Robustness of a Unidimensional Computerized Testing Mastery Procedure with Multidimensional Testing Data

By C. Allen Lau
1996

Doctoral dissertation
University of Iowa
Iowa City IA.

To my wife, Wing-Shan, for
her love and sacrifice

ACKNOWLEDGMENTS

I would like to express my appreciation to all those who have helped me in my dissertation.

First of all, thanks are given to members in my dissertation committee. I would like to thank Professor Robert A. Forsyth, my advisor and thesis supervisor, for his patient guidance. I would like to thank Professor Timothy N. Ansley and Professor Stephen B. Dunbar, who taught me item response theory and multivariate analysis respectively. I would like to thank other members of my committee, Professors Richard L. Dykstra, and Robert D. Ankenmann for their participation and suggestions.

Special thanks are given to Professor Leonard S. Feldt, whose scholarly work is a great motivation for me. I am so lucky that I could have taken two classes with him before he retired. I also thank all the professors in the Department of Measurement and Statistics. I always feel very proud of being a graduate student there.

Secondly, thanks are given to my colleagues in ACT. I would like to thank Dr. Judith A. Spray, my mentor, who enlightened me with research ideas; guided me to professional career development; taught me the sequential

probability ratio testing procedure; and helped me to develop computer programs. I also thank Dr. Abdel-fattah A. Abdel-fattah for his generously sharing research experience.

At last, thanks are given to my friends and family. I would like to thank Dr. Tianyou Wang, my brother in Christ, for his continuous encouragement and help. I am very grateful for my wife, Wing-Shan, for her sacrifice and ever lasting love. I also thank my mother and my family members for their support. I would like to thank Mrs. Hui Lee Ying, my foster-mother for her unconditional love and care.

ABSTRACT

Unidimensional IRT (UIRT) models are usually used in computerized mastery testing (CMT) to assist in item parameter calibration, theta estimation, item selection, test administration, and mastery decision making. However, in almost all practical testing situations, the unidimensionality assumption will be violated to some degree. When it is violated, the accuracy of the pass/fail decision procedure may be adversely affected. The primary purpose of this study was to investigate the accuracy of one specific procedure for making pass/fail decisions in CMT when the unidimensionality assumption is violated. Specifically, the accuracy of the sequential probability ratio testing (SPRT) procedure was of interest.

Monte Carlo simulation techniques were used to examine the robustness of the SPRT procedure. Two-dimensional dichotomous test data were generated and calibrated by UIRT models. Four factors (type of UIRT model, correlation between ability estimates, test length constraint, & cut-score) were manipulated and 60 combinations of conditions were examined. The outcomes of interest included classification accuracy (false positive, false negative, &

total classification error) and test efficiency (number of items used).

Based on the results of this study, it was concluded that: (1) the SPRT procedure was useful for making mastery decisions in CMT with parameters estimated by either the UIRT three-parameter logistic (3-PL) model or the UIRT one-parameter logistic (1-PL) model even when the unidimensionality assumption was violated; (2) the use of the UIRT 3-PL model leads to greater test efficiency than the use of the UIRT 1-PL model; (3) the impact of a test length constraint on classification accuracy and efficiency depends on which unidimensional model is used; and (4) violation of the unidimensionality assumption may cause bias in the estimation of the cut-score on the theta scale, which in turn may cause differential classification errors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Achievement tests serve a variety of educational purposes. They can be used to help determine if students have the prerequisite knowledge at the beginning of instruction, to monitor students' learning progress during instruction, to diagnose students' learning difficulties during instruction, and to evaluate students' accomplishment at the end of instruction (Gronlund & Linn, 1985).

In many professional fields, when students finish their training, their knowledge, skill and competence must be evaluated before they are allowed to practice their specialty. Achievement tests designed for this purpose are frequently labeled certification or licensure tests. Such tests are used to classify the test takers into one of two categories: qualified (pass) or unqualified (fail). The examinee who has the minimal required knowledge and/or skills deemed essential passes the test; otherwise the examinee fails. The main function of a certification or licensure test is to certify that the examinees who pass the test have the minimum required knowledge and skills to practice their specialty properly (Jaeger, 1988).

As stated in the Standards for Educational and Psychological Testing (American Psychological Association, 1985):

> The primary purpose of licensure or certification is to protect the public. Licensing requirements are imposed to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively. The purpose of certification is to provide the public with a dependable mechanism for identifying practitioners who have met particular standards (p. 63).

Certification or licensure testing is an important part of our society. It plays the role of "gatekeeper" to assure the quality of the professional practitioners in many professions. This type of testing is discussed in more detail in the next section.

Certification or Licensure Testing

Generally, there are four main components in certification or licensure testing: (1) test development, (2) setting the passing score, (3) test administration, and (4) mastery decision making. These components are related to each other and are discussed below.

Test Development

In developing a test, the purpose that the test will serve is always the primary concern. For certification or licensure testing, the purpose is to classify the examinees into one of the two categories: qualified (pass) or

unqualified (fail). In order to do this, test developers have to prepare test specifications, construct a related item pool, select a set of items, establish a passing score, and validate the test results.

Setting the Passing Score

Based on certain criteria, the passing score (or cut-score) is set at a point on the achievement continuum to separate the unqualified examinees from the qualified ones. Traditionally, an examinee whose score is at or above this score is classified as passing, while an examinee whose score is below the score is classified as failing.

There are two main types of procedures used to set passing or cut-scores: (1) procedures based on judgments about test questions, and (2) procedures based on judgments about individual test-takers. Nedelsky's method (Nedelsky, 1954), Angoff's method (Angoff, 1971), and Ebel's methods (Ebel, 1972) are based on judgments about test questions while the borderline-group (Nedelsky, 1954) method, the contrasting-groups method (Nedelsky, 1954), and the up-and-down method (Livingston & Zieky, 1982) are based on judgments about individual test-takers. Among these methods, Angoff's is one of the most commonly used. In the Angoff procedure, judges are asked to estimate the probability that a minimally qualified person can answer each item correctly. For each item, the average judges' rating is computed. The

cutting point is the sum of these average ratings across items (Crocker & Algina, 1986).

Test Administration

There are three main methods used to administer certification or licensure tests: (1) a conventional paper-and-pencil test format, (2) a computer-administered test format, and (3) a computerized adaptive test (CAT) format. Until recently, the conventional test format was used almost exclusively. However, in recent years, the CAT format has been used more frequently because of its convenience and efficiency, and because the costs of computer equipment have decreased markedly.

When the conventional test format is used, all examinees are administered the same test at the same time. The test length and time limit are fixed.

The computer-administered test format involves the administration of the conventional test by means of a computer instead of by paper and pencil.

The computerized adaptive test format also administers the test by means of a computer. However, when the CAT format is used, the test is tailored to fit either different individual examinees or decision points. The items for an individual examinee are either selected on the basis of the individual's responses to previous items or on the basis of the amount of information at the cutting point. The length

(number of items) of the CAT varies by examinee as does the time limit (Wainer, 1990). In certification or licensure testing, the computerized adaptive testing format is referred to as computerized mastery testing[1] (CMT) (Way, Lewis, & Smith, 1995). For convenience, this term (computerized mastery testing or CMT) will be used in this paper.

Mastery Decision Making in CMT

The mastery decision in CMT is based on either ability (theta) estimates or a likelihood ratio test. When either the examinee's estimated ability (theta) or the likelihood ratio test meets the preset criterion, the test is stopped. Thus, the length of CMT can vary for different examinees. Two procedures can be used to make this decision: (1) sequential Bayes (SB), or (2) sequential probability ratio test (SPRT). These procedures are briefly described below. Additional details about these procedures are presented in Chapter II.

Sequential Bayes Procedure

With the sequential Bayes procedure, item responses are scored using a sequential Bayes estimation procedure. A

---

[1] Computerized mastery testing is a generic term used in this study to refer to mastery tests administered in a computerized testing format.

confidence interval (e.g., 90%) is then computed for the theta estimation. If the passing score is greater than the upper bound of this confidence interval, the decision is "fail". If the passing score is less than the lower bound of this confidence interval, the decision is "pass". Otherwise, testing continues (Weiss, 1985).

Sequential Probability Ratio Testing Procedure

The sequential probability ratio test (Wald, 1947) is used to decide which of two simple hypotheses (i.e., fail or pass) is more likely to be correct. In this procedure the likelihood of a response to an item under each of two alternative hypotheses is determined. If the likelihood is sufficiently larger for one hypothesis than for the other, that hypothesis is accepted and the test stops (Spray, & Reckase, 1987).

Spray and Reckase (in press) concluded that the SPRT procedure required fewer test items to achieve the same level of classification accuracy as the sequential Bayes procedure. In other words, the SPRT procedure appears to be more efficient and more powerful than the sequential Bayes procedure for making the pass/fail decision.

Item Response Theory Models Used with CAT

In CMT, the item presented to an examinee, at any point after the first item, is either selected on the basis of the responses to previous items or on the information at the cutting point. The selection of items is typically done using item parameter estimates to find the optimal item to administer to the examinee. The item parameter estimates are obtained from fitting previous item responses to a particular IRT model.

IRT models can be distinguished in terms of the dimensionality of the domains or latent traits thought to be measured by the test. Unidimensional IRT (UIRT) models assume that a single ability adequately accounts for the item/test performance, while multidimensional IRT (MIRT) models assume that more than one ability is necessary to account for the item/test performance. Within each of these two domains, there are a variety of models that can be used.

Items used in CMT are usually calibrated with UIRT models rather than MIRT models due to the availability of the unidimensional software for item calibration and the simplicity of the unidimensional models.

The Problem

Certification or licensure testing serves an important role in professional fields. Such tests promise to maintain

the quality of the people who are certified to practice the profession. In recent years, computerized mastery testing has been found quite promising in such tests because it is convenient and efficient.

When the computerized mastery test format is used, the items in the item pool are usually calibrated using a unidimensional item response theory (UIRT) model. However, the unidimensionality assumption is usually violated to some degree in most applications. If the assumption of unidimensionality is violated, there will probably be some adverse impact on the accuracy of the pass/fail classifications. The issue is whether the CMTs with UIRT models are still useful for making a dichotomous classification decision when the unidimensionality assumption is violated.

Another issue of interest in the use of the CMT format is the range of the test length. The range of the test length is usually preset in order to cover the test content specifications on the one hand and control the item exposure rate on the other. That is, the examinees must respond to a minimum number of items and not exceed a maximum number of items. If the range of test length varies, what will be the impact on the dichotomous classification decision?

An additional factor of interest is the level of difficulty of CMT. Certification or licensure testing may be

easy, medium, or difficult. In other words, the passing score may be set at different points along with the ability continuum in CMT. If the passing score varies, what will be the impact on the dichotomous classification decision?

The unidimensionality assumption, the range of the test length, and the location of the passing score are important features that can impact the classification decision. Only a few studies have examined the use of the SPRT when the unidimensionality assumption is violated (e.g., Abdel-fattah, Lau, & Spray, 1995, 1996).

The Purpose of Study

This study focused on the robustness of the SPRT procedure using UIRT models when the unidimensionality assumption was violated. Specifically, the impact of the violation of the unidimensionality assumption on the classification errors was of concern in this study.

When a mastery decision is made, two types of classification errors can occur: (1) type one error - an unqualified examinee is classified as qualified (false positive); and (2) type two error - a qualified examinee is classified as unqualified (false negative).

In this study, the effects of the following factors on these error rates were considered:

1. The type of UIRT model used to calibrate the items.

2. The degree of correlation between the two dimensions assumed to underlie the test responses.
3. The range of the test length.
4. The level of the cut-score.

In general, the major purpose of this study was to consider the impact of a number of factors on the robustness of the UIRT models used in a computerized mastery testing situation.

CHAPTER II

LITERATURE REVIEW

An Overview

As indicated in Chapter I, certification or licensure tests are used to control the quality of professionals entering a particular field. Computerized adaptive testing has been found to be more efficient than paper-and-pencil conventional testing because it provides individualized item selection for examinees of different abilities (e.g., Green, 1983; Hambleton, Swaminathan, & Rogers, 1991). When a certification or licensure test is administered in computerized adaptive format, it is frequently called a computerized mastery test (CMT). Item response theory plays a central role in CMT. It is used for managing the item pool, selecting items, estimating ability, and making the mastery decision. Unidimensional IRT (UIRT) models are usually adopted in CMT because of their simplicity and the availability of software. However, the unidimensionality assumption of UIRT is always violated to some degree. Therefore, it is critical to gain knowledge about the robustness of UIRT models used in CMT.

This chapter discusses computerized mastery testing in some detail. Its rationale and components, in particular its relationship with item response theory, are considered. Two procedures, sequential Bayes and sequential probability ratio testing are discussed and compared. In addition, important issues in IRT modeling such as dimensionality are considered and previous studies investigating the robustness of UIRT models are reviewed.

This chapter contains six sections. In Section I, computerized mastery testing is introduced. In Section II, the relationships between CMT and unidimensional item response theory are discussed. The assumptions and models of UIRT are also described. Additional details about CMT components are described in Section III. In Section IV, two procedures commonly used in CMT: sequential Bayes and sequential probability ratio testing are described and compared. Issues about dimensionality are discussed in Section V. In this section, previous research about the robustness of UIRT models is reviewed. The final section is a summary. As a whole, this chapter provides the rationale and the background for this investigation.

Section I. Computerized Mastery Testing: An Introduction

Computerized mastery testing is a special kind of computerized adaptive testing (CAT). The primary purpose of

CMT is to determine if the examinee has reached a certain required level of achievement. CATs are characterized by two distinct features that are not present in paper-and-pencil conventional testing: (1) scoring occurs during the testing process, and (2) items are selected and/or administered according to the examinee's responses to the previously administered items.

Weiss and Kingsbury (1984) identified six basic components of a CAT application:

1. An item response model.
2. An item pool.
3. A method for selecting the first item to administer.
4. A method for scoring and ability estimation at each step.
5. A method for selecting the subsequent items at each step.
6. A method for terminating the test.

In the context of computerized mastery testing, the testing procedures are similar to CAT. Thus the components for designing a CMT are almost the same as those for designing a CAT. The additional component needed for CMT is a method to set the cut-score. In CMT, only pass-fail distinctions are required so that precise measurement across a wide range of proficiency is not necessary (Thissen & Mislevy, 1990).

Advantages and Disadvantages of Computerized Mastery Testing

An examinee is measured most effectively when the test items are neither too difficult nor too easy for her or him (Lord, 1980). The primary goal of CMT (and CAT in general) is to "tailor" the test to fit each individual. According to Linacre (1988), the major advantages of CAT are: (1) improved test security, (2) shorter testing times, (3) quicker availability of results, and (4) reduced guessing and other undesirable test behavior. These advantages are achieved as a result of the individualized administration procedure. As noted above, this procedure matches the difficulty level of the items to the ability level of the examinee.

On the other hand, some potential disadvantages of the CAT procedure relative to the traditional paper-and-pencil procedure include: (1) greater costs - both developmental and administrative, (2) the need for larger item pools and for larger samples to calibrate the items, and (3) test security problems if the CAT is not administered in a proper way.

Section II. Unidimensional Item Response Theory and Computerized Mastery Testing

In the context of computerized mastery testing, unidimensional item response theory models are usually

applied to calibrate the items, to estimate abilities, to select the items to administer, and to make the mastery decision. Thus, UIRT is an important element of CMT.

Item response theory is in fact a collection of mathematical models that define how the item response depends on the ability level of the examinee. It describes the interaction between an examinee and a test question using the probability of a correct response to the test question as the dependent variable. The relationship between the probability of correct response and the corresponding ability is described by the item response function. The performance of an examinee on a test question can be predicted on the basis of the item characteristics and the ability level of the examinee. Common unidimensional IRT models are discussed below.

Unidimensional IRT Models

For the unidimensional IRT models, the three-parameter logistic (UIRT 3-PL), two-parameter logistic (UIRT 2-PL) and one-parameter logistic (UIRT 1-PL) are commonly used (Hambleton & Swaminathan, 1985).

The UIRT 3-PL model defines the probability of a correct response as

$$P_i(\theta_j) = P_i\langle x_{ij} = 1 \mid \theta_j \rangle = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i(\theta_j - b_i)\right]},$$

where

$P_i(\theta_j)$ is the probability that an examinee with ability θ_j answers item i correctly,

x_{ij} is the dichotomous (1/0) response to item i by person j,

a_i is the item i discrimination parameter,

b_i is the item i difficulty parameter,

c_i is the item i "guessing" parameter, and

D is a scaling factor applied to make the logistic model as close as possible to normal ogive model, in general, D=1.7.

The UIRT 2-PL model and UIRT 1-PL model are special cases of the UIRT 3-PL model. For the UIRT 2-PL model, the "guessing" parameter is set equal to zero for all items. For the UIRT 1-PL model, the discrimination parameter is constant across all items and the "guessing" parameter is set equal to zero for all items.

Among these three models, the 1-PL model is the most restrictive: correct guessing is assumed not to occur and the discrimination parameters are assumed to be equal across items. Only the difficulty parameter varies for the 1-PL model. The 3-PL model is the least restrictive among the three models. This model allows for correct guessing and permits both the discrimination and the difficulty parameters to vary.

In CMT, items are selected to be administered according to the item information either at the previous estimated θ or at the cutting point: the greater the information, the greater the probability of selection. More information means more accuracy or less error in the ability estimation. However, the amount of information given by an item varies with ability level. Information at a given ability level varies directly as the square of the item discriminating power, a_i (Lord, 1980). The item information function, $I_i(\theta)$, can be stated as follows (Lord, 1980):

$$I_i(\theta) = \frac{P_i'^2}{P_i Q_i},$$

where

$P_i = P_i(\theta)$ is the item response function,

$Q_i = 1 - P_i$, and

P_i' is the derivative of P_i with respect to θ.

The Unidimensionality and Local Independence Assumptions

Unidimensionality means that only one ability is measured by a set of items in a test. Local independence means that the item responses of examinees with same ability level are statistically independent. Local independence can be obtained when all the ability dimensions influencing performance have been taken into account (Hambleton &

Swaminathan, 1985). The property of local independence can be stated as follows:

$$\text{Prob}(U_1, U_2, \ldots, U_n|\theta) = P(U_1|\theta)P(U_2|\theta)\ldots P(U_n|\theta),$$

where

U_i is the response of a randomly chosen examinee to item i(i=1,2,...,n), and

$P(U_i|\theta)$ is the probability of the response of a randomly chosen examinee with ability θ.

When the assumption of unidimensionality is true, local independence is also obtained. However, local independence can be obtained even when the data set is not unidimensional.

If the assumptions are satisfied and the data fit the specific UIRT model being used, two desirable features are obtained. One of these is the invariance of ability estimation (Lord, 1980; Hambleton & Swaminathan, 1985). This invariance feature implies that the ability parameters can be estimated with different items. In other words, examinees' ability estimates are not test-dependent. The other feature is the invariance of item parameters. This invariance feature implies that the item parameters can be estimated with different ability level groups. In other words, item parameter estimation is not group-dependent. If these two features hold, many practical measurement problems

can be easily solved. Specifically, in the CMT context, invariance of ability estimation means that no fixed set of items is needed to estimate an examinee's ability. Without this capacity, computerized adaptive testing would not be possible.

Section III. Components of CMT

Lord (1980) suggested the following steps in the design of a mastery test for a unidimensional skill with a single cut point, θ_c:

1. Obtain a pool of items for measuring the skill of interest.

2. Calibrate the items on some convenient group by determining the parameters a_i, b_i, c_i for each item.

3. Consider the entire item pool as a single test; determine what true-score level, ξ_c or levels ξ_1 and ξ_2, will be used to determine mastery. This decision is a matter of judgment for the subject-matter specialist.

4. Using the item parameters obtained in step 2, find θ_c (or θ_1 and θ_2) from ξ_c (or from ξ_1 and ξ_2) by means of the relation $\xi \equiv \sum_i P_i(\theta)$.

5. Compute $P_i(\theta_c)$ for each item.

6. Evaluate $I\{\theta_i, u_i\} \equiv P_i'^2/P_iQ_i$ at θ_c for each item.

7. Decide what length confidence interval for θ will be adequate at θ_c. Find the required $I_c\{\theta\}$.

8. Select items with the highest information, $I\{\theta, u_i\}$, at θ_c. Continue selecting until the sum $\sum^n I_c\{\theta, u_i\}$ equals the required $I_c\{\theta\}$.

9. Compute scoring weight $w_i^c = P_i'/P_iQ_{i|\theta=\theta c}$ for each selected item.

> 10. For each examinee, compute the weighted sum of item scores $Y = \sum_i w_i u_i$. (In practice, an unweighted score may be adequate.)
>
> 11. Compute the cutting score $Y_c \equiv \sum_i P_i'(\theta_c)/Q_i(\theta_c)$.
>
> 12. Accept each examinee whose score Y exceeds Y_c; reject each examinee whose score is less than Y_c (pp. 174-175).

In summary, according to Lord, there are four main components in computerized mastery testing: (1) developing the item pool and calibrating the items, (2) setting the cut-score, (3) selecting the items to administer, and (4) making mastery decision and test termination rules. Typically, item pool development and cut-score setting are done by the content-experts in the related field. Item selection is usually implemented according to the IRT information function. The termination of CMT is based on: (1) the accuracy of the ability estimation, or (2) the number of items that have been administered. These four components of CMT are discussed below.

Item Development and Calibration

Content experts are typically used to write test items for CMT situations. Usually, the specifications for the test are very well defined and most items are developed within a multiple-choice format.

After a sufficiently large number of items are written, items are administered to some representative groups and then, based on the responses, the items are calibrated. A

particular UIRT model is usually adopted for the item calibration (e.g., Hambleton, Swaminathan, & Rogers, 1991).

Setting the Cutting Point

A primary purpose of mastery testing is to determine whether or not an examinee has reached a certain required level of achievement. In latent trait test theory, the level of achievement can be denoted by θ. Presumably, there is a cutting point, θ_c on the θ continuum that divides certified from non-certified examinees. That is, if the θ for an examinee is equal to or greater than θ_c, the examinee is considered certified; however, if the θ is less than θ_c, the examinee is considered non-certified. The cutting point is typically set by the experts in the content area. Several methods can be adopted to set the cut-score and are briefly described below.

Nedelsky (1954) developed a method of setting the cut-score that applies only to tests using multiple-choice items. In this procedure, a panel of experts determines the number of distractors that a minimally successful student would know to be incorrect choices for each item. A cut-score is then computed from the expected scores based on these judgments about the distractors.

Angoff's (1971) method is similar to Nedelsky's but it can be applied to test formats other than multiple-choice. In the Angoff method, experts examine each item and judge

the probability that a minimally competent student would correctly respond to the item.

Ebel's (1972) method requires the experts to classify each item according to its relevance and its difficulty. This technique uses a two-dimensional grid (relevance and difficulty) to categorize each item. The items are first categorized into the cells of the grid, and then the judge estimates the percentage of these items that a minimally successful student would be expected to answer correctly. The cut-score, x_c, is calculated using the formula:

$$x_c = \sum P(m),$$

where

x_c is the cut-score in terms of the raw score scale,

P is the proportion of items in the cell that a minimally qualified student should response to correctly, and

m is the number of items in the cell.

If there is more than one judge, the final passing score is based on the average of all judges' cut-scores.

Nedelsky's (1954) method, Angoff's (1971) method, and Ebel's (1972) method are based on judgments about individual items. Other methods for setting the cut-score are based on judgments about individual examinees. Among the most common

of these methods are: (1) the borderline group method (Nedelsky, 1954), (2) the contrasting group method (Nedelsky, 1954), and (3) the up-and-down method (Livingston & Zieky, 1982).

With the borderline group method, a group of minimally successful students is identified and administered the test. The cut-score is set at the median of their scores.

The contrasting group method assumes that the test-takers can be divided into two contrasting groups, a qualified group and an unqualified group. After administering the test to both groups, the cut-score is set at the intersecting point of the smoothed score distributions of these two groups. Crocker and Algina (1986) have identified the specific steps in this procedure:

1. Select qualified judges who are familiar with the examinee population.

2. Allow the judges to discuss and, if possible, agree on what constitutes minimally competent performance.

3. Use the judges to identify examinees who are competent or incompetent performers (excluding any who appear to be borderline).

4. Test both groups of examinees.

5. Plot the score distribution for each group on the same continuum.

6. Set the performance standard at the intersection point of the two distribution curves.

The up-and-down method is a variation of the contrasting group method. An examinee whose test score is

thought to be at the proper passing score is selected and her/his competency is judged. If the examinee is judged to be qualified, then an examinee with a test score lower than this examinee is selected and her/his competency is judged. If the examinee chosen is judged to be not qualified, then an examinee with a test score higher than this examinee is selected and her/his competency is judged. This process is continued by choosing each examinee based on the judgment of the previous examinee. This process can be stopped when several direction-changes have been observed. That is, the direction is changed from up to down or vice versa. The cut-score is set as the average of the qualified-and-unqualified (up-and-down) scores.

Item Selection and Administration

In the context of computerized mastery testing, items may be selected based on different criteria, depending on the particular procedure being applied. Two procedures, sequential Bayes (SB) and sequential probability ratio testing (SPRT), are frequently used (e.g., Reckase, 1983; Kingsbury & Weiss, 1983).

With the SB procedure, the first item(s) for administration is/are usually of average difficulty because the ability level of the examinee is unknown. After obtaining the first estimate of ability, items with the greatest information at this estimation of the examinee's θ

are selected. With the SPRT procedure, items with the greatest information at the cutting point are selected.

Mastery Decision and Test Termination Rule

The mastery decision is based on the cut-score. The basic idea is as follows: if the examinee's θ estimate is at or above the cut-score, then the examinee is classified as a master; otherwise, the examinee is classified as a nonmaster. Because the scores or abilities of the examinees are estimates, the degree of accuracy of these estimates is usually considered.

Two kinds of errors can happen in such dichotomous classification systems: (1) classifying an examinee whose true θ is below θ_c as a master, and (2) classifying the examinee whose true θ is equal to or above θ_c as a nonmaster. The following notation is used to represent the probabilities of these two types of errors:

α = prob(classify as a master | $\theta < \theta_c$)

β = prob(classify as a nonmaster | $\theta \geq \theta_c$)

where

θ_c is the cutting point on the theta scale.

Generally, two criteria are imposed to terminate the CMT administration. First, if the degree of accuracy of estimation reaches a certain preset degree, a mastery/nonmastery decision can be made and the test

administration stops. Second, if the maximum permitted number of items has been reached, the test administration stops.

Section IV. SB and SPRT Procedures in CMT

Assuming that a cut-score has been established, two procedures can be used to select items to be administered, to make the mastery decisions, and to terminate the test. One of these is the sequential Bayes procedure and the other is the sequential probability ratio testing procedure. These two procedures are discussed in more detail below.

Sequential Bayes Procedure

The basic rationale of the SB procedure is to match the item difficulty with the examinee's ability so that the estimation can be more accurate. For this purpose, items are always ranked on the amount of information at a posterior estimate of an examinee's latent ability or theta. With the SB procedure, the first administered item(s) is/are usually of middle difficulty. The item selected next is the one which is predicted to provide the greatest amount of information about a particular examinee, given an estimate of that examinee's θ level. This procedure is called restricted Bayesian updating (Owen, 1975) and requires that the prior distribution of θ be assumed to be normal before the first item administration. In this procedure, the

posterior distribution of θ, following the administration of the *i*th item, is also assumed to be normal.

With this procedure, testing continues until the confidence interval of the examinee's estimate θ does not contain θ_c, the cutting point on the ability scale. If the lower bound of the interval is above θ_c, the examinee is categorized as a master. If the upper bound of the interval is below θ_c, the examinee is categorized as a nonmaster. If the confidence interval contains θ_c, the testing will continue with the next item using the restricted Bayesian updating procedure.

In summary, the procedures for selecting items, making a mastery decision and terminating the test with SB are summarized as followed:

1. Establish a cut-score.
2. Administer the first item(s) (middle difficulty).
3. Select items from the remaining items in the pool according to the information function. The item with the greatest information at the examinee's estimated θ is selected.
4. Administer the selected items and estimate the ability parameter of the examinee.
5. When the estimation reaches the preset degree of precision, compare the examinee's estimated θ with the cut-score.

If the lower bound of the interval is above θ_c, categorize the examinee as a master and terminate the test.

If the upper bound of the interval is below θ_c, categorize the examinee as a nonmaster and terminate the test.

If the confidence interval contains θ_c, continue testing with the next item using the restricted Bayesian updating procedure.

6. If the maximum permitted number of items has been reached, terminate the testing. When this occurs, the mastery/nonmastery decision must be made on the basis of the policies established by the professional organization administering the tests. One common practice in this situation is the following: if the estimated θ is equal to or greater than the θ_c, categorize the examinee as a master; and if the estimated θ is less than the θ_c, categorize the examinee as a nonmaster.

Sequential Probability Ratio Testing Procedure

The sequential probability ratio testing (SPRT) developed by Wald (1947) can be used to classify examinees into two categories (mastery/nonmastery). Ferguson (1969) was the one of the first researchers to apply the SPRT

procedure to criterion referenced testing. Reckase (1983) noted that the SPRT procedure could be modified and applied to adaptive mastery testing with UIRT models.

Wald's SPRT theory makes use of the local independence assumption of IRT through the formulation of the likelihood functions under H_0 and H_1 as products of probabilities. SPRT theory does not require the probabilities to be constant from item to item within the pool (Spray & Reckase, 1987).

The SPRT procedure is based on the binomial model. The practical advantages of this model are: (1) it is relatively simple to implement in a computer-based testing system; and (2) it requires no prior data collection on test item parameters (Frick, 1990).

With the SPRT procedure, the test items are selected to be administered to each examinee according to the amount of information at the cutting point. The test items are ranked at the cutting point with respect to the amount of information provided. That is, the greater the information, the greater the priority of administration. The use of the SPRT procedure in computerized mastery testing has not been widespread. The general procedure of SPRT is outlined below (Spray & Reckase, 1987).

Suppose an examinee has ability θ_j. The decision about the examinee's status (pass or fail) is made on the basis of a consideration of two simple hypotheses:

$$H_0: \theta_j = \theta_0$$

versus

$$H_1: \theta_j = \theta_1$$

where

θ_j is an unknown parameter, and

θ_0 and θ_1 represent decision points that correspond to lower and upper limits, respectively, of the passing criterion, δ, where $\theta_0 < \delta < \theta_1$.

If $P(\theta_j)$ is the probability that examinee j responds correctly to an item, and $Q(\theta_j) = 1 - P(\theta_j)$ is the probability that examinee j responds incorrectly to an item, then,

$$\pi(\theta_j) = \text{Prob}(X=x \mid \theta=\theta_j) = P(\theta_j)^x Q(\theta_j)^{1-x},$$

where

x = 1, correct response, and

x = 0, incorrect response.

The functions, $\pi(\theta_1)$ and $\pi(\theta_0)$, are called likelihood functions of x, and a ratio of these two functions, $L(x) = \pi(\theta_1) / \pi(\theta_0)$, is called a likelihood ratio and

$$L(x_1, x_2, \ldots, x_n \mid \theta_0, \theta_1) = \frac{\pi_1(\theta_1)\pi_2(\theta_1) \ldots \pi_n(\theta_1)}{\pi_1(\theta_0)\pi_2(\theta_0) \ldots \pi_n(\theta_0)} \quad .$$

The likelihood ratio is compared to the boundaries, A and B,

where

$A = (1-\beta)/\alpha$, and

$B = \beta/(1-\alpha)$,

where α and β are the error probabilities defined as follows:

Prob(choosing H_1 | H_0 is true) = α (false positive)

Prob(choosing H_0 | H_0 is true) = 1 - α (correct decision)

Prob(choosing H_0 | H_1 is true) = β (false negative)

Prob(choosing H_1 | H_1 is true) = 1 - β (correct decision)

For example, if α and β are preset at .05, then

A = (1-.05)/.05 = 19, and

B = .05/(1-.05) = .053.

The mastery decision and test termination rules in SPRT are as follow:

If $L(x_1,x_2,...,x_n \mid \theta_0,\theta_1) \geq A$, then H_1 is accepted. In this case, the examinee is classified as a master and the test administration is stopped.

If $L(x_1,x_2,...,x_n \mid \theta_0,\theta_1) \leq B$, then H_0 is accepted. In this case, the examinee is classified as a nonmaster and the test administration is stopped.

If $B < L(x_1,x_2,...,x_n \mid \theta_0,\theta_1) < A$, then no decision is made and another item is selected to be administered.

The region from θ_0 to θ_1 is the indifference region. The distance, $|\theta_0 - \theta_1|$ is the width of the indifference region. Test length is a function of this region.

The procedures for selecting items, making a mastery decision and terminating the test with SPRT are summarized below:

1. Establish a cut-score.
2. Administer the first item with the greatest information at the cutting point.
3. Compute the likelihood ratio.
   - If the likelihood ratio function is at or above the preset upper bound, classify the examinee as a master and terminate the test.
   - If the likelihood ratio function is below the preset lower bound, classify the examinee as a nonmaster and terminate the test.
   - If the likelihood ratio function is between the preset lower and upper bound, administer the next item according to the information functions of the remaining items at the cutting point.

4. If the maximum permitted number of items has been reached, terminate the test. When this occurs, the mastery/nonmastery decision must be made on the basis of the policies established by the professional organization administering the tests. One common practice in this situation is to make decision based on some "distance rule". That is, compare the logarithm value of the likelihood function to the logarithm values of the boundaries, A and B. If the $|\log\{L(x_1,x_2,...,x_{max})\}-\log A|$ smaller than $|\log\{L(x_1,x_2,...,x_{max})\}-\log B|$, categorize the examinee as a master; and vice versa (Spray, 1993).

Comparison between SB and SPRT Procedures

In 1980, Lord suggested that:

> when similar groups of examinees are tested year after year, the psychometrician knows, in advance of testing, the approximate distribution of ability in the group to be tested. In this case, a sequential Bayes approach is appropriate. If, on the other hand, the distribution of ability in the group to be tested is unknown, what is needed is a way of evaluating the testing procedure that does not depend on the unknown distributions of ability in the groups to be tested (pp. 162-163).

Compared with the SB procedure, the SPRT procedure does not necessarily require an assumption about the ability distribution. Usually, having fewer assumptions implies that

the model/procedure can be applied in more situations. From this perspective, SPRT seems a better procedure than SB.

Several studies have compared the accuracy of the SPRT procedure with that of the SB procedure. However, these studies have not produced consistent results.

A study by Kingsbury and Weiss (1983) supported the use of SB procedure relative to SPRT. However, seven years later, Frick (1990) observed that the SPRT formulation used by Kingsbury and Weiss (1983) could not be algebraically transformed into Wald's original formulation.

In addition, for the SPRT procedure, Kingsbury and Weiss (1983) applied the random item selection procedure instead of selecting items with the greatest information at the cutting point.

Frick (1990) also compared the SPRT and SB procedures and found that the SB procedure was superior. However, Frick also used a random item selection procedure as part of SPRT.

In order to compare SB and SPRT procedures in terms of efficiency, Spray and Reckase (in press) performed a simulation study. A unidimensional IRT 3-PL model was adopted. The decision criterion, δ_c, was set to -.5, .0, and 1.5 respectively. Simulated examinees with known θ_j were administered items from a previously calibrated pool of 200 items. The maximum number of items that could be administered was set to 50.

For the SPRT procedure, the items were ranked and administered according to the information at the cutting point. The nominal error rates, α and β, were set at .05 and the width of indifference region was fixed at 1.0. For the SB procedure, items were administered according to the information at the estimated theta level of a particular examinee. The results of this study indicated that to achieve approximately the same level of classification accuracy, the SPRT procedure usually required fewer items than the sequential Bayes procedure. Thus, they concluded that the SPRT procedure was more efficient than the SB procedure.

Section V. Dimensionality Issues

Item response theory models can deal with both single dimensional or multiple dimensional data. If there is only one single ability required to correctly respond to a test question, then the unidimensional IRT (UIRT) models can be adopted. If, on the other hand, there is more than one ability required to respond correctly to a test question, multidimensional IRT (MIRT) models should be applied. The UIRT models and their assumptions have been discussed previously. MIRT models are discussed below.

Multidimensional Item Response Theory Models

Multidimensional item response theory (MIRT) models are used with response data that occur as a result of more than one ability dimension. Basically, there are two major kinds of multidimensional IRT models: compensatory and noncompensatory. Compensatory models permit high ability on one dimension to compensate for low ability on another dimension. Noncompensatory models do not permit this. That is, for noncompensatory models, weakness in one ability dimension cannot be compensated for by strength in another dimension.

McKinley and Reckase (1983) proposed a compensatory multidimensional IRT (COMIRT) three-parameter model (COMIRT 3-PL) that permits high ability on one dimension to compensate for low ability on the others. This model defines the probability of a correct response as:

$$P_i\langle x_{ij} = 1 \mid \underline{a}_i, di, c_i; \underline{\theta}_j \rangle = c_i + \frac{1 - c_i}{1 + \exp\left[-D\left(\sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i\right)\right]},$$

where

x_{ij} is the dichotomous (1/0) response to item i by person j,

$\underline{a}_i = (a_{i1}, a_{i2}, \ldots, a_{im})$ is the vector of item discrimination parameters for item i,

d_i is a scale parameter that is related to the difficulty parameter for item i,

c_i is the guessing parameter for item i,

D is a scaling factor applied to make the logistic model as close as possible to normal ogive model, in general, D=1.7,

$\underline{\theta}_j = (\underline{\theta}_{j1}, \underline{\theta}_{j2}, \ldots, \underline{\theta}_{jm})$ is the vector of ability parameters for person j, and

m is the number of dimensions.

Sympson (1978) proposed the following noncompensatory multidimensional IRT 3-PL model:

$$p_i\langle x_{ij} = 1 \mid \underline{a}_i, \underline{b}i, c_i; \underline{\theta}_j \rangle = c_i + \frac{1 - c_i}{\prod_{k=1}^{m} (1 + \exp[-Da_{ik}\{\theta_k - b_{ik}\}])},$$

where

$\underline{a}_i$ is a vector of discrimination parameters, and

$\underline{b}_i$ is the vector of difficulty parameters.

Which one should be applied? If a test consists of several dimensions and if the items in the test can be grouped within each dimension, then a compensatory model might be appropriate. However, if the items require simultaneous application of several dimensions to be

answered correctly and if an examinee's ability is below a certain threshold on one dimension, then no amount of strength in other abilities can offset this weakness. In this situation, a noncompensatory model would seem to better fit the data.

A Special Case: $r_{\theta 1 \theta 2} = 1$

When the correlation between two sets of θ-values is equal to one, the compensatory MIRT 3-PL model is in fact, a unidimensional 3-PL model. The 2-D COMIRT 3-PL model can be written as:

$$p_i(\theta_{j1}, \theta_{j2}) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + d_i\right)\right]} \qquad \text{(a)},$$

If $r_{\theta 1 \theta 2} = 1$, then $\theta_{j1} = \theta_{j2}$, assuming $\bar{\theta}_1 = \bar{\theta}_2$ and $SD_{\theta 1} = SD_{\theta 2}$,

where

$\bar{\theta}_1$ and $\bar{\theta}_2$ are the means of θ_1 and θ_2 respectively, and

$SD_{\theta 1}$ and $SD_{\theta 2}$ are standard deviations of θ_1 and θ_2 respectively.

Then, equation (a) can be written as a unidimensional model:

$$p_i(\theta_{j1}, \theta_{j2}) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\left((a_{i1} + a_{i2})\left(\theta_j + \frac{d_i}{a_{i1}+a_{i2}}\right)\right)\right]} \qquad \text{(b)},$$

where

$(a_{i1}+a_{i2})$ is the item discrimination parameter, and

$\frac{d_i}{a_{i1}+a_{i2}}$ is the item difficulty parameter.

Relation between Dimensionality and Computerized Mastery Testing

As noted above, item response theory plays a central role in computerized mastery testing. Most computerized mastery tests available now are based on unidimensional item response theory models because of their simplicity. The unidimensionality assumption of UIRT will not be met absolutely in most testing situations. When this assumption is violated, it will have some negative impact on item parameter calibration and ability estimation (e.g., Ansley & Forsyth, 1985). Thus, the accuracy of the mastery/nonmastery decision may be impacted. It seems reasonable to assume that classification errors may increase as the "degree" of multidimensionality increases.

Robustness of UIRT Models

Unidimensional item response theory models are an important element of CMT. Compared with multidimensional IRT models, unidimensional IRT models are simpler and require smaller sample sizes to estimate the item parameters. However, the unidimensionality assumption of UIRT is probably violated to some degree in any application.

Ansley and Forsyth (1985) noted that the violation of the assumption of unidimensionality impacted the parameter estimation for the modified (fixed c-parameter) 3-PL logistic model. They suggested that parameter estimates had to be interpreted carefully when a unidimensional IRT model was applied to two-dimensional data.

On the other hand, Drasgow and Parsons (1983) found that unidimensional IRT models provided good descriptions of multidimensional data sets when the dominant latent trait was sufficiently "strong". They suggested that researchers should be more concerned with the robustness of estimation techniques to minor violations of dimensionality assumptions than with measuring all latent abilities in a particular content domain.

Reckase, Ackerman, and Carlson (1988), using both simulated and real data, demonstrated that sets of items which require more than one ability for a correct response still meet the unidimensionality assumption of most IRT models. These items can be selected to construct a test based on UIRT models. They showed that the unidimensionality assumption required only that the items in a test measure the same composite of abilities, rather than a single ability.

Correlation between θ_1 and θ_2

Ansley and Forsyth (1985) used two-dimensional item sets with Sympson's noncompensatory three-parameter multidimensional item response theory model. (The guessing parameter was fixed at .2.) They noted that, as the correlation between the ability dimensions increased, the average absolute differences between values of the true ability parameters and the estimated ability parameters decreased. They concluded that the unidimensional estimate of the a-value (discrimination parameter) seemed best considered as the average of the multidimensional a-values, that the estimated b-value (difficulty parameter) seemed to be an overestimate of the multidimensional b-values, and that the ability values were best considered as the average of the true multidimensional abilities. They also observed that sample size and test length had very little effect on these interpretations.

Folk and Green (1989) used two-dimensional simulated item sets with a compensatory MIRT model in both adaptive and nonadaptive testing format to study the effects of the correlation between thetas on item parameter estimates. The UIRT 3-PL model was used to do the calibration. They found that as the correlation between ability dimensions decreased, the estimated a-parameter increasingly emphasized either a_1 or a_2. The average estimated a-parameter values

were highest when the correlation between theta 1 and theta 2 was 1. They suggested that if the data did not deviate greatly from unidimensionality, the UIRT model can provide reasonable approximations. If, however, the correlation between θ_1 and θ_2 is low, the data sets are better characterized as multidimensional and they cannot be approximated with a unidimensional model.

Using simulated two-dimensional data with both compensatory and noncompensatory MIRT models and different correlations between the abilities, Ackerman (1987) found that as the correlation between the two-dimensional abilities increased, the response data appeared to become more unidimensional. Using simulated two-dimensional data with compensatory MIRT model, Ackerman (1991) again found that the orientation of the unidimensional scale in relationship to the two-dimensional ability plane appears to be a function of the multidimensional composition of the items administered in the test. (This conclusion is consistent with Wang (1986).) It was also found that it can be determined whether a test is measuring mostly ability one, mostly ability two, or both ability one and ability two equally by examining the relationship of the projected contours of the plane of two-dimensional estimated abilities and the corresponding angles. For instance, if both

dimension one and dimension two are equally dominant, the corresponding angle is 45 degrees.

In summary, on the basis of the above studies, it seems reasonable to conclude that as the correlation between thetas in the two-dimensional case increases, the response data "act" more like a unidimensional data set.

Studies about Dimensionality in Computerized Mastery Testing

Within the context of CMT, relatively few studies have investigated the impact of using unidimensional item response theory models when the item pool is multidimensional. One study by Abdel-fattah, Lau, and Spray (1995) used the UIRT 3-PL model to estimate the item parameters of two-dimensional data set and used the SPRT procedure to make the mastery decision. The following conditions were manipulated: (1) $r_{\theta 1 \theta 2}$ was set either equal to .00 or .50; (2) two test length constraints (min=40, max=360; min=1, max=360) were applied; and (3) the correlation between first and second discrimination parameters was either .25 or .50.

Abdel-fattah et al. concluded that: (1) under certain conditions, UIRT 3-PL models can be used in computerized mastery testing even when the unidimensionality assumption is violated; and (2) when the unidimensionality assumption is violated, the false negative error (type II error) rates

are consistently higher than false positive error (type I error) rates. More specifically, they observed the following:

1. The type I error and the number of items used to make the mastery decision were less when $r_{\theta 1\,\theta 2}$=.50 than when $r_{\theta 1\,\theta 2}$=.00. No specific pattern was apparent for the type II and total error rates. The difference in the number of items used was significant. However, the type I, type II, and total error rates were relatively similar in both cases.
2. The total error rates generally decreased when the minimum length of the test was increased from 1 to 40 items. However, the differences were very slight.
3. As the correlation between a_1 and a_2 increased, the total error rate decreased. However, this decrease was not significant.

Different Levels of the Cutting Point

Within the context of CMT, Abdel-fattah, Lau, and Spray (1996) investigated the impact of using an UIRT model when the item pool was multidimensional and when different difficulty levels were used for the cut-score.

In this study, the UIRT 3-PL model was adopted to calibrate two-dimensional data. The cut-score was set at .5,

.6, and .7 respectively. The other manipulated conditions were: (1) $r_{\theta 1 \theta 2}$ was set either equal to .00 or .50; (2) two test length constraints (min=40, max=360; min=1, max=360) were applied; and (3) the correlation between the first and second discrimination parameters was either .25 or .50. Again, the UIRT 3-PL model was found useful in mastery decision making with two-dimensional data. More specifically, Abdel-fattah et al. observed the following:

1. As the level of cut-score increased, the number of items used to make the dichotomous decision decreased.
2. As the level of cut-score increased, the type I error rates generally decreased. However, there was no consistent pattern for the total error rates.
3. Type II error rates were greater than type I error rates at each cutting point.
4. The results of the other manipulations were consistent with their 1995 study.

Section VI. Summary

Certification or licensure testing presumably helps to guarantee that professional practitioners have met a minimum standard. If the certification or licensure testing is administered in a computerized adaptive format, it is called computerized mastery testing. CMT is more efficient than conventional testing because of the individualized

administration procedure. The four main components of CMT are: (1) item pool development and calibration, (2) cut-score setting, (3) item selection to administer, and (4) mastery decision and test termination.

For item writing and cut-score setting, content experts are usually employed. For item selection, mastery decision and test termination, two procedures, sequential Bayesian and sequential probability ratio testing, can be applied in CMT. The SPRT procedure has been found more efficient and more powerful than the SB procedure (Spray & Reckase, in press).

Compared with multidimensional IRT models, unidimensional IRT models are simpler and require smaller sample sizes to estimate the item parameters. For these reasons, UIRT models are usually adopted in the context of CMT to do the item calibration and ability estimation. However, the unidimensionality assumption of UIRT is probably violated to some degree in any application. If the unidimensionality assumption is violated, there may be some negative impact on the item calibration, ability estimation, and mastery/nonmastery decisions.

Previous studies have noted the following about the use of unidimensional IRT models with multidimensional data: (1) unidimensional models provide a good description of the multidimensional data set when the dominant latent trait is

sufficiently potent (Drasgow & Parsons, 1983); (2) as the correlation between the dimensions or abilities increases, the usefulness of the unidimensional models increases (e.g., Ackerman, 1987; Folk & Green, 1989); and (3) if the items in a test measure the same composite of abilities, unidimensional models are robust (Reckase, Ackerman, & Carlson, 1988).

Only a few studies investigated the effects of violation of the unidimensionality assumption within the CMT context. Specifically, the impact of the violation of the unidimensionality assumption on the use of the SPRT procedure has not been extremely studied. Abdel-fattah, Lau, and Spray (1995, 1996) did provide some evidence that the UIRT 3-PL model could be useful.

The purpose of this study was to further examine the use of SPRT procedures for computerized mastery testing when the unidimensionality assumption is violated.

The simplest multidimensional situation is the one involving two dimensions. Two-dimensional data provide a good starting point to study the "robustness" of UIRT models. Within a two-dimensional situation, there are a variety of conditions that can be considered: the type of UIRT model used, the correlations between the two latent traits needed to respond correctly to the items, the range of test length, and the level of cutting point. This study

was designed to provide information about the effects of these factors on the accuracy and efficiency of decision making using the SPRT procedure in a CMT setting.

CHAPTER III

METHODOLOGY

The primary purpose of this study was to examine whether the unidimensional item response theory models used in making mastery decisions with the sequential probability ratio test (SPRT) procedure produced acceptable accuracy in terms of type I errors (false positive) and type II errors (false negative) when the assumption of unidimensionality was violated. Monte Carlo simulation techniques were used to assess the robustness of the UIRT models under a variety of testing conditions. In this study, a model is defined as robust if the model provides acceptable classification accuracy even when the assumptions of the model are violated.

There are four sections in this chapter. In Section I, the research questions are defined. The simulation procedures are described in Section II and the research design is given in Section III. The final section describes the data analysis procedures.

Section I. Research Questions

This study addressed the following research questions:

1. Does computerized mastery testing using the sequential probability ratio testing procedure based on UIRT models have acceptable accuracy in terms of type I and type II error rates when the true IRT models underlying the responses are two-dimensional?
2. Is the accuracy of the UIRT model-based procedures affected by the
   (1) use of different UIRT models such as the three-parameter model or the one-parameter model?
   (2) correlation between the two sets of ability parameters?
   (3) range of test length?
   (4) level of the cut-score?

Section II. Simulation Procedure

This study was based on simulated data rather than the responses of actual subjects. The reasons for using simulated data were: (1) the factors of interest could be manipulated more efficiently; (2) large response data sets could be generated; and (3) the true item parameters could be fixed.

The basic simulation procedure used in this study was to generate two-dimensional dichotomous data (i.e., 1/0) and then calibrate the items, estimate the thetas, and choose

items to administer with UIRT models under different conditions. The robustness of UIRT models was then investigated by comparing the false positive classification rates, the false negative classification rates, and the numbers of items administered for classification under the different conditions.

Computer Programs

As this study used simulation, several computer programs were needed. Some of the these programs had been developed for general uses, others were developed specifically for this study. A brief description of these programs follows.

NOHARM

Developed by Fraser (1983, 1986), Normal Ogive Harmonic Analysis Robust Method (NOHARM) applies the generalized multidimensional normal ogive item response model. This program was used to calibrate a set of actual items in order to identify a set of two-dimensional MIRT parameters that could be used in the simulation. (See Figure 1.)

GENMIRT

Developed by Spray (1995) and modified by the author, GENMIRT was used to generate dichotomous data (1/0) based on a two-dimensional compensatory three-parameter model (2-D COMIRT 3-PL). The correlation, ρ, between θ_1 (ability one)

and θ_2 (ability two) could be manipulated in this program. θ_1 and θ_2 were randomly selected from a bivariate normal distribution. Two thousand "examinees" were created. (See Figure 2.)

BILOG 3

With 1/0 data generated by GENMIRT on a sample of 2,000 "examinees", BILOG 3 (Mislevy & Bock, 1990) was used to calibrate the two-dimensional item parameters using either the UIRT 3-PL model or the UIRT 1-PL model. The marginal maximum likelihood estimation procedure was used in BILOG. One thousand randomly drawn respondents were used to do the item calibration.

UTCC

UTCC is a program developed by the author. With item parameters calibrated by BILOG, this computer program was used to map the cut-score (i.e., $P(\theta_c)=0.4$, 0.6, and 0.8) onto the UIRT theta scale. (See Figure 3.) The curve in Figure 3 is based on the test characteristic function. The formula is as following:

$$P_T(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} P_i(\theta),$$

where

$P_i(\theta)$ is the item characteristic function, and

n is the number of items.

MIRTSPRT

Developed by Spray (1995) and modified by the author, this computer program was used to determine the false positive rates, false negative rates, and average number of items used to make the mastery decision under different conditions. (See Figure 4.)

Data Simulation

Item Pool

In this study, the two-dimensional item parameters used to create the simulated data were based on actual test data. The actual test data were the examinees' responses for six test forms of ACT Assessment Mathematics tests. Each form contains 60 multiple-choice items that require students to use their reasoning skills to solve practical problems in mathematics. For each form, the responses of 2,000 randomly selected examinees were used to calibrate the items with the computer program NOHARM (Fraser, 1983) in which the generalized multidimensional normal ogive item response was applied. NOHARM provides estimates of a_1, a_2, d, and c for each item.

Dichotomous Response Data Generation

With the set of two-dimensional item parameters obtained from NOHARM and with pairs of thetas (θ_1, θ_2) drawn from a bivariate normal distribution with a specific correlation (.00, 0.30, 0.60, 0.90 or 1.00), the computer program GENMIRT was used to generate dichotomous scores using the two-dimensional compensatory three-parameter IRT (2-D COMIRT 3-PL) model:

$$P_i\langle x_{ij} = 1 \mid a_{i1}, a_{i2}, d_i, c_i; \theta_{j1}, \theta_{j2}\rangle$$

$$= c_i + \frac{1 - c_i}{1 + \exp\left[-D\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + d_i\right)\right]}.$$

If the value of the probability of correct response to the item with a specific pair of ability thetas, $P(\theta_1, \theta_2)$, was less than a value randomly drawn from a uniform (0,1) distribution, a wrong response (i.e., 0) would be recorded. If the value of $P(\theta_1, \theta_2)$ was equal to or greater than the value drawn from the uniform distribution, a correct response (i.e., 1) would be recorded. Two thousand "examinees" were generated to respond to the 360 items. Thus, a 2,000 (examinees) by 360 (items) 1/0 matrix was obtained.

Figures 1 and 2 summarize how the data were generated under these different conditions.

Section III. Research Design

In order to answer the research questions, a 2x5x2x3 crossed factorial design was used. The manipulated conditions are identified below:

1. UIRT models:
   (1) UIRT 3-PL model.
   (2) UIRT 1-PL model.
2. Correlation between θ_1-values and θ_2-values:
   (1) 0.00.
   (2) 0.30.
   (3) 0.60.
   (4) 0.90.
   (5) 1.00.
3. Range of test length:
   (1) min=15, max=50.
   (2) min=1, max=360 (no constraint).
4. Passing score:
   (1) p=0.4.
   (2) p=0.6.
   (3) p=0.8.

Thus, a total of 60 different combinations of conditions based on UIRT models were investigated.

Section IV. Data Analysis Procedures

Given the simulated 1/0 response data generated by GENMIRT, the computer program BILOG, which applies a

marginal maximum likelihood estimation procedure, was used to calibrate these two-dimensional dichotomous response data with both a UIRT 3-PL model and a UIRT 1-PL model. As mentioned previously, one thousand randomly drawn respondents were used for the item calibrations. The estimated item parameters from the UIRT 3-PL and the UIRT 1-PL models were obtained and used for the mastery tests.

With item parameters calibrated by BILOG, the computer program UTCC was used to obtain the different cut-scores on the theta scale: $P(\theta_c)=0.4$, 0.6, and 0.8.

Definition of True Master/Nonmaster

In order to distinguish between true pass and true fail in the true model, a threshold function, $f(\theta_1,\theta_2)$, was defined as:

$$f(\theta_1,\theta_2) = \frac{1}{360}\sum_{i=1}^{360} P_i\langle x_{ij} = 1 \mid a_{i1}, a_{i2}, d_i, c_i; \theta_{j1}, \theta_{j2}\rangle \text{ - cutting point,}$$

where

cutting point = 0.4, 0.6, and 0.8.

If $f(\theta_1,\theta_2) \geq 0$, the examinee truly passes.

If $f(\theta_1,\theta_2) < 0$, the examinee truly fails.

It should be noted that the function, $f(\theta_1,\theta_2)$ may or may not be linear, depending on the item parameters. Figure 5 shows the projected functions at different cutting points.

At the cutting point .6, the function was near linear. However, at the other two cutting points, .4 and .8, the functions were nonlinear.

Input of MIRTSPRT

After the threshold function was established, the "true" parameters together with the threshold function, $f(\theta_1,\theta_2)$ of the 2-D COMIRT 3-PL model, and the "false" parameters from the UIRT model together with the cut-score theta, θ_c mapped from $P(\theta_c)$= 0.4, 0.6 or 0.8 of UIRT (either UIRT 3-PL or UIRT 1-PL) were used as input to the computer program MIRTSPRT. MIRTSPRT was used to simulate a SPRT-CMT and to estimate the type I error rates, type II error rates, and number of items used to make the mastery decision under different combinations of conditions.

For each of the 60 combinations of conditions, the range of θ-value was set from -3 to +3 and 13 equally spaced values were used (-3.0, -2.5,..., 2.5, 3.0). A total of 169 pairs of θ_1 and θ_2 values were created (e.g., θ_1=-3.0 and θ_2=2.5) and 300 replications of the mastery decision procedure were performed for each pair of values using MIRTSPRT. Thus, for each combination of conditions, a total of 50,700 (169x300) mastery decisions were made.

Mastery Decisions Based on SPRT

In the MIRTSPRT, item responses were generated from the "true" 2-D COMIRT model with values of θ_1 and θ_2. However, the SPRT-CMT was carried out using estimates of UIRT parameters. Thus, items were ranked on the "false" values of information at a "false" cut-score.

The sequential probability ratio testing procedure was used to determine the pass/fail status using the UIRT models. The general SPRT procedure was described in Chapter II.

Definition of Type I and Type II Errors

The mastery decision based on the true model (i.e., 2-D COMIRT 3-PL) was defined as:

If $f(\theta_1,\theta_2) \geq 0$, the examinee truly passes.

If $f(\theta_1,\theta_2) < 0$, the examinee truly fails.

The decision (pass or fail) made by the SPRT procedure with UIRT models was then compared to the decision based on the true model. If the decision based on true item parameters and the corresponding θ_1 and θ_2 was consistent with the decision based on false (i.e., either UIRT 3-PL or UIRT 1-PL) item parameters and the corresponding thetas, there was no error. However, if the decision based on true item parameters was to fail the examinee while the decision based on UIRT item parameters was to pass the examinee, a type I error (false positive) occurs. If the decision based

on true item parameters was to pass the examinee while the decision based on UIRT item parameters was to fail the examinee, a type II error (false negative) occurred. (See Figure 7.) The expected errors and test lengths were calculated based on the expectations relative to the density of ability. The density function is described below.

Density Function

As was mentioned earlier, 300 replications were calculated at each of 169 pairs of θ-values. It was assumed that the examinee population had a bivariate normal distribution with respect to these two theta scales. In order to calculate the expected error rates and expected number of items used to make the mastery decision in such a population, a set of weights was needed to reflect the relative frequencies of each pair of thetas. These weights were calculated according to the density function formula:

$$\frac{1}{2\pi\sqrt{1 - r_{\theta 1 \theta 2}^{\ 2}}} \exp\{[-.5/(1-r_{\theta 1 \theta 2}^{\ 2})][\theta_1^{\ 2} - 2(r_{\theta 1 \theta 2})\,\theta_1\,\theta_2 + \theta_2^{\ 2}]\},$$

where

θ_1 is the value of theta 1,

θ_2 is the value of theta 2, and

$r_{\theta 1 \theta 2}$ is the correlation value between theta 1 and theta 2.

It can be seen that as the θ-correlation is changed, the value of the weight is also changed. The expected type I, type II error rates and expected number of items used to make the mastery decision were then calculated using these weights.

Test Length and Cut-Score Setting

In the program MIRTSPRT, a test length constraint (i.e., minimum and maximum test length) and level of cut-score could be manipulated. The expected false positive rates, expected false negative rates, and expected number of items used for making the decision can be calculated based on 300 replications with each pair of thetas. The robustness of the models was examined by comparing the average false positive rates, average false negative rates, and average number of items used under different combinations of conditions. Figures 4, 6, and 7 summarize the procedures used in this study.

Criterion Indices

For each simulation, the outcomes of interest were: (1) expected type I error (false positive) rates, (2) expected type II error (false negative) rates, and (3) expected numbers of items used to make the mastery decision via computerized mastery testing.

Fixed Factors

The fixed factors for all simulations of this design are listed below:

1. The number of dimensions (2).
2. The number of items in the pool (360).
3. The means and standard deviations of the generated thetas.
4. The means and standard deviations of a_1, a_2, d, and c in the 2-D COMIRT 3-PL model.
5. The number of simulation replications (300).
6. The width of the SPRT indifference region ($|\theta_0 - \theta_1|=0.5$) (i.e., $\theta_0 = \delta - 0.25$, $\theta_1 = \delta + 0.25$, where δ is the passing criterion).
7. The nominal error rates, α and β in the SPRT procedure (0.05).

CHAPTER IV

RESULTS AND DISCUSSION

The purpose of this study was to investigate the usefulness of the SPRT procedure with UIRT models in computerized mastery testing when the unidimensionality assumption is violated. The study was conducted using the design described in Chapter III.

This chapter contains six sections. Section I describes the characteristics of the simulated data. Section II and Section III present and discuss the main and interaction effects of the three manipulated factors within the UIRT 3-PL and UIRT 1-PL models, respectively. The results for the UIRT 3-PL model and the 1-PL model are compared in Section IV. Section V discusses the results when the correlation between θ_1 and θ_2 (hereafter referred to as the θ-correlation or $r_{\theta 1 \theta 2}$) was equal to one. A summary of the results is provided in Section VI.

Section I. Description of Simulated Data

Table 1 shows the descriptive statistics for the original two-dimensional item parameters based on the 360 items calibrated by NOHARM. The mean of the estimates of the first a-parameter was .93 and the mean of the estimates of

the second a-parameter was .64. The mean of difficulty ($d = -a_1b_1-a_2b_2$) estimate was -.79 and the mean of the estimates of the guessing parameter was .18. The mean of the estimates of the first item discrimination parameter was about 1.5 times the mean of the estimates of the second item discrimination parameter. This suggests that the first dimension is more dominant than the second dimension. The standard deviations, skewness indices, and kurtosis indices of each of the distributions of estimated item parameters are also given in Table 1.

Table 2 gives the descriptive statistics for the bivariate theta distributions used to generate the dichotomous data. The means and standard deviations of the theta distributions were close to zero and one respectively. The skewness and kurtosis were relatively similar to those for a normal distribution.

Table 3 displays the first three eigen-values associated with five different theta correlations. For each correlation, 2000 pairs of thetas were drawn from a specific bivariate normal distribution. These values were entered into the 2-D COMIRT 3-PL equation with the original two-dimensional parameters of 360 items. For each pair of thetas, 360 probabilities of correct responses (like “true scores”) were recorded. The eigen-values of these 360-by-360 “true score” variance-covariance matrices were calculated.

It was found that as the θ-correlation increased, the first eigen-value increased and the third eigen-value decreased.

Table 4 presents the means and standard deviations of the item parameter estimates calibrated by the UIRT models. For the UIRT 3-PL model, as the θ-correlation increased, the mean of the discrimination parameters increased, the mean of the difficulty parameter decreased, and the mean of the guessing decreased. The same pattern was exhibited for the UIRT 1-PL model.

Table 5 provides the cut-scores on the theta scales (θ_c) associated with each of the cut-scores on the p-value scale ($P(\theta_c)$). As $r_{\theta 1 \theta 2} = 1$ was a special case of 2-D COMIRT 3-PL model, the θ_c could be calculated based on the original two-dimensional item parameters with the equation derived in Chapter II. These θ_cs are shown in parentheses in Table 5. Compared with the thetas derived from 2-D COMIRT 3-PL model (true item parameters), the estimated thetas based on the UIRT 3-PL model seemed to overestimate the cut-score on the theta scale, suggesting that the estimated item parameters may yield a biased estimate of θ_c.

As can be seen in Table 5, as the θ-correlation increased, the value of θ_c generally increased when $P(\theta_c)=.4$, except when $r_{\theta 1 \theta 2}=1$ and $P(\theta_c)=.4$ with the UIRT 3-PL model. However, at $P(\theta_c)=.6$ or .8, the estimated θ_c decreased as the θ-correlation increased.

Table 6 summarizes the accumulated item information values for each UIRT model under different conditions. As was mentioned earlier, the items were ranked at the cut-score according to their information: the higher the item information, the higher the priority for administration. The equation for the item information calculation was given in Chapter II. Table 6 lists the accumulated item information of the first 50 most informative items (max=50). That is:

$$\sum_{i=1}^{50} I_i(\theta_c) .$$

Also, the accumulated item information for the entire 360 items (max=360) is shown in Table 6. That is:

$$\sum_{i=1}^{360} I_i(\theta_c) .$$

It was found that as the θ-correlation increased, the amount of accumulated information generally increased. The only exception to this trend was when the θ-correlation was equal to one, the cut-score was .8 and the UIRT 3-PL model was used.

For the UIRT 3-PL model, when the maximum test length was set to 50, accumulated information increased as the cut-score increased. When the maximum test length was set to 360 (i.e., no test length constraint), the accumulated

information was similar at cut-scores of .6 and .8 and the least accumulated information was observed at .4.

For the UIRT 1-PL model, when the maximum test length was set to 50, the amount of accumulated information was similar across the three cut-scores. When the maximum test length was set to 360, the accumulated information was similar at cut-scores of .4 and .6 and the least information was found at .8.

The amount of accumulated information was greater for the UIRT 3-PL model than for the UIRT 1-PL model. For the UIRT 3-PL model, item discrimination parameters in the test are allowed to vary. However, for the UIRT 1-PL model, item discrimination parameters in the same test are assumed to be equal. Because item discrimination has a dominant effect on item information (the higher the discrimination, the higher the information (Lord, 1980)), the difference between the accumulated information for the two models was expected.

Section II. Results for UIRT 3-PL Model

In this section, the results for the UIRT 3-PL model are presented and discussed. The next section presents the results and discussion for the UIRT 1-PL model.

The main effects and interaction effects of either the UIRT 3-PL model or the UIRT 1-PL model are described and discussed in the following order:

1. Three main effects:

(1) magnitude of $r_{\theta 1 \theta 2}$.

(2) test length constraint.

(3) level of cut-score.

2. Three 2-way interaction effects:

(1) magnitude of $r_{\theta 1 \theta 2}$ and test length constraint.

(2) magnitude of $r_{\theta 1 \theta 2}$ and level of cut-score.

(3) test length constraint and level of cut-score.

3. One 3-way interaction effects:

(1) magnitude of $r_{\theta 1 \theta 2}$ and test length constraint and level of cut-score.

Outcomes of Interest

For the results of the UIRT 3-PL and UIRT 1-PL models, the four outcomes of interest in this study are:

1. False positive error (type I error) rates.
2. False negative error (type II error) rates.
3. Total error rates (1 + 2).
4. Number of items used to make the mastery decision (NI).

In order to facilitate the discussion, differences of .01 or less in error rates (type I, type II, or total error) and differences of 4 or less in number of item used (NI) to make the mastery decision are considered negligible and will not be discussed. It should be noted that the results with $r_{\theta 1 \theta 2} = 1$ are presented in a separate section.

The results of applying the UIRT 3-PL model when the response data were simulated using a two-dimensional 3-PL compensatory model are discussed in this section. Table 7 presents the results for each combination of conditions within the UIRT 3-PL model. Over all the conditions, the average false positive rate, average false negative rate, average total error rate, and average NI were .0171, .0209, .0380, and 14.1476 respectively.

Magnitude of $r_{\theta_1 \theta_2}$

As the θ-correlation was changed, a clear pattern was found: as the θ-correlation increased, the type I error rates decreased, the total error rates decreased, and the NI value decreased. In addition, the type II error rate was consistently greater than the type I error rate within each θ-correlation. (See Table 8.)

Although there was a clear pattern, the range of the error rates was relatively small: false positive rates ranged from .0138 to .0195; false negative rates ranged from .0200 to .0220; and total error rates ranged from .0346 to .0415. As mentioned earlier, differences of .01 or less in error rates were considered negligible.

The changes in the θ-correlation had a more obvious impact on the NI index. When the θ-correlation was .00, 16.7663 items were needed on average for the mastery

decision. When the θ-correlation was .90, the average number of items needed decreased to 12.1158.

Thus, for the UIRT 3-PL model, the magnitude of the θ-correlation seemed to have an obvious impact only on the number of items used. The testing procedure became more efficient as the θ-correlation increased. Recall from Table 6 that as the magnitude of θ-correlation increased, the accumulated information also increased. When the maximum test length was set to 50 items, the correlations between accumulated information and number of items used to make mastery decision were -.8590 and -.4239 for UIRT 3-PL and 1-PL models, respectively. This negative correlation suggests that the θ-correlation affects the accumulated information, which in turn affects the item-consumption.

Test Length Constraint

As shown in Table 9, the impact of the test length constraint was relatively small. The average type I, type II, and total error rates were slightly lower (.0168 vs. .0175, .0206 vs. .0212, .0374 vs. .0387) for the constraint condition (min=15, max=50) than for the unconstrained condition (min=1, max=360). However, the average NI was greater (17.4571 vs. 10.8382).

These results seem to indicate that the test length constraint only had an impact on the number of items used. If test efficiency is the main concern, these results

indicated that it would be better not to apply a constraint on the minimum number of items when using the UIRT 3-PL model.

Level of Cut-Score

As the cut-score level increased, the type I error rate and the NI value decreased. (See Table 10.) The most obviously impact of changing the cut-score level was observed for NI. NI was 18.5651 at $P(\theta_c)=.4$ and 10.5864 at $P(\theta_c)=.8$. Also, the type I error rate decreased from .0268 at $P(\theta_c)=.4$ to .0096 at $P(\theta_c)=.8$.

At $P(\theta_c)=.4$, the type I error rate was greater than the type II error rate (.0268 vs. .0178). However, at the $P(\theta_c)=.6$ or .8, the type II error rate was greater than the type I error rate. This pattern suggests an estimated θ_c bias. That is, θ_c may be underestimated at $P(\theta_c)=.4$ and overestimated at $P(\theta_c)=.6$ and $P(\theta_c)=.8$.

Combined Effects of the Magnitude of $r_{\theta_1\theta_2}$ and Test Length Constraint

Table 11 shows the combined effects of the θ-correlation and the test length constraint on the four outcome measures. As can be seen in Table 11, the general trends for the error rates are highly similar to the main effect trends observed in Tables 8 and 9. However, for the NI index, the effect of the θ-correlation seemed to be a

function of the test length constraint. The effect of the θ-correlation on the NI index appeared to be greater for the no test length constraint (min=1, max=360) than for the test length constraint (min=15, max=50). For the constraint condition, the NI decreased from 18.5 to 16.2 as the θ-correlation changed from .00 to .90. However, for the no constraint condition, the NI decreased from 15.1 to 7.6 as the θ-correlation increased from .00 to .90.

These results indicate that the 2-way interaction effects ($r_{\theta 1\,\theta 2}$ x cut-score) had an impact mainly on test efficiency and not on error rates.

Combined Effects of the Magnitude of $r_{\theta 1\,\theta 2}$ and Level of Cut-Score

Table 12 shows the combined effects of the θ-correlation and the level of cut-score on the four outcome measures. As can be seen in Table 12, the general trend observed for the four outcome measures are similar to main effect trends observed in Tables 8 and 10. The one exception to these trends was for the false positive error rates when $r_{\theta 1\,\theta 2}$=.9. This error rate did not consistently decreased as the cut-score increased. However, these results also indicated that the effect of the θ-correlation varied with the cut-score level. The nature of this interaction is described below.

At $P(\theta_c)=.4$ and .6, as the θ-correlation increased, the type I error rate decreased, but at $P(\theta_c)=.8$, the opposite happened. At $P(\theta_c)=.4$ and .8, as the θ-correlation increased, the type II error rate decreased but at $P(\theta_c)=.6$, the opposite happened. At $P(\theta_c)=.4$, as the θ-correlation increased, the total error rate decreased but the total error rates were approximately equal at $P(\theta_c)=.6$ and at $P(\theta_c)=.8$. At $P(\theta_c)=.4$, as the θ-correlation increased, the NI index decreased from 24.1 to 14.5, but at $P(\theta_c)=.6$ and .8, the NI-values were similar.

These results suggest that there were 2-way interaction effects (θ-correlation and level of cut-score) on both the classification accuracy and test efficiency. These results seem to indicate that regardless of the degree of correlation between the two θs, tests that have relatively high cut-scores (e.g., .8) will perform better than tests with lower cut-scores.

Combined Effects of Test Length Constraint and Level of Cut-Score

Table 13 shows the combined effects of the test length constraint and the level of cut-score on the four outcome measures. As can be seen in Table 13, the general trends observed for the four outcome measures are similar to the main effect trends observed in Tables 9 and 10. However, the

effect of the cut-score on the NI index appeared to be greater for the unconstrained test length (min=1, max=360) than for the constrained test length (min=15, max=50). For the constrained test length, the NI decreased from 19.5 to 15.7 as the cut-score changed from .4 to .8. However, for the unconstrained test length, the NI decreased from 17.6 to 5.4 as the cut-score changed from .4 to .8.

These results indicate that the 2-way interaction effects (test length constraint x cut-score) had an impact mainly on test efficiency. A relatively smaller number of items is needed to make the mastery decisions when a test length constraint is not applied and when a high cut-score is used.

Combined Effects of Magnitude of $r_{\theta_1\theta_2}$, Test Length Constraint, and Level of Cut-Score

The 3-way interaction effects are reflected in the means reported in Table 7. The general trends for all three error rates are consistent with the main effect trends noted in Tables 8, 9, and 10.

However, the means reported in Table 7 suggest 3-way interaction effects relative to the NI index. For the constrained test length (min=15, max=50), across different θ-correlations and within each cut-score, the NI values were similar. (That is, the differences were less than 4.) But

for the unconstrained test length (min=1, max=360), across different θ-correlations and within each cut-score, the NI value decreased at $P(\theta_c)=.4$. The decrease was from 26.6 when $r_{\theta 1\,\theta 2}=.00$ to 11.2 when $r_{\theta 1\,\theta 2}=.90$. At $P(\theta_c)=.6$ and .8, however, the NI values were relatively similar. With a test length constraint, the variation among the NI values was relatively small. Without a test length constraint, the test is more efficient but the NI values showed greater variation.

Summary of the Results for the UIRT 3-PL Model

Table 14 summarizes the effects observed in this study. The major impact of the factors of interest ($r_{\theta 1\,\theta 2}$, test length constraint, & level of cut-score) was on test efficiency (number of items required for the mastery decision). The NI index generally decreased as $r_{\theta 1\,\theta 2}$ increased and the level of the cut-score increased. The NI was also less for the no test length constraint as compared to the constrained condition. However, the rate of decrease accounted for by a factor was not always the same at levels of the other factors.

With respect to error rates, the SPRT procedure seemed robust to violations of the unidimensionality assumption under all conditions when the UIRT 3-PL model was used.

Section III. Results for UIRT 1-PL Model

The results of applying the UIRT 1-PL model when the response data were simulated using a 2-D COMIRT 3-PL model are presented and discussed in this section. Like the last section, the main effects and interaction effects of UIRT 1-PL model under different conditions are described and discussed. Table 15 presents these results for each combination of conditions within the UIRT 1-PL model. The average false positive rate, false negative rate, and total error rate were .0156, .0264, and .0420, respectively. The average NI was 31.2765.

Magnitude of $r_{\theta 1 \theta 2}$

Table 16 presents the averages for the four outcome variables for each of the four θ-correlations. In general, the three error rates were similar. However, the average NI value decreased as the θ-correlation increased. When the $r_{\theta 1 \theta 2}$=.00, the average NI was 38.0. This average NI value decreased to 25.5 when $r_{\theta 1 \theta 2}$=.90.

Thus, for the UIRT 1-PL model, the magnitude of θ-correlation seemed to impact only the test efficiency. This outcome was consistent with the outcome observed for UIRT 3-PL model. However, compared with UIRT 3-PL model, UIRT 1-PL model required about twice the number of items to make the classification decision.

Test Length Constraint

Table 17 presents the results for the test length constraint condition. As can be seen in Table 17, the error rates for the no constraint condition (min=1, max=360) were slightly less than for the constraint condition (min=15, max=50). However, the constraint condition used fewer items to make the mastery decision, which means the maximum test length constraint (max=50) forced the test to stop.

Based on these results, the use of a test length constraint with the UIRT 1-PL model had an impact on the total error rates and the number of items used to make the decision. If classification accuracy is the main concern, it would be better not to apply a test length constraint when using the UIRT 1-PL model.

Level of Cut-Score

Table 18 presents the averages of the four outcome measures for each cut-score level. The type I error rates (false positive) tended to decreased as the cut-score increased. This trend was also true for the total error rate. However, the greatest type II error rate (false negative) occurred at $P(\theta_c)=.6$. The average NI value decreased as the level of cut-score increased.

At $P(\theta_c)=.4$, the type I error rate was greater than the type II error rate. However, at $P(\theta_c)=.6$ or .8, the type II error rate was greater than type I error rate. The same

pattern also occurred when the UIRT 3-PL model was used. This pattern may suggest an estimated θ_c bias. That is, θ_c might be underestimated at $P(\theta_c)=.4$ and overestimated at $P(\theta_c)=.6$ and .8.

Combined Effects of the Magnitude of $r_{\theta_1\theta_2}$ and Test Length Constraint

Table 19 gives the means for the θ-correlation and test length constraint condition. The trends observed in Table 19 were consistent with the main effect trends described in Tables 16 and 17, except for the NI index.

When the test length was constrained (min=15, max=50), the average NI value decreased from 30.2 ($r_{\theta_1\theta_2}=.00$) to 23.5 ($r_{\theta_1\theta_2}=.90$). However, in the unconstrained condition (min=1, max=360), the NI decrease was even greater. The average NI value decreased from 45.8 ($r_{\theta_1\theta_2}=.00$) to 27.6 ($r_{\theta_1\theta_2}=.90$).

Thus, the θ-correlation seems to have a greater impact on NI in the no constraint condition than in the constraint condition. These results are consistent with that of UIRT 3-PL model.

Combined Effects of the Magnitude of $r_{\theta_1\theta_2}$ and Level of Cut-Score

The means of the four outcome measures for the twelve combinations of θ-correlations and cut-score levels are given in Table 20. As shown in Table 20, the general trends

for the four outcome measures were relatively similar to the main effect trends noted in Tables 16 and 18. However, when the θ-correlation was .90, a slightly different trend was seen for the type II and total error rates. When $r_{\theta 1 \theta 2}$=.90, the total error rate for a cut-score of .6 was actually greater than the total error rate for a cut-score of .4 (.0638 vs. .0504). Similarly, the difference between the type II error rates for these two cut-scores(.0613-.0099=.0514) was relatively greater when $r_{\theta 1 \theta 2}$=.90 than for the other θ-correlations.

With respect to the NI index, at $P(\theta_c)$=.8 it decreased from 25.2 when $r_{\theta 1 \theta 2}$=.00 to 19.3 when $r_{\theta 1 \theta 2}$=.90. However, at $P(\theta_c)$=.4 the NI value decreased from 48.2 when $r_{\theta 1 \theta 2}$=.00 to 31.8 when $r_{\theta 1 \theta 2}$=.90. Thus, the decrease in NI was greater at $P(\theta_c)$=.4 than at $P(\theta_c)$=.6.

Based on these results, a combined effect of θ-correlation and the level of cut-score on type II error rates, total error rates, and NI was suggested. The violation of unidimensionality had an impact on both error rates and the NI at lower cut-scores (i.e., .4 & .6) but a relatively smaller impact on the higher cut-score. This suggests that if the unidimensionality assumption is violated, CMT might perform better when used for differentiating among high ability examinees.

Combined Effects of Test Length Constraint and Level of Cut-Score

The average values of the four outcome measures for the six combinations of test length and cut-score level are shown in Table 21. The general trends noted in Tables 17 and 18 also hold for most of the outcome measures. One exception was that for the unconstrained test length condition. In this condition, the total error rates at the cut-scores of .4 and .6 were almost identical.

The results in Table 21 also seemed to indicate that the effect of the cut-score level varies as a function of the test length constraint. For the test length constraint condition (min=15, max=50), the decreases in the type I error rates, the total error rates and NI from $P(\theta_C)=.4$ to $P(\theta_C)=.8$ were .0473, .0486, and 10.4, respectively. However, for the no constraint condition (min=1, max=360), the decreases were .0266, .0266, and 22.8, respectively.

Based on these results, a combined effect of test length constraint and cut-score level on type I error rates, total error rates, and NI was suggested. The test length constraint had a relatively small impact on classification accuracy and test efficiency at a higher cut-score level (e.g., .8).

Combined Effects of Magnitude of $r_{\theta 1\,\theta 2}$, Test Length Constraint, and Level of Cut-Score

The means of the four outcome variables for the twenty-four combinations of θ-correlation values, cut-score levels, and test length constraints are given in Table 15. In general, the main effect trends observed for type I error rates and NI values in Tables 16, 17, and 18 were also observed in this table.

The data in Table 15 were used to examine the three 2-way interaction effects identified previously ((1) θ-correlation by test length constraint on NI (Table 19), (2) θ-correlation by cut-score level on error rates and NI (Table 20), and (3) test length constraint by cut-score level on error rates and NI (Table 21)) at levels of the third variable.

The test length constraint by cut-score interaction appeared to vary as a function of the θ-correlation. When $r_{\theta 1\,\theta 2}$=.90, the total error rate at a cut-score of .6 was greater than the total error rate at .4 for both test lengths. However, for the other correlation values, the highest total error rate was at .4.

Summary of the Results for the UIRT 1-PL Model

As was true when the UIRT 3-PL model was used, the SPRT procedure using parameters estimated by UIRT 1-PL model seemed relatively robust with respect to classification errors. The range of error rates across the manipulated conditions was relatively small. Although these factors affected the classification accuracy and the test efficiency, the error rates and NI values would probably be considered reasonable in most testing situations. The main and combined effects of these manipulated factors are summarized in Table 22.

Section IV. UIRT 3-PL Versus UIRT 1-PL Model

In this section, the results for the UIRT 3-PL model and those for the 1-PL model are compared in terms of type I error rates, type II error rates, total error rates, and NI. The average values of these indices for two models are summarized in Tables 7 and 15.

The average type I, type II, and total error rates were similar for both the UIRT 3-PL and the 1-PL models (.0171 vs. .0156, .0209 vs. .0264, .0380 vs. .0420). However, the average NI value was somewhat different. The average NI for the 3-PL model was 14.1476, while for the 1-PL model it was 31.2765. In other words, the UIRT 1-PL model required about

twice number of the items as the 3-PL model to make mastery decisions with similar accuracy.

Overall, it was found that the test efficiency of UIRT 3-PL model was better than 1-PL model in all conditions. That is, the average NI values were less for the 3-PL model than for 1-PL model. Also, the error rates for both models were similar. These results are consistent with the fact that the accumulated information value based on UIRT model was consistently higher for 3-PL model than for the 1-PL model.

Magnitude of $r_{\theta 1 \theta 2}$

The effect of the magnitude of the θ-correlation was similar for both models. As the θ-correlation increased, the NI decreased. (See Tables 8 & 16.)

Test Length Constraint

The SPRT procedure using the UIRT 3-PL model had similar classification precision under either the test length constraint condition or the no test length constraint condition (average total error: .0374 vs. .0387). However, without the test length constraint, the UIRT 3-PL model required fewer items to make the classification (10.8382 vs. 17.4571) than with the constraint. (See Table 9.)

For UIRT 1-PL model, however, the test length constraint had the opposite effect. Without the test length

constraint, the UIRT 1-PL model made fewer average errors (.0326 vs. .0514) than with the constraint, but used more items (35.9858 vs. 26.5673) than with the constraint. (See Table 17.)

Thus, the test length constraint had little effect on error rates for the 3-PL but did have some effect for the 1-PL model. For the 1-PL model, greater mastery decision accuracy was gained when no test length constraint was applied than when a test length constraint was applied.

In common test administration situations, test length constraints are always applied for practical concerns (e.g., content coverage, item exposure rate control). Thus, it appears that the UIRT 3-PL model might be better in such situations. Moreover, the range of the test length constraint can be set differently. It would appear to be better to set a higher maximum number, if the UIRT 1-PL model is used.

Level of Cut-Score

Both models showed similar patterns across different cut-scores: (1) the greatest false positive error rates were at $P(\theta_C)=.6$ and the least were at $P(\theta_C)=.8$; (2) the greater the level of the cut-score, the less the false positive error and the less the NI; and (3) the total error rates were similar at $P(\theta_C)=.4$ and 6.

Although the average type I, type II, and total errors were similar for both models, within each cut-score, the 1-PL model constantly required more items (over 2 times) to make the mastery decision. In other words, the UIRT 1-PL model was not as efficient as the 3-PL model.

Also, for both models, the highest type I error rates were always at $P(\theta_c)=.4$ and the highest type II error rates were always at $P(\theta_c)=.6$. This results suggest a bias in the estimation of θ_c.

Section V. Results for $r_{\theta_1\theta_2} = 1$

The simulated data in this study were based on the two-dimensional compensatory three-parameter model. Dichotomous data were generated with different θ-correlations. Then, either the UIRT 3-PL model or the 1-PL model was used to do the item calibration and θ_c estimation. When the correlation between θ_1 and θ_2 is 1, the θ_c scale could be calculated directly from the equation derived in Chapter III. (See Table 5 for the θ_c values.) However, θ_c based on the derived equation for 1-PL model could not be calculated because the "true" model underlying the data was 2-D COMIRT 3-PL.

UIRT 3-PL Model

Tables 23 and 24 summarize the results for 3-PL model with the two different sets of θ_cs for $r_{\theta_1\theta_2}=1$. The results given in Table 23 are derived from the θ_cs based on

estimated item parameters (UIRT 3-PL model.) The results given in Table 24 are derived from on the θ_cs based on true item parameters (2-D COMIRT 3-PL model.)

The results presented in Table 23 are generally consistent with the trends displayed in Table 7. However the magnitude of the type II error rate was surprisingly high. It is hypothesized that the high type II error rates at $P(\theta_c)=.6$ and the relatively high type I error rates at $P(\theta_c)=.8$ were caused in part by a bias in the estimated θ_c.

To investigate this hypothesis, the simulation was repeated based on the true θ_cs and the results are summarized in Table 24. It was found that the type II error rate at $P(\theta_c)=.6$ was reduced dramatically from about .09 to about .04, which is still relatively high. These results suggest that the bias associated with the cut-score does have an effect on the error rates, but there may also be other factors that contribute to those high errors. At this time, it is difficult to interpret this result. Further research is needed to gain insights into this issue.

UIRT 3-PL Model versus UIRT 1-PL Model

Table 25 presents the results based on the item calibration and θ_c estimation with the UIRT 1-PL model when $r_{\theta 1 \theta 2}=1$. Compared with the results of the UIRT 3-PL model reported in Table 23, the average type I error, type II

error, and total error rates are similar but the average NI values are significantly different. In the test length constraint condition, the average NI for UIRT 3-PL and 1-PL models were 16.7094 and 23.0230, respectively. Without the test length constraint condition, the average NI for UIRT 3-PL and 1-PL models were 7.2085 and 26.7559, respectively. Almost 4 times more items were needed for the mastery decision when the UIRT 1-PL model was used than when the UIRT 3-PL model was used.

Section VI. Summary

In the previous sections, the simulated data were described and the results of applying the SPRT procedure in computerized mastery testing with UIRT 3-PL and 1-PL models when the item response were two-dimensional were presented. Three variables, θ-correlation, test constraint, and cut-score level, were manipulated to examine the robustness of the UIRT models. The main effects, as well as interaction effects were discussed in terms of type I error rate, type II error rate, total error rates, and NI.

Generally, it was found that the manipulated conditions had an impact mainly on test efficiency rather than on classification accuracy. For example, it was found that as the θ-correlation increased, the number of items needed to make the mastery decision decreased while the error rates remained constant. These results also suggest that these

conditions had an effect on item and/or test information, which in turn affects the item consumption. These findings were consistent with those reported in previous studies (e.g., Abdel, Lau, & Spray, 1995, 1996).

Both UIRT models showed similar patterns across the manipulated conditions. However, the UIRT 3-PL model was more efficient than the UIRT 1-PL model. Both models had similar classification accuracy. However, the 3-PL model only used about half items as of the 1-PL model to make the mastery decisions.

Also, it appeared that there was bias in the estimate of θ_c if the true data set was two-dimensional but calibrated by UIRT models. For some cut-scores, the UIRT-estimated θ_c was underestimated and more type I errors occurred; for other cut-scores, the UIRT-estimated θ_c was overestimated and more type II errors occurred. It appeared that the θ_c bias was dependent on the magnitude of the θ-correlation.

CHAPTER V

CONCLUSIONS AND IMPLICATIONS

Certification or licensure tests represent an important type of achievement test. Such tests are used to help ensure that the quality of a profession is maintained. Computerized mastery testing (CMT) is the label given to certification tests that are administered in a computerized adaptive format. For convenience and availability, unidimensional IRT models are almost always adopted in CMT to assist item parameter calibration, θ-estimation, item selection/administration, and mastery decision making. However, the unidimensionality assumption will probably be violated in most, if not all, situations.

The purpose of this study was to investigate whether the UIRT models used in making mastery decisions with the sequential probability ratio test (SPRT) procedure produced acceptable classification accuracy when the assumption of unidimensionality was violated. Monte Carlo simulation techniques were used to examine the robustness of these procedures.

The specific procedures used to study robustness were outlined in Chapter III and the results were presented and

discussed in Chapter IV. In this chapter, conclusions based on these results are presented and their implications to practical testing situations are discussed. The strengths and limitations of this study and directions for further research are also considered. Four sections are contained in this chapter: (1) Conclusions, (2) Practical Implications, (3) Limitations and Strengths, and (4) Recommendation for Future Study.

Section I. Conclusions

The major conclusions about the SPRT procedure, the UIRT models, the test length constraint, and the cut-score levels are presented below.

SPRT Procedure in CMT

In this study, the SPRT procedure was found useful and robust for making mastery decisions in CMT with parameters estimated by either the UIRT 3-PL or the UIRT 1-PL models even when the unidimensionality assumption was violated. The actual average (across the two UIRT models) type I and type II error rates were .0164 and .0237, respectively. These results suggest that SPRT is an effective method to control error rates even when the unidimensionality assumption is violated.

The Appendix displays one of the outputs of the MIRTSPRT program. In this run, the cut-score was .4 (the

corresponding θ_c was -.132). In this Appendix, it can be seen that as the function (FUNC) based on the true model departed more from the cut-score, fewer classification errors occurred and fewer items were needed to make the mastery decision. In other words, type I and type II errors happened mainly in these "marginal" examinees. Passing a borderline examinee would seem to be a less serious error than passing an examinee whose ability level is far away from the required ability.

As indicated above, the average type I error rate was less than the average type II error rate. In general, a type II error (false negative) is not as serious as type I error (false positive) in the context of certification testing because there are usually other opportunities to retake the test. However, passing an unqualified candidate is potentially more harmful.

In view of the above, it can be concluded that the SPRT procedure using unidimensional item calibration models is a promising procedure for making mastery decisions, even when the test response data are two-dimensional.

Usefulness of UIRT Models

Generally speaking, both the UIRT 3-PL and 1-PL models were robust with respect to error rates. The violation of the unidimensionality assumption primarily had an impact on test efficiency and not on classification accuracy. This was

also true for the other factors of interest in this study. The error rates were generally within a reasonable range. False positive rates were lower than false negative rates in most of the conditions.

Although the underlying response data were based on a two-dimensional COMIRT model, the classification accuracy of UIRT 1-PL model was similar to that of the UIRT 3-PL model. However, the item-consumption of UIRT 1-PL model was twice that of the UIRT 3-PL model.

Based on these results, it is concluded that both unidimensional IRT models are robust and useful for parameter estimation and for assisting in the mastery classification decision making. With respect to test efficiency, the UIRT 3-PL model performs better than UIRT 1-PL model. With respect to classification accuracy, both models perform adequately.

Test Length Constraint

The two unidimensional models showed a different pattern of error rates and item-consumption when the test length constraint was changed.

The classification accuracy of the UIRT 3-PL model was similar under both test length conditions. However, for the UIRT 1-PL model, when the maximum number of items was set to 50, the classification errors increased relative to the no constraint condition.

Also, the test length constraint had different effects in terms of test efficiency for these two UIRT models. For the UIRT 3-PL model, if a test length constraint was applied, the test efficiency was less than when there was no constraint. For the UIRT 1-PL model, the test efficiency was greater with the constraint condition as compared to the no constraint condition.

Based on these results, it is concluded that the impact of a test length constraint on classification accuracy and efficiency depends on which unidimensional model is used.

Location of Cut-Score

When the unidimensionality assumption was violated, the results of this study suggested an existence of a bias in the determination of θ_c. That is, if the response data were two-dimensional but the item parameters and θ_c were estimated by UIRT models, θ_c might be overestimated at some points and underestimated at other points. If the θ_c is overestimated, the false negative error rate would be greater than false positive error rate. If the θ_c is underestimated, the false positive error rate would be greater than negative positive error rate.

Based on these results, it is concluded that violation of the unidimensionality assumption will cause bias in the estimation of θ_c, which in turn will cause differential classification errors.

Section II. Practical Implications

In almost all practical testing situations, the unidimensionality assumption does not hold absolutely. The results of this study indicate that violations of the unidimensionality assumption (in terms of changing the θ-correlation) do not significantly increase error rates. It may cause the testing procedure to use more items to make the mastery decisions but the average number of items used is still within a reasonable range for most testing situations. Based on the results of this study, test practitioners should feel confident about adopting an UIRT-SPRT procedure for certification testing even when the response data are two-dimensional.

UIRT Models

Theoretically, the UIRT 3-PL model is less restrictive than the UIRT 1-PL model. (That is, it allows for guessing and for different discrimination parameter values within a test). Therefore, the UIRT 3-PL model should fit the item data better than the UIRT 1-PL model, especially with a multiple choice item format. Practically, compared with the UIRT 3-PL model, the UIRT 1-PL model is simpler and usually requires smaller sample sizes for accurate item calibration. For this reason, the Rasch model is often adopted in certification testing situations. The results of this study

suggest that 1-PL model is a useful option, even when the unidimensionality assumption is violated to some degree.

Although the underlying response data was based on a two-dimensional COMIRT 3-PL model, both the UIRT 3-PL and the 1-PL models were found robust and yielded acceptable classification accuracy across different conditions with the SPRT procedure was used. Thus, if the primary concern is to control error rates, using either the UIRT 3-PL or 1-PL model can be defended.

However, the UIRT 3-PL model seems to perform more efficiently and is more stable than the UIRT 1-PL model. If other conditions are the same, the UIRT 3-PL model should have the first priority to be used. However, if the available calibration sample size is small and test length is not a major concern, it may be preferable to use UIRT 1-PL model.

Test Length Constraint

To control the exposure rate and test security, a test length constraint is commonly implemented in CMT. Based on the results of this study, if the unidimensionality assumption is violated, the effects of using a test length constraint could be different depending on the UIRT model used. Both test efficiency and classification accuracy can be influenced. The results of this study suggest that: (1) if the UIRT 3-PL model is adopted, set a lower minimum

number of items to gain higher test efficiency; (2) if the UIRT 1-PL model is adopted, set a higher maximum number of items to gain higher classification accuracy.

Test Difficulty

Tests differ in their difficulty. In CMT, the level of difficulty determines the location of θ_c. Based on the results of this study, if the unidimensionality assumption is violated, the location of θ_c could increase type I error rates at some points but increase type II error rates at other points. However, for tests with high cut-score (e.g., .8), the impact is relatively small.

Section III. Strengths and Limitations

Strengths of this Study

This study addressed important questions related to applications of computerized mastery testing. It was based on simulation procedures. Four factors: (1) type of UIRT model, (2) degree of θ-correlation, (3) presence or absence test length constraint, and (4) level of cut-score within CMT were investigated. These represent important considerations that are of interest when using CMT. This study provides substantial information about the impact of each of these factors on classification accuracy and test efficiency.

Because this study used simulation procedures, the true item parameters were fixed and the factors of interest could be manipulated directly. Also, because this study used "large response data sets", sampling error concerns were minimal.

Limitations of this Study

Two important considerations in CMT, content balance and item exposure control, were not considered in this study. The SPRT procedure used here selects items that have maximum information at the cutting point and every "examinee" was exposed to the same sequence of items. Item exposure rate is a serious concern in many CAT situations.

The other limitations of this study are mainly due to the research design. Only four factors were examined in this study. These limitations are discussed below.

Data Set

The two-dimensional item parameters used to generate dichotomous responses were based on six test forms (60 items each form) of ACT Assessment Mathematics tests. The test content and test characteristics are limited to these 360 item parameters and the representativeness of these conditions is unknown. Because the results were based on two-dimensional data set, which is the simplest

multidimensional case, the conclusions might not be generalized to more complex multidimensional situations.

Test Conditions

One level of the test length constraint condition and three levels of cut-score were examined. Thus, the conclusions are restricted to these limited conditions.

SPRT Procedure

Within the SPRT procedure, the width of indifference region and the nominal error rates were fixed for all 60 conditions. No knowledge was gained in this study concerning how the changes in the indifference region and nominal error rates would impact the actual error rates and item consumption.

Section IV. Recommendation for Future Study

It is said that you cannot do everything at one time. In this study, four factors (UIRT model, θ-correlation, test length constraint, and cut-score) and sixty conditions were investigated. This study should be expanded. More factors and/or variations should be investigated in the future so that greater generalizability can be achieved.

Data Set

In order to increase representativeness, other kinds of tests (e.g., Vocabulary, Reading) and different item pool characteristics (e.g., mean, standard deviation, number of items) should be used to see if the results are replicated. Also, three or more dimensional response data should be studied to check the robustness of the UIRT-SPRT procedure in more complex situations.

Test Conditions

Test length constraint interacted with type of UIRT model. More variations of the test length constraint (e.g., min. = 1, 15, 30, 45,... by max. = 75, 100, 125, 150,...) should be examined to observe this relationship in more detail.

There seems to be an interaction between the θ-correlation and the θ_c bias. More cutting points should be set so that this relationship can be observed more precisely.

Variation of SPRT Procedure

With the SPRT procedure, the width of indifference region and nominal error rates can be varied. The width of the indifference region affects item consumption while the nominal error rate influences the observed error rates. In

future investigations, different indifference regions and nominal error rates should be studied.

Item Exposure Control

No item exposure control was adopted in this study. There are different methods for item exposure control (e.g., parallel booklets) that can be applied in the item selection and administration procedures. Item exposure control should be included in future studies to make the testing situation more realistic and the results more applicable.

SB versus SPRT

Sequential Bayes is another procedure employed for item selection/administration, mastery decision making, and test termination. Few studies have been performed to investigate whether the SB procedure can be used in computerized mastery testing when the unidimensionality assumption is violated. The SPRT and SB procedures could be also compared in terms of classification accuracy and efficiency in such studies.

Different MIRT Models

In this study, a two-dimensional compensatory MIRT three-parameter model (2-D COMIRT 3-PL) was adopted to generate response data. There are other COMIRT models such as 2-D COMIRT 1-PL, and 2-D COMIRT 2-PL. Also, there are other MIRT models such as the noncompensatory (NOCOMIRT) model (Sympson, 1978) (2-D NOCOMIRT 1-PL, 2-D NOCOMIRT 2-PL,

& 2-D NOCOMIRT 3-PL). Different MIRT models for response data sets can be adopted and compared to see if the results are similar.

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.

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 Educational Research Association, New York City, New York.

Ackerman, T. (1987). *A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data* (ACT Research Report Series 87-12). Iowa City, IA: American College Testing.

Ackerman, T. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement*, *15*(1), 13-24.

Ackerman, T. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement*, *18*(3), 257-275.

Ackerman, T. (1994). Using multidimensional item response to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4), 255-278.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D. C.: American Psychological Association.

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.

Bergstrom, B. A. (1992). Confidence in pass/fail decisions for computer adaptive and paper and pencil examinations. *Evaluation and The Health Professions*, *15*(4), 435-464.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, *5*(2), 137-149.

Berk, R. A. (1976). Determination of optional cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, *45*, 4-9.

Bowers, J. J., & Shindoll, R. R. (1989). *A comparison of the Angoff, Beuk, and Hofstee methods for setting a passing score*. (ACT Research Report Series 89-2). Iowa City, IA: American College Testing.

Cangelosi, J. S. (1984). Another answer to the cut-off score question. *Educational Measurement: Issues and Practice*, *v3*(4) 23-25.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (ACT Research Report Series 87-19). Iowa City, IA: American College Testing.

Cascio, W. F. (1988). Setting cutoff scores: Legal psychometric, and professional issues and guidelines. *Personnel Psychology*, *41*, 1-24.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: legal, psychometric, and professional issues and guidelines. *Personnel Psychology*, *41*.

Cizek, G. J., Webb, L. C., & White, A. S. (1990). *Criterion-Referenced Standard Setting: A User's Guide*.

Crock, L., & Algina J. (1986). *Introduction to classical and modern test theory*. Orlando, Florida: Harcourt Brace Jovanovich, Inc.

Davey, T. (1994). *Scoring an innovative writing assessment: Option weighting data meiosis and the SPRT*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

De Ayala, R. J. (1992). The influence of dimensionality on CAT ability estimation. *Educational and Psychological Measurement*, *52*, 513-528.

Doody, E. N. (1985). *Examining the effects of multidimensional data on ability and item parameter estimation using the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7(2), 189-199.

Eignor, D. R., Way, W. D., Amoss, K. E. (1994). *Establishing the comparability of the NCLEX using CAT with traditional NCLEX examinations*. Paper presented at the annual meeting of NCME, New Orleans.

Ferguson, R. (1969). *Computer-Assisted Criterion-Referenced Measurement*. University of Pittsburgh Learning Research and Development Center, Pittsburgh.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, *13*(4), 373-389.

Forsyth, R., Saisangjan, U., & Gikmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological measurement*, *5*(2), 175-186.

Frick, T. W. (1986). *An investigation of the validity of the sequential probability ratio test for mastery decisions in criterion-referenced testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice*, *10*(2) 17-22.

Glass, G. V. (1978). Standards and Criteria. *Journal of Educational Measurement*, *15*(4), 237-261.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publication, Inc.

Hambleton, R. K., & De Gruijter, D. N. M. (1989). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement*, *20*(4) 355-366.

Hambleton, R. K., & Rogers, H. J. (1989). Solving criterion-referenced measurement problems with item response models. *Instructional Journal of Educational Researcher*.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principle and application*. Boston, MA: Kluwer Nijhoff Publishing.

Haynie, K. A., & Way, W. D. (1994). *The effects of item pool depth on the accuracy of pass/fail decisions for the NCLEX using CAT*. Paper presented at the annual meeting of the NCME, New Orleans.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika*, *41*(1), 65-78.

Huynh, H. (1979). *An empirical Bayes approach to decisions based on multivariate test data* (Publication Series in Mastery Testing). South Carolina, Columbia: University of South Carolina.

Huynh, H. (1982). A Bayesian procedure for mastery decisions based on multivariate normal test data. *Psychometrika*, *47*(4), 309-319.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.)(pp. 485-514). New York: Macmillan.

Junker, B. W., & Stout, W. F. (1991). *Robustness of ability estimation when multiple traits are present with one trait dominant*. Paper presented at the 1991 International Symposium on Modern Theories in Measurement: Problems and Issues, Montebebello, Quebec.

Kingsbury, G. G., & Houser, R. L. (1990). *Assessing the utility of item response models: Computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Kingsbury, G. G., & Weiss, D. J. (1980a). *A comparison of ICC-based adaptive mastery testing and the Waldian probability ratio method*. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference (pp. 120-139). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

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

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

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: Academic Press.

Linacre, J. M. (1988). *Simple and effective algorithms: Computer-adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Lewis, C., & Sheehan, K. (1990). Using Bayesian theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*(4), 367-386.

Lim, C. (1993). *An application of the joint maximum likelihood estimation procedure to a two-dimensional case of Sympson's non-compensatory IRT model*. Unpublished doctoral dissertation, University of Iowa, 1993.

Lim, C., & Ansley, T. N. (1994). *An investigation of the characteristics of multidimensional item response theory models*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Livingston S. A., & Zieky, M. J. (1982). *Passing Score: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service.

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

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

Luecht, R. M. (1994). *A few more issues to consider in multidimensional computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lunz, M. E., & Stone, G. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education*, 7(3), 211-222.

McKinley, R. L., & Reckase, M. D. (1982*). The use of the general Rasch model with multidimensional item response data* (Research Report ONR 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.

McKinley, R. L., & Reckase, M. D. (1984). *An investigation of the effect of correlation abilities on observed test characteristics* (Research Report ONR 84-1). Iowa City, IA: American College Testing.

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

Morrison, C. A., & Nungester, R. J. (1995). *Computer adaptive testing in a medical licensure setting: A comparison of outcomes under the one- and two- parameter logistic models*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*, *16*(3), 237-248.

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement*, *28*(3), 249-256.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor test: results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

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. 237-255). New York: Academic Press.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401-412.

Reckase, M. D. (1989). *Controlling the psychometric snake; or how I learned to love multidimensionality*. Paper presented at the annual meeting of the American Educational.

Reckase, M. D. (1989). *The interpretation and application of multidimensional item response theory model; and computerized testing in the instructional environment final report* (Research Report ONR 89-2). Iowa City, IA: American College Testing.

Reckase, M. D. (1989, Fall). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice*, *3*, 11-15.

Reckase, M. D. (1990). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

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

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*(3), 193-203.

Reckase, M. D., Carlson, J, E., Ackerman, T. A., & Spray, J. (1986). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the meeting of the Psychometric Society, Toronto.

Samejima, F. (1990). *Validity study in multidimensional latent space and efficient computerized adaptive testing* (ONR/Final Report).

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, *16*(1), 65-76.

Spray, J. (1993). *Multiple-category classification using a sequential probability ratio test* (ACT Research Report Series 93-7). Iowa City, IA: American College Testing.

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 ONR 90-8). Iowa City, IA: American College Testing.

Spray, J. A., Reckase, M. D.(in press). Comparison of SPRT and Sequential Bayes procedures for classifying examinees into Two Categories Using a Computerized Test. *Journal of Educational and Behavioral Statistics*.

Spray, J., Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (ACT Research Report Series 87-1). Iowa City, IA: American College Testing.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika*, *48*(2), 259-267.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, *12*(2), 87-98.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 103-135). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H. (1990). Introduction and History. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 1-20). Hillsdale, NJ: Lawrence Erlbaum.

Wald, A. (1947). *Sequential Analysis*. New York: Dover Publications, Inc.

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT*. Unpublished ONR research proposal. University of Iowa.

Way, W. D. (1994) *Psychometric models for computer-based licensure testing*. Paper presented at the 1994 Annual Meeting of CLEAR, Boston.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1986) *The effects of two-dimensional data on unidimensional IRT parameter estimates*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Way, W. D., Lewis, C., & Smith, R. L. (1995). *A comparison of two IRT-based models for computerized mastery testing when item parameter estimates are uncertain*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Weiss, D. J. (1985). *Computerized Adaptive Measurement of Achievement and Ability* (Final Report). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

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

Wetherill, G. B. (1975). *Sequential Methods in Statistics*. NY: John Wiley & Sons, Inc.

Table 1. Summary Statistics for Original Two-Dimensional Item Parameters

parameter	mean	SD	skewness	kurtosis
a_1	0.93	0.38	0.98	2.59
a_2	0.64	0.53	1.25	2.30
d	-0.79	1.20	-0.90	0.96
c	0.18	0.08	0.71	0.21

Note: SD is standard deviation.

Table 2. Descriptive Statistics of θ_1 and θ_2 with Different Correlations

$r_{\theta 1 \theta 2}$	theta	mean	SD	skewness	kurtosis
.00	θ_1	-.0221	1.0024	-.0078	-.1611
	θ_2	.0022	1.0018	.0721	.0173
.30	θ_1	-.0221	1.0078	-.0078	-.1611
	θ_2	-.0045	.9958	.0540	.0822
.60	θ_1	-.0221	1.0024	-.0078	-.1611
	θ_2	-.0114	.9919	.0034	.0657
.90	θ_1	-.0221	1.0024	-.0076	-.1611
	θ_2	-.0187	.9940	-.0398	-.0602
1.00	θ_1	-.0221	1.002	-.0078	-.1611
	θ_2	-.0221	1.002	-.0078	-.1611

Note: θ_1 and θ_2 are theta 1 and theta 2.

Table 3. The First Three Eigen-Values

$r_{\theta 1 \theta 2}$	eigen 1	eigen 2	eigen 3
.00	277.67	42.06	22.19
.30	291.35	34.16	19.01
.60	303.21	32.32	11.88
.90	313.68	32.97	7.05
1.00	316.93	33.27	7.00

Table 4. Means and Standard Deviations of Unidimensional Item Parameters

	UIRT 3-PL			UIRT 1-PL		
$r_{\theta 1 \theta 2}$	a	b	c	a	b	c
.00	1.0813 (.4665)	.6678 (.9496)	.1970 (.0776)	.5312	.1727 (1.0274)	.000
.30	1.1811 (.4907)	.5808 (.8731)	.1906 (.0747)	.6057	.1276 (.8942)	.000
.60	1.2819 (.5166)	.5340 (.8223)	.1880 (.0749)	.6749	.0952 (.7998)	.000
.90	1.4113 (.6216)	.4908 (.7742)	.1859 (.0759)	.7396	.0687 (.7265)	.000
1.00	1.4418 (.6271)	.4636 (.7536)	.1839 (.0756)	.7661	.0572 (.7000)	.000

Note: Standard deviations are shown in parentheses.

Table 5. Cut-Scores and Their Corresponding Thetas

		UIRT 3-PL	UIRT 1-PL
$r_{\theta 102}$	cut-score	cut-score on θ scale	
.00	.4	-.1600	-.3560
.00	.6	.7920	.7040
.00	.8	1.6360	1.9640
.30	.4	-.1480	-.3320
.30	.6	.6960	.5960
.30	.8	1.4640	1.6960
.60	.4	-.1320	-.3200
.60	.6	.6440	.5120
.60	.8	1.3520	1.5000
.90	.4	-.1160	-.3120
.90	.6	.5960	.4480
.90	.8	1.2480	1.3520
1.00	.4	-.1240 (-.1640)	-.3080
1.00	.6	.5680 (.4905)	.4240
1.00	.8	1.2080 (1.0794)	1.2920

Note: The corresponding theta estimates based on 2-D COMIRT 3-PL model are shown in parentheses.

Table 6. Accumulated Information Value at Each Condition

		accumulated information			
		UIRT 3-PL		UIRT 1-PL	
$r_{\theta 1 \theta 2}$	cut-score	max=50	max=360	max=50	max=360
.00	.4	27.9562	64.2432	10.1665	59.6267
.00	.6	52.8911	135.8011	10.1662	59.7696
.00	.8	63.3480	141.1518	9.8029	41.5512
.30	.4	37.2529	83.7545	13.2217	77.8227
.30	.6	64.4188	167.0143	13.2210	77.8612
.30	.8	73.7314	166.9306	12.7553	54.0386
.60	.4	45.1929	100.8731	16.4185	96.7734
.60	.6	76.0380	198.6984	16.4109	96.8255
.60	.8	88.9715	197.3844	15.8277	67.0963
.90	.4	55.2764	120.1003	19.7152	116.4226
.90	.6	91.6467	234.4135	19.7052	116.4356
.90	.8	114.8079	241.5028	18.9875	80.3927
1.00	.4	58.0903	127.7852	21.1499	125.1073
1.00	.6	98.9188	249.2493	21.1489	124.9657
1.00	.8	111.9223	244.9277	20.3591	86.4154

Table 7. UIRT 3-PL Model: Error Rates and NI

$r_{\theta 1 \theta 2}$	test length	cut-score	false positive	false negative	total error	NI
.00	15,50	.4	.0357	.0245	.0602	21.6876
.00	15,50	.6	.0189	.0284	.0473	17.8908
.00	15,50	.8	.0064	.0140	.0204	15.8457
.30	15,50	.4	.0311	.0196	.0507	19.9046
.30	15,50	.6	.0178	.0295	.0473	17.2173
.30	15,50	.8	.0071	.0143	.0214	15.8401
.60	15,50	.4	.0245	.0162	.0407	18.8116
.60	15,50	.6	.0152	.0293	.0445	16.6911
.60	15,50	.8	.0082	.0140	.0222	15.7345
.90	15,50	.4	.0212	.0111	.0323	17.8631
.90	15,50	.6	.0048	.0404	.0452	16.4223
.90	15,50	.8	.0106	.0057	.0163	15.5767
.00	1,360	.4	.0302	.0224	.0506	26.5764
.00	1,360	.6	.0192	.0292	.0484	12.4604
.00	1,360	.8	.0067	.0135	.0202	6.1368
.30	1,360	.4	.0249	.0166	.0415	18.4102
.30	1,360	.6	.0197	.0297	.0494	10.2431
.30	1,360	.8	.0094	.0152	.0246	6.5742
.60	1,360	.4	.0251	.0159	.0410	14.0345
.60	1,360	.6	.0173	.0310	.0483	8.1602
.60	1,360	.8	.0111	.0138	.0249	4.6293
.90	1,360	.4	.0214	.0156	.0370	11.2323
.90	1,360	.6	.0077	.0449	.0526	7.2471
.90	1,360	.8	.0168	.0070	.0238	4.3533
		average:	.0171	.0209	.0380	14.1476

Note: False positive is type I error. False negative is type II error. NI is the number of items used. $r_{\theta 1 \theta 2}$ is the value of correlation between the two thetas.

Table 8. UIRT 3-PL Model: Average Error Rates and NI for Different Theta Correlations

$r_{\theta 1 \theta 2}$	false positive	false negative	total error	NI
.00	.0195	.0220	.0415	16.7663
.30	.0184	.0208	.0392	14.6983
.60	.0169	.0200	.0369	13.0102
.90	.0138	.0208	.0346	12.1158

Table 9. UIRT 3-PL Model: Average Error Rates and NI for Different Test Length Constraints

test length	false positive	false negative	total error	NI
15,50	.0168	.0206	.0374	17.4571
1,360	.0175	.0212	.0387	10.8382

Table 10. UIRT 3-PL Model: Average Error Rates and NI for Different Cut-Scores

cut-score	false positive	false negative	total error	NI
.4	.0268	.0178	.0445	18.5651
.6	.0151	.0328	.0479	13.2916
.8	.0096	.0122	.0218	10.5864

Table 11. UIRT 3-PL Model: Average Error Rates and NI for Different Theta Correlations and Test Length Constraints

$r_{\theta 1 \theta 2}$	test length	false positive	false negative	total error	NI
.00	15,50	.0203	.0223	.0426	18.4747
.30	15,50	.0187	.0211	.0398	17.6540
.60	15,50	.0160	.0198	.0358	17.0791
.90	15,50	.0122	.0191	.0313	16.6207
.00	1,360	.0187	.0217	.0404	15.0579
.30	1,360	.0180	.0205	.0385	11.7425
.60	1,360	.0178	.0202	.0380	8.9413
.90	1,360	.0153	.0225	.0378	7.6109

Table 12. UIRT 3-PL Model: Average Error Rates and NI for Different Theta Correlations and Cut-Scores

$r_{\theta 1 \theta 2}$	cut-score	false positive	false negative	total error	NI
.00	.4	.0330	.0235	.0564	24.1320
.00	.6	.0191	.0288	.0479	15.1756
.00	.8	.0066	.0138	.0203	10.9913
.30	.4	.0280	.0181	.0461	19.1574
.30	.6	.0188	.0296	.0484	13.7302
.30	.8	.0083	.0148	.0230	11.2072
.60	.4	.0248	.0161	.0409	16.4231
.60	.6	.0163	.0302	.0464	12.4257
.60	.8	.0097	.0139	.0236	10.1819
.90	.4	.0213	.0134	.0347	14.5477
.90	.6	.0063	.0427	.0489	11.8347
.90	.8	.0137	.0064	.0201	9.9650

Table 13. UIRT 3-PL Model: Average Error Rates and NI for Different Cut-Scores and Test Length Constraints

cut-score	test length	false positive	false negative	total error	NI
.4	15,50	.0281	.0179	.0460	19.5667
.6	15,50	.0142	.0319	.0461	17.0554
.8	15,50	.0081	.0120	.0201	15.7493
.4	1,360	.0254	.0176	.0430	17.5634
.6	1,360	.0160	.0337	.0497	9.5277
.8	1,360	.0110	.0124	.0234	5.4234

Table 14. Summary of the Results for the UIRT 3-PL Model

factors of interest	effects
$r_{\theta 1 \theta 2}$	NI
TLC	NI
$P(\theta_c)$	I, NI
$r_{\theta 1 \theta 2}$ x TLC	NI
$r_{\theta 1 \theta 2}$ x $P(\theta_c)$	I, II, total error, NI
TLC X $P(\theta_c)$	NI
$r_{\theta 1 \theta 2}$ x TLC x $P(\theta_c$	NI

Note: TLC is test length constraint. I is type I error. II is type II error. Total is I + II. NI is number of items used.

Table 15. UIRT 1-PL Model: Error Rates and NI

$r_{\theta 1 \theta 2}$	test length	cut-score	false positive	false negative	total error	NI
.00	15,50	.4	.0441	.0393	.0834	35.4067
.00	15,50	.6	.0110	.0448	.0558	32.5814
.00	15,50	.8	.0030	.0190	.0220	22.4648
.30	15,50	.4	.0531	.0231	.0762	32.0791
.30	15,50	.6	.0137	.0369	.0506	28.9062
.30	15,50	.8	.0019	.0246	.0265	21.0503
.60	15,50	.4	.0485	.0224	.0709	29.3455
.60	15,50	.6	.0073	.0478	.0551	26.5075
.60	15,50	.8	.0017	.0278	.0295	19.9266
.90	15,50	.4	.0518	.0108	.0626	27.3215
.90	15,50	.6	.0033	.0607	.0640	24.1686
.90	15,50	.8	.0017	.0189	.0206	19.0490
.00	1,360	.4	.0270	.0155	.0425	61.0561
.00	1,360	.6	.0053	.0257	.0310	48.3970
.00	1,360	.8	.0016	.0121	.0137	27.8600
.30	1,360	.4	.0265	.0140	.0405	49.4524
.30	1,360	.6	.0039	.0307	.0346	39.5341
.30	1,360	.8	.0020	.0158	.0178	25.3111
.60	1,360	.4	.0306	.0131	.0437	41.4159
.60	1,360	.6	.0029	.0350	.0379	33.6958
.60	1,360	.8	.0012	.0163	.0175	22.4135
.90	1,360	.4	.0292	.0089	.0381	34.3713
.90	1,360	.6	.0018	.0618	.0636	28.7407
.90	1,360	.8	.0018	.0076	.0094	19.5811
		average:	.0156	.0264	.0420	31.2765

Table 16. UIRT 1-PL Model: Average Error Rates and NI for Different Theta Correlations

$r_{\theta 1 \theta 2}$	false positive	false negative	total error	NI
.00	.0154	.0261	.0415	37.9610
.30	.0169	.0242	.0411	32.7222
.60	.0154	.0271	.0425	28.8842
.90	.0149	.0281	.0430	25.5387

Table 17. UIRT 1-PL Model: Average Error Rates and NI for Different Test Length Constrtaints

test length	false positive	false negative	total error	NI
15,50	.0201	.0313	.0514	26.5673
1,360	.0112	.0214	.0326	35.9858

Table 18. UIRT 1-PL Model: Average Error Rates and NI for Different Cut-Scores

cut-score	false positive	false negative	total error	NI
.4	.0389	.0184	.0573	38.8061
.6	.0062	.0430	.0491	32.8194
.8	.0019	.0178	.0197	22.2071

Table 19. UIRT 1-PL Model: Average Error Rates and NI for Different Theta Correlations and Test Length Constraints

$r_{\theta 1 \theta 2}$	test length	false positive	false negative	total error	NI
.00	15,50	.0194	.0344	.0538	30.1510
.30	15,50	.0229	.0282	.0511	27.3452
.60	15,50	.0192	.0327	.0519	25.2599
.90	15,50	.0189	.0301	.0490	23.5130
.00	1,360	.0113	.0178	.0291	45.7710
.30	1,360	.0108	.0202	.0310	38.0992
.60	1,360	.0116	.0215	.0331	32.5084
.90	1,360	.0109	.0261	.0370	27.5644

Table 20. UIRT 1-PL Model: Average Error Rates and NI for Different Theta Correlations and Cut-Scores

$r_{\theta 1 \theta 2}$	cut-score	false positive	false negative	total error	NI
.00	.4	.0356	.0274	.0630	48.2314
.00	.6	.0082	.0353	.0434	40.4892
.00	.8	.0023	.0156	.0179	25.1624
.30	.4	.0398	.0186	.0584	40.7658
.30	.6	.0088	.0338	.0426	34.2202
.30	.8	.0020	.0202	.0222	23.1807
.60	.4	.0396	.0178	.0573	35.3807
.60	.6	.0051	.0414	.0465	30.1017
.60	.8	.0015	.0221	.0235	21.1701
.90	.4	.0405	.0099	.0504	30.8464
.90	.6	.0026	.0613	.0638	26.4547
.90	.8	.0018	.0133	.0150	19.3151

Table 21. UIRT 1-PL Model: Average Error Rates and NI for Different Cut-Scores and Test Length Constraints

cut-score	test length	false positive	false negative	total error	NI
.4	15,50	.0494	.0239	.0733	31.0382
.6	15,50	.0088	.0476	.0564	28.0409
.8	15,50	.0021	.0226	.0247	20.6227
.4	1,360	.0283	.0129	.0412	46.5739
.6	1,360	.0035	.0383	.0418	37.5979
.8	1,360	.0017	.0130	.0146	23.7914

Table 22. Summary of the Results for the UIRT 1-PL Model

factors of interest	effects
$r_{\theta_1\theta_2}$	NI
TLC	total error, NI
$P(\theta_c)$	I, total error, NI
$r_{\theta_1\theta_2}$ x TLC	NI
$r_{\theta_1\theta_2}$ x $P(\theta_c)$	II, total error, NI
TLC X $P(\theta_c)$	I, total error, NI
$r_{\theta_1\theta_2}$ x TLC x $P(\theta_c)$	II, total error, NI

Table 23. $P(\theta_c)$ Esimation Based on UIRT 3-PL Model: Average Error Rates and NI for $r_{\theta 1 \theta 2}=1$

test length	cut-score	false positive	false negative	total error	NI
15,50	.4	.0018	.0286	.0304	17.5820
15,50	.6	.0000	.0922	.0922	16.7648
15,50	.8	.0266	.0000	.0266	15.7814
	average:	.0095	.0403	.0497	16.7094
1,360	.4	.0035	.0260	.0295	9.9665
1,360	.6	.0007	.0946	.0953	7.5881
1,360	.8	.0326	.0000	.0326	4.0709
	average:	.0123	.0402	.0525	7.2085

Table 24. $P(\theta_c)$ Calculation Based on 2-D COMIRT 3-PL model: Average Error Rates and NI for $r_{\theta 1 \theta 2}=1$

test length	cut-score	false positive	false negative	total error	NI
15,50	.4	.0035	.0160	.0195	17.5935
15,50	.6	.0007	.0388	.0394	16.3329
15,50	.8	.0723	.0000	.0723	16.1449
	average:	.0255	.0183	.0438	16.6904
1,360	.4	.0029	.0173	.0202	9.9961
1,360	.6	.0013	.0488	.0501	6.0488
1,360	.8	.0778	.0000	.0778	4.3439
	average:	.0273	.0220	.0494	6.7963

Table 25. $P(\theta_c)$ Esimation Based on UIRT 1-PL Model: Average Error Rates and NI for $r_{\theta 1 \theta 2}=1$

test length	cut-score	false positive	false negative	total error	NI
15,50	.4	.0151	.0346	.0497	26.5567
15,50	.6	.0007	.1081	.1088	23.8232
15,50	.8	.0057	.0002	.0059	18.6891
	average:	.0072	.0476	.0548	23.0230
1,360	.4	.0070	.0153	.0223	30.9595
1,360	.6	.0000	.1187	.1187	31.0374
1,360	.8	.0044	.0002	.0046	18.2709
	average:	.0038	.0447	.0485	26.7559

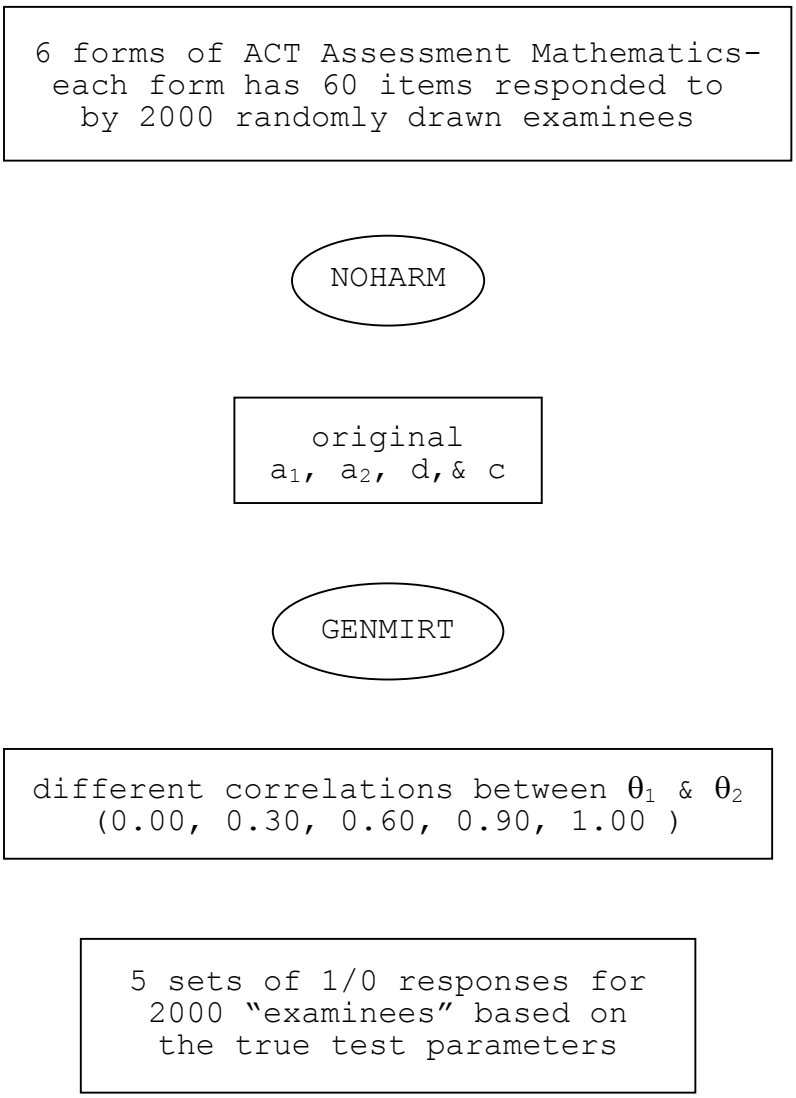

Figure 1. Dichotomous response data generation procedure

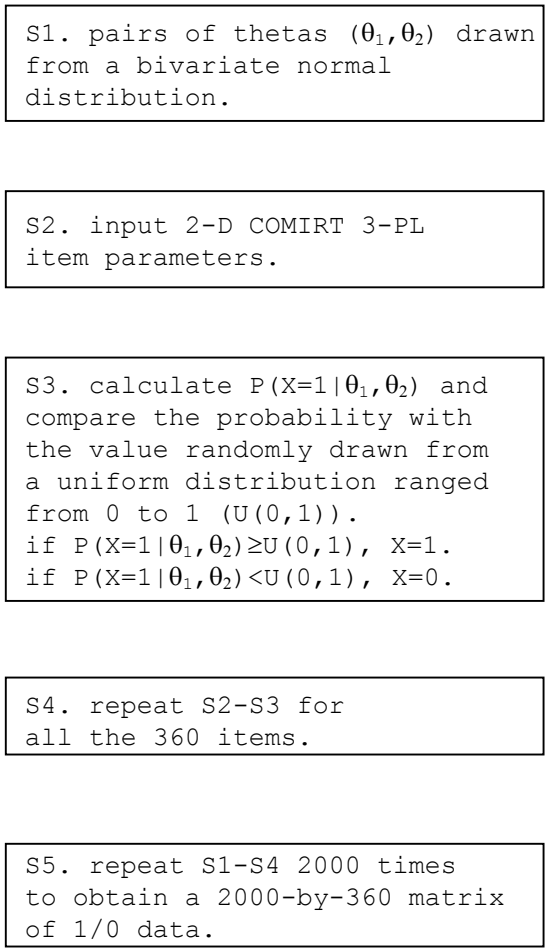

Figure 2. Method for generate dichotomous responses in GENMIRT

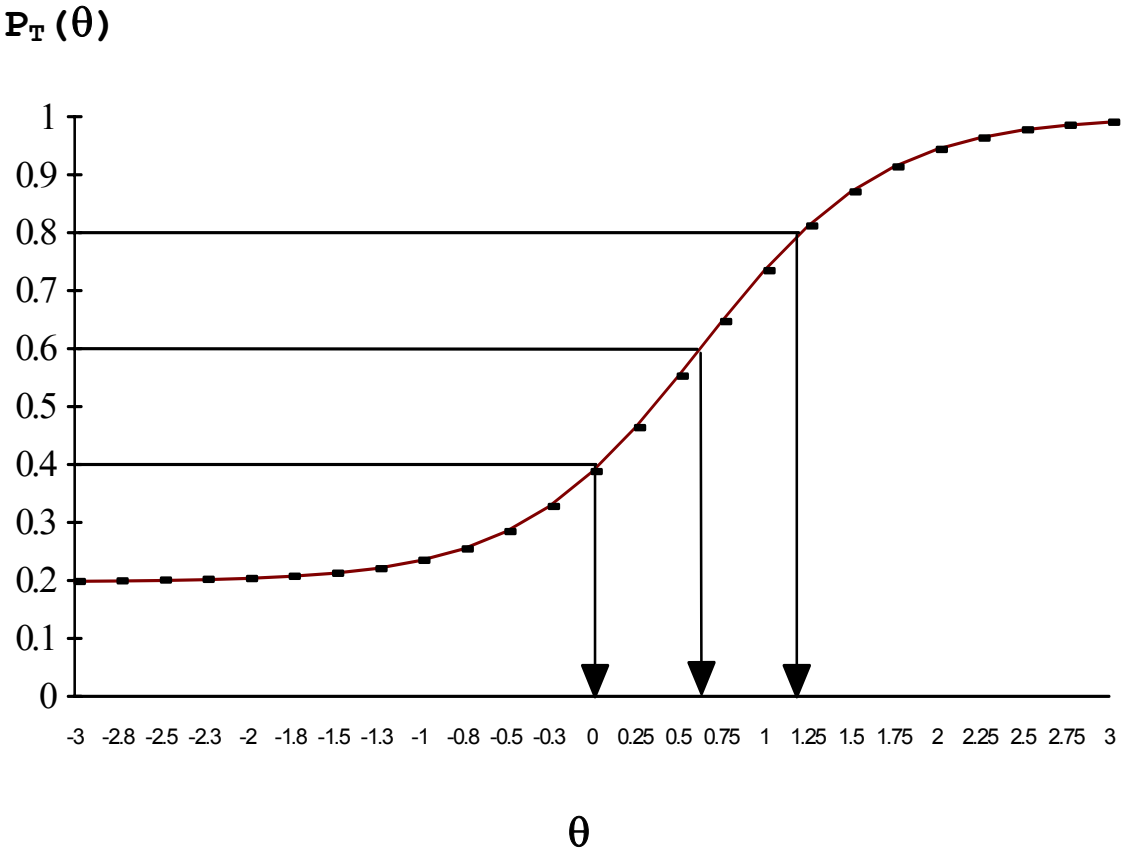

Figure 3. Illustration of the mapping of $p_T(\theta_c)$ to θ_c

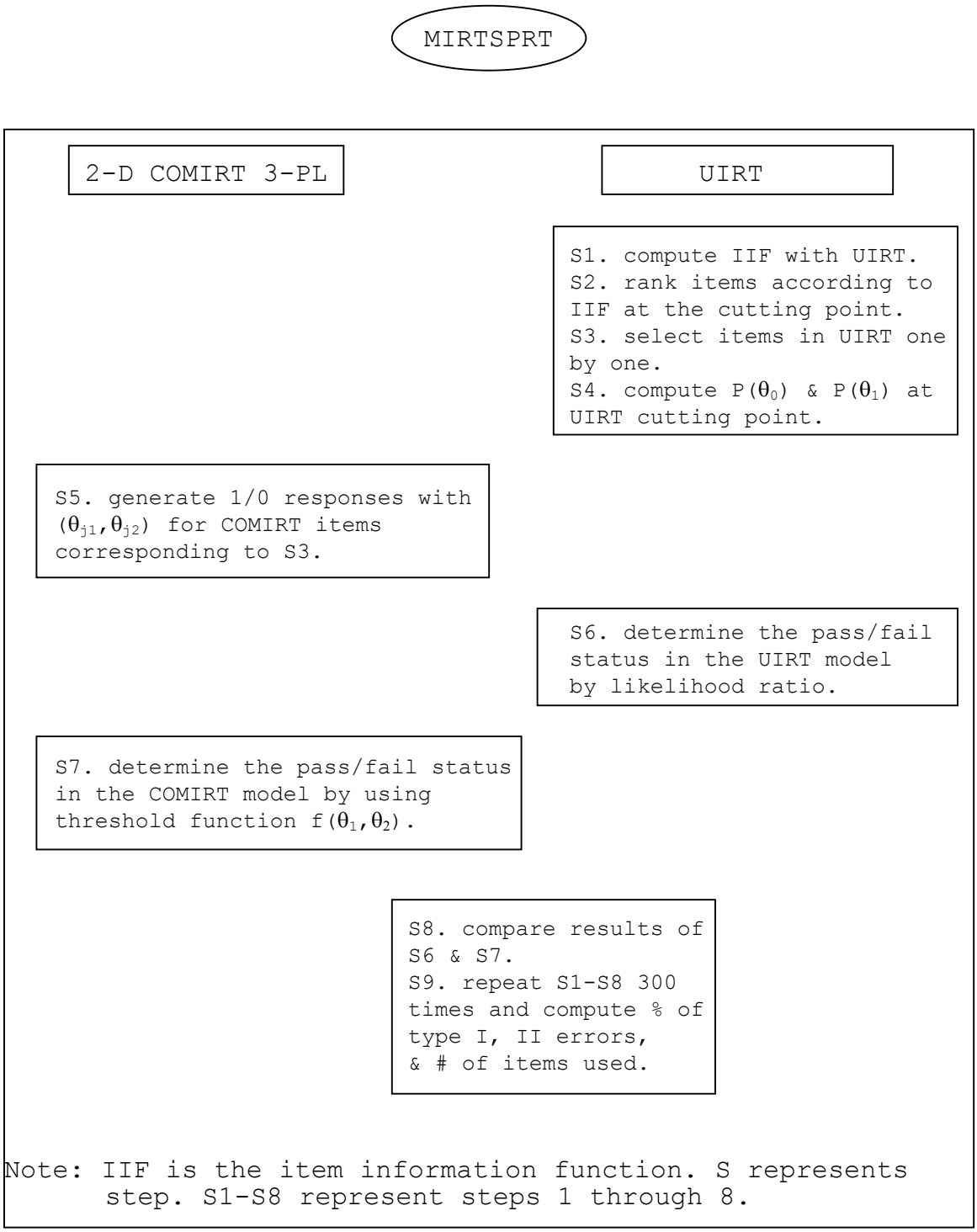

Figure 4. Type I and type II error calculation using MIRTSPRT

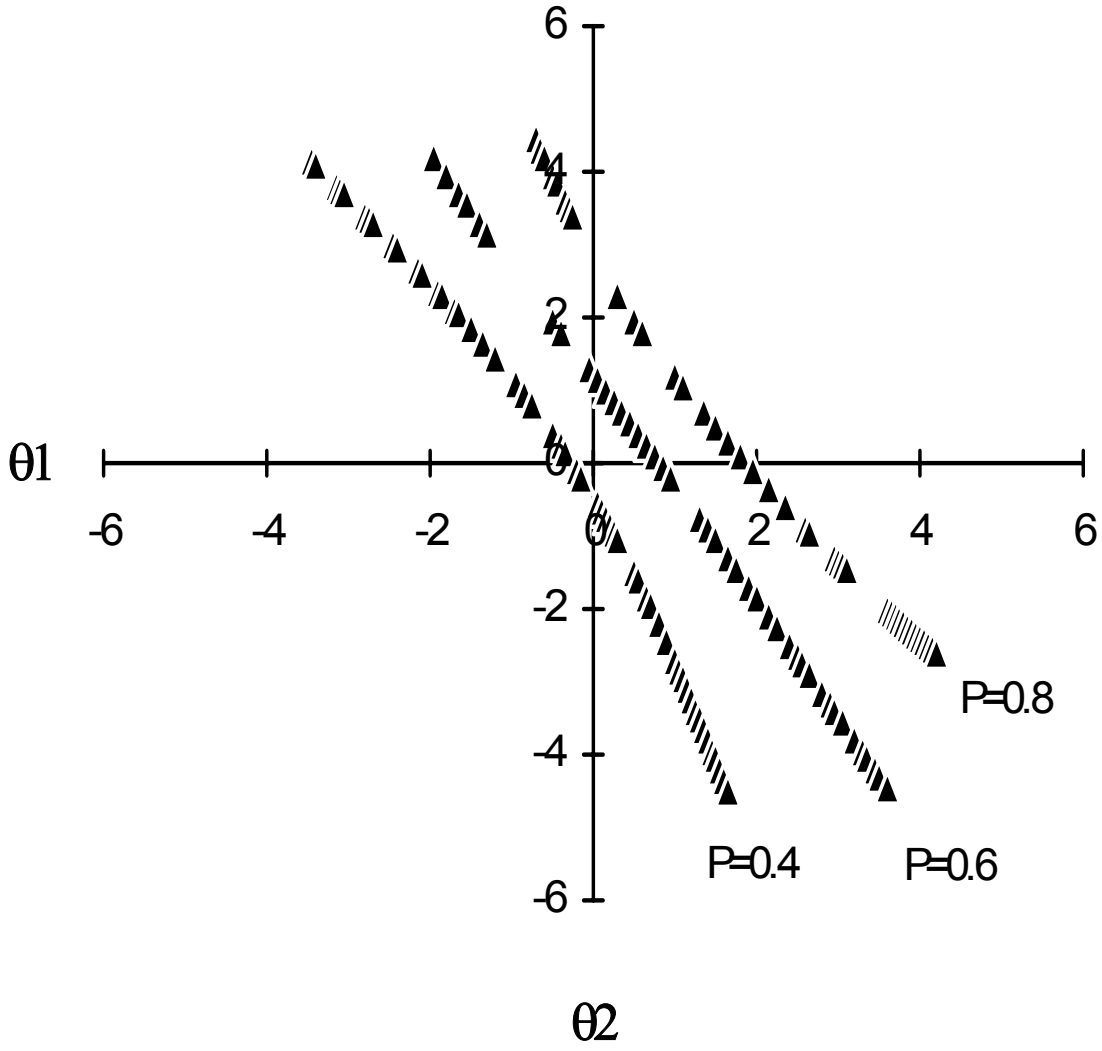

Figure 5. Threshold function contour

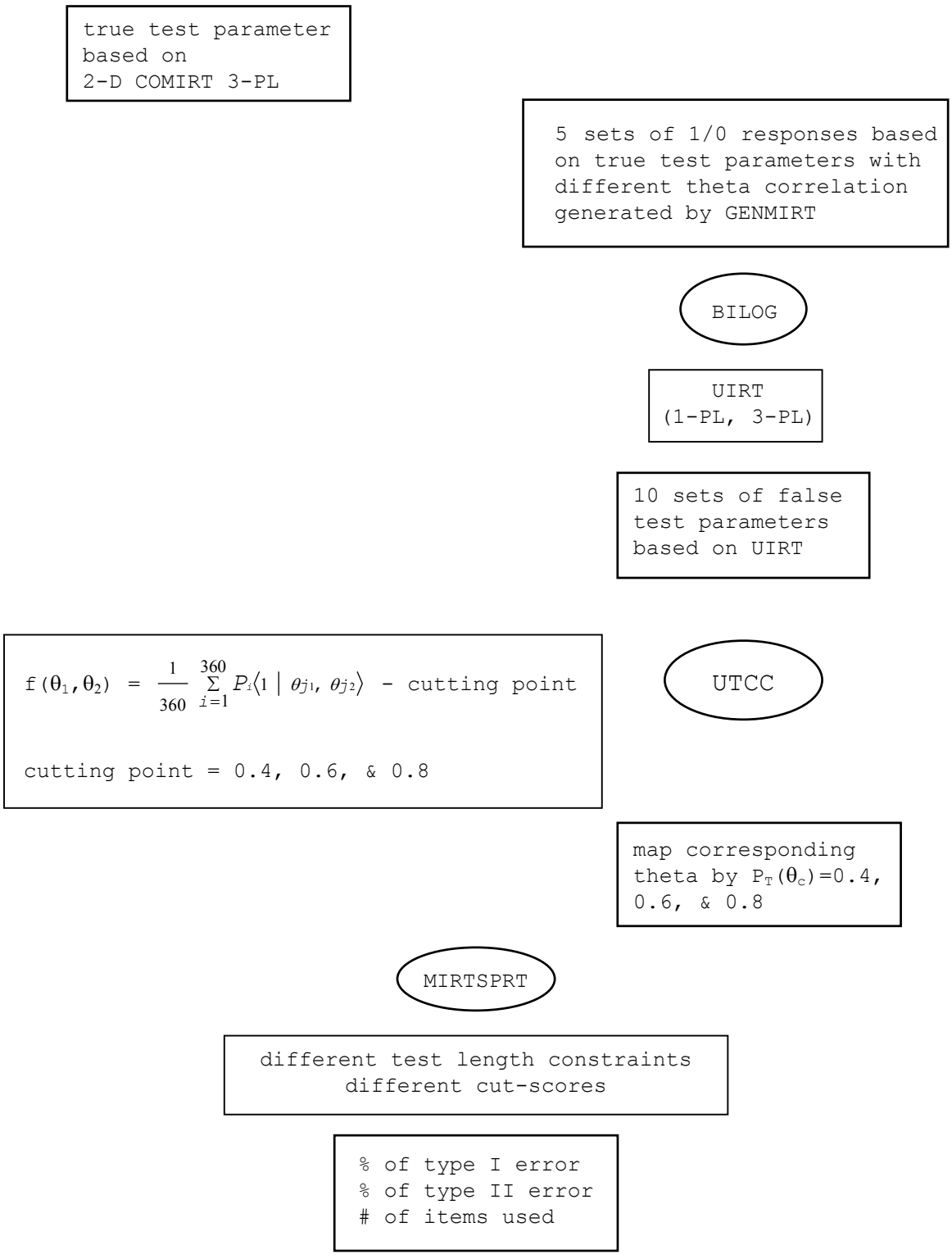

Figure 6. Data analysis flow chart

		UIRT (false)	
		pass	fail
MIRT (true)	pass		type II error
	fail	type I error	

Figure 7. Type I and type II error definition in MIRTSPRT

APPENDIX

MIRTSPRT OUTPUT

***** MIRTSPRT OUTPUT *****

FUNC	Theta1	Theta2	NI	H0	H1	TI	TII
0.187	-3.00	-3.00	5.950	1.000	0.000	0.000	0.000
0.188	-3.00	-2.50	5.570	1.000	0.000	0.000	0.000
0.189	-3.00	-2.00	5.640	1.000	0.000	0.000	0.000
0.191	-3.00	-1.50	6.037	1.000	0.000	0.000	0.000
0.194	-3.00	-1.00	6.063	1.000	0.000	0.000	0.000
0.198	-3.00	-0.50	5.797	1.000	0.000	0.000	0.000
0.204	-3.00	0.00	5.877	1.000	0.000	0.000	0.000
0.211	-3.00	0.50	5.957	1.000	0.000	0.000	0.000
0.222	-3.00	1.00	6.287	1.000	0.000	0.000	0.000
0.239	-3.00	1.50	7.143	0.997	0.003	0.003	0.000
0.263	-3.00	2.00	9.787	0.997	0.003	0.003	0.000
0.296	-3.00	2.50	17.903	0.980	0.020	0.020	0.000
0.339	-3.00	3.00	24.727	0.910	0.090	0.090	0.000
0.192	-2.50	-3.00	5.693	1.000	0.000	0.000	0.000
0.194	-2.50	-2.50	6.073	1.000	0.000	0.000	0.000
0.196	-2.50	-2.00	5.990	1.000	0.000	0.000	0.000
0.198	-2.50	-1.50	5.867	1.000	0.000	0.000	0.000
0.202	-2.50	-1.00	5.970	1.000	0.000	0.000	0.000
0.208	-2.50	-0.50	5.847	1.000	0.000	0.000	0.000
0.216	-2.50	0.00	5.817	1.000	0.000	0.000	0.000
0.227	-2.50	0.50	6.403	1.000	0.000	0.000	0.000
0.242	-2.50	1.00	6.493	1.000	0.000	0.000	0.000
0.265	-2.50	1.50	9.727	1.000	0.000	0.000	0.000
0.298	-2.50	2.00	16.140	0.977	0.023	0.023	0.000
0.341	-2.50	2.50	27.813	0.877	0.123	0.123	0.000
0.393	-2.50	3.00	37.893	0.633	0.367	0.367	0.000
0.201	-2.00	-3.00	5.503	1.000	0.000	0.000	0.000
0.203	-2.00	-2.50	5.907	1.000	0.000	0.000	0.000
0.206	-2.00	-2.00	6.133	1.000	0.000	0.000	0.000
0.210	-2.00	-1.50	6.150	1.000	0.000	0.000	0.000
0.216	-2.00	-1.00	6.113	0.997	0.003	0.003	0.000
0.223	-2.00	-0.50	5.990	1.000	0.000	0.000	0.000
0.234	-2.00	0.00	6.053	1.000	0.000	0.000	0.000
0.250	-2.00	0.50	6.653	1.000	0.000	0.000	0.000
0.273	-2.00	1.00	8.947	0.997	0.003	0.003	0.000
0.305	-2.00	1.50	15.657	0.967	0.033	0.033	0.000
0.348	-2.00	2.00	29.460	0.813	0.187	0.187	0.000
0.403	-2.00	2.50	42.930	0.453	0.547	0.000	0.453
0.463	-2.00	3.00	25.030	0.070	0.930	0.000	0.070
0.214	-1.50	-3.00	5.977	1.000	0.000	0.000	0.000
0.217	-1.50	-2.50	5.740	1.000	0.000	0.000	0.000
0.222	-1.50	-2.00	6.050	1.000	0.000	0.000	0.000
0.228	-1.50	-1.50	5.847	1.000	0.000	0.000	0.000
0.236	-1.50	-1.00	6.247	1.000	0.000	0.000	0.000
0.247	-1.50	-0.50	6.703	1.000	0.000	0.000	0.000

0.263	-1.50	0.00	7.240	1.000	0.000	0.000	0.000
0.286	-1.50	0.50	8.977	1.000	0.000	0.000	0.000
0.318	-1.50	1.00	16.410	0.977	0.023	0.023	0.000
0.362	-1.50	1.50	40.897	0.770	0.230	0.230	0.000
0.418	-1.50	2.00	33.697	0.213	0.787	0.000	0.213
0.482	-1.50	2.50	15.303	0.033	0.967	0.000	0.033
0.546	-1.50	3.00	10.780	0.007	0.993	0.000	0.007
0.233	-1.00	-3.00	5.953	0.997	0.003	0.003	0.000
0.238	-1.00	-2.50	6.010	1.000	0.000	0.000	0.000
0.245	-1.00	-2.00	6.397	1.000	0.000	0.000	0.000
0.254	-1.00	-1.50	6.227	0.997	0.003	0.003	0.000
0.266	-1.00	-1.00	7.080	1.000	0.000	0.000	0.000
0.282	-1.00	-0.50	7.857	1.000	0.000	0.000	0.000
0.305	-1.00	0.00	9.753	1.000	0.000	0.000	0.000
0.338	-1.00	0.50	21.987	0.950	0.050	0.050	0.000
0.383	-1.00	1.00	42.960	0.573	0.427	0.427	0.000
0.439	-1.00	1.50	24.913	0.090	0.910	0.000	0.090
0.506	-1.00	2.00	10.410	0.017	0.983	0.000	0.017
0.576	-1.00	2.50	7.577	0.000	1.000	0.000	0.000
0.638	-1.00	3.00	5.730	0.000	1.000	0.000	0.000
0.259	-0.50	-3.00	6.210	1.000	0.000	0.000	0.000
0.266	-0.50	-2.50	6.447	1.000	0.000	0.000	0.000
0.276	-0.50	-2.00	6.717	1.000	0.000	0.000	0.000
0.289	-0.50	-1.50	7.570	1.000	0.000	0.000	0.000
0.307	-0.50	-1.00	9.760	0.993	0.007	0.007	0.000
0.331	-0.50	-0.50	13.340	0.990	0.010	0.010	0.000
0.365	-0.50	0.00	33.217	0.897	0.103	0.103	0.000
0.410	-0.50	0.50	46.097	0.270	0.730	0.000	0.270
0.467	-0.50	1.00	12.240	0.010	0.990	0.000	0.010
0.535	-0.50	1.50	7.650	0.000	1.000	0.000	0.000
0.608	-0.50	2.00	5.673	0.000	1.000	0.000	0.000
0.675	-0.50	2.50	4.657	0.000	1.000	0.000	0.000
0.729	-0.50	3.00	4.480	0.000	1.000	0.000	0.000
0.292	0.00	-3.00	7.423	1.000	0.000	0.000	0.000
0.303	0.00	-2.50	8.213	1.000	0.000	0.000	0.000
0.318	0.00	-2.00	9.103	1.000	0.000	0.000	0.000
0.338	0.00	-1.50	14.447	1.000	0.000	0.000	0.000
0.363	0.00	-1.00	25.510	0.930	0.070	0.070	0.000
0.397	0.00	-0.50	56.527	0.627	0.373	0.373	0.000
0.443	0.00	0.00	31.000	0.090	0.910	0.000	0.090
0.501	0.00	0.50	9.617	0.007	0.993	0.000	0.007
0.568	0.00	1.00	5.863	0.000	1.000	0.000	0.000
0.641	0.00	1.50	4.950	0.000	1.000	0.000	0.000
0.711	0.00	2.00	4.547	0.000	1.000	0.000	0.000
0.768	0.00	2.50	4.347	0.000	1.000	0.000	0.000
0.810	0.00	3.00	4.123	0.000	1.000	0.000	0.000
0.335	0.50	-3.00	10.493	0.990	0.010	0.010	0.000
0.352	0.50	-2.50	14.397	0.993	0.007	0.007	0.000
0.373	0.50	-2.00	26.433	0.953	0.047	0.047	0.000

0.400	0.50	-1.50	54.233	0.750	0.250	0.250	0.000
0.434	0.50	-1.00	53.610	0.270	0.730	0.000	0.270
0.479	0.50	-0.50	17.183	0.017	0.983	0.000	0.017
0.537	0.50	0.00	8.440	0.000	1.000	0.000	0.000
0.603	0.50	0.50	5.283	0.000	1.000	0.000	0.000
0.675	0.50	1.00	4.443	0.000	1.000	0.000	0.000
0.746	0.50	1.50	4.203	0.000	1.000	0.000	0.000
0.804	0.50	2.00	4.120	0.000	1.000	0.000	0.000
0.847	0.50	2.50	4.077	0.000	1.000	0.000	0.000
0.876	0.50	3.00	4.020	0.000	1.000	0.000	0.000
0.387	1.00	-3.00	22.427	0.960	0.040	0.040	0.000
0.410	1.00	-2.50	52.020	0.783	0.217	0.000	0.783
0.438	1.00	-2.00	55.900	0.437	0.563	0.000	0.437
0.473	1.00	-1.50	33.257	0.060	0.940	0.000	0.060
0.517	1.00	-1.00	14.277	0.000	1.000	0.000	0.000
0.572	1.00	-0.50	7.910	0.000	1.000	0.000	0.000
0.638	1.00	0.00	4.997	0.000	1.000	0.000	0.000
0.708	1.00	0.50	4.360	0.000	1.000	0.000	0.000
0.776	1.00	1.00	4.150	0.000	1.000	0.000	0.000
0.835	1.00	1.50	4.090	0.000	1.000	0.000	0.000
0.878	1.00	2.00	4.043	0.000	1.000	0.000	0.000
0.906	1.00	2.50	4.017	0.000	1.000	0.000	0.000
0.924	1.00	3.00	4.010	0.000	1.000	0.000	0.000
0.446	1.50	-3.00	56.233	0.410	0.590	0.000	0.410
0.475	1.50	-2.50	41.847	0.117	0.883	0.000	0.117
0.510	1.50	-2.00	20.707	0.007	0.993	0.000	0.007
0.553	1.50	-1.50	10.567	0.000	1.000	0.000	0.000
0.606	1.50	-1.00	6.833	0.000	1.000	0.000	0.000
0.669	1.50	-0.50	4.803	0.000	1.000	0.000	0.000
0.737	1.50	0.00	4.307	0.000	1.000	0.000	0.000
0.800	1.50	0.50	4.097	0.000	1.000	0.000	0.000
0.857	1.50	1.00	4.037	0.000	1.000	0.000	0.000
0.900	1.50	1.50	4.020	0.000	1.000	0.000	0.000
0.928	1.50	2.00	4.010	0.000	1.000	0.000	0.000
0.944	1.50	2.50	4.000	0.000	1.000	0.000	0.000
0.953	1.50	3.00	4.007	0.000	1.000	0.000	0.000
0.509	2.00	-3.00	27.187	0.037	0.963	0.000	0.037
0.544	2.00	-2.50	15.907	0.003	0.997	0.000	0.003
0.586	2.00	-2.00	9.233	0.000	1.000	0.000	0.000
0.637	2.00	-1.50	7.007	0.000	1.000	0.000	0.000
0.696	2.00	-1.00	5.247	0.000	1.000	0.000	0.000
0.759	2.00	-0.50	4.367	0.000	1.000	0.000	0.000
0.818	2.00	0.00	4.127	0.000	1.000	0.000	0.000
0.871	2.00	0.50	4.037	0.000	1.000	0.000	0.000
0.913	2.00	1.00	4.007	0.000	1.000	0.000	0.000
0.941	2.00	1.50	4.000	0.000	1.000	0.000	0.000
0.958	2.00	2.00	4.020	0.000	1.000	0.000	0.000
0.967	2.00	2.50	4.010	0.000	1.000	0.000	0.000
0.972	2.00	3.00	4.000	0.000	1.000	0.000	0.000

0.573	2.50	-3.00	12.200	0.000	1.000	0.000	0.000
0.614	2.50	-2.50	8.673	0.000	1.000	0.000	0.000
0.662	2.50	-2.00	6.717	0.000	1.000	0.000	0.000
0.716	2.50	-1.50	4.957	0.000	1.000	0.000	0.000
0.775	2.50	-1.00	4.300	0.000	1.000	0.000	0.000
0.830	2.50	-0.50	4.073	0.000	1.000	0.000	0.000
0.878	2.50	0.00	4.043	0.000	1.000	0.000	0.000
0.919	2.50	0.50	4.013	0.000	1.000	0.000	0.000
0.947	2.50	1.00	4.000	0.000	1.000	0.000	0.000
0.966	2.50	1.50	4.000	0.000	1.000	0.000	0.000
0.975	2.50	2.00	4.000	0.000	1.000	0.000	0.000
0.981	2.50	2.50	4.000	0.000	1.000	0.000	0.000
0.983	2.50	3.00	4.000	0.000	1.000	0.000	0.000
0.637	3.00	-3.00	8.303	0.000	1.000	0.000	0.000
0.682	3.00	-2.50	6.237	0.000	1.000	0.000	0.000
0.731	3.00	-2.00	5.033	0.000	1.000	0.000	0.000
0.784	3.00	-1.50	4.320	0.000	1.000	0.000	0.000
0.835	3.00	-1.00	4.040	0.000	1.000	0.000	0.000
0.880	3.00	-0.50	4.007	0.000	1.000	0.000	0.000
0.919	3.00	0.00	4.013	0.000	1.000	0.000	0.000
0.948	3.00	0.50	4.013	0.000	1.000	0.000	0.000
0.968	3.00	1.00	4.007	0.000	1.000	0.000	0.000
0.979	3.00	1.50	4.000	0.000	1.000	0.000	0.000
0.985	3.00	2.00	4.000	0.000	1.000	0.000	0.000
0.989	3.00	2.50	4.000	0.000	1.000	0.000	0.000
0.990	3.00	3.00	4.000	0.000	1.000	0.000	0.000

E(TYPE1) = .0251 E(TYPE2) = .0159 E(ERR) = .0410
E(NI) = 14.03448

DECPT = -.1320
INDIF REGION = (-.3820 .1180)

Rho = .6000
P = .4000

Alpha = .0500 beta = .0500
SEED = .98765D+00

Maximum Test Length = 360 Minimum Test Length = 1
Simulation is based on 300 replications for 360 items.

Note: FUNC is computed based on $\frac{1}{360}\sum_{i=1}^{360} P_i\langle 1 \mid \theta_{j1}, \theta_{j2}\rangle$. Theta1 is the first theta, θ_1. Theta2 is the second theta, θ_2. NI is number of items used. H0 is the null hypothesis in SPRT. H1 is the alternative hypothesis in SPRT. TI is type I error rate. TII is type II error rate.

E(TYPE1) is the expected type I error rate. E(TYPE2) is the expected type II error rate. E(NI) is the expected number of items used. DECPT is the cut-score on theta scale. INDIF REGION is the range of lower and upper limits set in SPRT. Rho is the θ-correlation. P is the cut-score. Alpha is the nominal type I error rate. Beta is the nominal type II error rate.