40	N=60
Conventional-Proportional Correct									
Conventional-EAP									
Threshold loss									
Bayesian L=2M									
Bayesian 2L=M									
Bayesian L=M									
Linear loss									
Bayesian L=2M									
Bayesian 2L=M									
Bayesian L=M									

Table 5. Simulation Conditions for Comparing the 1PL and 3PL Models Using the Bayesian Procedures

	3PL Dataset			1PL Dataset		
Bayesian Procedure	N=20	N=40	N=60	N=20	N=40	N=60
Threshold Loss L=2M						
3PL						
1PL						
Threshold Loss 2L=M						
3PL						
1PL						
Threshold Loss L=M						
3PL						
1PL						
Linear Loss L=2M						
3PL						
1PL						
Linear Loss 2L=M						
3PL						
1PL						
Linear Loss L=M						
3PL						
1PL						

Criterion Indices

Evaluation of Testing Procedures

For each simulation, the outcomes of interest were (1) the percentages of correct classifications, (2) false positive error rates, (3) false negative error rates, and (4) phi correlations between the true classification status and observed classification status. In order to calculate these indices, the true masters and true non-masters needed to be defined first. The examinee's true theta level was compared with the cut score. If his/her true theta level was equal to or larger than the cut score, the examinee was truly a master; otherwise, the examinee was truly a nonmaster.

When an examinee finished the pre-specified number of items, he/she was determined as a master or a non-master by the procedure outlined previously. The estimated mastery status was compared to the true mastery status. If the estimated mastery status matched the examinee's true status, a correct decision was made. On the other hand, if a mastery decision was made but the examinee's true status was a non-master, a false positive error was recorded. If a non-master decision was made but the examinee's true status was a master, a false negative error was recorded. For each mastery procedure, the percentages of false positive errors, false negative errors and the correlation between the true mastery status and observed mastery status were calculated. Table 6 gives a summary for these conditions.

Table 6. True Status and Decision Made by Mastery Procedures

	True Status	
Decision	Master	Non-master
Master	Correct Masters Identified	False Positive Error
Non-master	False Negative Error	Correct Non-masters Identified

The values in Table 6 were used to calculate an odds ratio. The odds ratio can be estimated as

$$o = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Where n_{11} represents the proportion of masters correctly identified;

n_{12} represents the proportion of false positive errors;

n_{21} represents the proportion of false negative errors;

n_{22} represents the proportion of correct non-masters correctly identified.

Based on the odds ratio, chi-square tests were implemented to test the overall degree of homogeneity among different test procedures. Because there were 8 testing procedures, the degrees of freedom for the chi-square statistics was 7. If the value of a chi-square test statistic was significant, post hoc pairwise tests were used to identify those groups in which the association was different from that in the other groups.

CHAPTER IV

RESULTS

In this chapter, the results of the simulations are summarized and presented. In this study, the Bayesian decision-theoretic procedure with different loss functions was compared with two conventional procedures (conventional- Proportion Correct and conventional- EAP) with regard to different simulation conditions. A series of simulation conditions were investigated, including three item pools (high discrimination item pool, moderate discrimination item pool and real item pool), three test lengths (20, 40, and 60 items) and eight testing procedures (conventional-Proportional Correct, conventional-EAP, Bayesian decision-theoretic procedure with threshold loss L=2M, Bayesian procedure with threshold loss 2L=M, Bayesian decision-theoretic procedure with threshold loss L=M, Bayesian decision-theoretic procedure with linear loss L=2M, Bayesian decision-theoretic procedure with linear loss 2L=M, Bayesian decision-theoretic procedure with linear loss L=M). These simulations were carried out using simulated responses generated from a unidimensional IRT model for 5000 simulated examinees.

This chapter contains four sections. In the first section, the descriptive statistics of the three item pools are presented. In the second section, the convergence of the WinBUGS models for Bayesian decision-theoretic procedures is examined. Four diagnostic measures (see chapter III for detailed descriptions) were used to evaluate the sampler performance: (1) Monte Carlo errors; (2) history plots; (3) autocorrelation plots; and (4) Brooks, Gelman, and Rubin diagnostic plots. In the third section, the results of

the conventional-Proportion Correct, conventional-EAP, and Bayesian decision-theoretic procedures are compared. To compare these three methods, four outcomes of interest were described: (1) the percentages of correct classifications; (2) false positive error rates; (3) false negative error rates; and (4) phi correlations between the true classification status and observed classification status. The overall differences between these methods were examined using chi-square tests. In the fourth section, comparisons of the 1PL and 3PL IRT models are made in terms of the percentages of correct classifications and phi correlations. A summary is provided at the end of this chapter.

Description of Item Pools

In order to evaluate the effect of item pool characteristics on mastery testing, three types of item pools were included in this study. Each of these three pools contained 60 items. For the high discrimination item pool, the mean of the a-parameters was 1.922, and the values ranged from 1.747 to 2.122; the mean of the b-parameters was 0.274; and the c-parameters were fixed at 0.15. The mean of the a-parameters in the moderate discrimination item pool was equal to 0.982 and the distributions for the b- and c-parameters were exactly the same as those in the high discrimination item pool. The distribution of the real item pool was different from the two simulated item pools. The mean of the a-parameters for the real item pool was slightly lower, and the range of the b-parameters was narrower than the corresponding values for the two simulated item pools. Table 7 shows the descriptive statistics for these three item pools.

Table 7. Descriptive Statistics for the Item Parameters of the Three Item Pools

	Mean	St. Dev	Min	Max
High Discrimination Item Pool				
a- parameter	1.922	0.082	1.747	2.122
b- parameter	0.274	2.036	-4.685	5.149
c- parameter	0.150	0	0.150	0.150
Moderate Discrimination Item Pool				
a- parameter	0.982	0.082	0.847	1.222
b- parameter	0.274	2.036	-4.685	5.149
c- parameter	0.15	0	0.150	0.150
Real Item Pool				
a- parameter	0.892	0.247	0.419	1.481
b- parameter	0.228	0.987	-1.812	1.949
c- parameter	0.157	0.065	0.084	0.425

Convergence of WinBUGS Models

The Bayesian decision-theoretic procedure was implemented in WinBUGS 1.4.1 (Spiegelhalter, et al., 2003). Although a number of MCMC convergence diagnostic methods have been proposed in the literature, none is universally accepted. Cowles and Carlin (1996) provided some examples to illustrate where different diagnostic criteria failed to detect a lack of convergence. Thus, it is suggested that a variety of diagnostic and visual inspections of the trace plots should be used to evaluate the MCMC convergence (Cowles et. al., 1996). In this section, MC error, history plots, autocorrelation plots, and BGR plots are examined as bases for evaluating the convergence of the three parameters in the Bayesian model: theta, mu and tau.

For each simulation condition, there were 5,000 thetas, one tau and one mu that needed to be assessed in the evaluation of convergence. As a result, it is impossible to present all of the diagnostic measures used in this section. For ease of presentation, only some of the diagnostic measures assessed with the threshold loss function L=M (which

represents the costs of making a false positive error and a false negative error are equal) and test lengths equal to 20 in the high discrimination item pool are shown. However, the other simulation conditions produced essentially the same results.

MC Error

Table 8 shows the descriptive statistics for the Bayesian procedure using the threshold loss function L=M with the high discrimination item pool. In this table, "Node" defines the variable that was monitored in WinBUGS. Three variables were monitored: theta, mu and tau. Theta represents the estimated ability level for each examinee. For example, theta [1] represents the estimated ability level of the first examinee (1.699, see Table 8). In total, five thousand thetas were evaluated in this study. For convenience of illustration, only the information of the first ten examinees is shown. Table 8 also contains the values of mu and tau, which represent the mean and precision of the estimated ability levels of the five thousand examinees, respectively. For each monitored variable, three statistics are presented: mean, standard deviation and MC error. "Mean" is the mean of the posterior distribution, "standard deviation" is the standard deviation of the posterior distribution, and MC error is the Monte Carlo standard error of the mean.

MC errors are usually used to evaluate the convergence for the variable. MC error is purely technical (like the standard error of the mean but adjusted for autocorrelation) and can be made as small as desired by increasing the number of iterations. This value was estimated using Roberts's (1996) batch means methods. Generally, as the number of MC samples increases, the MC error decreases. For this study, the number of samples for burn-in (the number of initial iterations that needed to be discarded in order to ensure that the remaining samples were drawn from a distribution close enough to the true stationary

distribution to be usable for estimation and inference) was 1000, and another 4000 samples were used to compute the posterior statistics. An ad hoc rule of thumb is that the number of samples is adequate when MC error is less than 1/20 of a standard deviation.

As shown in Table 8, the MC errors were all less than 1/20 of a standard deviation. This indicates that the 4000 samples were sufficient for computing the posterior statistics.

Table 8. Descriptive Statistics for the Bayesian Procedure Using the Threshold Loss Function L=M

Node	Mean	Standard Deviation	MC Error
theta[1]	1.699	0.467	0.016
theta[2]	0.821	0.391	0.016
theta[3]	0.372	0.390	0.016
theta[4]	1.023	0.389	0.015
theta[5]	-0.449	0.365	0.012
theta[6]	2.451	0.612	0.024
theta[7]	1.622	0.455	0.015
theta[8]	0.376	0.357	0.015
theta[9]	1.340	0.426	0.017
theta[10]	0.569	0.349	0.015
mu	0.103	0.017	0.000
tau	0.860	0.021	0.001

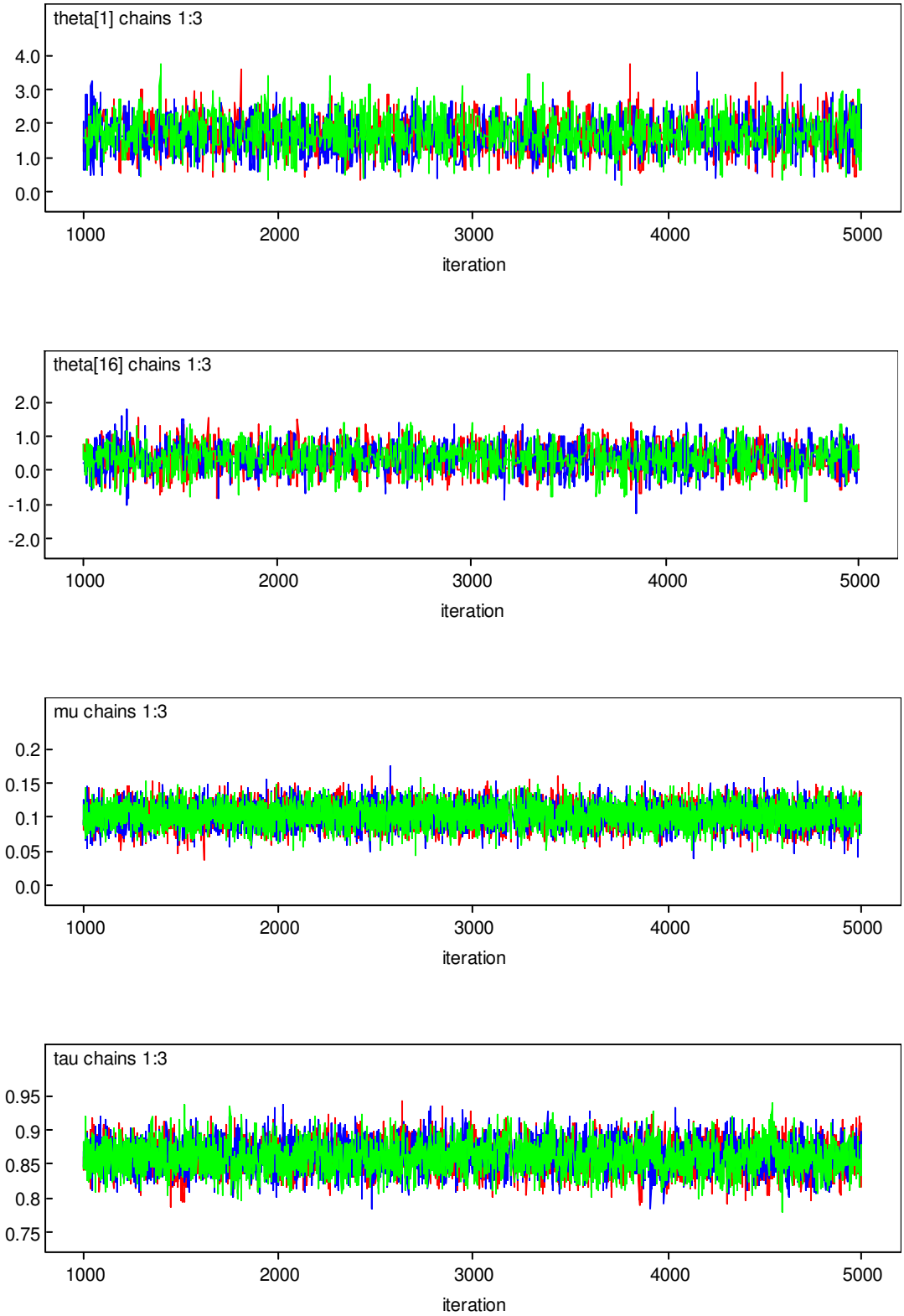
History Plot

Figure 5 contains the history plots for these three parameters: theta, mu and tau. As defined previously, theta represents the estimated ability level for each examinee. Mu and tau represent the mean and precision of the estimated ability level of all examinees, respectively. The horizontal axis in Figure 5 represents the number of MC iterations, while the vertical axis represents the range of values for the monitored parameter.

There are four history plots in Figure 5. In the upper left corner of the first history plot, the expression " theta [1] chains 1:3 " is presented. This means that this history plot was for estimating theta [1] and three chains are shown in this plot. As specified in Chapter III, these three chains use the same Bayesian model but different starting values. There were 5000 iterations and the range of these iterations was from 0 to 4. Similarly, the second history plot represents the iteration history for another example of an examinee's ability estimate (16th examinee). The third and fourth history plots show the iterative history for mu and tau. As shown in Figure 5, the Bayesian models in these history plots appear to have converged since the three chains essentially overlap each other and could not be easily differentiated. These overlapping chains provided evidence of convergence for those monitored variables. More examples of history plots for the Bayesian procedure using the threshold loss function L=M are presented in Appendix A. In total, there were 5,002 history plots (5000 thetas, 1 mu and 1 tau).

There are many other similar history plots that were examined; however, the remaining history plots were very similar to the history plots presented.

Figure 5. History Plots for the Bayesian Procedure Using the Threshold Loss Function L=M and Test Length =20



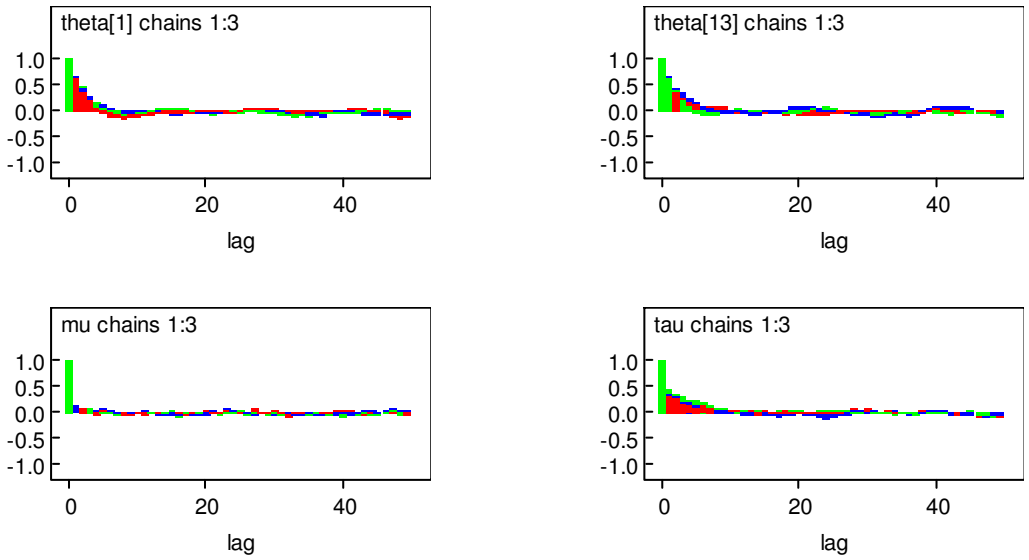
Note: Each shade represents one chain in the history plot

Autocorrelation Plots

Figure 6 shows the autocorrelation plots of the Bayesian Procedure using the threshold loss function L=M. The vertical axis represents autocorrelation, while the horizontal axis represents the lags. The lags were obtained by partitioning the long chains of iterations into continuous batches of equal length, such that the sample means between batches were approximately independent. Typically, the levels of autocorrelation decrease with increasing numbers of lags in the chain. WinBUGS plots the level of autocorrelation for up to 50 lags. Figure 6 shows some examples of the autocorrelation plots. In the upper left corner of the first autocorrelation plot, the expression "theta [1] chains 1:3" is presented. This means that this autocorrelation plot was for estimating theta [1] and three chains are shown in this plot. These three chains use the same Bayesian model but different starting values. Each chain is represented by one shade in the autocorrelation plot. Although it is hard to differentiate these three chains, it can be seen that for each chain, the autocorrelation dropped close to zero soon after beginning of sampling. Similarly, the second autocorrelation plot was for estimating the theta value for another examinee (13th examinee). The third and fourth history plots show the autocorrelation plots for the mu and tau. Each of the autocorrelations in these four plots dropped quickly after the beginning of the sampling, implying that the model converged quite well.

For each simulation, there were 5,002 autocorrelation plots. Again, it is not possible to show all of the autocorrelation plots here; however, all of the other plots were similar to the ones presented. More examples of autocorrelation plots for the Bayesian procedure using the threshold loss function L=M are presented in Figure A.2.

Figure 6. Autocorrelation Plots for the Bayesian Procedure Using the Threshold Loss Function L=M and Test Length = 20



Note: Each shade represents one chain in the autocorrelation plot

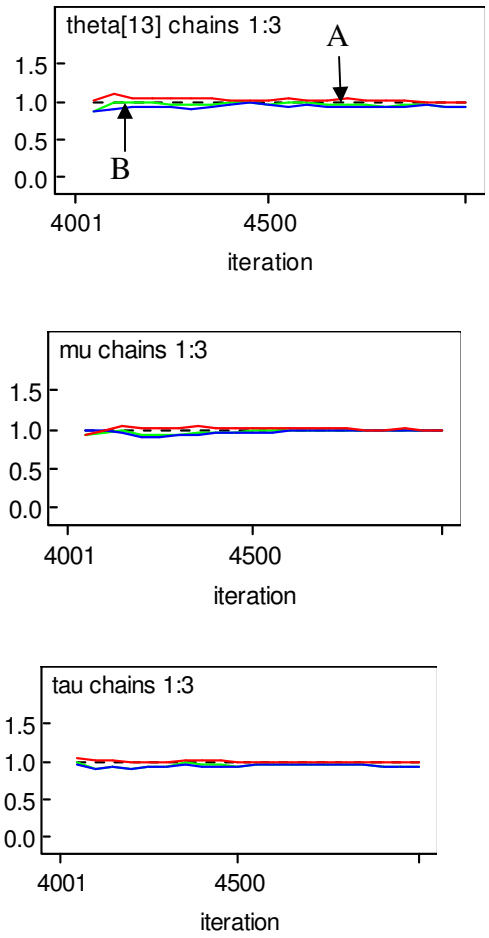
BGR Plots

A BGR plot (Brooks and Gelman, 1998; Gelman and Rubin,1992) is a diagnostic tool that is used to compare the variances among and within multiple chains. There are three horizontal lines in each BGR plot. These three lines represent: (1) the normalized width of the central 80% interval constructed from pooled runs; (2) the normalized average width of the 80% intervals constructed for each individual run; and (3) the ratio R which is equal to the ratio of values (1) and (2). The vertical axis represents the ranges for the values of intervals or ratio of intervals. The MCMC samplers are considered converged if the ratio R is close to 1, and both the pooled and within interval widths are stabilized. There are three examples of BGR plots presented in Figure 7. For all BGR plots presented in Figure 7, it can be seen that after 4000 iterations, the line for R was

very close to 1, and both the pooled and within intervals were stabilized and close to each other.

There were 5002 BRG plots for each simulation. It is not possible to show all of the BGR plots here; however, all of the other plots were similar to the ones presented here. More examples of BGR plots for the Bayesian procedure using the threshold loss function L=M are presented in Figure A.3.

Figure 7. BGR Plots for the Bayesian Procedure Using the Threshold Loss Function L=M and Test Length = 20



Note: The dashed line represents the ratio equal to 1. A: this line represents the ratio R. B: The two almost overlapping lines represents the intervals of pooled and within runs.

Comparison of Mastery Procedures

In this section, the following results are presented: percentages of correct classifications, false positive error rates, false negative error rates, and phi correlations between the true classification status and estimated classification status. For each of these four indices, the Bayesian decision-theoretic method and two conventional methods were compared for different test lengths and item pools. A series of figures illustrates these findings. For each figure, the horizontal axis represents either the length of the test (20, 40, or 60 items) or the type of item pool (high discrimination item pool, moderate discrimination item pool, or real item pool). The vertical axis is either the percentage of correct classifications, false positive error rates, false negative error rates, or phi correlations. The figures illustrate the impact of the factors examined in this study, including test length, item pool quality, mastery decision procedures, and loss functions.

These figures were examined to address three factors of interest. The first factor was the overall impact of test length and item pool quality on the testing procedures. The second factor was the difference between the Bayesian decision-theoretic methods and the conventional methods. The third factor was the impact of different loss functions within the Bayesian decision-theoretic methods on the accuracy of decision making. As mentioned in Chapter III, linear loss functions and threshold loss functions were investigated in this study. Each type of loss function was investigated with three different combinations of relative weights of false positive errors and false negative errors.

Percentage of Correct Classifications

Test Lengths

As seen in Figures 8 to 10, generally for a given test length, the range of percentages of correct classifications across different mastery testing procedures was approximately 6%. Figures 8 to 10 compare the percentages of correct classifications for each mastery testing procedure with different test lengths and types of item pools. The horizontal axis represents three different test lengths (20, 40, and 60 items) and the vertical axis is the percentage of correct classifications. In comparing the percentages of correct classifications for these three different test lengths, as expected, the longest tests were found to have the highest percentages of correct classifications. The next highest percentages were for the 40-item tests, followed by the 20-item tests. In most cases, when the test length increased from 20 to 40, the percentages of correct classifications increased dramatically. However, when the test length increased from 40 to 60, there was not much increase in the percentages of correct classifications. For high discrimination item pools, the impact of test length on the percentage of correct classifications was relatively small compared to the moderate and real item pools for most of the mastery procedures.

Item Pool

Figures 11 to 13 compare the percentages of correct classifications with different types of item pools for different testing procedures. These figures show that overall, the high discrimination item pool had the highest percentages of correct classifications, followed by the real item pool, and then the moderate discrimination item pool. The

differences among the percentages of correct classifications for these three item pools were smaller for test length of 60.

Figure 8. Percentages of Correct Classifications at Each Level of Test Length for the High Discrimination Item Pools

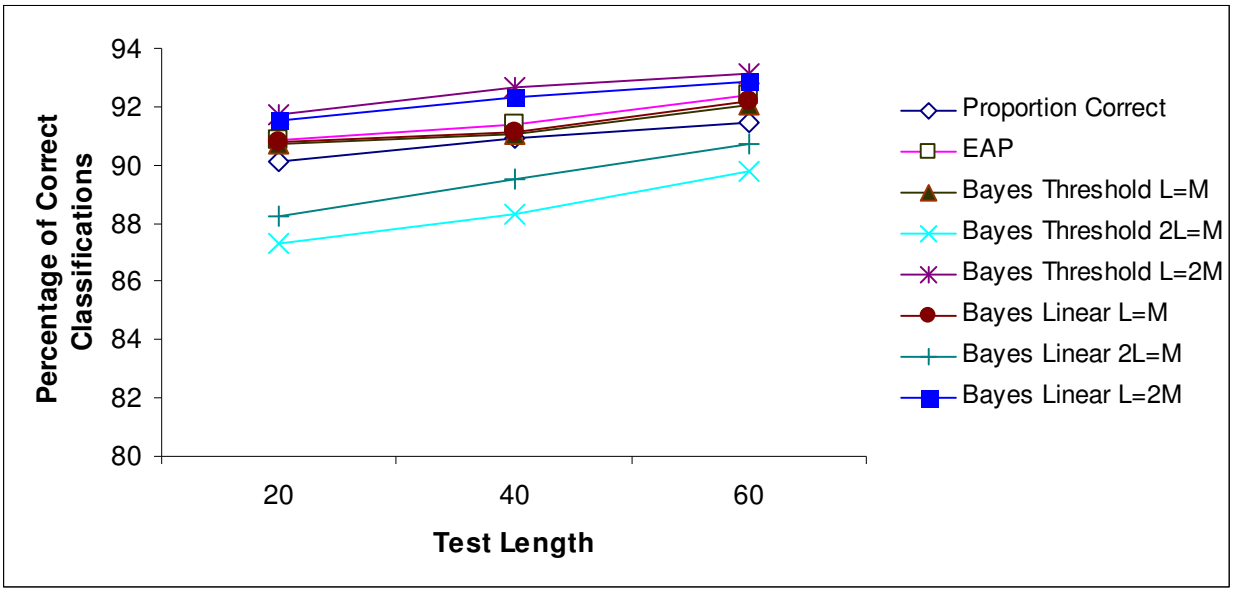


Figure 9. Percentages of Correct Classifications at Each Level of Test Length for the Moderate Discrimination Item Pools

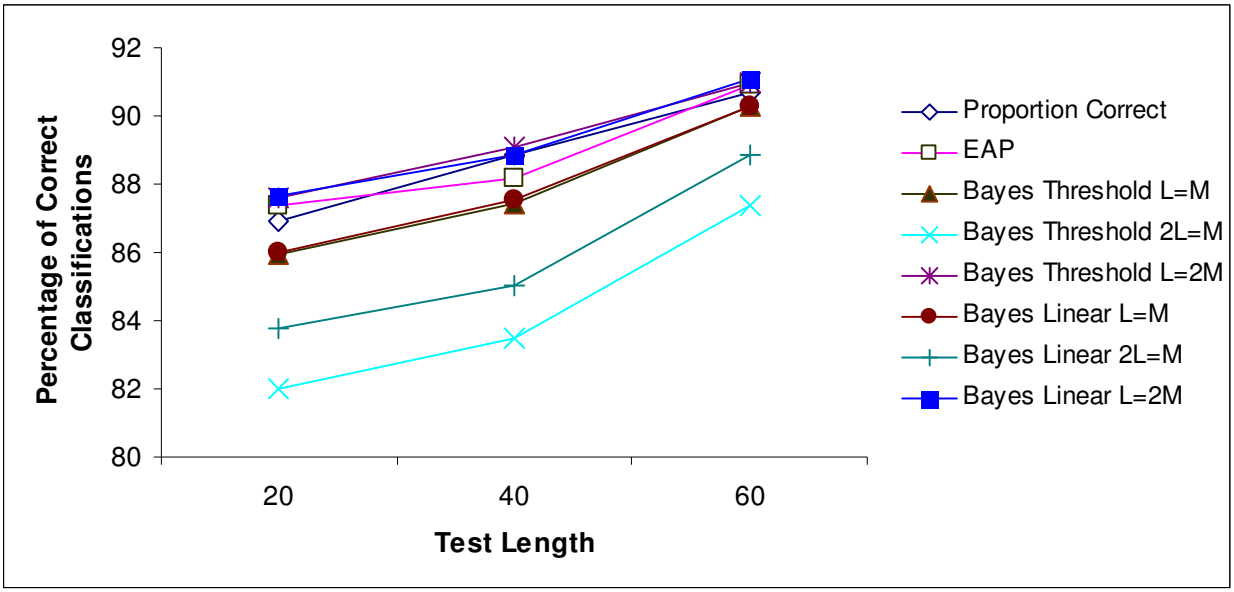


Figure 10. Percentages of Correct Classifications at Each Level of Test Length for the Real Item Pools

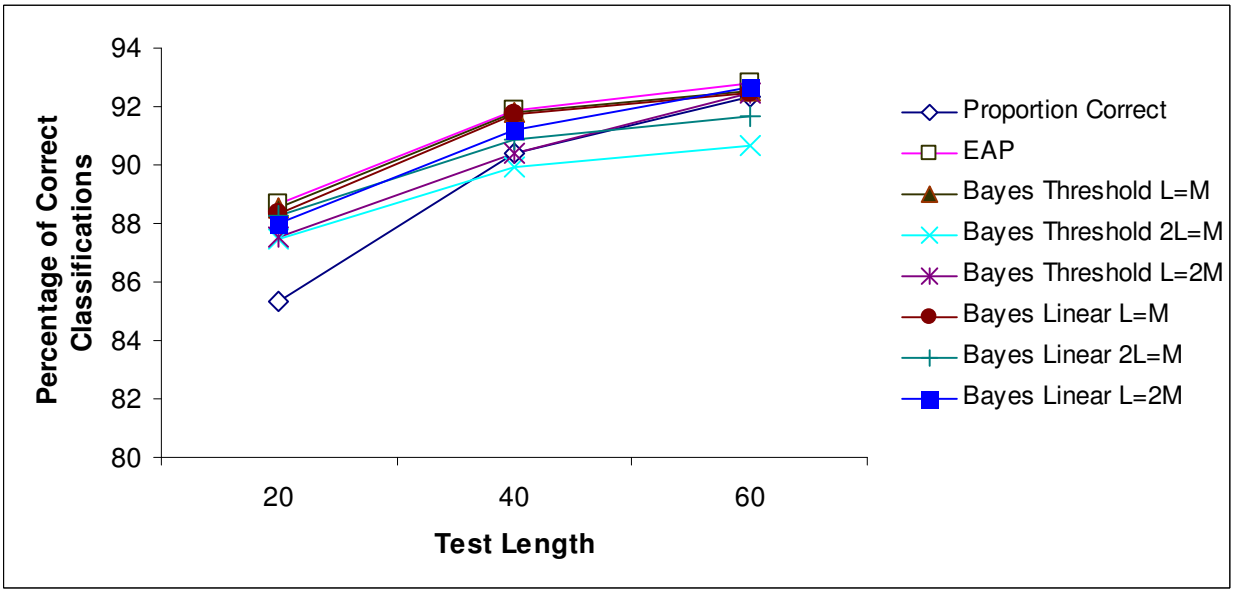


Figure 11. Percentages of Correct Classifications for Test Length Equal to 20 for Different Types of Item Pools

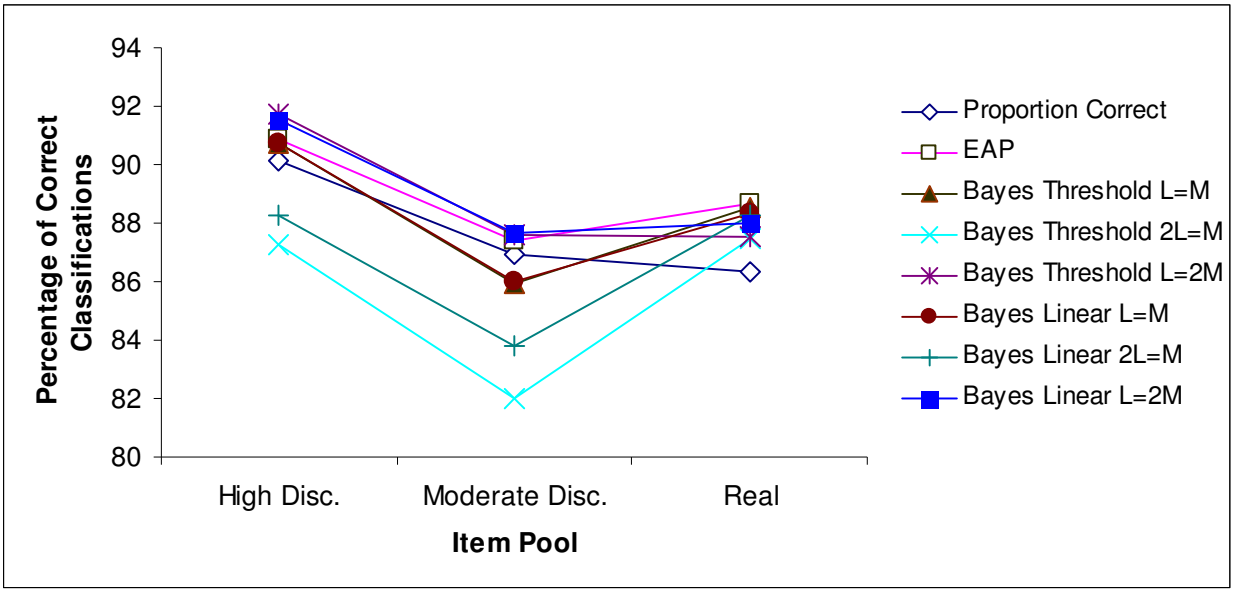


Figure 12. Percentages of Correct Classifications for Test Length Equal to 40 for Different Types of Item Pools

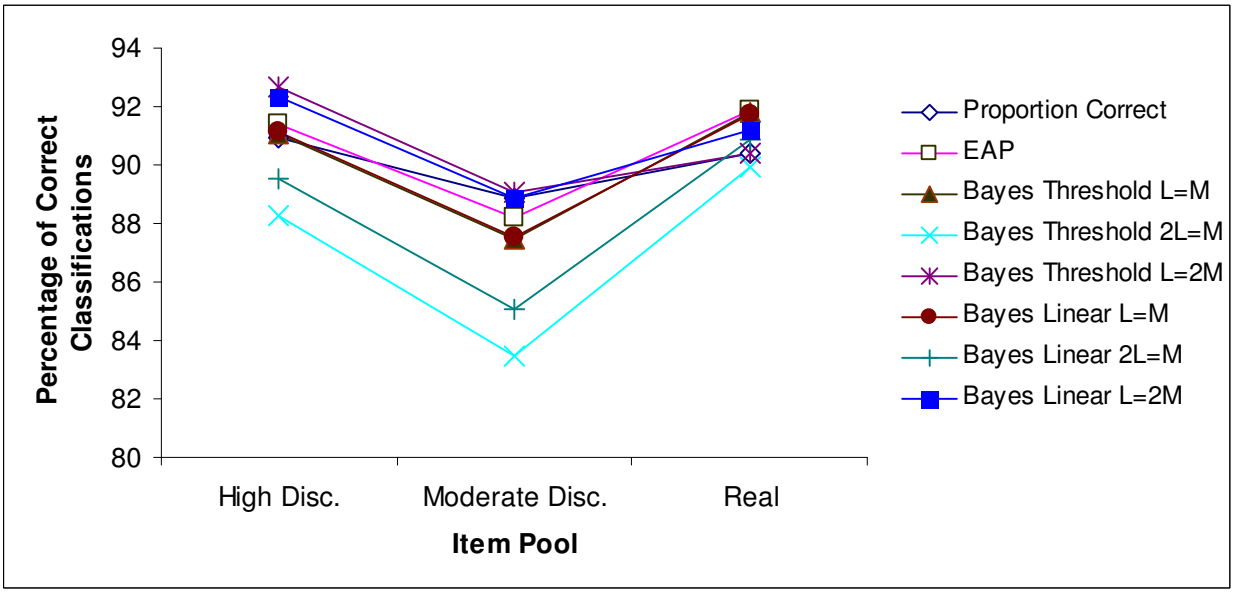
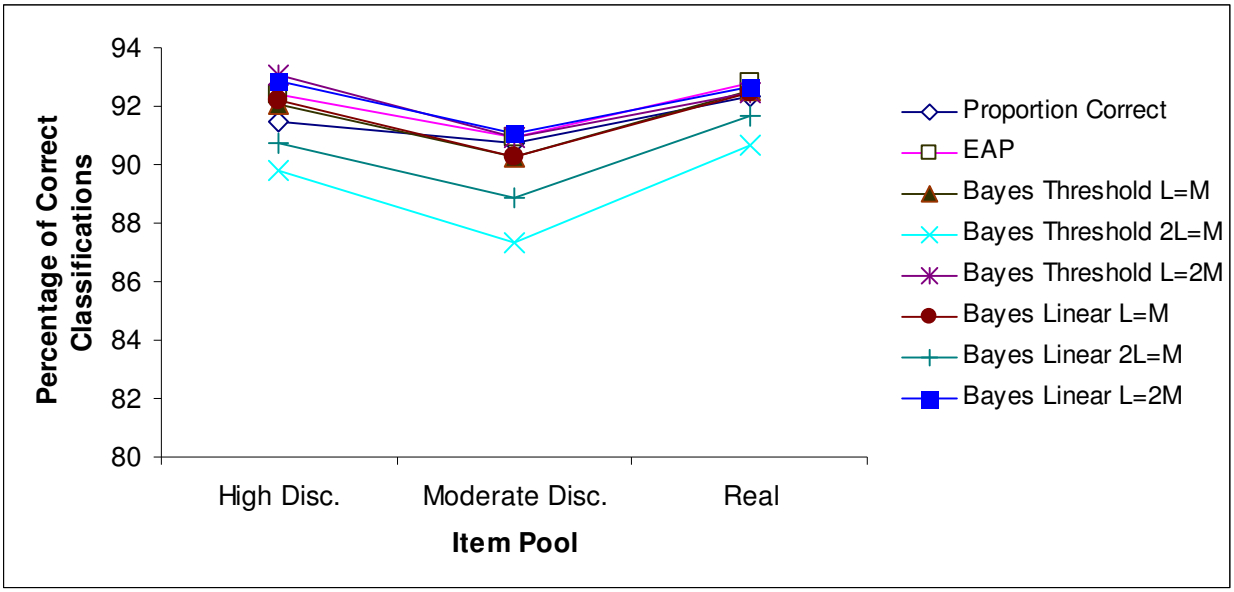


Figure 13. Percentages of Correct Classifications for Test Length Equal to 60 for Different Types of Item Pools



Comparison of Testing Procedures

Generally, the Bayes threshold 2L=M (which represents the cost of making a false negative error being twice as serious as making a false positive error) and the Bayes linear 2L=M yielded the lowest percentages of correct classifications. The only exception was for the real item pool with test length equal to 20. In this case, the conventional-Proportion Correct method had the lowest percentages of correct classification (see Figures 8 to 13).

The Bayesian decision-theoretic methods were compared to the two conventional methods. The figures showed that the Bayes threshold and Bayes linear with L=2M had higher percentages of correct classification rates, compared to the conventional EAP and the conventional- Proportion Correct methods for high and moderate discrimination item

pools. For the real item pool, however, the conventional-EAP method had a slightly higher percentage of correct classification rates than those of any of the Bayesian decision-theoretic methods. The differences among these mastery methods became smaller with increasing test lengths.

Among the Bayesian methods, the highest percentages of correct classification rates were associated with the Bayes threshold L=2M, and the lowest percentages of correct classification rates were associated with the Bayes threshold 2L=M. No obvious pattern was found to show whether threshold or linear loss functions worked better. However, these results reveal that in most cases, for both the Bayes threshold and the Bayes linear method, L=2M tended to have the higher percentage of correct classification rates, followed by L=M and 2L=M. The overall levels of these percentages are quite encouraging for the MCMC techniques applied here.

False Negative Error Rates

Test Length

Figures 14 to 16 compare the false negative error rates for different testing procedures with different test lengths. It was found that the test length equal to 60 had the lowest false negative error rates, followed by the test length equal to 40, and then the test length equal to 20. For the high discrimination item pools, the impact of test length on false negative error rates was relatively small compared to the moderate and real item pools in most of the mastery procedures.

Item Pool

Figures 17 to 19 compare the false negative error rates with different types of item pools for different mastery testing procedures. From these figures, in most cases, the high

discrimination item pool had the lowest false positive error rates, followed by the moderate and the real item pools. The differences in false negative error rates among item pools became smaller with longer test lengths.

Figure 14. False Negative Error Rates at Each Level of Test Length for the High Discrimination Item Pools

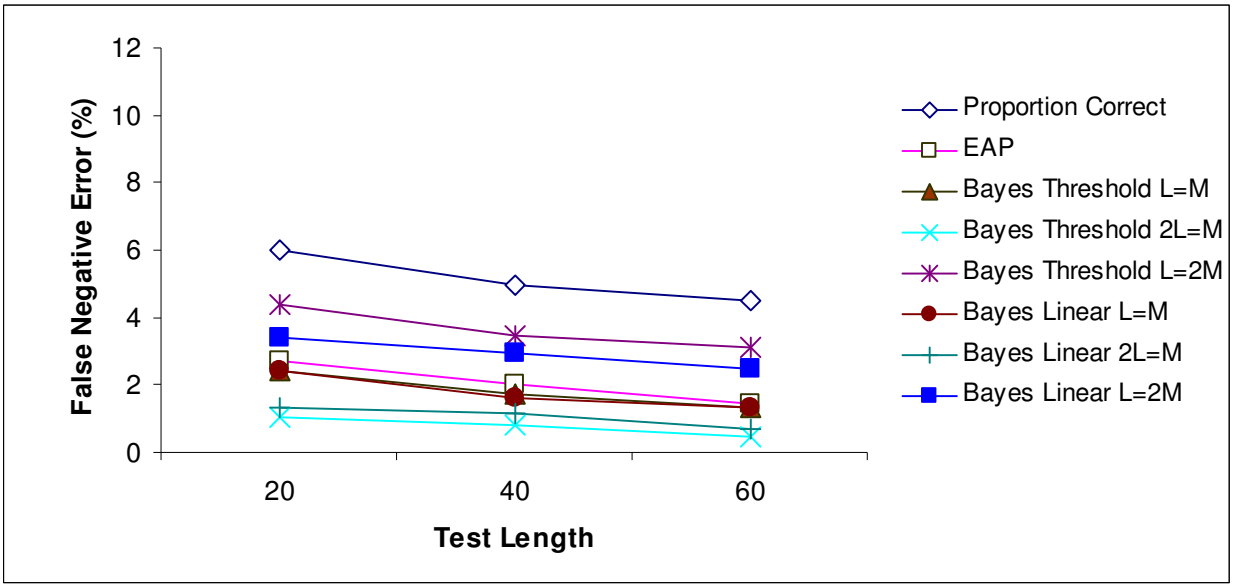


Figure 15. False Negative Error Rates at Each Level of Test Length for the Moderate Discrimination Item Pools

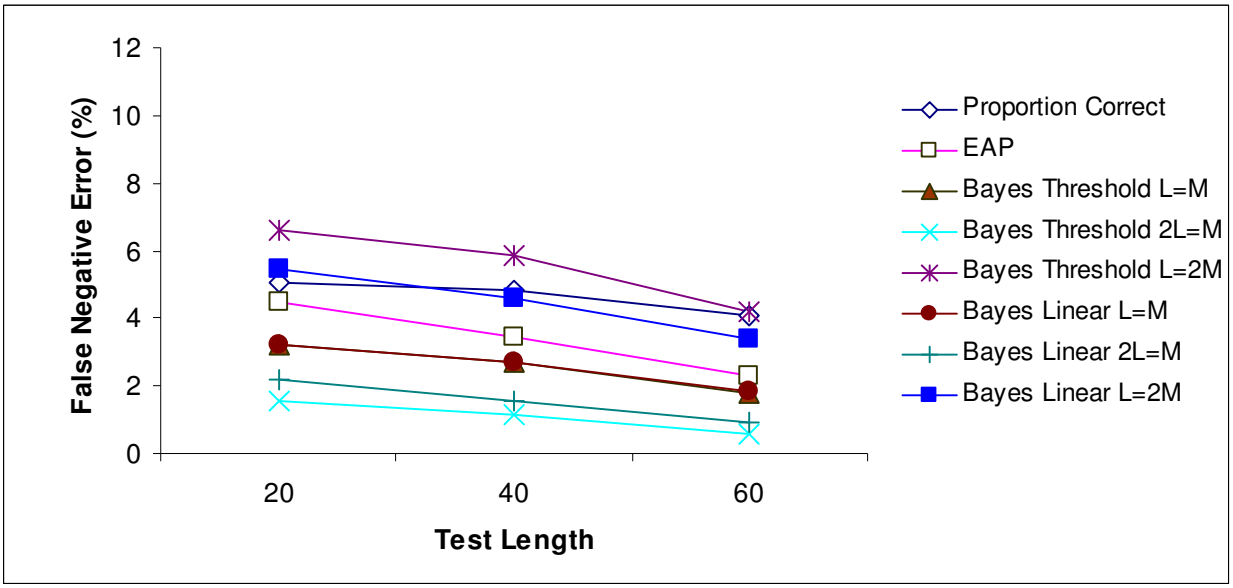


Figure 16. False Negative Error Rates at Each Level of Test Length for the Real Item Pools

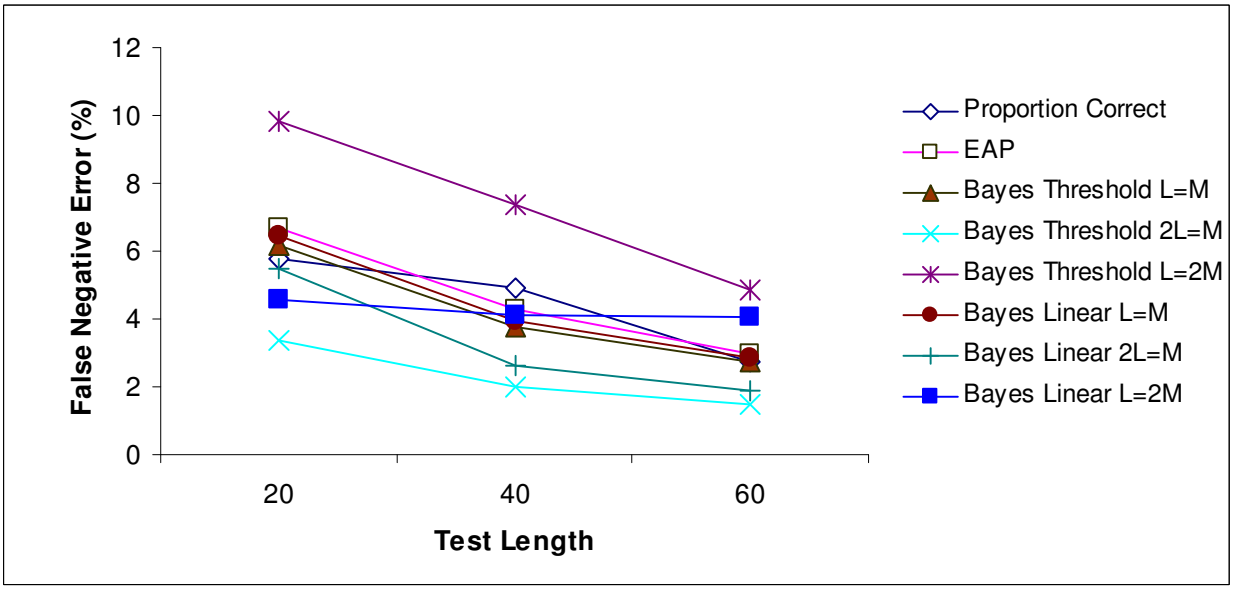


Figure 17. False Negative Error Rates for Test Length Equal to 20 for Different Types of Item Pools

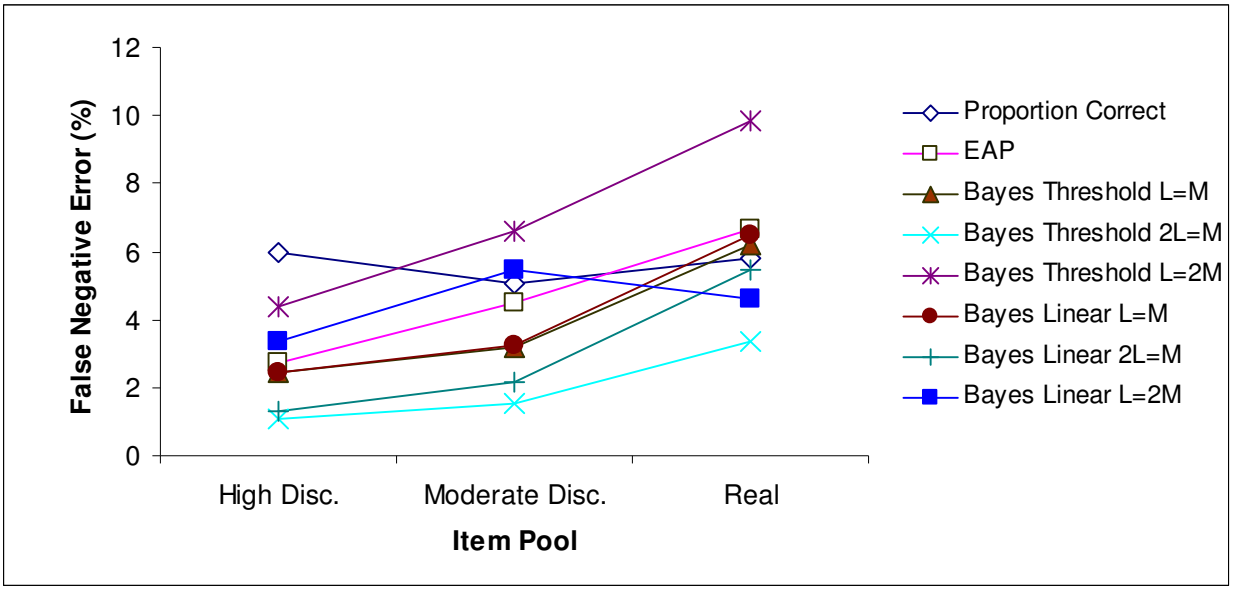


Figure 18. False Negative Error Rates for Test Length Equal to 40 for Different Types of Item Pools

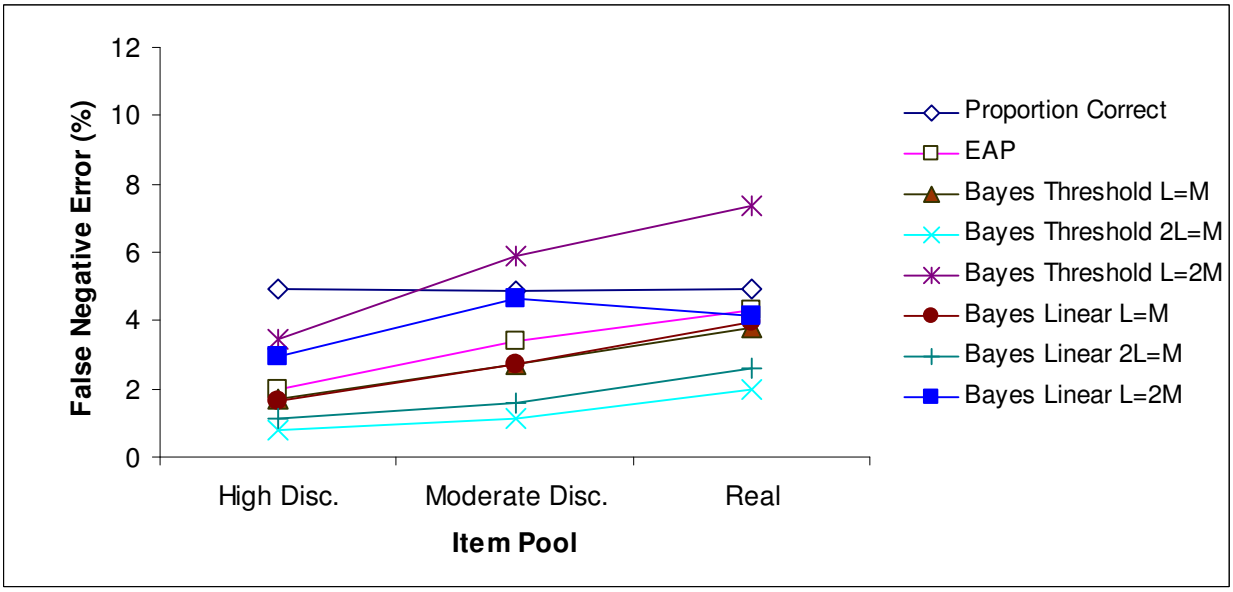
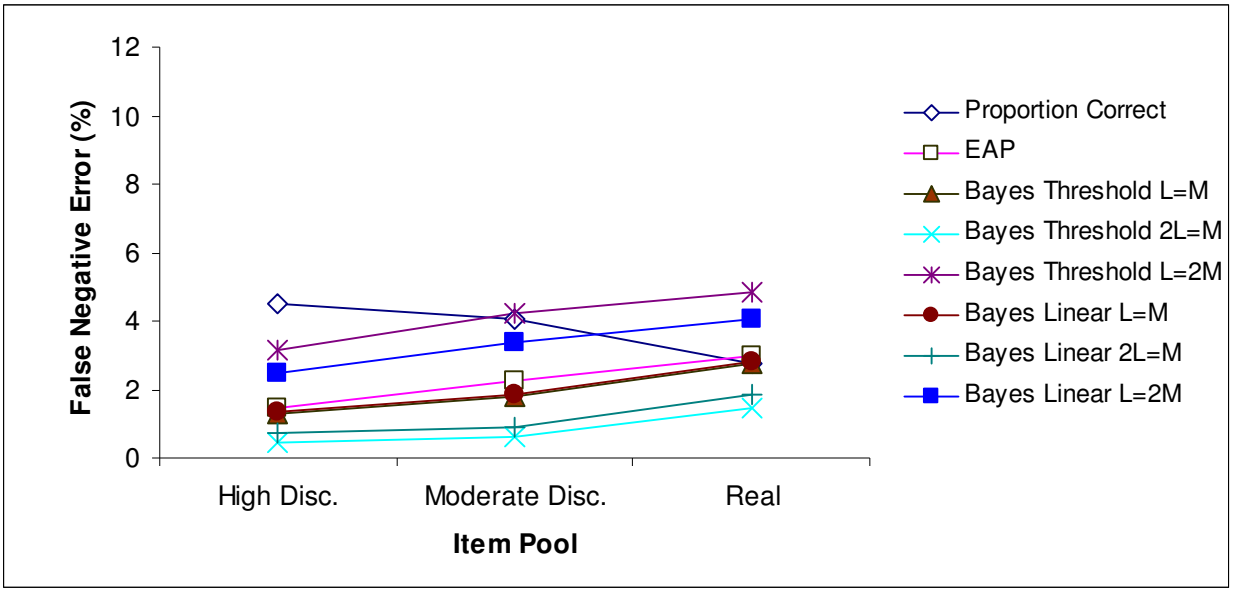


Figure 19. False Negative Error Rates for Test Length Equal to 60 for Different Types of Item Pools



Comparison of Testing Procedures

For all item pools used in this study, as expected, the Bayes threshold 2L=M and Bayes linear 2L=M yielded the lowest false negative error rates. The Conventional-Proportion Correct method had the highest false negative error rates with the high discrimination item pool, and, as expected, the Bayes threshold L=2M had the highest false negative error rates for the moderate discrimination and real item pools.

In comparing the Bayesian decision-theoretic methods to the conventional methods, the Bayes threshold loss and Bayes linear loss functions with 2L=M had lower false negative error rates than the conventional-EAP and the conventional-Proportion Correct methods. The difference between these mastery methods became smaller as the test length increased.

Within the Bayesian methods, Bayes threshold 2L=M had the lowest and Bayes threshold L=2M had the highest false negative error rates. No obvious pattern was observed to indicate whether linear or threshold loss functions performed better. For example, for 2L=M, the threshold loss function yielded lower false negative error rates, compared to the linear loss function. However, for L=2M, the results were the opposite. For L=M, the results for these two types of loss functions were very similar to each other. These results revealed that for both Bayes threshold and Bayes linear methods, 2L=M tended to have the lowest false negative error rates, followed by L=M, and then L=2M. These results were in accordance with expectations.

False Positive Error Rates

Test Length

Figures 20 to 22 compare the false positive error rates for different testing procedures with different test lengths. In comparing the false positive error rates for these three different test lengths, the figures show that test length equal to 60 had the lowest false positive error rates, followed by the test length equal to 40 and the test length equal to 20. For high discrimination item pools, the impact of test length on false positive error rates was relatively small compared to the moderate discrimination and real item pools in most of the mastery procedures.

Item Pool

Figures 23 to 25 compare the false positive error rates for different types of item pools for different mastery testing procedures. These figures show that overall, the real item pool tended to have the lowest false positive error rates, followed by the high discrimination, and then the moderate discrimination item pools. However, the difference

between testing procedures in false positive error rates for the moderate discrimination and the real item pools became smaller as the test length increased.

Figure 20. False Positive Error Rates at Each Level of Test Length for the High Discrimination Item Pools

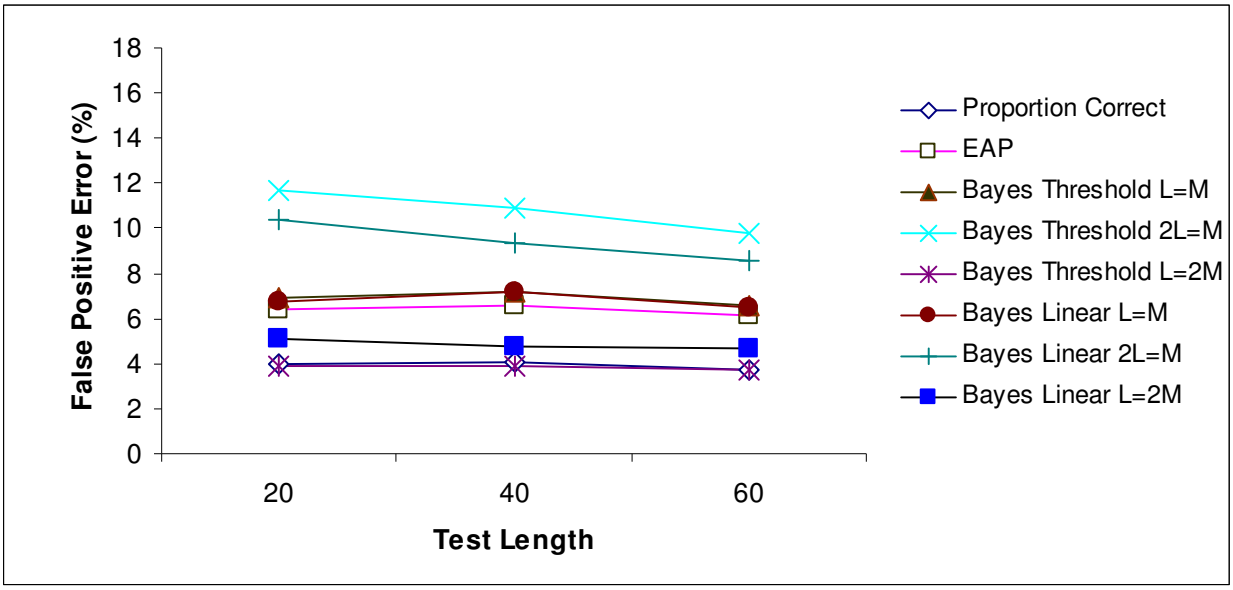


Figure 21. False Positive Error Rates at Each Level of Test Length for the Moderate Discrimination Item Pools

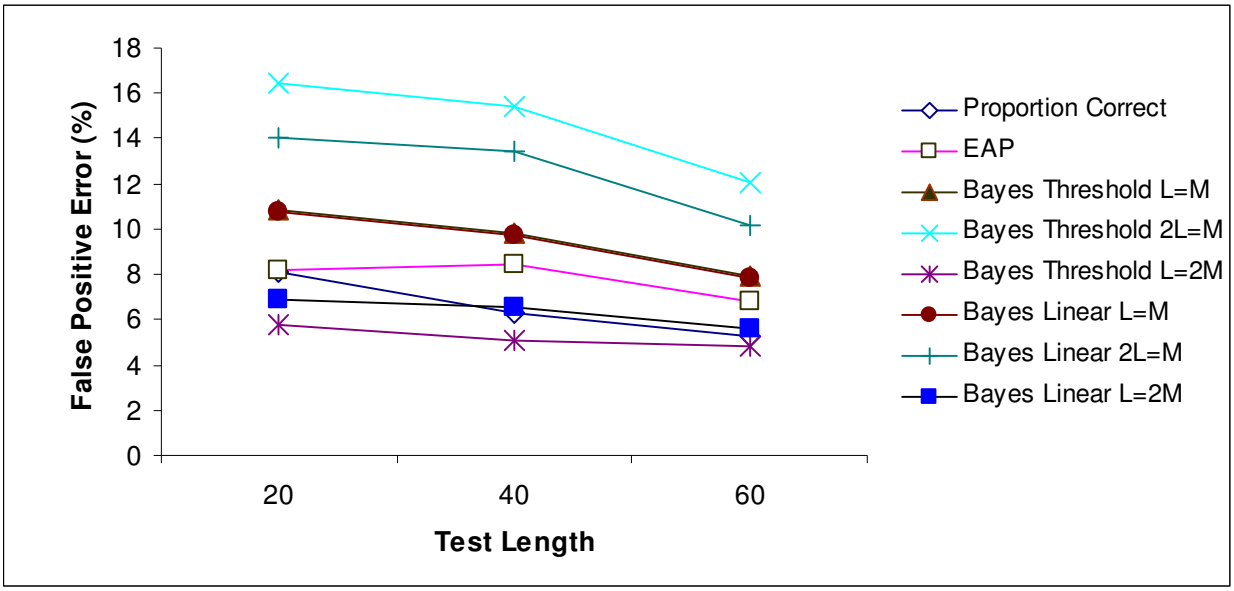


Figure 22. False Positive Error Rates at Each Level of Test Length for the Real Item Pools

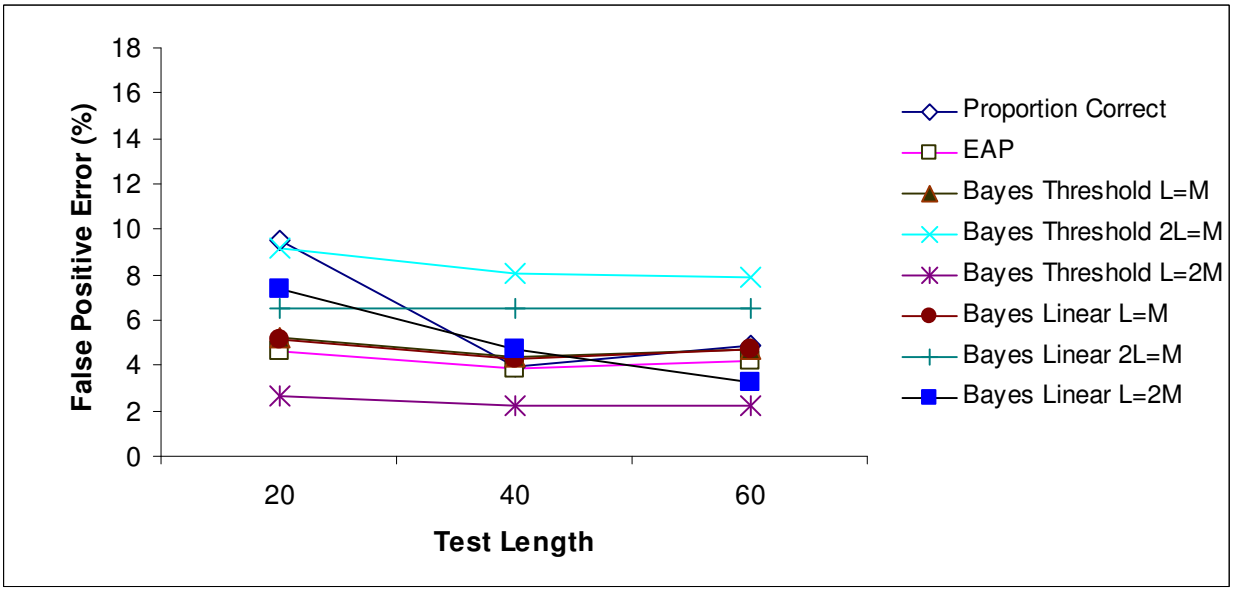


Figure 23. False Positive Error Rates for Test Length Equal to 20 for Different Types of Item Pools

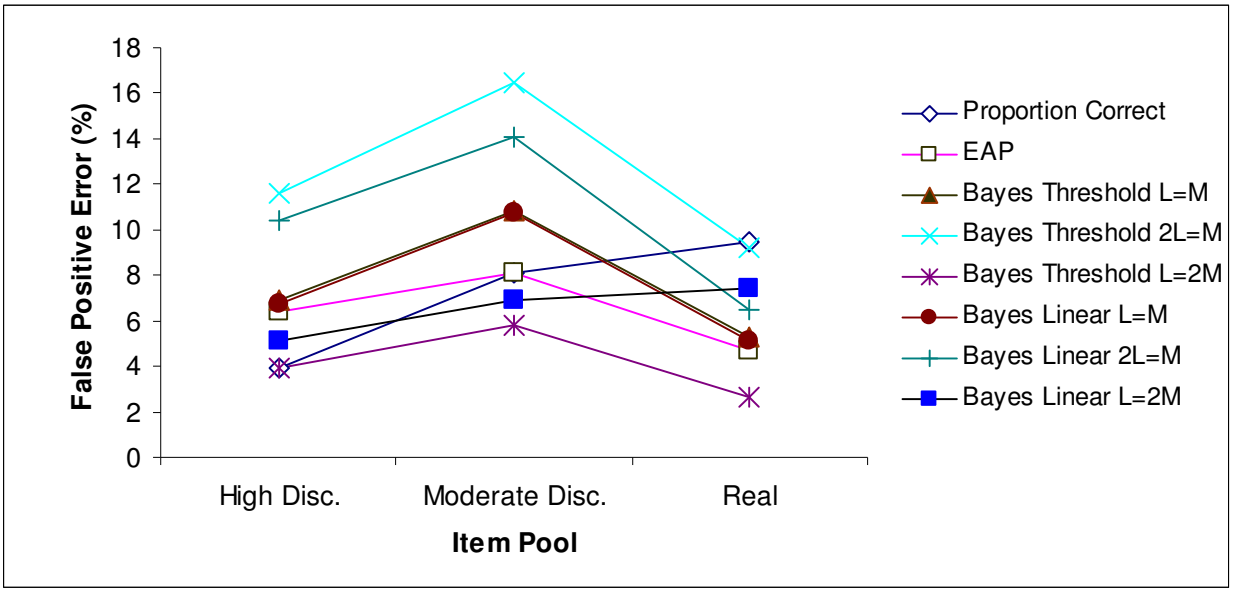


Figure 24. False Positive Error Rates for Test Length Equal to 40 for Different Types of Item Pools

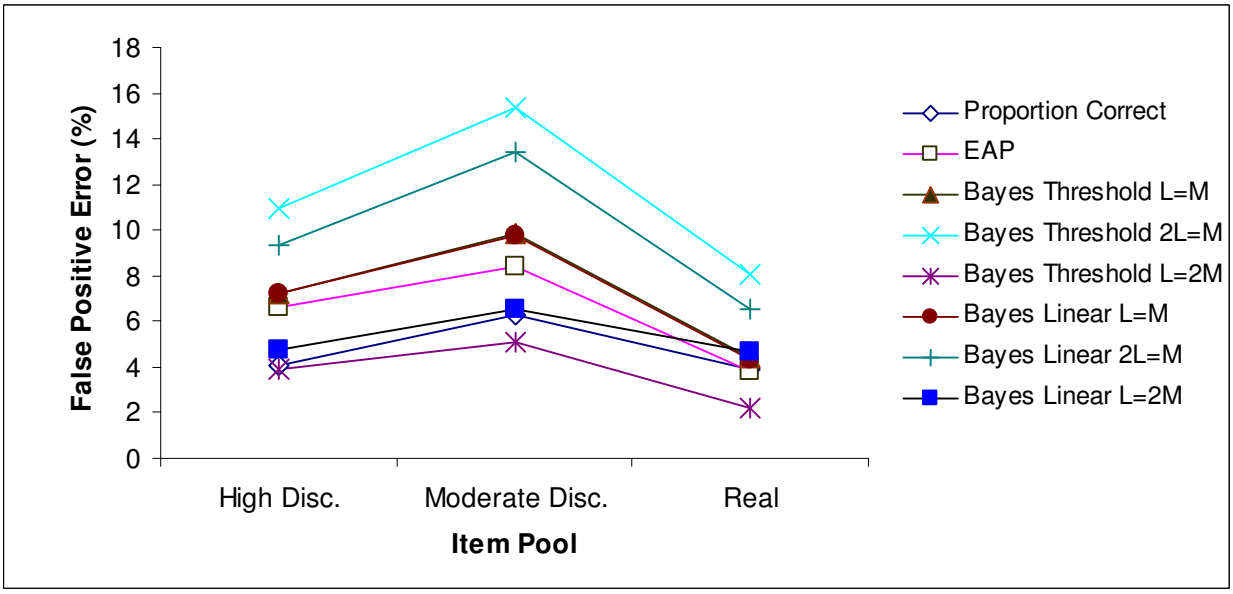
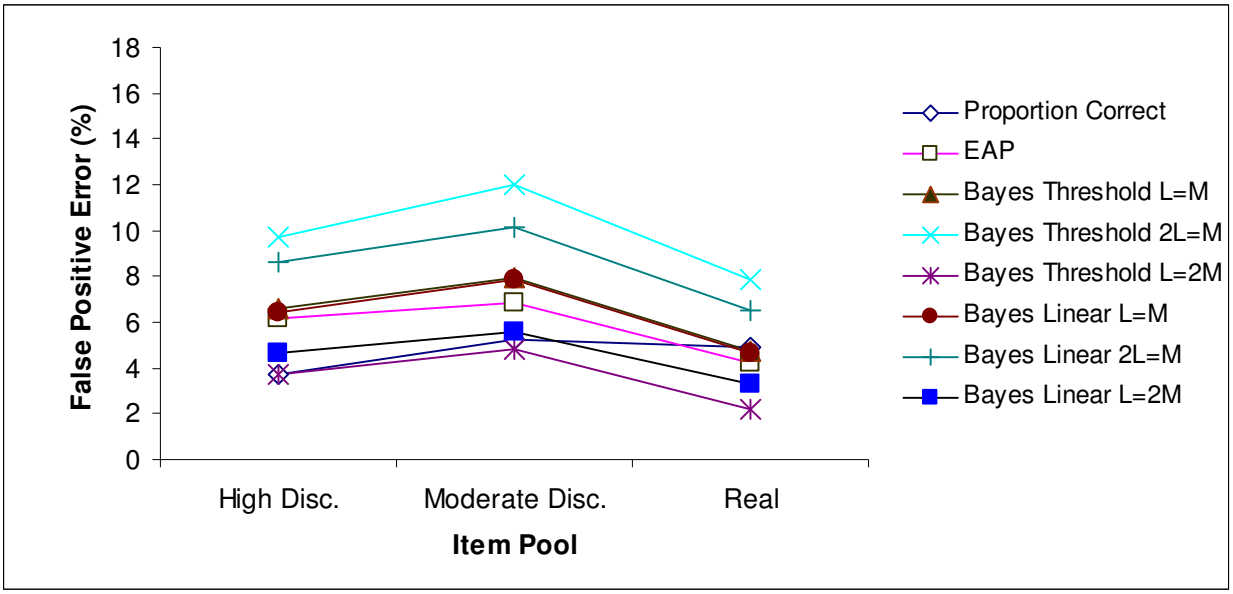


Figure 25. False Positive Error Rates for Test Length Equal to 60 for Different Types of Item Pools



Comparison of Testing Procedures

In general, and as expected, for all types of item pools used in this study, the Bayes threshold L=2M yielded the smallest and the Bayes threshold 2L=M yielded the highest false positive error rates.

In comparing the Bayesian decision-theoretic methods with the conventional methods, the Bayes threshold L=2M had lower false positive error rates than the conventional-EAP and conventional- Proportion Correct methods for all types of item pools. However, note that with the high discrimination item pool, the false positive errors for the conventional-Proportion Correct procedure were consistently small across test lengths. Generally, the differences between these mastery methods became smaller as the test length increased.

Within the Bayesian methods, Bayes threshold L=2M had the lowest and Bayes threshold 2L=M had the highest false positive error rates. There is no obvious pattern as to whether threshold or linear loss function yielded better outcomes. For example, for L=2M, the threshold loss function yielded the lower false positive error rates, compared to the linear loss function; however, for 2L=M, the results were the opposite. For L=M, however, the results from these two types of loss functions were very similar to each other. These results revealed that for both the Bayes threshold and the Bayes linear methods, L=2M tended to have the lowest false positive error rates, followed by L=M and 2L=M. These results support the utility of the loss functions.

Phi Correlations between True and Predicted Mastery Status

Test Length

Generally, the phi correlations between the true and predicted mastery status differed by no more than 0.08 across mastery testing procedures (see Figures 26 to 28). Figures 26 to 28 compare the phi correlations between true and predicted mastery status for different testing procedures with different test lengths. These figures show that as test length increased, the corresponding phi correlations also increased. The more distinct increases occurred in both the moderate discrimination item pools and the real item pools.

Item Pool

Figures 29 to 31 compare the phi correlations between true and predicted mastery status for different types of item pools for different mastery testing procedures. The results show that overall, the high discrimination item pools tended to have the highest phi correlations, followed by the real item pools, and then the moderate discrimination item pools.

Figure 26. Phi Correlations Between True and Predicted Mastery Status at Each Level of Test Length for the High Discrimination Item Pools

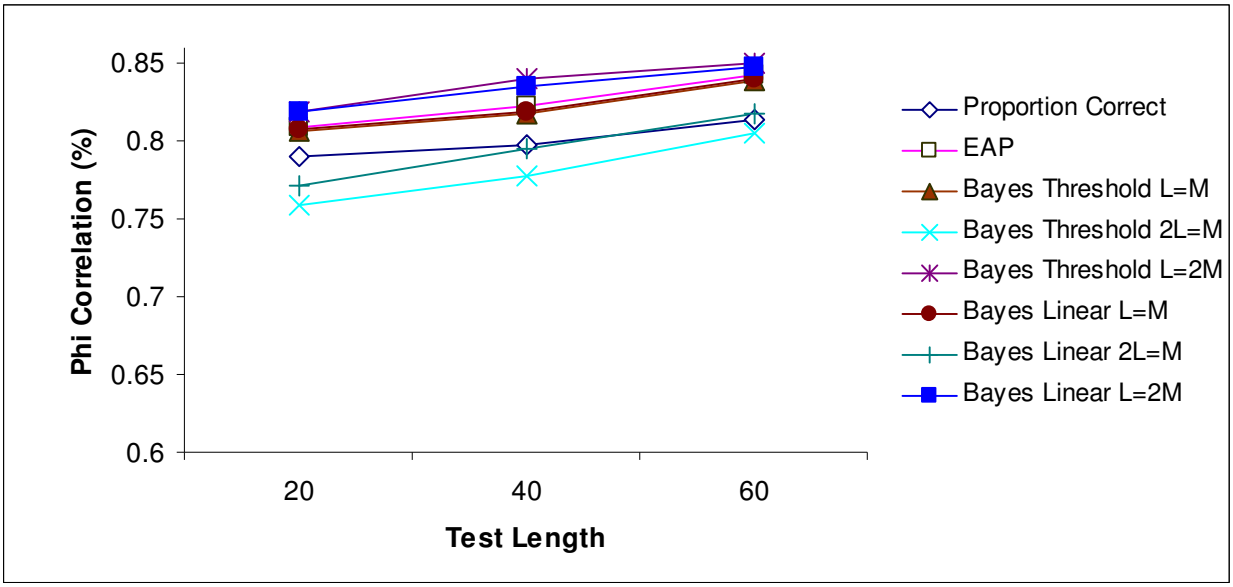


Figure 27. Phi Correlations Between True and Predicted Mastery Status at Each Level of Test Length for the Moderate Discrimination Item Pools

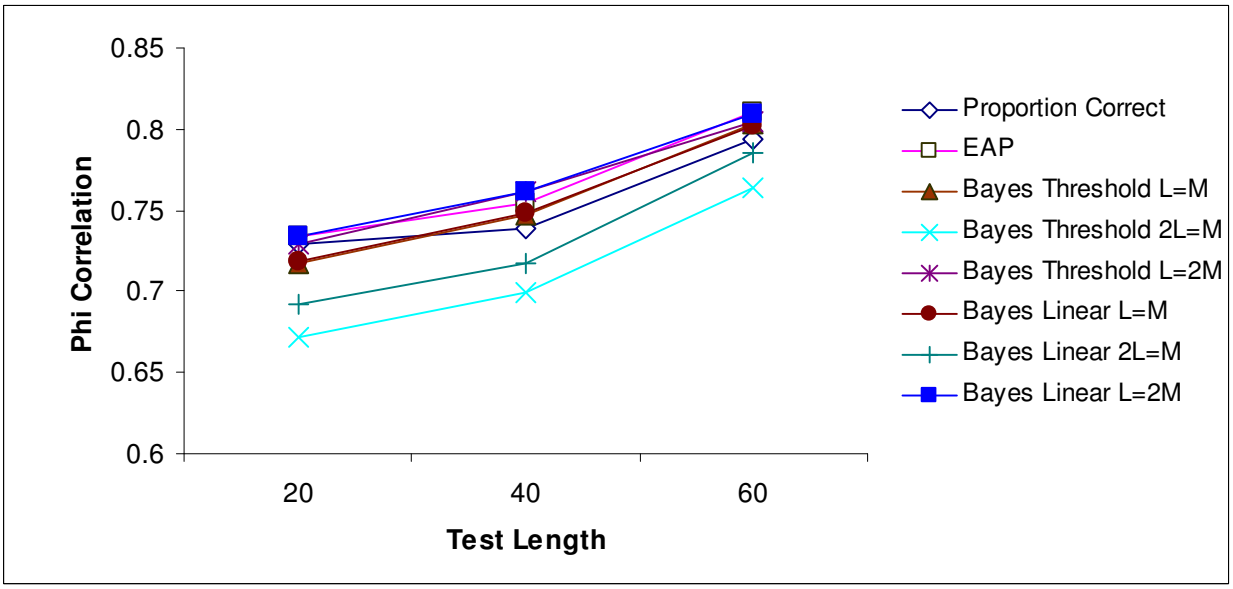


Figure 28. Phi Correlations Between True and Predicted Mastery Status at Each Level of Test Length for the Real Item Pools

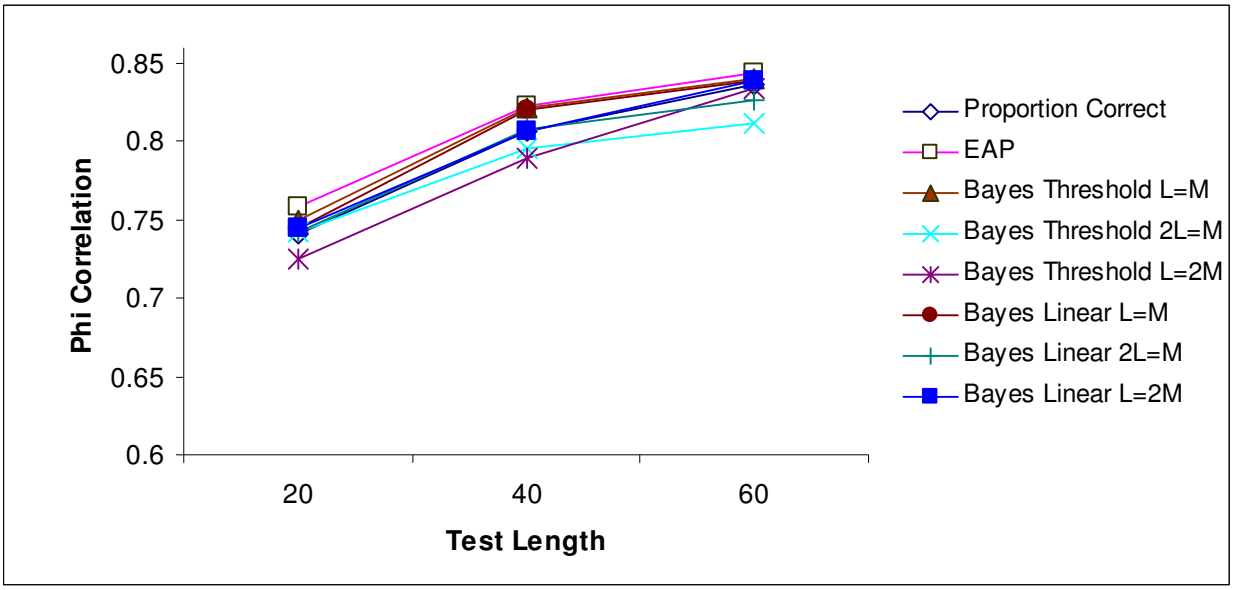


Figure 29. Phi Correlations Between True and Predicted Mastery Status for Test Length Equal to 20 for Different Types of Item Pools

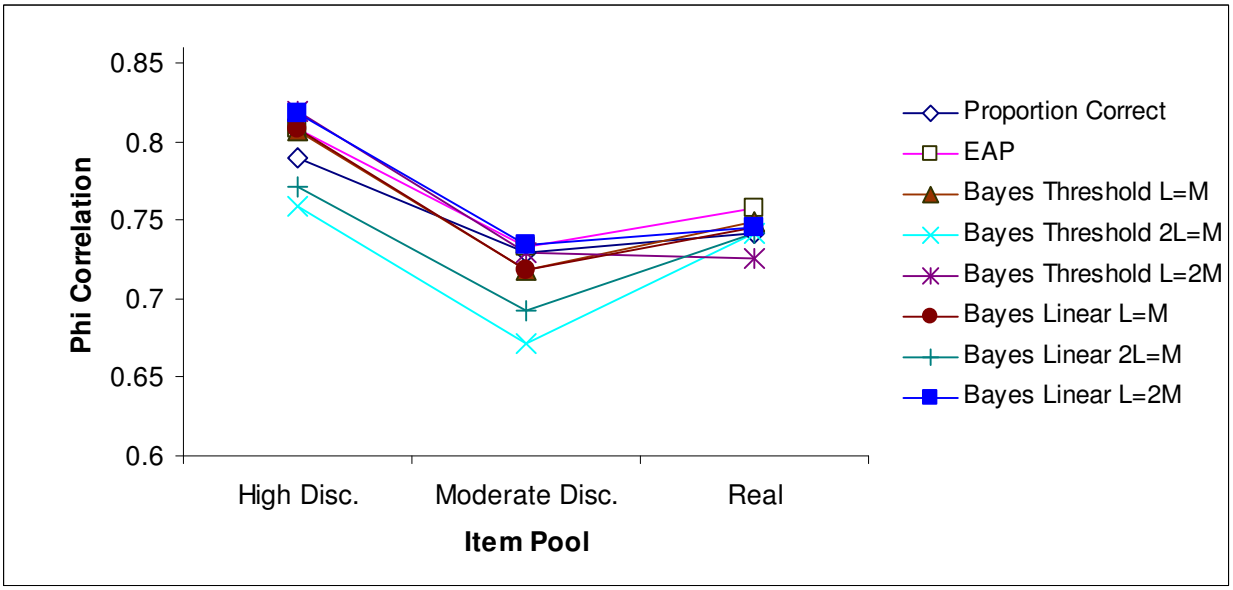


Figure 30. Phi Correlations Between True and Predicted Mastery Status for Test Length Equal to 40 for Different Types of Item Pools

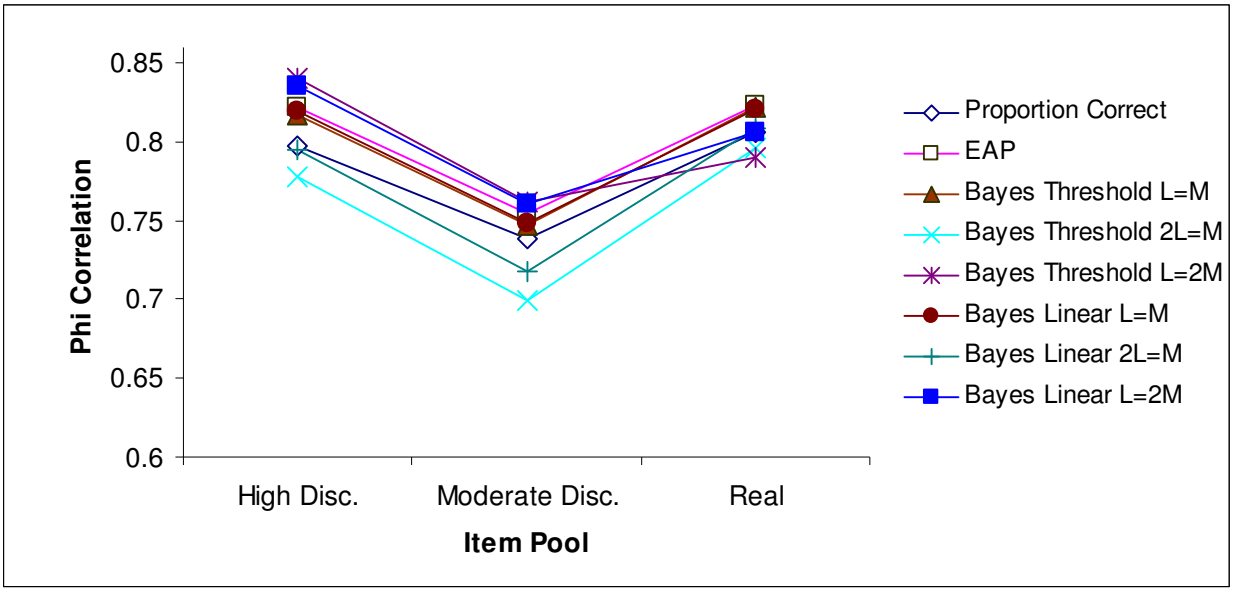
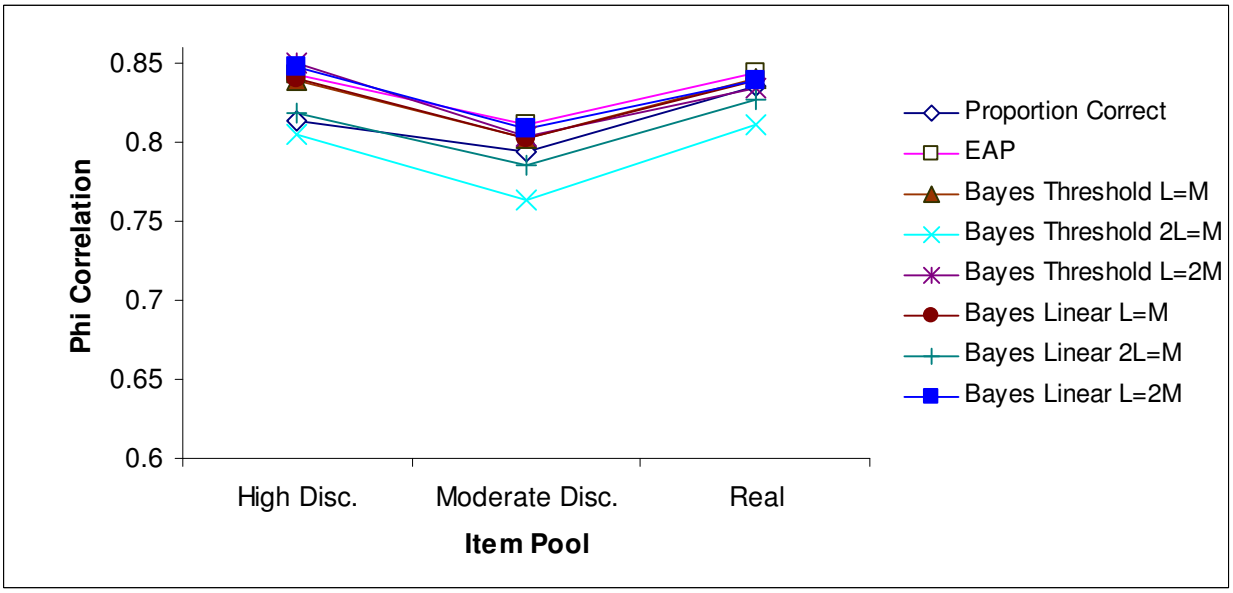


Figure 31. Phi Correlations Between True and Predicted Mastery Status for Test Length Equal to 60 for Different Types of Item Pools



Comparison of Testing Procedures

The Bayes threshold 2L=M yielded the lowest phi correlations for most situations. The procedure that yielded the highest phi correlations depended on the test length and type of item pool. For example, the Bayes threshold L=2M yielded the highest phi correlations with the high discrimination item pool and test length equal to 20, but the conventional-EAP method had the highest phi correlations for the real item pool with the same test length. Among all the simulation conditions, the Bayesian threshold 2L=M (with the moderate discrimination item pool and test length equal to 20) had the lowest phi correlation (0.68), and the Bayesian threshold L=2M (with the high discrimination item pool and test length equal to 60) had the highest phi correlation (0.84). Thus, the magnitude of the differences between testing procedures across all conditions ranged from 0.68 to 0.84. These results were comparable to those reported by Vos (2000, 0.59 to 0.79).

In comparing the Bayesian methods with the conventional methods, the results show that the Bayes threshold L=2M and the conventional-EAP had similar patterns in terms of phi correlations, and the two procedures had higher phi correlations compared to the other procedures. Nevertheless, the differences between these mastery methods became smaller as the test length increased.

Within the Bayesian methods, the Bayes threshold L=2M had the highest phi correlations for the high and the moderate discrimination item pools, but for the real item pools, the Bayes threshold L=M had the highest. The Bayes threshold 2L=M had the lowest phi correlations. These results also show that in most cases, for both the Bayes

threshold and the Bayes linear methods, L=2M tended to have the highest phi correlations, followed by L=M and then 2L=M.

Chi-square Statistics for Different Testing Procedures

In order to test the overall differences among testing procedures, a chi-square test was implemented. Based on true versus observed mastery status, an odds ratio was calculated for each of the eight testing procedures (See Tables in Appendix B for the odds ratio data). The chi-square statistics are the weighted sums of squares of the deviation of each odds ratio from the mean of the eight odds ratio. The weight for each odds ratio is the reciprocal of the squared standard error of odds ratio. The chi-square statistics reflect the degree of homogeneity of the eight procedures. The larger the chi-square value, the more differences there are among the procedures. Because there are 8 testing procedures, the degrees of freedom for the chi-square statistics was 7. Table 9 shows the chi-square statistics for each condition. For the high discrimination item pool, the test procedures were significantly different across all test lengths with an alpha level of 0.05. For the moderate discrimination item pool, the procedures were significantly different for tests with length equal to 40 and 60. As for the real item pool, there were no significant differences.

Table 9. Chi-square Statistics for Different Testing Procedures

Item Pool	Test Length		
	N=20	N=40	N=60
High Disc. Item Pool	19.44*	70.99*	169.70*
Moderate Disc. Item Pool	9.64	38.01*	90.87*
Real Item Pool	8.68	2.93	2.68

Note: * represents the chi-square value was significant at an alpha level of 0.05.
The critical value for chi-square with degrees of freedom equal to 7 is 14.07.

Table 10 provides pairwise tests for those conditions with significant chi-square values in Table 9. The pairwise tests were also implemented with a chi-square test; however, these chi-square statistics had degrees of freedom equal to 1 since there were only 2 procedures being compared. There are a total of 28 pairwise comparisons for eight testing procedures. However, in order to simplify the presentation, only the comparisons related to the research questions are shown, including the comparisons between loss functions given the same loss parameters and the comparisons between the Bayesian threshold loss methods with the two conventional methods. When comparing the Bayesian threshold loss with the conventional methods, only L=2M and 2L=M were presented since these two loss functions represented the highest and lowest overall percentages of correct classifications and phi correlations indices among all the Bayesian procedures in most cases. Table 10 indicates that there were no significant differences between the threshold loss and the linear loss, however, the threshold loss function L=2M was significantly better than the conventional-EAP and conventional-Proportion Correct procedures for all the test lengths in the high discrimination item pool and for test length equal to 40 and 60 in the moderate discrimination item pool.

Table 10. Post Hoc Pair-wise Comparison between Different Testing Procedures

Pairwise Comparison	High Disc. Item Pool			Moderate Disc. Item Pool	
	Test Length			Test Length	
	N=20	N=40	N=60	N=40	N=60
Threshold L=2M vs. Linear L=2M	1.53	0.91	1.42	0.83	1.01
Threshold L=M vs. Linear L=M	0.00	0.08	0.00	0.00	0.04
Threshold 2L=M vs. Linear 2L=M	3.16	0.05	0.03	0.00	0.44
Threshold L=2M vs. EAP	8.07*	4.35*	25.90*	4.23*	9.00*
Threshold 2L=M vs. EAP	0.92	0.07	0.03	2.07	0.04
Threshold L=2M vs. Proportion Correct	15.95*	63.86*	238.33*	37.84*	114.56*
Threshold 2L=M vs. Proportion Correct	14.08*	15.34*	11.65*	2.88	0.58

Note: * represents the chi-square value was significant at an alpha level of 0.05.
The critical value for chi-square with degrees of freedom equal to 1 is 3.84.

Comparison of the 1PL Model and the 3PL Model using Bayesian Decision-theoretic Procedures

One of the purposes of this study was to extend the Bayesian procedures from the 1PL to the 3PL model. To demonstrate the efficacy of the 3PL procedure compared to that of the 1PL procedure, two types of datasets were used. The first dataset was generated to fit the 3PL model using item parameters from the high discrimination item pool. This dataset was used for the analyses reported in the previous sections. The second dataset was generated to fit the 1PL model using the b-parameter from the high discrimination item pool. The results of applying the Bayes threshold procedure for the 3PL model and those for the 1PL model were compared in terms of the percentages of correct classifications and phi correlations between the true and predicted mastery status using these two datasets.

Table 11 compares the percentages of correct classifications using the 3PL dataset with the Bayes threshold procedure. In Table 11, the first column lists the different

simulation conditions with different types of loss functions and test lengths. Given the simulation conditions specified in the first column, the second column represents the percentages of correct classifications estimated by the 3PL model and the third column represents the percentages of correct classifications estimated by the 1PL model. The last column represents the differences between the second and third columns.

As expected, the percentages of correct classifications under the 3PL model were higher than those produced under the 1PL IRT model because the dataset was generated to fit the 3PL model. However, this table shows that the percentages of correct classifications estimated by the 1PL model were often much lower than those produced under the 3PL model. For thresholds L=M and 2L=M in particular, the differences between the 1PL model and the 3PL model ranged from approximately 9% to 22%. As for the threshold L=2M, the differences in percentages of correct classifications were relatively small.

Similar to Table 11, Table 12 compares the phi correlations between the true and predicted mastery status using Bayes threshold procedures under both the 1PL and 3PL models. For thresholds L=M and L=2M, the differences between phi correlations for the 1PL and 3PL were relatively large. The differences of phi correlations ranged from 0.14 to 0.29. As for threshold L=2M, the differences of the phi correlations were generally lower. In addition, the differences were larger for the 20-item test than for the 40-item and 60-item test.

Table 11. Comparison of Percentages of Correct Classification Rates for the 1PL and 3PL Models Using the Bayes Threshold Procedure with the 3PL Dataset

Simulation Conditions	% Correct		
	3PL	1PL	Difference
L=M			
N=20	90.70	74.26	16.44
N=40	91.06	81.80	9.26
N=60	92.80	82.38	10.42
L=2M			
N=20	91.72	81.60	10.12
N=40	92.64	86.34	6.30
N=60	93.10	86.48	6.62
2L=M			
N=20	87.30	64.88	22.42
N=40	88.31	76.52	11.28
N=60	89.80	78.64	11.16

Table 12. Comparison of Phi Correlations Between the True and Predicted Mastery Status for the 1PL and 3PL Models Using the Bayes Threshold Procedure with the 3PL Dataset

Simulation Conditions	Phi Correlations		
	3PL	1PL	Difference
L=M			
N=20	0.81	0.57	0.24
N=40	0.82	0.68	0.14
N=60	0.84	0.69	0.15
L=2M			
N=20	0.82	0.64	0.18
N=40	0.84	0.74	0.10
N=60	0.85	0.74	0.11
2L=M			
N=20	0.76	0.47	0.29
N=40	0.78	0.64	0.14
N=60	0.80	0.64	0.16

For comparison purposes, the same analyses were conducted using the second dataset which was generated to fit the Rasch model. All the testing procedures referenced in Tables 13 and 14 used the second dataset to compare the percentages of correct classifications and phi correlations using the Bayes threshold procedure under both the 1PL and 3PL models. In Table 13, the first column lists the different simulation conditions with different types of loss functions and test lengths. Given the simulation conditions specified in the first column, the second column represents the percentages of correct classifications estimated by the 1PL model and the third column represents the percentages of correct classifications estimated by the 3PL model. The last column represents the differences between the second and third columns.

As expected, the percentages of correct classifications under the 1PL model were higher (in all but one case) than those produced under the 3PL model because the dataset was generated to fit the Rasch model. However, Table 13 shows that the percentages of correct classification rates estimated by the 3PL model were quite close to those produced under the 1PL model. The differences between using the 1PL model and the 3PL model across the manipulated conditions were quite small in most of the situations, even with the 20-item test.

Similar to Table 13, Table 14 compares the phi correlations between the true and predicted mastery status using the Bayes threshold procedures under both the 1PL and 3PL models. For threshold L=2M, the differences in phi correlations between the 1PL and 3PL were larger than those from the other two thresholds. However, these differences were considerably less than those for the 3PL datasets.

Comparisons were also examined with the Bayesian linear procedures. The results obtained were quite similar to those presented in this section. The results of applying the Bayes linear procedure for the 3PL model and the 1PL model are presented in the Appendix C.

Table 13. Comparison of Percentages of Correct Classification Rates for the 1PL and 3PL Models Using the Bayes Threshold Procedure with the 1PL Dataset

Simulation Conditions	% Correct		
	3PL	1PL	Difference
L=M			
N=20	83.00	88.72	5.72
N=40	89.52	92.34	2.82
N=60	90.84	93.84	3.00
L=2M			
N=20	79.42	89.02	9.60
N=40	86.30	91.40	5.10
N=60	87.08	92.96	5.88
2L=M			
N=20	86.92	88.72	1.80
N=40	91.86	92.10	0.24
N=60	93.06	92.48	-0.58

Table 14. Comparison of Phi Correlations Between True and Predicted Mastery Status for the 1PL and 3PL Models Using the Bayes Threshold Procedure with the 1PL Dataset

Simulation Conditions	Phi Correlations		
	3PL	1PL	Difference
L=M			
N=20	0.70	0.76	0.06
N=40	0.76	0.83	0.07
N=60	0.79	0.87	0.08
L=2M			
N=20	0.64	0.76	0.12
N=40	0.71	0.81	0.10
N=60	0.73	0.85	0.12
2L=M			
N=20	0.75	0.76	0.01
N=40	0.81	0.83	0.02
N=60	0.83	0.84	0.01

Summary of the Comparison of Mastery Procedures

The results of this study showed that test length had an impact on the classification accuracy for both conventional methods and Bayesian decision-theoretic procedures. When test length increased, the percentages of correct classifications and the corresponding phi correlations became higher while the false negative error rates and the false positive error rates became lower. As for the effect of the item pool characteristics, the results showed that the high discrimination items tended to have higher accuracy indices (phi correlations and percentages of correct classifications) and lower error rates (false positive error rates and false negative error rates) for all mastery procedures. When comparing these indices for the moderate discrimination item pool and the real item pool, however, the real item pool yielded higher accuracy indices and lower false positive error rates, compared to the results for the moderate discrimination pool, in most cases.

Nevertheless, the moderate discrimination item pool produced lower false negative error rates than did the real item pool.

For the percentages of correct classifications and phi correlations, the differences in these values between Bayesian procedures and the conventional procedures were quite small (approximately 5 to 8%). However, the chi-square tests showed that overall, the loss function L=2M was significantly better than conventional-EAP and conventional-Proportion Correct in most of the cases for the high discrimination item pool and the moderate discrimination item pool.

Two types of loss functions-threshold loss functions and linear loss functions-were considered in this study. The results from this study showed that there was no clear advantage for either loss function for a fixed-length mastery test. For example, under the same simulation conditions, the Bayesian threshold L=2M yielded the smallest false positive error rates but higher false negative error rates than the Bayesian linear L=2M. The threshold loss function seemed to control the false positive errors better, but in terms of controlling both false negative errors and false positive errors simultaneously, the linear loss function performed somewhat better than the threshold loss function.

One of the purposes of this study was to extend the Bayesian procedure from the 1PL to the 3PL model. Two datasets, which fit the 1PL model and the 3PL model, were simulated for comparing the Bayesian procedures using either the 1PL or 3PL IRT model. The results showed that when the datasets were simulated with the 3PL IRT model, using the 1PL model in the Bayesian procedures yielded less accurate results especially when L=M and L=2M. However, when the datasets were simulated with the 1PL model, using the 3PL model in Bayesian procedure yielded reasonable classification accuracies in most

cases. Thus, the use of MCMC procedures with the 3PL model offers a methodology that appears to be very useful in the context of mastery testing.

CHAPTER V

SUMMARY AND DISCUSSION

In this chapter, a summary of this study is presented, which includes the mastery procedures, major findings and general conclusions. Some issues are also discussed with regard to the implementation of the Bayesian procedure in real testing situations. Finally, this chapter provides current limitations and possible directions for future research.

Summary

There are many goals associated with the No Child Left Behind (NCLB) act, such as improving student achievement, closing the achievement gap, improving teacher quality, and improving literacy by establishing comprehensive reading programs, among others. One of the most prominent specific purposes of NCLB is to ensure that all students will attain proficiency or better in reading and mathematics by 2013-2014. Classifying students as proficient or nonproficient is usually based on the students' performance on a state assessment. Thus, finding an optimal mastery procedure to accurately classify students has become very important.

Over the past few decades, many researchers investigated different procedures for deciding if an examinee has passed or failed a certain subject domain. A few procedures, such as adaptive mastery testing (AMT), sequential probability ratio testing (SPRT) and various Bayesian models, have been proposed and implemented in some testing programs. However, no consensus has been reached about which procedure is preferable.

Among those procedures, the Bayesian procedure has drawn some attention for its flexibility in incorporating prior knowledge and taking the costs of each possible

classification outcome explicitly into account. This Bayesian procedure was originally developed by Lewis and Sheehan (1990) and expanded by Glas and Vos (1998) with the Rasch model. Based on Glas and Vos's (1998) discussion, the most difficult part in implementing the Bayesian procedure for mastery testing is calculating the expected loss function from the posterior distribution. When mastery testing involves many examinees and items, a large number of response patterns can make the posterior distribution function quite complex. Glas and Vos developed a general form to estimate the posterior loss function by incorporating the sufficient statistics into the exponential function of the Rasch model. However, their derivation cannot be extended to the 3PL IRT model because of the lack of sufficient statistics.

This study proposed an alternative procedure to estimate the expected loss function for the 3PL IRT model. This modified Bayesian decision-theoretic procedure adopts the logic from Glas and Vos's (1998) work. However, a Markov Chain Monte Carlo method (MCMC) was employed to solve the expected loss function problem.

The purpose of this study was to investigate the performance of the Bayesian decision-theoretic procedure using the 3PL model. This procedure, with different loss functions, was compared with two conventional procedures (conventional- Proportion Correct and conventional-EAP) with regard to test length and item pool quality. Four criteria were used for comparison: percentage of correct classifications, false positive error rates, false negative error rates, and phi correlations between the true classification status and the observed status. In addition, the performance of the Bayesian decision-theoretic procedure under the 1PL and 3PL models was also compared.

The major findings of this study are organized based on the research questions in Chapter III. The research questions are repeated here for reference. The results follow each question.

(1) To what extent are the classification errors controlled by setting constraints for false positive and false negative errors in the Bayesian decision-theoretic procedure for the 3PL?

By employing the appropriate loss parameters, the Bayesian decision-theoretic procedure can effectively control false positive error rates and false negative error rates. For example, when comparing the Bayesian decision-theoretic methods to the conventional methods, the Bayesian threshold loss and Bayesian linear loss functions with 2L=M (which represents the cost of making a false negative error is twice as serious as making a false positive error) had lower false negative error rates than the conventional-EAP and the conventional-Proportion Correct methods for all types of item pools. Similarly, the Bayesian threshold loss and Bayesian linear loss function with L=2M had lower false positive error rates than other methods in most of the simulation conditions.

There were two types of loss functions investigated in this study, linear loss and threshold loss functions. However, no obvious pattern was found to show which type of loss function was preferable. For example, for 2L=M, the threshold loss function yielded lower false negative error rates, compared to the linear loss function. However, for L=2M, the results were the opposite. In addition, the differences in phi correlations between true and predicted mastery status resulting from these two types of loss

functions were quite small. In most of the cases, the differences of the phi correlations between these two types of the loss functions were around 0.01-0.02.

(2) Does this Bayesian decision-theoretic procedure perform better than the conventional procedures in terms of the degree of accuracy for the 3PL?

For the percentages of correct classifications and phi correlations, the ranges of these values between different procedures were quite small. However, based on an overall evaluation of the testing procedures, the loss function L=2M yielded better results than the conventional-EAP and the conventional-Proportion correct procedures for all the test lengths in the high discrimination item pool and for the test length of 40 and 60 in the moderate discrimination item pool.

(3) How does the item pool quality affect the performance of the testing procedures for making mastery decisions?

The quality of the item pool did impact the performance of the testing procedures. The results showed that the high discrimination items tended to have higher accuracy indices (phi correlations and percentages of correct classifications) and lower error rates (false positive error rates and false negative error rates) for all mastery procedures. When comparing these indices for the moderate discrimination item pools and the real item pools, however, the real item pool yielded higher percentages of correct classifications, higher phi correlations and lower false positive error rates compared to the results for the moderate discrimination pools, in most cases. Nevertheless, the moderate discrimination item pools produced lower false negative error rates than did the real item pools.

(4) Is the accuracy of the testing procedures affected by the length of the test?

As expected, the test length had an impact on the classification accuracy for all of the mastery procedures considered in this study. Generally, when test length increased, the percentages of correct classifications and the corresponding phi correlations became larger, while the false negative error rates and the false positive error rates became smaller. However, depending on the outcome of interest, the impact of increasing the test lengths from 20 to 40 and 40 to 60 appeared to be different. For example, when the test length increased from 20 to 40, the percentages of correct classifications increased considerably. However, when the test length increased from 40 to 60, there was not much increase in the percentages of correct classifications. As for the false negative error rates and false positive error rates, when the test length increased from 20 to 40, the error rates decreased about the same amount as when the test length increased from 40 to 60 (around 0.01 to 0.02).

The impact of test length on the accuracy rates also depended on the item pool quality. For example, for high discrimination item pools, the impact of test length on the percentage of correct classifications was relatively small compared to the moderate and real item pools in most of the mastery procedures.

(5) Is the accuracy of the Bayesian decision-theoretic procedure affected by use of different IRT models such as the three-parameter model or the one-parameter model?

The results showed that when the datasets were simulated using the 3PL model, using the 1PL model in the Bayesian procedure yielded less accurate results, especially when L=M and L=2M. However, when the datasets were simulated with the 1PL model, using the 3PL model in the Bayesian procedure yielded reasonable classification

accuracies in most of the cases. Thus, the Bayesian decision-theoretic procedure with the 3PL model seemed quite useful in the context of mastery tests.

Discussion

Item Pool

Based on the descriptive statistics of the item pools (as shown in Table 7), the means of a-parameters for the two simulated item pools (high and moderate discrimination item pool) were higher than the mean of the a-parameters for the real item pool, but the distribution of b-parameters in the simulated pools was more variable than the distribution of b-parameters in the real item pool. When examining the percentages of correct classifications and phi correlations between true and observed mastery status, the results showed that the percentages of correct classifications and phi correlations for the real item pool were quite comparable with those of the high discrimination item pool and were better than those of the moderate discrimination item pool. This might imply that when applying the Bayesian procedures in mastery tests, especially in the application of dichotomous classification decisions, broad ranges of b-parameters in the item pool may decrease the classification accuracies somewhat. An ideal item pool for implementing Bayesian procedures might consist of items with not only high a-parameters but also with b-parameters more narrowly distributed around the cut score.

However, classification decisions often need to be made for more than two groups. For example, the NCLB exams may need to identify different subgroups of students such as advanced, proficient, average or basic level. Grades in the classroom are often at least classified as A, B, C, D, F categories. Making these classifications accurately is also highly important. Previous studies have shown that using item pools

with the b-parameters distributed throughout the range of all possible theta levels may be preferable in those situations (Bleiler, 1998; Kingsbury & Weiss, 1983); such situations were not included in this study.

Loss Function

A loss function is used to evaluate the total costs and benefits of all possible outcomes for an examinee at a given true level of ability. The costs may consider all relevant psychological and economic consequences which the decision entails. In making mastery decisions, two kinds of errors need to be considered: passing examinees who are true non-masters and failing examinees who are true masters. A loss function can be attached to these two types of errors to minimize the expected losses based on the available response data.

This study considered two types of loss functions: threshold loss functions and linear loss functions. Generally, the threshold loss function is less desirable since it assumes a constant loss for all examinees. Some authors (van der Linden et.al, 1977; Vos, 1999) noted that this assumption is probably unrealistic in some applications. It seems more realistic to assume that loss is an increasing function of theta for nonmasters and a decreasing function of theta for masters (van der Linden, 1996, Vos, 1997). Moreover, the threshold loss function is discontinuous at the cutoff point. The loss for correct and incorrect decisions should change smoothly rather than abruptly (van der Linden, 1981). To overcome this problem, van der Linden and Mellenbergh (1977) proposed a continuous loss function that is a linear function of test takers' proficiency levels for fixed-length mastery tests.

Theoretically, linear loss functions seem more desirable in real testing situations. However, the results from this study showed that the two loss functions performed similarly for a fixed-length mastery test. For example, under the same simulation conditions, the Bayesian threshold L=2M yielded the smallest false positive error rates but higher false negative error rates than the Bayesian linear L=2M. The threshold loss function seems to control the false positive errors better, but in terms of controlling both false negative errors and false positive errors simultaneously, the linear loss function performed somewhat better than the threshold loss function.

In general, for the four criteria used in this study, the values produced by these two different loss functions were very similar. This could imply that using the linear loss or threshold loss function does not have much impact on making binary decisions, at least under the conditions manipulated in this study. This result was consistent with the study conducted by Vos (Vos, 1999). In his study, Vos employed the beta prior and binomial model as the likelihood function for ability. He used both linear and threshold loss structures to decide the optimal number of items that needed to be administered in a variable-length format mastery test. He also concluded that there was not much difference between these two types of loss functions.

The Bayesian decision-theoretic procedure is a relatively new testing procedure for mastery tests, and this procedure is especially appropriate when the test user wants to minimize a specific kind of classification error. For instance, when selecting physicians, the test user may want to reduce false positive error rates. On the other hand, making false negative errors may be relatively less serious. In this case, the test user can set a higher loss weight for false positive errors and a smaller loss weight for false negative

errors. In such situations, the Bayesian decision-theoretic approach might be considered as an alternative to testing strategies such as SPRT, AMT, or other conventional methods.

Cut Score

The Bayesian procedure with loss function L=2M (the cost of making a false positive error was twice as serious as making a false negative error) had higher accuracy indices than other Bayesian procedures in this study. This could be an artifact of the cut score used in this study. As stated in Chapter III, the cut score was set at a theta level of 0.4. With this cut score, approximately 65% of the simulees were nonmasters and 35% were masters since the population was simulated using a normal distribution. Thus, false positive errors can be considered less likely than false negative errors.

It is quite likely that if the cut score were changed, the results would also change. For example, if the cut score was set at -0.4, it is reasonable to predict that the Bayesian procedures with loss function 2L=M (the cost of making a false negative error was twice as serious as making a false positive error) would yield the highest accuracy indices instead. This is because with this cut score, approximately 35% of simulees would be nonmasters and 65% would be masters. In this case, false negative errors would be less likely than false positive errors.

Priors

For this study, vague priors were selected for the examinee's ability level. The reason for selecting the vague prior was to "let the data speak for themselves" so that the inferences were not affected by any other information besides the available dataset. This made the data play the main role in estimation. Selecting vague priors for theta for this simulation study was reasonable since there was no way to assume any information

regarding the parameters for the population. However, if the information for the population parameters were available, an informative prior could be used to increase the classification accuracy. The flexibility of incorporating different kinds of priors is another advantage of using the Bayesian decision-theoretic procedure.

IRT Models

There are three commonly used unidimensional IRT models: the three-parameter logistic (3PL), the two-parameter logistic (2PL), and the one-parameter logistic (1PL) models. Among the three, the 1PL model is regarded as the most restrictive. In the 1PL, it is assumed there is no guessing and that the discrimination parameters are the same across all items in the test. Only the difficulty parameters are different across items for the 1PL model. On the other hand, the 3PL model is less restrictive since it allows for guessing and different discrimination parameter values within the test. In general, the 3PL IRT model should fit the item data better than the 1PL IRT model.

However, the 1PL model is simpler and requires a smaller sample size for accurate item calibration compared to the 3PL model. Thus, the 1PL model is often used in real testing situations. The results in this study showed that when the datasets were simulated to fit the 3PL model, using the 1PL model in the Bayesian procedure yielded less accurate results.

Although the 1PL model is a special case of the 3PL IRT model (with constant a-parameters and c-parameters equal to 0), using the 3PL model in the Bayesian procedure with the 1PL datasets yielded slightly less accurate results than using the 1PL model. This is likely because when employing the 3PL model in Bayesian procedures, a, b, c and theta must all be estimated. The estimates of the c- parameters in the 1PL datasets are

clearly biased. In addition, as indicated in many studies, it is quite difficult to estimate the a- and c- parameters precisely (Baker, 1987; Skaggs et. al, 1989). These factors would very likely result in decreasing the accuracy of the ability estimates. Thus, when the dataset was generated based on the 1PL model, employing the 1PL model in the Bayesian methods yielded slightly better accuracy rates than using the 3PL model. However, in this study, using the 3PL model with the Bayesian procedure still produced quite reasonable classification accuracies in most cases. The use of the Bayesian decision-theoretic procedure with the 3PL model seemed quite promising in the context of mastery tests.

Implementing MCMC in WinBUGS

WinBUGS was used for the MCMC sampling in this study. WinBUGS is well-established software used for Bayesian-related analyses. It is very easy to specify the models in WinBUGS and get the MCMC outputs. However, using WinBUGS to do the MCMC sampling was time-consuming and computer-intensive. For the conventional methods, it only took few minutes to estimate the classification errors in R. However, with the same computer, it took more than 4 hours to run the Bayesian decision-theoretic procedure with a test length equal to 60 for 5000 examinees. In a real testing situation, there could be a very large number of students taking the exam. In addition, the score reports normally need to be finished within one to two weeks after receiving students' raw responses from the scanning center. Using the MCMC method in WinBUGS may not be realistic in operationally-based situations, since it could be difficult to meet the deadlines in some real testing situations. Until advances in hardware and programming are achieved, it might be more reasonable to use alternative numerical methods for making mastery decisions in such situations.

Strengths, Limitations and Future Research

Strengths of this Study

Correctly classifying students into accurate categories is a very important issue in many educational applications. Many mastery testing procedures have been proposed and utilized, and this study explored a relatively new mastery procedure using a Markov Chain Monte Carlo sampling algorithm in the Bayesian procedure. Several factors (test length, item pool quality, and loss functions) related to the application of this procedure were addressed. These are important considerations that are of interest when implementing Bayesian procedures with mastery tests. This study provided some useful information about the impact of each of these factors on classification accuracy and classification error rates.

Since this was a simulation study, the item parameters used for this study could be fixed and thus, the factors of interest could be examined and manipulated directly. In addition, because the responses were simulated, the concern with sampling errors was reduced.

Limitations and Future Research

This study investigated some features that influence Bayesian decision-theoretic procedures in the context of fixed-format mastery testing using the 3PL IRT model. There were some limitations of this study. First, this study only considered one cut score. Different locations of cut scores on the theta scale should be considered to examine the impact of the two types of loss functions on the classification accuracy. Second, in this study, the b-parameters in the two simulated item pools were generated in a relatively broad range. As mentioned previously, the broad range of b-parameters may decrease the

accuracy of classification decisions in a dichotomous categorization somewhat. It would be informative to see how much improvement in classification accuracy could be obtained by using an item pool with b-parameters centered at the cut score. Furthermore, different types of item pools, such as uniform item pools, b-variable item pools, a-and b-variable item pools, a-, b-, c- variable item pools, should be investigated in the future to examine the influence of discrimination, difficulty and guessing parameters on the Bayesian procedure. Readers who are interested in how to generate those different types of item pools can consult Kingsbury and Weiss (1983).

Third, this study only considered fixed-format mastery tests. It might be desirable to develop a variable-length format procedure to enhance the efficiency of test administration. Also, Vos (2000) indicated that an optimal situation for the sequential rules would be to choose an action (declaring pass, declaring fail, or continuing testing) that minimizes posterior expected loss at each stage of testing, using dynamic programming. This technique would consider the expected loss at the final stage of testing and then estimate backwards to the first stage of testing. In doing so, the action chosen would be optimal with regards to the entire sequential testing process. Currently, the implementation of this variable-length procedure would not be realistic using WinBUGS, since the processing speed is too slow.

Fourth, item responses were generated by a simulation based on a unidimensional IRT model in this study. However, the unidimensionality assumption is not easily satisfied in real testing situations. The item types for mastery testing can be independent dichotomous items, blocks of test items (sometimes referred to as testlets), open-ended questions or performance-type tasks. These diverse item types could possibly lead to

another distinct dimension in the test. In addition, the item pool for mastery testing is often not unidimensional since many professional exams involve different content areas, cognitive categories, or skill domains. Glas and Vos (2000) investigated this topic, and they noticed that the violation may be a concern for noncompensatory multidimensional models. More research is needed to examine the performance of the Bayesian decision-theoretic procedures when the assumptions of IRT are not completely satisfied.

Fifth, this study used four criteria (percentages of correct classifications, false positive error rates, false negative error rates and phi correlations between the true and observed classification status) to evaluate the results. Although these four indices are commonly used to evaluate the performance of testing procedures in mastery tests, some other criteria might reveal different trends. For example, Glas and Vos (2006) used the average of actual loss for all examinees (mean loss) to evaluate the performance of different test procedures.

Sixth, this study only examined the binary classification situation (pass/fail). However, real testing situations sometimes involve multi-category classifications. For example, for some state assessments, students are classified as either beginner, intermediate, advanced or full proficiency. It would be of interest to develop an algorithm for a Bayesian procedure that could be effective in making multi-category decisions. The performance of the Bayesian decision-theoretic procedure in the multi-category situation should be carefully examined in future studies.

Seventh, this study only considered linear and threshold loss functions for the Bayesian decision-theoretic procedure. However, there are other types of loss structures that could be applied. For example, Novick and Lindley (1979) presented procedures for

specifying nonlinear loss functions for estimating examinees' ability levels in terms of cumulative distribution functions and using least square fitting techniques. This type of nonlinear loss function does not only reflect realistic situations but also can be incorporated with the standard normal distribution for the psychometric model.

Finally, it would be of interest to apply the Bayesian decision-theoretic procedure in a mixed-format mastery test. This study only considered dichotomous item responses. However, many test forms contain not only multiple choice items, but also open-ended questions. In order to apply the Bayesian methods in realistic assessment situations, it would be necessary to develop a Bayesian procedure that can handle mixed-format test forms.

REFERENCES

Abdel-fattah, A. A., Lau, C. A., & Spray, J. A. (1995). *The effect of model misspecification on classification decisions made using a computerized test: MIRT versus UIRT*. Paper presented at the meeting of the Psychometric Society, Minneapolis, MN.

Abdel-fattah, A. A., Lau, C. A., & Spray, J. A. (1996). *Effect of altering passing score in computer adaptive classification testing when unidimensionality is violated*. Paper presented at the annual meeting of the American Education Research Association, New York City, NY.

ACT Assessment Technical Manual (1997). Iowa City, IA: American College Testing.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Angoff, W. H. (1971). *Scales, norms and equivalent scores*. In R.L. Thorndike (2nd ed.), Educational Measurement (pp.508-600).Washington DC: American Council on Education.

Ansley, T. N. (1984). *An empirical investigation of the effects of applying a unidimensional latent trait model to two-dimensional data*. Unpublished doctoral dissertation, The University of Iowa.

Ansley, T. N. & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*, 37-48.

Beguin, A. A., & Glas, C. A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

Binns, M. R. (1994). Sequential sampling for classifying pest status. In Pedigo, L. P. & G. D. Buntin (Eds.), *Handbook of sampling methods for arthropods in agriculture* (pp. 137-174). Boca Raton, FL: CRC Press.

Bleiler, T. L. (1998). *Adaptive classification strategies within an extendible program framework*. Unpublished doctoral dissertation, The University of Iowa.

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

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

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167-174.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883-904.

Cowles, M. K. (2004). Review of WinBUGS 1.4. *The American Statistician*, *58*(5), 330-336.

Crocker, L, & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Davey, T, & Parshall, C. G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Downing S. M., Tekian A., & Yudkowsky, R.(2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine*, *18*, 50-57.

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

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computer assisted branched test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh.

Florida Department of Education. (2006). Florida teacher certification testing. Available from the Florida Department of Education Web site. http://www.fldoe.org/edcert/.

Fraser, C. (1986). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer program]. New South Wales, Australia: Center for Behavioral Studies, The University of New England.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavior Research*, *23*, 267-269.

Frick, T. W. (1990). A comparison of three decision models for adapting the length computer-based mastery tests. *Journal of Educational Computing Research*, *6*(4), 479-513.

Frick , T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research*, *8*(2), 187-213.

Garren, S. T., & Smith, R. L. (1993). *Convergence diagnostics for Markov Chain samplers*. Technical Report, Department of Statistics, University of North Carolina.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd. ed.). Boca Raton, FL: Chapman and Hall.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & A. F. M. Smith (4th ed.), *Bayesian statistics* (pp.237-288). Oxford, UK: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London, UK: Chapman and Hall.

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

Glas, C. A. W., & Beguin, A. A. (1996). *Appropriateness of IRT observed score equating* (Research Report 96-04). The Netherlands: University of Twente.

Glas C. A. & Vos H. J.(1998). *Adaptive mastery testing using the Rasch model and Bayesian sequential decision theory* (Research Report 98-15). The Netherlands: University of Twente.

Glas C. A. W., & Vos, H. J. (2000). *Adaptive mastery testing using a multidimensional IRT model and Bayesian sequential decision theory* (Research Report 00-06). The Netherlands: University of Twente.

Glas, C.A.W., & Vos, H.J. (2006). *Testlet-based adaptive mastery testing* (Computerized Testing Report 99-11). Newtown, PA: Law School Admission Council.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *48*, 1-47.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In Berk, R.A. (Ed.), *Criterion-referenced measurement: The state of the art* (pp.80-123), Baltimore, MD: John Hopkins University Press.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Hingham, MA: Kluwer.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*(2), 91-115.

Kass, R. E., Carlin, B. P., Gelman, A.,& Neal, R. M. (1998). Markov Chain Monte Carlo in practice: a roundtable discussion. *American Statistician*, *52*, 93-100.

Kao, S., & Reckase, M. D. (2006). *Comparisons between parametric and nonparametric procedures in recovery of the dimensionalities of test data*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Kingston, N. M., & McKinley, R. L. (1988). *Assessing the structure of the GRE General Test using confirmatory multidimensional item response theory*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Kingsbury, G. G., & Weiss, D. J.(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 test theory and computerized adaptive testing* (pp.257-283). New York City, NY: Academic Press.

Johnson, V. E. (1994). Studying convergence of Markov Chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, *91*, 154-166.

Lau, C. A. (1996). *Robustness of a unidimensional computerized mastery testing procedure with multidimensional testing data*. Unpublished doctoral dissertation, The University of Iowa.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems (with discussion). *Journal of the Royal Statistical Society*, Series B, *50*, 157-224.

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

Liu, C., Liu, J., & Robin, D. B. (1992). *A variational control variable for assessing the convergence of the Gibbs sampler*. In the proceedings of the American Statistical Association, Statistical computing section, 74-78.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*(1), 95-100.

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

Lunz M. E., & Bergstrom B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, *31*(3), 251-263.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp.223-236). New York: Academic Press.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report 82-1). Iowa City, IA: American College Testing.

McKinley, R. L.,& Reckase, M. D. (1983). *The use of IRT analysis on dichotomous data from multidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Medina-Diaz, M. (1992). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement*, *17*, 117-130.

Mengersen, K. L., Robert, C. O., & Guihenneuc-Jouyaux, C. (1999). MCMC convergence diagnostics: a "reviewww". In J. M. Bernardo, J.O. Beger, A. P. Dawid & A. F. M., Smith (6th ed.), *Bayesian Statistics* (pp.415-440), London, UK: Oxford University Press.

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

Mills C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2001). *Computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Miller, T. R. (1991). *Empirical estimation of standards errors of compensatory MIRT model parameters obtained from the NOHARM estimation program* (Research Report Series 91-2). Iowa City, IA: American College Testing.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* [computer program]. Chicago: Scientific Software.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Mykland, P., Tierney, L., & Yu, B. (1995). Regeneration in Markov Chain samplers. *Journal of the American Statistical Association*, *90*, 233-241.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

Norcini, J. J.,& Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, *10*, 39-59.

Novick M. R. & Lindley D.V. (1978). On the use of the cumulative distribution as a utility function in educational or employment selection. *Journal of Educational Measurement*, *15(3)*, 181-191

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.

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

Patz, R. J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-366.

Raftery, A. E., & Lewis, S. (1992). *How many iterations in the Gibbs sampler? In*. J. M. Bernardo, J. Berger, A. P. Dawid, & A. F. M. Smith (4th ed.), *Bayesian statistics* (pp. 763-773*)*. Oxford, UK: Oxford University Press.

Reckase, M. D., & McKinley, R. L. (1983). *The definition of difficulty and discrimination for multidimensional item response theory models*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp.237-257). New York, NJ: Academic Press.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15(4)*, 361-373.

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, *87*, 861-868.

Robert, G. O. (1992). *Convergence diagnostics of the Gibbs sampler.* In. J. M. Bernardo, J. Berger, A. P. Dawid, & A. F. M. Smith (4th ed.). *Bayesian statistics* (pp. 775-782) Oxford, UK: Oxford University Press.

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

Smith, R., & Lewis, C.(1995). *A search procedure to determine sets of decision points when using testlet-based Bayesian sequential testing procedures*. Paper presented at the annual meeting of National Council on Measurement in Education, New York City, NY.

Spiegelhalter, D., Grigg, O., Kinsman, R., & Treasure, T. (2003) Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *International Journal of Quality Health Care*, *15*, 7-13

Spiegelhalter, D. J., Thomas, A., Best, N.G., & Lunn, D. (2003). *WINBUGS 1.4. User Manual*. MRC Biostatistics Unit, Cambridge. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/.

Spray, J. A. Davey, T. C. Reckase, M.D. Ackerman, T. A., & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (Research Report 90-8). Iowa City, IA: American College Testing.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (Research Report 87-17). Iowa City, IA: American College Testing.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computer adaptive tes*t. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

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

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An international review*, *36*, 263-277.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp.82-98) Minneapolis: University of Minnesota.

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

Thissen, D., & Wainer. H. (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum Publisher.

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

van der Linden, W.J. & Mellenbergh, G.J.(1977) Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, *1*, 593-599.

van der Linden, W. J. (1981). Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. *Psychometrika, 46*, 257-274.

van der Linden, W. J. (1990). Applications of decision theory to test based decision making. In R. K. Hambleton & J. N. Zaal (Eds.) *New developments in testing: theory and applications* (pp.129-155). Boston, MA: Kluwer.

van der Linden, W. J. & Vos, J. H. (1996). A compensatory approach to optimal selection with mastery scores. *Psychometrika, 61*, 155-172.

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

Verhelst, N. D., & Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In G. H. Fischer & I. W. Molenaar (Ed.). *Rasch models: Their foundations, recent developments and applications*. New York: Springer.

Verstralen, H.H.E.M. (1996). Optimal integer category weights in the OPLM and GPCM. (Measurement and Research Department Reports, 95-2). Cito: Amhem

Vos, H. J. (1997). *Applications of Bayesian decision theory to sequential mastery testing*. (Research Report 97-06). The Netherlands :University of Twente.

Vos, H. J.(1999). Application of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Research*, *32*, 403-433.

Vos, H. J.(2000). A Bayesian's procedure in the context of sequential mastery testing. *Psicologica*, *21*, 191-211.

Vos, H. J., & Glas, C.A. W. (2000). Testlet based adaptive mastery testing. In W. J. van der Linden & C. A. W. Glas (Ed.), *Computerized adaptive testing: Theory and practice* (pp.289-310). Boston, MA: Kluwer.

Vos, H. J. (2002). Applying the minimax principle to sequential mastery testing. *Metodološki zvezki*, *18*, Ljubljana: FDV.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Publisher.

Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A.W. Glas (Ed.), *Computerized adaptive testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wang, T. (1995). *The precision of ability estimation methods in computerized adaptive testing*. Unpublished doctoral dissertation, The University of Iowa

Wald, A. (1947). *Sequential analysis*. New York City, NY: John Wiley and Sons.

Yi, Q., Hanson, B. A., Widiatmo, H., & Harris, D. J. (2001). *Comparison of the SPRT and CMT procedures in computerized classification tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Yu, B. (1994). *Monitoring the convergence of Markov samplers based on estimated errors* (Technical Report 409). University of California at Berkeley, Department of Statistics.

Yu, B., and Mykland, P. (1994). *Looking at Markov samplers through Cusum path plots: A simple diagnostic idea* (Technical Report, 413). University of California at Berkeley, Department of Statistics.

Zellner, A., & Min, C. K. (1995). Gibbs sampler convergence criteria. *Journal of the American Statistical Association*, *90*, 921-927.

APPENDIX A.

FIGURES FOR EVALUATING THE CONVERGENCE OF THE MCMC METHODS

Figure A.1. Examples of History Plots of the Bayesian Decision-theoretic Method with Threshold Loss L=M and Test Length =20

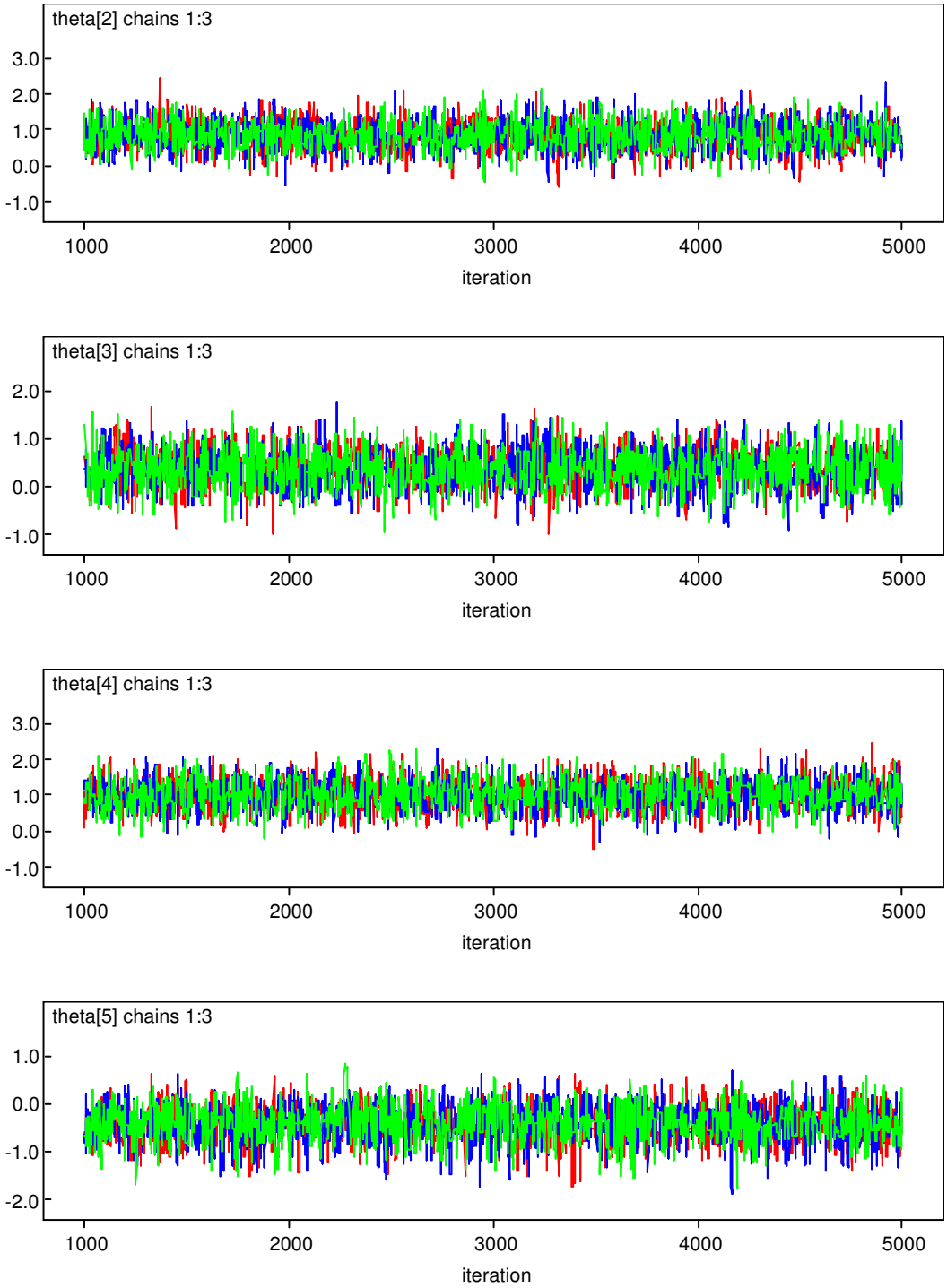


Figure A.1. Examples of History Plots of the Bayesian Decision-theoretic Method with Threshold Loss L=M and Test Length =20 (Continued)

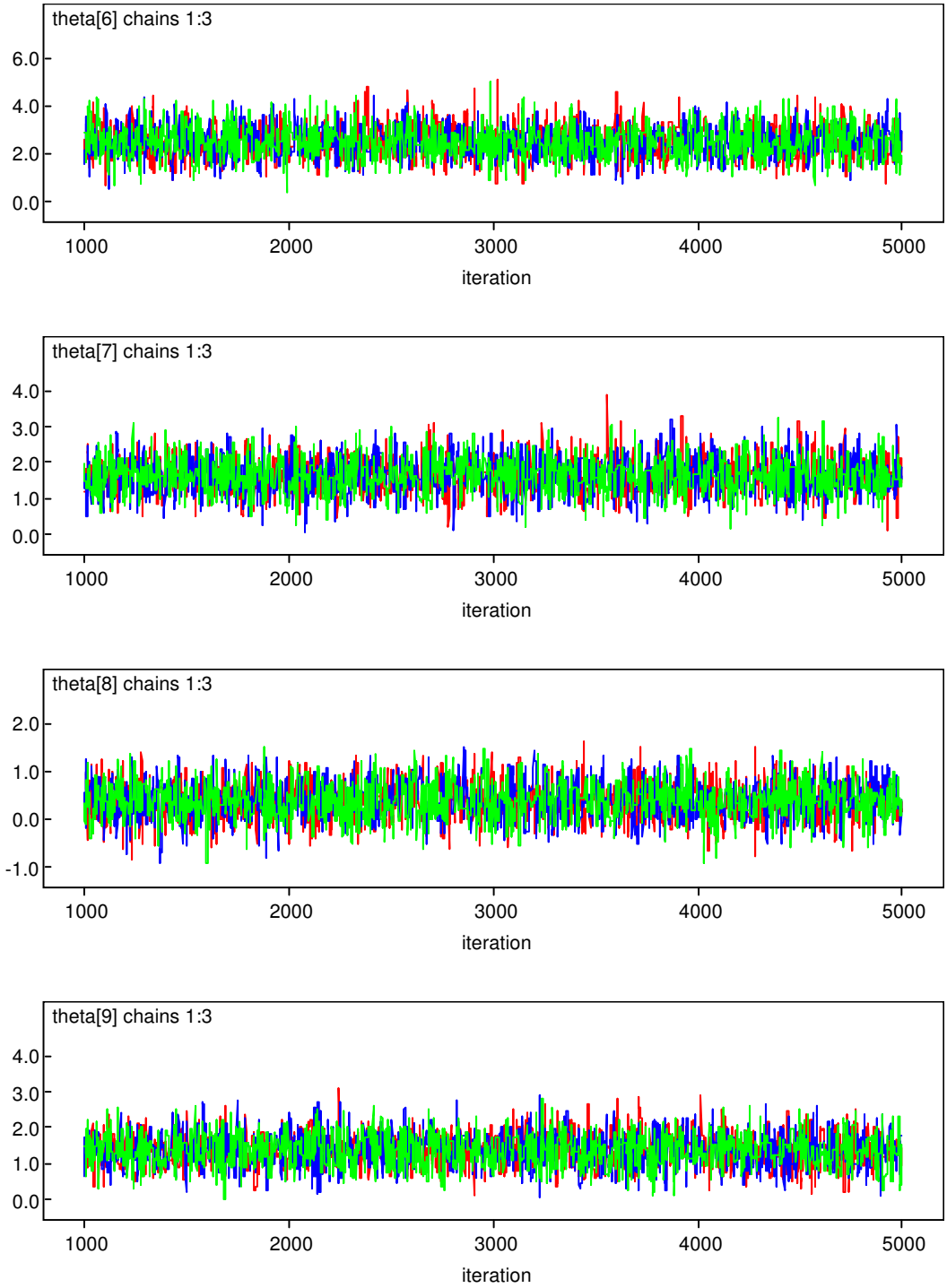


Figure A.2. Examples of Autocorrelation Plots of the Bayesian Decision-theoretic Method with Threshold Loss L=M and Test Length =20

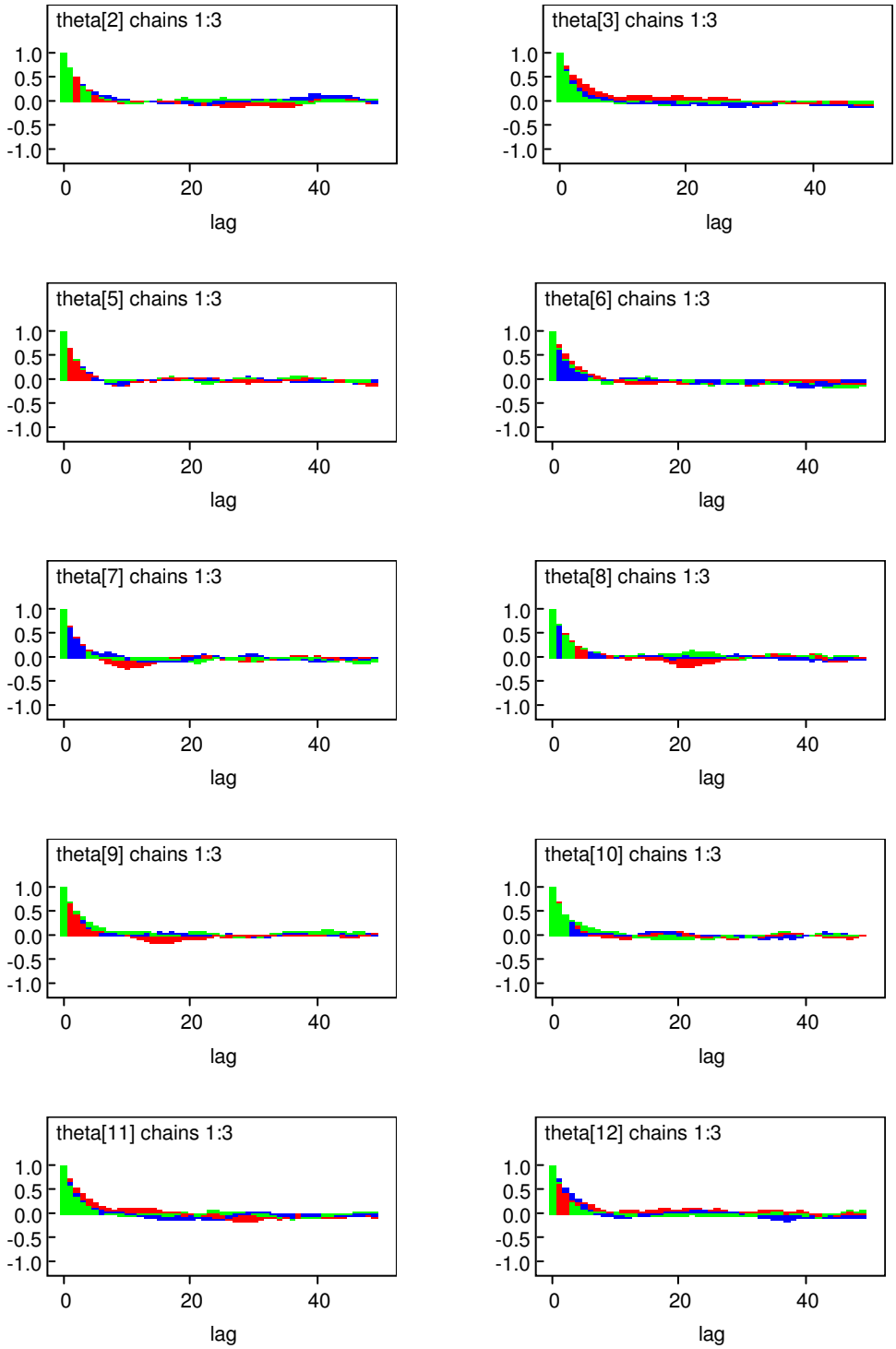
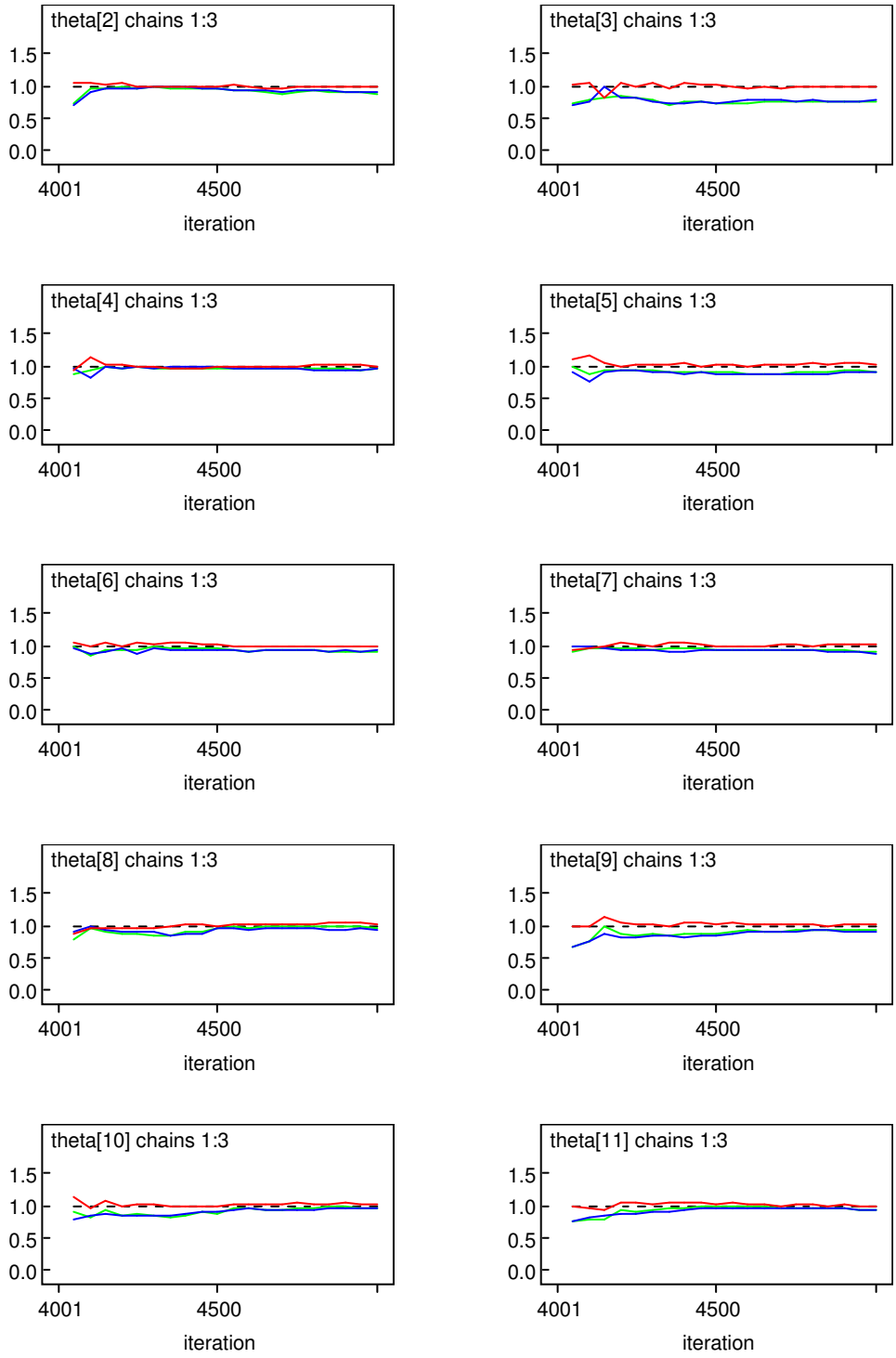


Figure A.3. Examples of BGR Plots of the Bayesian Decision-theoretic Method with Threshold Loss L=M and Test Length =20



APPENDIX B.

ODDS RATIOS FOR DIFFERENT TESTING PROCEDURES IN EACH SIMULATION CONDITION

Table B.1. Odds Ratios for Different Testing Procedures in Each Simulation Condition

Testing Procedures	True Masters	True Nonmasters	False Negative	False Positive	Odds Ratio
High Discrimination Item Pool, Test Length =20					
Threshold L=M	0.33	0.57	0.07	0.02	114.86
Threshold 2L=M	0.31	0.60	0.04	0.04	110.85
Threshold L=2M	0.35	0.53	0.12	0.01	147.93
Linear L=M	0.33	0.57	0.07	0.02	115.69
Linear 2L=M	0.32	0.59	0.05	0.03	111.56
Linear L=2M	0.34	0.54	0.10	0.01	134.76
EAP	0.33	0.58	0.07	0.03	96.42
Proportion Correct	0.30	0.61	0.04	0.06	84.88
High Discrimination Item Pool, Test Length =40					
Threshold L=M	0.34	0.57	0.07	0.02	156.30
Threshold 2L=M	0.32	0.60	0.04	0.03	144.43
Threshold L=2M	0.35	0.53	0.11	0.01	218.97
Linear L=M	0.34	0.57	0.07	0.02	164.31
Linear 2L=M	0.35	0.55	0.09	0.01	178.50
Linear L=2M	0.33	0.60	0.05	0.03	139.73
EAP	0.34	0.57	0.07	0.02	150.74
Proportion Correct	0.31	0.60	0.04	0.05	86.54
High Discrimination Item Pool, Test Length =60					
Threshold L=M	0.34	0.58	0.07	0.01	228.55
Threshold 2L=M	0.33	0.65	0.04	0.03	179.04
Threshold L=2M	0.35	0.55	0.10	0.00	429.39
Linear L=M	0.34	0.58	0.06	0.01	229.75
Linear 2L=M	0.35	0.56	0.09	0.01	315.71
Linear L=2M	0.33	0.60	0.05	0.02	171.57
EAP	0.34	0.58	0.06	0.02	191.58
Proportion Correct	0.31	0.60	0.04	0.05	104.06
Moderate Discrimination Item Pool, Test Length =20					
Threshold L=M	0.33	0.53	0.11	0.03	50.22
Threshold 2L=M	0.29	0.58	0.06	0.07	44.69
Threshold L=2M	0.34	0.48	0.16	0.02	63.74
Linear L=M	0.33	0.54	0.11	0.03	50.31
Linear 2L=M	0.30	0.60	0.04	0.05	78.14
Linear L=2M	0.34	0.50	0.14	0.02	55.62
EAP	0.31	0.56	0.08	0.05	45.92
Proportion Correct	0.31	0.57	0.08	0.05	45.14

Table B.1. Odds Ratio Tables for Different Testing Procedures (Continued)

Testing Procedures	True Masters	True Nonmasters	False Negative	False Positive	Odds Ratio
Moderate Discrimination Item Pool, Test Length =40					
Threshold L=M	0.33	0.54	0.10	0.03	67.53
Threshold 2L=M	0.30	0.59	0.05	0.06	59.68
Threshold L=2M	0.35	0.49	0.15	0.01	98.26
Linear L=M	0.33	0.55	0.10	0.03	67.97
Linear 2L=M	0.31	0.58	0.07	0.05	59.65
Linear L=2M	0.34	0.51	0.13	0.02	83.18
EAP	0.33	0.56	0.08	0.03	72.15
Proportion Correct	0.31	0.57	0.07	0.05	48.31
Moderate Discrimination Item Pool, Test Length =60					
Threshold L=M	0.34	0.56	0.08	0.02	135.31
Threshold 2L=M	0.32	0.59	0.05	0.04	92.22
Threshold L=2M	0.35	0.52	0.12	0.01	254.06
Linear L=M	0.34	0.56	0.08	0.02	130.73
Linear 2L=M	0.32	0.59	0.06	0.03	101.37
Linear L=2M	0.35	0.54	0.10	0.01	200.99
EAP	0.33	0.57	0.07	0.03	95.61
Proportion Correct	0.32	0.59	0.06	0.04	83.00
Real Item Pool, Test Length =20					
Threshold L=M	0.30	0.59	0.05	0.06	53.46
Threshold 2L=M	0.26	0.62	0.03	0.10	61.46
Threshold L=2M	0.32	0.55	0.09	0.03	57.69
Linear L=M	0.29	0.59	0.05	0.06	51.94
Linear 2L=M	0.31	0.57	0.07	0.05	52.19
Linear L=2M	0.27	0.61	0.03	0.08	60.96
EAP	0.33	0.49	0.15	0.02	44.82
Proportion Correct	0.30	0.58	0.06	0.06	49.26
Real Item Pool, Test Length =40					
Threshold L=M	0.32	0.60	0.04	0.04	114.38
Threshold 2L=M	0.28	0.62	0.02	0.07	108.44
Threshold L=2M	0.34	0.56	0.08	0.02	118.92
Linear L=M	0.32	0.60	0.04	0.04	111.38
Linear 2L=M	0.30	0.62	0.03	0.06	110.39
Linear L=2M	0.33	0.58	0.07	0.03	111.94
EAP	0.34	0.55	0.09	0.02	117.03
Proportion Correct	0.31	0.60	0.04	0.05	96.39

Table B.1. Odds Ratio Tables for Different Testing Procedures (Continued)

Testing Procedures	True Masters	True Nonmasters	False Negative	False Positive	Odds Ratio
Real Item Pool, Test Length =60					
Threshold L=M	0.33	0.60	0.05	0.03	151.47
Threshold 2L=M	0.31	0.62	0.03	0.05	145.61
Threshold L=2M	0.34	0.56	0.08	0.01	165.66
Linear L=M	0.33	0.60	0.05	0.03	147.52
Linear 2L=M	0.32	0.61	0.03	0.04	144.31
Linear L=2M	0.34	0.58	0.07	0.02	161.90
EAP	0.33	0.59	0.06	0.03	131.95
Proportion Correct	0.33	0.59	0.05	0.03	144.16

APPENDIX C.

TABLES FOR COMPARING THE 1PL AND 3PL MODELS USING THE BAYES LINEAR LOSS PROCEDURE

Table C.1. Comparison of Percentages of Correct Classification Rates for the 1PL and 3PL Models Using the Bayes Linear Loss Procedure with the 3PL Dataset

Simulation Conditions	% Correct		
	3PL	1PL	Difference
L=M			
N=20	90.76	80.24	10.52
N=40	91.14	83.22	7.92
N=60	92.20	83.78	8.42
L=2M			
N=20	91.54	80.16	11.38
N=40	92.32	82.40	9.92
N=60	92.86	86.48	6.38
2L=M			
N=20	88.26	64.88	23.38
N=40	89.52	76.52	13.00
N=60	90.70	78.64	12.06

Table C.2. Comparison of Phi Correlations Between the True and Predicted Mastery Status for the 1PL and 3PL Models Using the Bayes Linear Loss Procedure with the 3PL Dataset

Simulation Conditions	Phi Correlations		
	3PL	1PL	Difference
L=M			
N=20	0.81	0.66	0.15
N=40	0.82	0.71	0.11
N=60	0.84	0.72	0.12
L=2M			
N=20	0.82	0.67	0.15
N=40	0.84	0.69	0.15
N=60	0.85	0.73	0.12
2L=M			
N=20	0.77	0.70	0.07
N=40	0.8	0.74	0.06
N=60	0.82	0.75	0.07

Table C.3. Comparison of Percentages of Correct Classification Rates for the 1PL and 3PL Models Using the Bayes Linear Loss Procedure with the 1PL Dataset

Simulation Conditions	% Correct		
	3PL	1PL	Difference
L=M			
N=20	85.92	88.72	2.80
N=40	88.93	92.34	3.41
N=60	91.46	93.86	2.40
L=2M			
N=20	87.21	88.72	1.51
N=40	90.30	92.11	1.81
N=60	91.46	93.86	2.40
2L=M			
N=20	83.40	89.02	5.62
N=40	87.21	92.30	5.09
N=60	88.28	93.90	5.62

Table C.4. Comparison of Phi Correlations Between the True and Predicted Mastery Status for the 1PL and 3PL Models Using the Bayes Linear Loss Procedure with the 1PL Dataset

Simulation Conditions	Phi Correlations		
	3PL	1PL	Difference
L=M			
N=20	0.69	0.76	0.07
N=40	0.76	0.83	0.07
N=60	0.83	0.87	0.04
L=2M			
N=20	0.74	0.76	0.02
N=40	0.79	0.83	0.04
N=60	0.82	0.87	0.05
2L=M			
N=20	0.65	0.76	0.11
N=40	0.73	0.83	0.10
N=60	0.75	0.87	0.12