RUNNING HEAD: Polytomous CCT for Multiple Cutscores

A COMPARISON OF TWO METHODS OF POLYTOMOUS COMPUTERIZED CLASSIFICATION TESTING FOR MULTIPLE CUTSCORES

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL OF
THE UNIVERSITY OF MINNESOTA BY

NATHAN A. THOMPSON

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

DAVID J. WEISS, PH.D., ADVISOR

JULY 2007

Abstract

The sequential probability ratio test (SPRT: Wald, 1947) and adaptive mastery testing (AMT: Kingsbury & Weiss, 1983) have been shown to be effective termination criteria for computerized classification testing (CCT: Kingsbury & Weiss, 1983; Reckase, 1983; Spray & Reckase, 1994; Lin & Spray, 2000). The SPRT and AMT were originally designed for only two classifications, such as “pass” and “fail” (Kingsbury & Weiss, 1983), but can be expanded to three or more classifications (Weiss & Kingsbury, 1984; Spray, 1993; Eggen & Straetmans, 2000). However, all research that has been done with the multiple cutscore CCT has utilized only dichotomous item response theory models. A procedure has been proposed (Lau & Wang, 1998; 1999; 2000) to modify the SPRT to accommodate ordered polytomous IRT models such as the partial credit model (Masters, 1982) and the generalized partial credit model (Muraki, 1992). These models potentially increase the level of information provided by an item across the ability scale, which is increasingly important as the number of classifications increases. This study investigated the applicability of polytomous item response theory methods to CCT for multiple cutscores. The relative efficiency of the procedure as compared to the dichotomous multiple cutscore CCT, as well as a comparison of termination criteria and item selection criteria, was investigated in a monte carlo simulation study. It was found that the advantages of testing efficiency, evaluated by the average test length and percentage of correct classifications, for two-classification CCT are even greater for three-classification CCT. CCTs with the SPRT termination criterion utilized fewer items than AMT while being matched on observed accuracy.

Table of Contents

Chapter 1: Introduction

Ability and achievement tests are an important part of modern society, serving many functions. One of the chief functions is the use of an ability or achievement test to make a decision regarding examinees with regards to a cutscore, a specific score on which the decision is based. Testing applications that are intended to classify examinees into mutually exclusive categories are known as *classification tests*. Usually, this is a dichotomous decision with only two possible outcomes, such as the ubiquitous case of pass/fail, which requires a single cutscore to make the classification, though some applications call for multiple cutscores. Many classification tests are high-stakes tests, where the decision carries important consequences such as graduating high school or being able to practice one's chosen profession. Because of the extensive amount of high-stakes classification testing, it is important to ensure that the tests are as efficient and accurate as possible. This study investigated methods of designing computerized tests that maximize efficiency and accuracy for both two classifications and three classifications.

Classification tests are most often administered to assess the extent of learning or mastery. The accompanying nomenclature reflects the aim of the test, which is to determine if a student has learned enough of the material to pass the test and be deemed a "master," or fail and be classified as a "nonmaster." Yet classification tests are applicable in any case where the purpose is to divide a sample of examinees into ordered groups. For example, job applicants might be described as having a distribution of qualifications, with higher qualified applicants having greater education, experience, intelligence, or job knowledge. It is beneficial for the hiring organization to easily classify applicants along this continuum into those applicants worth offering positions, and those for whom it would not be as sensible to offer a position. Likewise, a psychological clinic might wish to divide applicants along the depression continuum into those who are clinically depressed and those who are not. However, as the most common application is the simple pass-fail or master-nonmaster case, this will be the example terminology generally used herein for the two classification situation.

Two of the most visible types of classification tests, which provide a ready example, are licensure and certification tests. These are pass/fail tests that assess a person's ability to perform certain tasks or work in a certain profession. Licensure and certification tests are similar in form and construction, and are therefore often referred to collectively as *credentialing tests*, but have an important distinction in legal context and purpose. Licensure refers to requirements imposed by law for entry into a profession. Such tests are mandated in an effort to protect the public in situations where lack of qualifications could lead to public harm. Medical doctors require licensure for just such a reason, as well as many other professions. However, a situation is possible where licensure is required by law even though malpractice of the profession has no detrimental effect on the public. In such case, government has established licensure status by a profession that wishes to restrict entry of others into the profession. For example, a test must be passed to obtain licensure to be a florist in the State of Louisiana (Sullum, 2004), though the consequences of an incompetent florist are much less dire than an incompetent doctor. Certification tests, on the other hand, are voluntary. In this case, the credential is administered by a board or commission representing the best interests of a profession, but the credential is not required to work in the profession. Certification is intended to provide public knowledge of a higher level of knowledge and skills.

An even larger portion of testing in society can be considered classification testing if tests are included which are designed to result in scores that are later classified into groups. For example, many norm-referenced tests are later used to make decisions, such as a college establishing a minimum score on an entrance exam to distinguish students that meet expectations for admission.

Multiple Cutscore Classification Testing

While many classification tests are designed for only one cutscore with two classifications, some are designed to classify examinees into three or more groups ordered on an underlying continuum, which entails multiple cutscores. For instance, in the two examples used above, a third group might be defined in between the two classifications, such as “partial master,” or “moderately depressed.” There are several reasons for using multiple cutscores. The “partial master” classification would be useful for identifying a borderline group that does not have requisite mastery of the material, but would likely achieve that level with a small amount of retraining. For example, students could be classified with a placement test into those students who need remedial training, those who belong in a mainstream class, and those who warrant accelerated instruction. A human resource director in charge of hiring decisions might wish to classify examinees into those who would likely fail at the job, those who would likely perform adequately if put on the job immediately, and those who might perform adequately after some training.

Another type of multiple-cutscore test is for evaluation purposes, both of examinees and educational programs. The National Assessment of Educational Progress (NAEP: Loomis & Bourque), also known as the Nation’s Report Card, is an example of this type of test. The NAEP classifies examinees into four categories, often called *competence levels*: Below Basic, Basic, Proficient, and Advanced. This obviously serves to identify characteristics of each student, but not all students are tested because its primary purpose is to evaluate the quality of education in certain schools, districts, cities, or states by assessing the percentage of students in each classification.

An important issue in testing for multiple cutscores is the strain that multiple cutscores introduce to the testing procedure. In two-classification testing, the region of ability where the test needs the greatest number of items to make a decision, and still has reduced accuracy in classifying examinees, is near the cutscore (Spray and Reckase, 1994). If a test has three or four classifications (two or three cutscores), the number of examinees that need a high number of items increases (Spray, 1993). This therefore increases the average number of items needed across the population, which in turn increases the number of items needed in a bank for effective test administration without overexposure of items.

The development of classification tests for multiple cutscores is similar to that of two-classification tests. For example, the establishment of content areas or writing of items is not necessarily affected by how many cutscores there are. However, the two situations differ in terms of specific psychometric methods used to administer and score the test. While methods such as the sequential probability ratio test (Spray, 1993) are generally applicable to both cases, certain characteristics must be adapted to the multiple-cutscore situation. The aspect of classification testing for which this is most true is the algorithms employed in computerized classification testing.

Computerized Classification Testing

The proliferation of the personal computer in the 1990s has led to a large portion of examinations, especially high-stakes classification examinations, being administered by computer. This includes education (Eggen & Straetmans, 2000) and professions as varied as nurses (O’Neill, Marks, & Liu, 2006), architects (Braun, Bejar, & Williamson, 2006), and accountants (Buchanan, Vucinic, Rigos, & Gleim, 2004). While some classification tests delivered by computer are merely computerized administrations of conventional fixed-form tests, where each examinee receives a fixed number of items administered in a fixed order, the application of the computer allows for the design of much more technologically sophisticated tests. The technological advancement that is available with computers, but not with conventional paper-and-pencil testing, is *variable-length testing*. In variable-length testing, not every examinee receives a test of the same length. Instead, an effort is made to administer fewer items by administering items one at a time and terminating the test after a specific criterion has been reached. Those examinees who quickly distinguish themselves as easily classifiable after a small number of items are then classified without the administration of more items. This saves time and effort for both the examinee and the computer, allowing the testing of more examinees in a given amount of time. This benefit has important financial consequences when considered across thousands or hundreds of thousands of examinees.

The most widely known form of variable-length testing is *computerized adaptive testing* (CAT: Weiss & Kingsbury, 1984; Thissen, 2000). CAT not only adjusts the length of the test for each examinee, but also uses information on that individual examinee, in the form of their past responses to items with known parameters, to adaptively select each item. Item difficulty is matched with examinee ability, so that a highly able examinee does not waste time with very easy items, and an examinee with lower ability is not faced with items of high difficulty. The test is often terminated when the examinee's conditional standard error of measurement (CSEM) falls below a predetermined value, though other criteria may be applied. This makes for tests that typically use 50% as many items (Weiss & Kingsbury, 1984) as a fixed-form conventional test.

CAT is a general term that refers to variable-length tests that use adaptive item selection, whether the test is used for classification or not. A *computerized classification test* (CCT; Lin & Spray, 2000) is a variable-length computerized test that is specifically designed for examinee classification. CAT and CCT overlap. Some CCTs are adaptive; however, there are efficient CCTs that do not use adaptive item selection. Variable-length tests can also be constructed with algorithms for random item selection and sequential item selection. Sequential selection is a nonrandom algorithm that intelligently selects items, such as the selection of items to maximize information at a cutscore point, but does not make use of information for each individual examinee.

The construction of a variable-length CCT requires the specification of three main characteristics:

1. An item bank calibrated with a selected psychometric model
2. An item selection algorithm
3. A termination criterion.

CCTs can be categorized by the methodologies they use within the test to address these three design characteristics. Since CCT research began, there have been many procedures proposed, varying in the methods specified for each characteristic. The large number of possible permutations, as well as the interaction of characteristics, precludes the statement of one procedure as the most efficient.

The primary method of CCT categorization is the termination criterion. There are three available termination criteria: adaptive mastery testing (AMT; Kingsbury & Weiss, 1984), the sequential probability ratio test (SPRT; Wald, 1947), and Bayesian decision theory (BDT; van der Linden, 1990). AMT constructs confidence intervals around the current trait estimate using an estimate of the examinee's conditional standard error of measurement after each item. The SPRT makes a decision between simple competing hypotheses, such as mastery and nonmastery. BDT gives the test user choices of various loss structures and functions, and classifications are made by attempting to minimize loss.

The item selection method provides a secondary type of categorization. Variable-length tests can be divided in two ways, based on how the next item is selected: random vs. intelligent, and sequential vs. adaptive. These two dichotomies delineate methods on the same criterion – item selection – in an overlapping fashion. Random item selection assumes that the items are more or less equal and therefore randomly selects the next item to administer to an examinee. A more complex approach is to acknowledge that item characteristics, such as item difficulty, might vary, and an attempt is made to instruct the computer to intelligently select the "best" item to administer next to the examinee. The definition of "best" is often defined by the maximization of some psychometric function (e.g., item information) chosen by the test user, but there are many strategies for this. These include the use of prior information on the examinee or population (Rudner, 2002), item information in a region (Eggen, 1999), item information at a point such as the cutscore (Spray & Reckase, 1994), and the current trait estimate (Reckase, 1983).

With CAT (Weiss & Kingsbury, 1984), items are selected to match the examinee's trait level as estimated during the test and therefore must take into account *individual examinee information,* such as the response vector on all items administered, at a given stage of the test. The psychometric function to be maximized includes parameters for the examinee. Sequential tests, on the other hand, do not use examinee information, and item selection is either random (Ferguson, 1969) or based on information not related to the examinee. Sources of this information for intelligent sequential tests include Bayesian loss and utility structures (Rudner, 2002; see also Lewis and Sheehan, 1990; Vos, 1999), item information at the cutscore point (Spray and Reckase, 1994), global or regional (Kullback-Liebler) item information (Eggen, 1999; Rudner, 2002), and mutual information (Weissman, 2004). Therefore, CAT involves only those methods of intelligent item selection that utilize information on a given single examinee; whereas all other intelligent methods are subsumed under sequential testing, along with random item selection.

Certain item selection methods are often used with certain termination criteria. Of the three commonly used families of termination methods, sequential Bayesian procedures are traditionally used with the assumption of random item selection (e.g., Vos, 1998) and AMT procedures are traditionally adaptive (e.g., Kingsbury & Weiss, 1983). Although the SPRT was originally random (Ferguson, 1969), it is now used with intelligent item selection, both sequentially and adaptively (Eggen & Straetmans, 2000). It is a misconception to believe that the three families must be completely adaptive or sequential, random or intelligent; they are defined by the termination criteria involved, not by item selection.

The choice of psychometric model also presents an important option to the test user. While an efficient CCT can be designed with classical test theory (Frick, 1992), most CCT research has focused on the application of dichotomous item response theory (IRT: Embretson & Reise, 2000). The exception to this is Lau and Wang (1998; 1999; 2000), who used a polytomous item response theory model, which allows for partial credit in examinee responses. IRT is based on the premise that the relationship between an examinee's ability or achievement level and the probability of a response can be described by a logistic or cumulative normal mathematical function known as the item response function (IRF). With dichotomous IRT, the only response considered in the function is the correct or keyed response. With polytomous IRT, all responses are considered, such as an item where each of the incorrect responses receives some form of partial credit.

Because polytomously calibrated items provide more information across a wider range of ability than dichotomously scored items, fewer items are generally needed to make classifications with polytomous IRT than with dichotomous IRT. This early research on CCT with polytomous IRT examined only the two-category case, but the multiple-cutscore case offers the possibility of even greater superiority of polytomous IRT CCT over dichotomous IRT. This is because the multiple cutscore case requires more information across a greater range of ability, because examinees are evaluated in reference to two or more cutscores rather than a single point.

Polytomous IRT offers just that, presenting an alternative to increasing item bank size, which has great practical importance. Rather than face the expense of developing a larger item bank to perform multiple-cutscore CCT, a smaller item bank that is polytomously calibrated might be just as effective. This also reduces the average number of items needed to classify examinees, which saves testing time, an important issue in large-scale testing programs such as the NAEP, which tests millions of students across the United States.

Other choices that represent practical issues are often encountered during the development process. For instance, many CCTs have a truncation rule that prevents the CCT from administering every item in the bank if the termination criterion is never able to make a classification. Additionally, many testing programs are concerned with the overexposure of their items to the candidate pool, and institute some form of exposure controls into the item selection process. Options such as these are ancillary rather than necessary, but do serve very important purposes for the testing program.

How CCT Works

A CCT can be conceptualized as a test that operates in “rounds” rather than being given in one large block, as occurs with a paper-and-pencil test where the examinee is simply given a large number of items and must complete them all. The general idea is that an item or group of items (testlet) is selected at the beginning of each round, the examinee responds to what is presented, and the computer uses the responses to evaluate if the examinee can be classified. The termination criterion provides the quantitative basis for this evaluation. If the examinee can be classified, the test is terminated. If the termination criterion cannot make a decision, the process repeats itself with another round.

The following is an example of the interaction between an examinee and a computer in the administration of a CCT. The examinee is presented one item, which is answered. The computer immediately scores the item and checks to see if the termination criterion leads to a decision regarding the examinee’s classification. This is not likely to occur after only one item, so the computer proceeds to the next round and selects another item to administer. The examinee responds to this item, and the termination criterion is evaluated again. If the examinee is not able to be classified, a third item is administered. This process continues until the termination criterion classifies the examinee, the item bank is exhausted, or a truncation rule such as a maximum test length is reached. Because the termination criterion and item selection process operate efficiently, the termination criteria are sometimes able to make classifications after only a few items (Spray & Reckase, 1994).

It should be noted that CCTs considered in this study assume that the test is composed of single-best-answer multiple-choice items that are selected one at a time. Some research has used testlets (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992), or pre-bundled sets of items, but these are used less often for CCT. There also exist computerized tests for examinee classification that use sophisticated item formats such as performance scenarios (Braun, Bejar, and Williamson, 2006), but the methodologies used in constructing such tests are substantially different. In the future, procedures might be proposed that combine the two approaches, but currently no research has explored the use of alternative item formats with CCT methodologies such as IRT-based item selection and termination criteria. Such procedures would be highly efficient.

Purpose

The general purpose of CCT research is to identify specific methods that maximize the efficiency of the CCT. While a CCT can make a classification after only a few items, such as for examinees whose ability is far above or below the cutscore in the two-classification case, this is not true for the majority of examinees. In fact, examinees whose ability level is near the cutscore might respond to every item in the bank, and the termination criterion will still not able to make a classification. Therefore, the development of more efficient methods helps to reduce the average number of items per examinee while retaining high levels of precision and accuracy. This is important because, as mentioned previously, a reduction in testing time per examinee can have substantial financial and resource implications multiplied across large numbers of examinees.

The purpose of this study was to investigate the efficiency and accuracy of simulated CCTs under several different specifications, similar to previous research. Five independent variables were considered, each with two levels. Three variables reflect the three characteristics of CCT: termination criterion, item selection algorithm, and psychometric model. The two termination criteria that were compared were the SPRT and AMT. Two approaches to item selection were utilized: estimate-based (adaptive) and cutscore-based (sequential). The advantages of using a polytomous psychometric model over a dichotomous model were also examined. The remaining two independent variables were item bank shape and the number of cutscore.

What separates this research from previous research is the application of polytomous IRT as the psychometric model for a multiple-cutscore CCT. Only Lau and Wang (1998; 1999; 2000) have applied polytomous IRT to CCT, and that initial research was limited to the two-classification case. Although they found the polytomous model to perform more efficiently, the advantage of polytomous IRT should be even greater for multiple-cutscore CCT. This was the primary purpose of the study.

Chapter 2: Development of CCTs

To design a CCT, a test developer must specify a method to address each of the three characteristics. First, a calibrated item bank must be developed, which requires the choice of a psychometric model to calibrate the items. A psychometric model is necessary for a CCT because item selection and termination criteria utilize a psychometric model to specify parameters and perform calculations. Next, a termination criterion is necessary to both determine when to stop the test and how to classify the examinee. In CAT, these two characteristics are separate, but the termination criteria used for CCT do both simultaneously: the test is terminated when the examinee is classified. Lastly, because a CCT dynamically selects items or groups of items throughout the test, a rule for performing this selection must be specified.

Psychometric Model

There are three main options for the psychometric model to use in designing a CCT: classical test theory, dichotomous IRT, and polytomous IRT. CCTs were first designed with classical test theory (Ferguson, 1969). While it is possible to construct efficient CCTs using classical test theory (Frick, 1992), they are more often designed with IRT. The first application of IRT to CCT was Kingsbury and Weiss (1979), who utilized it in the development of the AMT termination criterion. Reckase (1983) later applied it to the SPRT termination criterion.

Recent CCT research has focused on the use of dichotomous IRT (e.g., Eggen, 1999; Eggen & Straetmans, 2000; Thompson & Weiss, 2006). This is due to several reasons. First, item banks for current-day large-scale testing programs are often calibrated with dichotomous IRT. Second, IRT item parameters provide an efficient method for specifying the parameters of the SPRT and are necessary for the estimation of ability (θ) and the conditional standard error of measurement for AMT (Kingsbury & Weiss, 1983). Moreover, IRT enables the use of item information functions, which are applied in highly efficient algorithms for item selection in CCT (Spray & Reckase, 1994).

There are three commonly used dichotomous IRT models. This study used the three-parameter logistic model (3PL), an IRT model commonly applied to multiple-choice data for achievement or ability tests when examinee guessing is likely. With the 3PL, the probability of an examinee with a given θ correctly responding to an item is (Hambleton & Swaminathan, 1985, Eq. 3.3)

$$P_i(X=1|\theta_j)=c_i+(1-c_i)\frac{\exp[Da_i(\theta_j-b_i)]}{1+\exp[Da_i(\theta_j-b_i)]} \qquad (1)$$

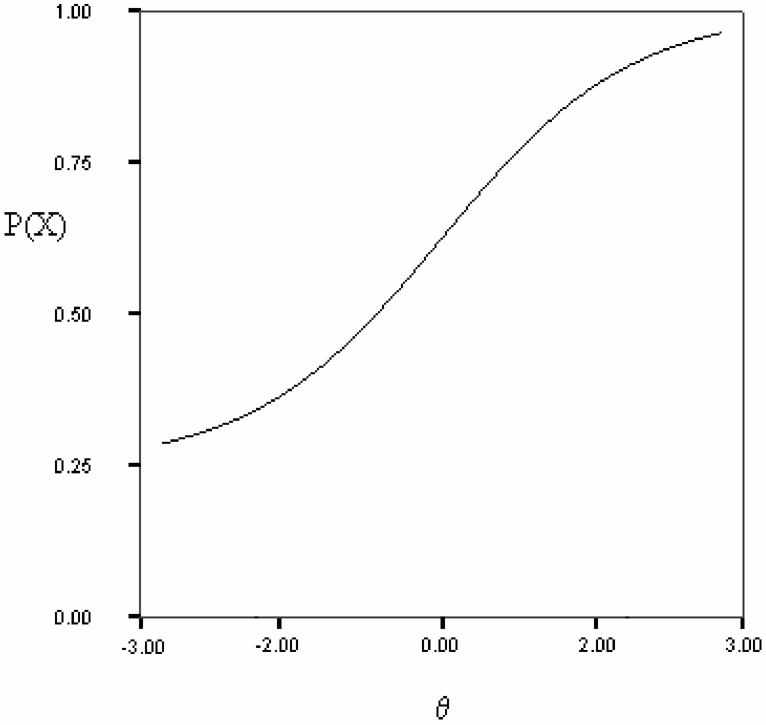
where

a_i is the item discrimination parameter,
b_i is the item difficulty or location parameter,
c_i is the lower asymptote, or pseudoguessing parameter, and
D is a scaling constant equal to 1.702 (in this study) or 1.0.

The three dichotomous IRT models are described by the number of parameters used in the equation. The 3PL is as presented above. The two-parameter logistic model assumes that c_i = 0.0, meaning that there is no guessing involved in the response to the item. Such a model is appropriate for measuring personality or other psychological traits when guessing is not applicable. The one-parameter model assumes that all items have equal discrimination, and uses only the b_i parameter to describe items. For achievement and ability type data, where examinee guessing is a near-certain possibility, the three-parameter model is the most appropriate of the dichotomous IRT models when sample size allows, hence its selection for this study.

The 3PL models the interaction between an examinee and a single item by describing the probability of correctly responding as a function of θ. The function that describes this relationship is an IRF. An example of a 3PL IRF is shown in Figure 1, with *a* = 0.99, *b* = 0.0, and *c* = 0.25. The probability of correctly responding is low for examinees with low θ_j, for instance -2.0, but does not drop below *c* = 0.25, as this is the probability of guessing the correct response on a four-option multiple-choice item. The probability increases with θ, but cannot surpass the upper asymptote of 1.0.

Figure 1: Three-Parameter Model IRF



While there might be four or five options presented to the examinee, the item is still scored only as incorrect or correct with dichotomous IRT, ignoring the possibility that information could be provided by responses to the incorrect options. Polytomous models make use of this possibility by modeling the response for each option, rather than simply the probability of responding correctly or incorrectly. Many polytomous models exist, and most assume a definite order of response options. Some of them are intended for Likert-type rating scales (e.g., Andrich, 1978), while some were designed to model achievement items with discrete steps, awarding partial credit for the steps (e.g., Masters, 1982; Muraki, 1992). These imply a strict ordering of response categories, they might not be appropriate for all multiple-choice data, in which case a model such as the nominal response model (Bock, 1972) is appropriate.

The generalized partial credit model (GPCM; Muraki, 1992), which allows discrimination values to vary, defines the probability of an observed response *X* out of the possible responses *x* as (Embretson & Reise, 2000, Eq. 5.8)

$$P_i(x = X \mid \theta) = \frac{\exp \sum_{j=0}^{X} a_i(\theta - b_{ij})}{\sum_{r=0}^{M} [\exp \sum_{j=0}^{r} a_i(\theta - b_{ij})]} . \qquad (2)$$

where

a_i is the item discrimination parameter of item *i*,
b_{ij} is the category boundary parameter for boundary *j* of item *i*,
r is the number of response categories, and
M is the number of (*r* -1) boundaries between the response categories.

Note that the number of boundaries between response categories *M* is always one less than the number of categories *r*.

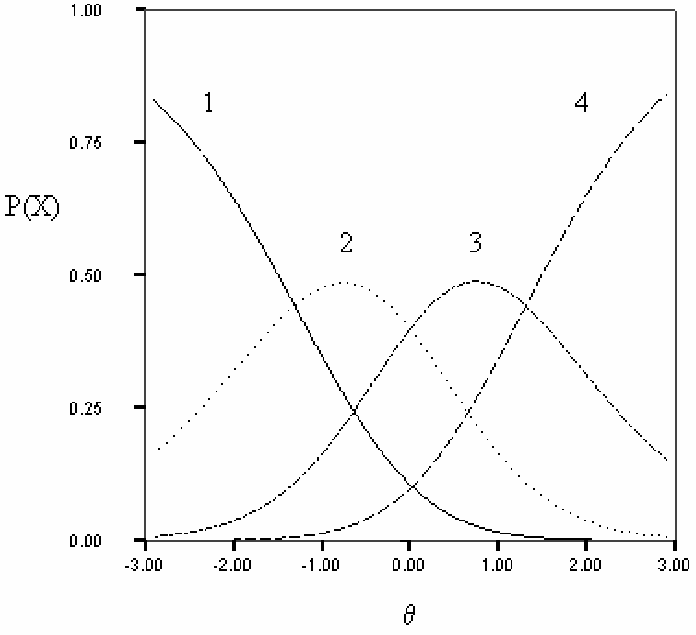
The GPCM is based on the assumption that the responses are ordered such that examinees with increasing ability will respond with an option higher in the order. For example, a math item might require several different calculations to be performed in a specific order, and the options presented to the examinee reflect the numbers that an examinee would arrive at after completing each step. This is advantageous because not only does it provide the opportunity for partial credit scoring, but also makes for highly plausible item distractors.

The GPCM is a divide-by-total model, where the probability of each category is calculated by dividing by a total summed value; obviously, the probabilities for each of the response categories conditional on θ must sum to 1. It can also be interpreted as an adjacent categories model. The boundary parameters b_{ij} represent the points on θ where the category response functions of adjacent categories cross, or where it becomes more probable that the response will be in the next category.

An example of a GPCM IRF is shown in Figure 2, with $a = 0.99$, $b_1 = -1.35$, $b_2 = 0.0$, $b_3 = 1.35$. The IRF models the probability of responding to one of the four options as a function of θ. Examinees with very low θ such as -3.00 or -2.00 are likely to respond to option 1, while examinees with very high θ such as 2.00 or 3.00 are likely to respond to option 4. The adjacent categories conceptualization is demonstrated by the first boundary; for θ below $b_1 = -1.35$ it is more likely that the response will be for option 1, while above it is more likely to be for option 2.

For instance, suppose the item was a mathematics item that involved the completion of several steps to obtain the correct final answer. Examinees of very low θ (up to -1.35) are not likely to be able to complete any steps, and are therefore likely to respond with a response (option 1) that reflects this stage. More able examinees (-1.35 to 0.0) are likely only able to complete the first step, and are therefore more likely to respond to the option (option 2) that contains the answer after that first step. Examinees with θ between 0.00 and 1.35 are likely to get past the second step, or select option 3, and examinees with θ above 1.35 are likely to obtain the correct answer (option 4).

Figure 2: GPCM CRFS for item with *a* = 0.99, b_1 = -1.35, b_2 = 0.0, b_3 = 1.35



As is evident from the previous paragraph, the advantage of the GPCM over the 3PL is that the item can differentiate among examinees of a wide range of θ levels. The ability of the item to differentiate peaks at the *b* values, such as the item in Figure 2 differentiating between examinees likely to choose option 1 vs. option 2 at b_1 = -1.35. There is a specific function that quantifies this differentiating ability, called the item information function, also known as Fisher information (FI). FI, the traditional conceptualization of item (and test) information in IRT, is broadly defined as the conditional slope squared divided by the conditional variance, or (Lord, 1980)

$$I_i(\theta) = \left[\frac{\partial P_i(\theta)}{\partial \theta} \right]^2 \Big/ P_i(\theta) Q_i(\theta). \qquad (3)$$

The item information function for the 3PL is specifically defined as (Embretson & Reise, 2000, Eq. 7 A.2)

$$I(\theta) = \left[a_i^{\,2} \, \frac{1 - P_i(\theta)}{P_i(\theta)} \right] \left[\frac{(P_i(\theta) - c_i)^2}{(1 - c_i)^2} \right] \qquad (4)$$

and the item information function for the GPCM is defined as (Ostini & Nering, 2005)

$$I(\theta) = \sum_{j=1}^{m} T_j^{\,2} P_{ij} - \left(\sum_{j=1}^{m} T_j P_{ij} \right)^2 \qquad (5)$$

where T_j is the scoring function (e.g., 1, 2, 3, 4).

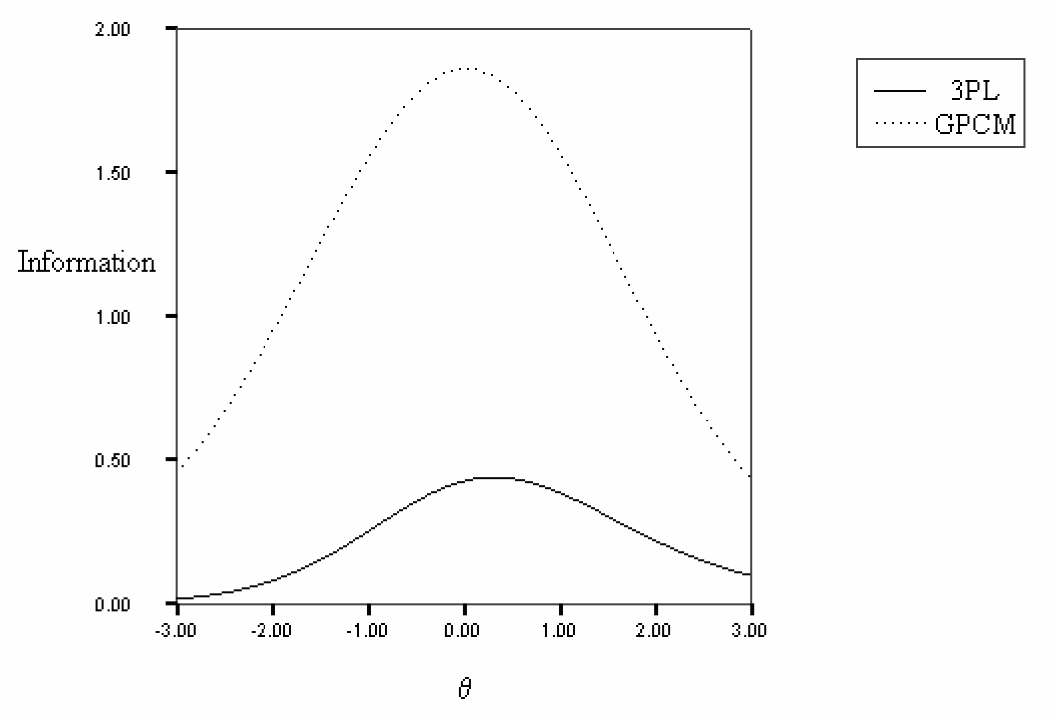
FI is an index of the amount of scoring precision contributed by an item or test at a given θ level. It is inversely proportional to the conditional standard error of measurement function (CSEM; Embretson & Reise, 2000),

$$CSEM = 1 \Big/ \sqrt{\sum_{i=1}^{n} I(\theta)} \qquad (6)$$

which is an index of the amount of error at a given θ level after *n* items. Because information is based on the function's slope, for a dichotomously scored item it peaks at b_i, where the slope is the greatest. For the same reason, items that have higher discrimination parameters will have more information at that point.

The IIFs for the two example items are shown in Figure 3, with a = 0.99, b = 0.0 for the 3PL, and a = 0.99, b_1 = -1.35, b_2 = 0.0, and b_3 = 1.35 for the GPCM. This figure demonstrates how much more information is offered by a polytomous model over a dichotomous model (Ostini & Nering, 2005). It is for this reason that the GPCM should offer substantial advantages for multiple cutscore CCTs.

Figure 3: IIFs for Dichotomous (3PL) and Polytomous (GPCM) models



The primary reason that item information is useful in the design of a CCT is that it provides a criterion by which to evaluate items and gauge to what extent they will help the CCT make a decision. The exact method by which this evaluation is performed is what distinguishes item selection methods. However, since item selection methods also depend on termination criteria to some degree, CCT termination criteria will be discussed first.

Termination Criteria

As discussed previously, there are three termination criteria commonly used with variable-length CCT: the SPRT, AMT, and BDT. Because not as much research has investigated BDT, it involves a greater amount of arbitrariness, and has not been used with intelligent item selection, it was not considered in this study. The SPRT and AMT have been more often used as termination criteria (Parshall, Spray, Kalohn, & Davey, 2006; Thompson, 2007a).

The Sequential Probability Ratio Test

The SPRT (Wald, 1947) is a simple test of the likelihood ratio between two competing hypotheses such as the mastery or nonmastery of an examinee. The SPRT was originally developed for use in quality control studies. The purpose was to sequentially sample single items from a lot of products (e.g., light bulbs) until the researcher was sufficiently confident that the lot would pass or fail quality control standards. Parameters had to be specified concerning the probability of drawing defective or working light bulbs, given that the lot was good or bad. This was structured as a simple hypothesis test:

$$H_0: \; p = p_0 \qquad (7)$$

$$H_1: \; p = p_1 \qquad (8)$$

where p is the proportion of defective light bulbs in the given lot, p_0 is the proportion of defective light bulbs below which the quality level is considered acceptable, and conversely p_1 is the proportion of defective light bulbs above which the quality level is deemed unacceptable. The region from p_0 to p_1 is known as the “indifference region” (IR). It is so named because the bounds are chosen so that it does not matter which decision is made for true proportions within this region – the costs and consequences associated with misclassification error in either direction are minimal.

The two bounds of the IR are used to formulate a ratio of likelihoods, the likelihood of observing d defective light bulbs out of a bad lot with a total sample t, to the likelihood of observing d defective light bulbs out of a good lot:

$$LR = \frac{L(d \mid bad)}{L(d \mid good)} = \frac{\prod_{i=1}^{t} p_1^{(t-d)} q_1^{\,d}}{\prod_{i=1}^{t} p_0^{(t-d)} q_0^{\,d}} \, . \qquad (9)$$

This ratio is then compared to two decision points, *A* and *B*. The complete computations of *A* and *B* are very complex, so Wald (1947) suggested as approximations

$$\text{Lower decision point} = B = \beta \,/\, (1 - \alpha) \qquad (10)$$

$$\text{Upper decision point} = A = (1-\beta)/\alpha \qquad (11)$$

where α is the nominal probability of accepting H_1 when H_0 is true, and β is the probability of accepting H_0 when H_1 is true. If the ratio is less than or equal to *B*, H_0 is accepted with error rates approximately $\beta* = \beta/(1-\alpha)$. If the likelihood ratio is greater than or equal to *A*, H_1 is accepted with error rates approximately $\alpha* = \alpha/(1-\beta)$. If the likelihood ratio is somewhere between *A* and

B, another light bulb is sampled. Although Wald claimed these to be valid approximations, the original equations (Wald, 1947) might now be of practical use, given the speed of computers.

Of course, if *p*, the true proportion, is somewhere between p_0 and p_1, it is then difficult to statistically show it is below p_0 or above p_1. It might not be possible to make a decision either way with error rates equal to the given α and β. Therefore a truncation rule must be imposed to keep the procedure from infinitely sampling without ever making a decision. Normally, this is done by just choosing the closer hypothesis after sampling a certain number *t* of light bulbs (e.g., 50) which is assumed to be large enough that adding to the sample would add an insignificant amount of information, or at least small enough that it would not offset the cost of increasing the sample. For instance, suppose that a decision could usually be made after sampling 20 bulbs, with $p_0 = 0.10$ and $p_1 = 0.20$. If 50 bulbs had been sampled with 9 defective, then *p* would be 0.18. At this point, a decision would be made to simply choose the 0.20 hypothesis because it is closer than 0.10, even if the likelihood is not sufficiently high for the SPRT to make a decision. Another option is to truncate the procedure, but not make a classification.

The terms p_0 and p_1 were selected to correspond with the hypothesis testing notation of H_0 and H_1. However, the SPRT has been extended to multiple cutscores, which entails more than two probability values being used. The points can be numbered sequentially starting at 1. This numbering is done on the θ metric, as well as the proportion metric. Therefore, this notation will generally be used in the current study.

Ferguson (1969) applied this procedure to variable-length classification testing. Instead of taking light bulbs one at a time and making the decision to pass, reject, or continue testing the lot, test items could be selected randomly, one at a time, and a dichotomous decision made concerning the examinee based on the examinee’s responses. In this application, correctly answering an item is analogous to drawing a defective light bulb from the lot. The hypothesis H_1 is that the examinee is a master and knows the correct answer to *p* proportion of the items, with *p* equal to p_2, the minimal proportion of items that a true master should know. The hypothesis H_0 is that the examinee is a nonmaster and that *p* is equal to p_1, the maximum proportion of items that a true nonmaster will know. The likelihood ratio is expressed after *n* items:

$$LR = \frac{L(master \mid u)}{L(nonmaster \mid u)} = \frac{\prod_{i=1}^{n} {p_{2i}}^{X_i} {q_{2i}}^{1-X_i}}{\prod_{i=1}^{n} {p_{1i}}^{X_i} {q_{1i}}^{1-X_i}}. \qquad (12)$$

where

p_{2i} is the probability of a master correctly responding to item *i*,
q_{2i} is $(1 - p_{2i})$,
X_i is the observed response, and
u is the observed response vector.

The value for each item *i* can be the same, or allowed to vary from item to item. The situation is inverted from the original application because a lot with a low proportion of defectives is “good,” but an examinee with a high proportion of correct answers is “good.”

Ferguson (1969) was the first to apply the SPRT to a classification testing situation. Ferguson’s research investigated the development of a computer-assisted test for a program of individually prescribed instruction for young mathematics students. Modules of instruction on a topic were adaptively administered, with a mastery test administered after each module to determine the next step in the branching procedure. Rather than have a conventional test for each module, these sequential tests randomly drew items from a bank until the SPRT made a pass-fail decision. Although this procedure makes the questionable assumption that all the items are equivalent or parallel, the module branching procedure and the sequential tests at the end of each module together reduced testing time and number of items by about two-thirds, as compared to completing conventional tests on all the modules.

Epstein and Knerr (1977) performed real-data simulations of sequential tests with the SPRT on data from military performance testing, finding that non-intelligent sequential testing reduced average test length (ATL) by about two thirds. Real-data (or post hoc) simulation is a research methodology that uses responses of real examinees to real items to simulate variable-length tests, as opposed to monte carlo simulation, which uses completely artificial (model-generated) data. Similar to Epstein and Knerr (1977), Kingsbury and Weiss (1983) made the comparison of the SPRT to conventional tests in a parameter recovery real-data simulation with three conditions of maximum test length: 10, 25, and 50 items. The random SPRT procedure used provided a mean reduction in ATL of 12% for the 10-item case, 48% for 25 items, and 69% for 50 items. Phi correlations between conventional test classifications and observed variable-length test classifications with the SPRT method were comparable with conventional tests; with a 50-item maximum, the correlation was 0.867 with the SPRT and 0.875 with a conventional test. However, the accuracy of the random SPRT method in the 50-item maximum condition decreased to 0.571 when IRT item parameters varied between items, demonstrating the relative inefficiency of the procedure when items are not equivalent but equivalence is assumed. Frick (1989) also did a real-data simulation, finding that the sequential tests agreed with the conventional tests 98% of the time, while using only one-fifth as many items. The differences between the results of these three studies of randomly sequential tests might possibly be attributed to variance in item characteristics and practical constraints, but it is evident that CCTs with random item selection and the SPRT criterion are quite efficient.

SPRT Parameter Specifications

These findings suggest that even when items are assumed to be parallel, sequential testing with the SPRT is considerably more efficient than conventional tests. However, in most cases it is not appropriate to assume that items are more or less equivalent. Furthermore, more advanced methods of specifying the parameters might give values of p_1 and p_2 farther apart for each item, causing the procedure to converge to a decision at a faster rate. For instance, suppose the cutscore score is 60%, with $p_1 = .50$ and $p_2 = .70$. In addition to allowing these two parameters to vary between items, other methods might estimate $p_1 = .40$ and $p_2 = .80$ for a given item simply by making use of the data from a pilot or calibration study to get a better estimate, as discussed below. Items such as these will cause the likelihood ratio to diverge more quickly. Therefore, most of the research in the last two decades has used more sophisticated specifications of the p_1 and p_2 parameters.

There are two alternative methods for specifying p_0 and p_1 for each item, one using the IRF from IRT, and the other utilizing classical test theory difficulty parameters for classified subgroups of masters and nonmasters in a standardization sample. Linn, Rock, and Cleary (1972) were the first to explore this idea of subgroup classical difficulty parameters with a real-data simulation using random item selection. Linn et al. used half of their sample to calibrate classical difficulty parameters for two subgroups on each item. Instead of p_1 and p_2 being the same for each item, they were now the proportion of the “nonmasters” (lower half of the calibration group) and “masters” (top half) who correctly answered each item. Their results showed that sequential testing with the approach would reduce the number of items required for making a mastery decision by about fifty percent. Unfortunately, no information was provided regarding the distribution of either the items or the examinees, so no conclusion can be drawn concerning whether the non-intelligent item selection impeded the efficiency. If the item characteristics did not vary to a great extent, the selection procedure would have had little impact.

Frick (1989; 1990; 1992) and Rudner (2002) also used a simple two-subgroup design of Linn et al. (1972) to estimate the SPRT parameters. However, Frick extended the SPRT to an adaptive testing procedure conceptually similar to that of Reckase (1983), but based it on item statistics from classical test theory rather than IRT. The item selection criterion was an item utility index that employed the difference between p_1 and p_2, as estimated by empirically sampling true masters and nonmasters for each item.

Weitzman (1982a, b) followed Linn et al. and estimated the parameters by taking the proportion of examinees in each subgroup that correctly answered each item. However, these studies were different in that they used four subgroups rather than two for the calculation of the statistic, even though the final classification was only pass/fail. The likelihood ratio statistic used was

$$L_n = \frac{\left[(K - K^* + 1)^{-1} \sum_{k=K^*}^{K} \prod_{i=1}^{n} P_{ik}^{x_i} (1 - P_{ik})^{1-x_i} \right]}{\left[(K^* + 1)^{-1} \sum_{k=1}^{K^*-1} \prod_{i=1}^{n} P_{ik}^{x_i} (1 - P_{ik})^{1-x_i} \right]} \qquad (13)$$

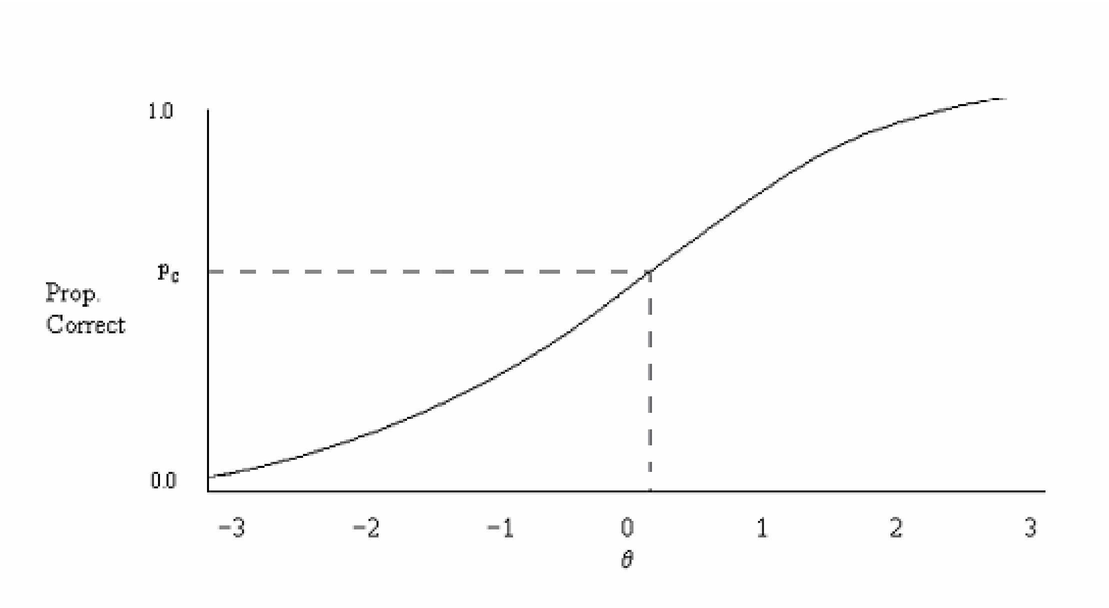
where P_{ik} is the proportion of examinees in the standardization sample within quantile k who correctly responded to item i and K^* is the quantile group directly above the cutscore point. The term X_i is the observed item response: 1 if correct, 0 if incorrect. This equation simply represents an expansion of the SPRT to having more than one parameterization subgroup above or below the cutscore, while still using a single cutscore. The numerator is the mean likelihood of the subgroups above the cutscore, while the denominator is the mean likelihood of the subgroups below the cutscore. As will be seen later, the SPRT can also be expanded from multiple subgroups for parameterization to having more than one cutscore, providing classification into multiple ordered categories.

The most important aspect of this research was that Weitzman (1982b) employed three methods of intelligently sequential item selection. Two of these selected items based on classical discrimination indices, while the third chose the next item by determining which remaining item was most unrelated to those already administered. These will be discussed below in the section on item selection strategies. This was the first attempt at nonrandom item selection with the SPRT. Unfortunately, Weitzman did not capitalize on the most important observation of this research--that the likelihood ratio will diverge to a decision quickly when the items maximize the difference between the probabilities used by the likelihood ratio. Later research on item selection (e.g., Frick, 1992; Lin & Spray, 2000; Eggen, 1999) addressed this.

This subdivision into quantiles was done in an attempt to eliminate the IR, since Weitzman suggested that one reason Linn et al. (1972) had erratic error rates was that the SPRT was only meant to classify examinees outside the IR, but the examinees inside the IR were used in the calculation of observed error rates. A real-data simulation (Weitzman, 1982a) was conducted and led to the conclusion that the differences between observed and nominal error rates were small enough to be due mostly to sampling error. However, observed error was still greater than nominal error. One possibility for this was that only four quantiles were used, providing little more specificity than two groups. Reckase's (1983) approach applies a continuum by utilizing the IRF, which much more accurately reflects the relationship between ability and performance on items than quantiles, though Reckase's method still oversimplifies the assumed model by examining only two points on the θ continuum.

Reckase (1983) proposed the currently most sophisticated method of specifying SPRT parameters using IRFs. He suggested that the IR should be conceptualized on the latent trait (θ) metric rather than the proportion-correct metric. A cutscore must be chosen on the θ metric, or calculated by converting the proportion-correct cutscore p_c to its equivalent cutscore θ_c using the test response function. The test response function (TRF) models the expected proportion of items in a test that an examinee would respond to correctly as a function of θ. This is analogous to averaging the IRF for each item in the test. To convert p_c to θ_c, suppose $\theta = 0.1$ corresponds to a proportion correct of 0.60 with the TRF, as shown in Figure 4. If the proportion of correct items required to pass the test in conventional test is 0.60, then $\theta_c = 0.1$.

Figure 4: Transforming p_c to θ_c

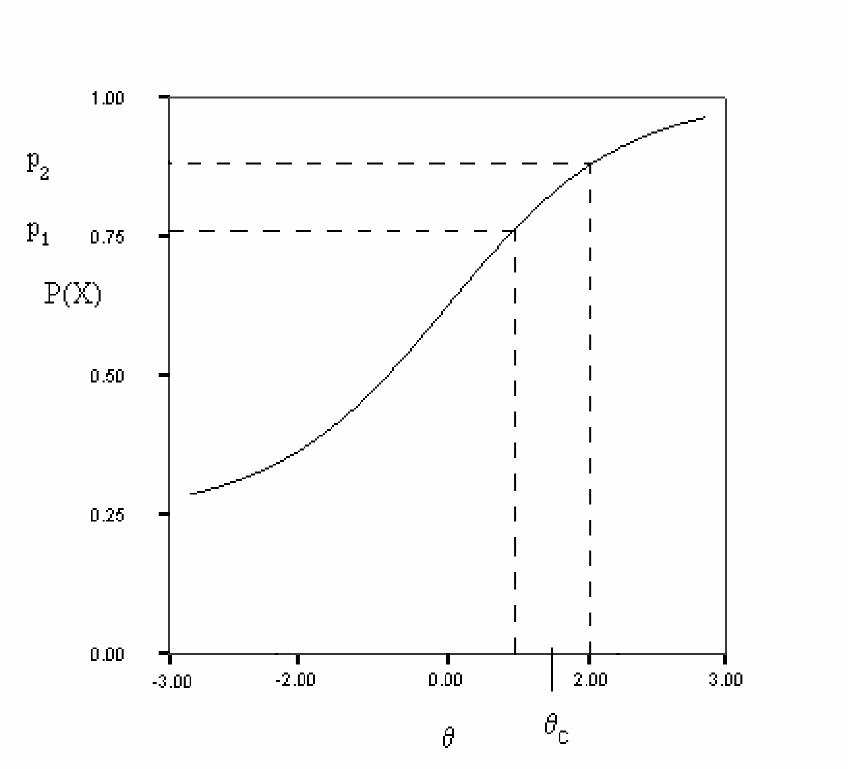


After a cutscore point θ_c is specified, two values are chosen above (θ_2) and below (θ_1) the cutscore to delineate the indifference region. These are often equal to θ_c plus or minus some arbitrary constant δ, such as 0.5, though it is not necessary for the region to be symmetrical. Some confusion exists in the literature as to the notation of the IR and the meaning of δ, as some researchers consider it the symbol for the cutscore. For example, Lau and Wang (1999) use the two symbols θ_c and δ interchangeably, rather than defining a single symbol, while what should be δ is given no symbol at all.

The size of the IR does matter, however, as the results in Reckase (1983) found that a larger IR, as compared to a smaller one, requires fewer items and commits errors closer to the nominal rates. This is due to the fact that a greater disparity between points on the θ metric transforms to greater disparity between points on the proportion-correct metric, as observed by Weitzman (1982b). Greater differences between p_1 and p_2 cause the likelihood ratio to diverge quickly. Therefore, the value of δ should not be chosen completely arbitrarily. Yet relatively so little attention has been paid to the size of the IR that some researchers have actually introduced a confound by setting different IR widths for different methods (Jiao, Wang, & Lau, 2004), which can severely and adversely affect ATL for those methods with smaller IRs.

The values for the p_1 and p_2 parameters are the probability of an examinee with a true θ level equal to θ_1 or θ_2 correctly answering the item, which corresponds to the probability of a minimum true master or a maximum true nonmaster. These are computed by means of the IRF, as is illustrated in Figure 5.

Figure 5: Transforming θ_0 and θ_1 to p_1 and p_2 with a dichotomous item and a single cutscore



In this example, the cutscore is 1.5 and δ is 0.5, so $\theta_1 = 1.0$ and $\theta_2 = 2.0$. The probability of a correct response is calculated at these two points on θ, with the resulting probabilities in this example $p_1 = 0.75$ and $p_2 = 0.875$. These two probabilities are utilized by the SPRT, and the SPRT makes a decision more quickly if there is a greater disparity between the two. The effect of increasing IR width is obvious in the example; inputting two more disparate points on θ will always produce two more disparate points on *P(X).*

This use of IRT provides more accurate specification of the parameters, allows direct application of large IRT-calibrated item banks, and indirectly enables the item selection method to utilize IRT. Reckase (1983) selected the next item by choosing the most informative item at the current examinee θ estimate, and demonstrated that this extension of the SPRT worked efficiently by means of a monte carlo simulation where the examinees were generated but the item parameters came from a real pool.

In conclusion, three methods are available for the specification of the p_1 and p_2 parameters. The original SPRT testing method defines these as the proportion of items that should be answered correctly by a maximally incompetent and minimally competent examinee, respectively. This assumes that items are equivalent. Alternatively, the researcher could take a calibration sample of defined masters and nonmasters, and then calculate the classical difficulty parameter separately for the two subgroups. Lastly, the indifference region can be originally defined on the θ, or latent trait, metric, and then converted to the proportion metric using the item response function. The last of these methods, possibly because of the predominance of IRT in recent measurement research, has superseded the others in use. Unless explicitly stated otherwise, all the SPRT research reviewed in the present paper makes use of Reckase's IRT-based method for specifying p_1 and p_2.

Robustness of the SPRT

Using IRT parameters for the SPRT requires relatively large samples (250 examinees or more) and can result in parameter estimates that vary due to the parameter estimation procedures and competing IRT models that might or might not be appropriate for a given set of data. Reckase (1983) was the first to investigate the effect of possible IRT-related errors on classifications with the SPRT. One of the conditions in that study was to determine simulee responses under the 3PL model, but to use the 1PL for the simulated testing procedure. This was to determine if oversimplification of the procedure and assumed model could adversely affect the performance of Reckase's IRT-based adaptive test with the SPRT. He concluded that using an incorrect IRT model did matter, because the number of items required to make a decision was roughly twice as large as the case in which the testing procedure and examinee responses were both modeled by the. Moreover, the specification of an incorrect model lowered the cutscore point by 1.5 θ units, making the test much easier to pass. Additionally, even when the same model was used to simulate and score the SPRT, the 3PL performed better than the 1PL, suggesting that selection of a model should be justified. Kalohn and Spray (1999) also used the 3PL to simulate responses in a monte carlo study but the 1PL to make classifications with the SPRT, finding similar results to Reckase (1983).

Jiao and Lau (2003) investigated the same effect, but expanded the scope. Instead of only assuming the 3PL as the true and the 1PL as the misspecified model, both of these along with the 2PL were crossed to obtain all comparisons. The research was divided into three studies, one with each of the three models defined as the true model. The effect of misspecifying the remaining two as the correct model was then simulated, along with two levels of item exposure constraints and two levels of length constraints. The item exposure constraint used was referred to as stratum depth, where items are selected from a sample of the most informative items rather than the single most informative item. For example, if the stratum depth is five, then the next item is randomly selected from the five most informative items with respect to the item selection criteria.

The only case with observed error rates above nominal levels was when the 3PL was the true model, supporting the findings of Kalohn and Spray (1999). Error rates were below nominal levels when the true 3PL model was used, but when this was relaxed to a 2PL, Type I errors shrank disproportionately while Type II error rates rose above nominal levels. When the 1PL was imposed, Type II errors increased to over three times the nominal values. Error rates were unaffected by the four study conditions, which only increased ATL. This research (Reckase, 1983; Kalohn & Spray, 1999; Jiao & Lau, 2003) supports the use of 3PL and cautions against the use of the 1PL unless it is strongly justified empirically.

One other choice to be made when IRT is used, in addition to the number of parameters, is whether to use a unidimensional or multidimensional model. Employing a unidimensional model when a multidimensional model is appropriate oversimplifies the testing procedure and could have detrimental effects on its efficiency or accuracy. Spray, Abdel-fattah, Huang, and Lau (1997) explored how to formulate a multidimensional SPRT to examine this issue. The problem with this endeavor was how to specify p_0 and p_1. With an IRF on a single dimension, this is not difficult (e.g., Reckase, 1983). But a two-dimensional IRF imposes the question of whether the parameters should be specified along one of the axes of the dimensions or one of the infinite number of lines that pass through the origin with an angle between the axes. Using one of these lines through the origin for SPRT purposes imposes a new cutscore line, perpendicular to the vector and intersecting it at whatever point on the vector is chosen for the cutscore (Spray et al., 1997). The procedure is then biased against examinees with vectors quite different than the one chosen to calculate the SPRT parameters, and the cutscore point/line changes with the vector that is selected.

Given these problems, if a unidimensional SPRT is used to approximate two-dimensional data and still performs efficiently, the procedure is then quite robust. To investigate this, Spray et al.(1997) generated item responses with a two-dimensional IRT model, then calibrated the data and simulated CCTs with a unidimensional IRT model. The mean overall (α and β) misclassification rates ranged from 0.042 to 0.056, which is near the nominal rates employed ($\alpha = \beta = 0.05$). Introducing a length constraint (minimum = 60, maximum = 120 items) and an exposure constraint (a random selection stratum of 10 items) reduced this already small difference to error rates of 0.041, 0.039, and 0.039. Similarly, error rates for two levels of correlation between the latent dimensions (0.0 and 0.5) differed by only 0.010, on average. An even smaller difference was found between conditions that attempted to manipulate dimensionality by multiplying discrimination parameters on one dimension to give it more weight.

The authors suggested that these results imply that the issue of dimensionality becomes less important when test constraints are imposed, which is often the case in real-world applications. Since observed error rates were consistently below nominal values, it appears that it might be safe to use a unidimensional CCT with the SPRT even when the true latent space is two-dimensional. However, note that ATL was not evaluated.

In the same line of research, Lau (1998) applied the same research method with a slightly different formulation of the independent variables. The specification of the cutscore was the same, with intended passing rates of 0.4, 0.6, and 0.8. The test length constraint was shifted from a minimum of 60 and a maximum of 120 to a minimum of 15 and a maximum of 50. Instead of having only two levels of correlation between the underlying dimensions, there were five: 0.0, 0.3, 0.6, 0.9, and 1.0. As the multiplication of discrimination parameters had little effect in Spray et al. (1997), this was replaced with a unidimensional model as an independent variable. The multidimensional data were calibrated and simulated with both the 1PL unidimensional model and the 3PL unidimensional model in an effort to determine if the robustness of the unidimensionality violation was moderated by the unidimensional model.

Lau found that, consistent with earlier work, the error rates were indeed robust to violation of unidimensionality; decreasing the correlation between the dimensions merely required more items to make a decision. Similarly, both unidimensional models showed similar patterns of results and error rates across other conditions, though the 1PL required about twice as many items to make decisions. The inclusion of a test length constraint did not affect error rates, but merely increased ATL.

Of course, even if the correct model is chosen, there is a possibility that the item calibration procedure did not accurately estimate item parameters. This could adversely affect the errors for the SPRT through inaccurate specification of p_1 and p_2. Spray and Reckase (1987) conducted a monte carlo simulation study in which item banks were created with known parameters, responses were simulated using a data set of 2,500 simulated examinees, and item parameters estimated from the responses. When averaged across all the other conditions, Type I errors decreased from 0.036 to 0.033 when parameter estimation error was present, while Type II errors only decreased from 0.032 to 0.031. The ATL rose from 17.6 to 18.7, suggesting that item parameter estimation error in the testing procedure might simply require a few more items, but not cause error rates to diverge from the nominal rates set *a priori*.

CCTs with an SPRT termination criterion can perform efficiently even when the parameters being used are not actually estimated. Huang, Kalohn, Lin, and Spray (2000) investigated the situation where a large item pool is available, but only a small subset has been calibrated with IRT while the majority of the items have CCT statistics. CCTs with classical parameters transformed to IRT performed just as accurately as CCTs with the known IRT parameters that were generated for the simulation, while maintaining the same ATL, demonstrating that the efficiency of SPRT-based CCTs is quite robust.

In summary, when computerized classification testing with the SPRT is based on IRT item parameter estimates, it seems to be rather robust with respect to the assumptions of the underlying IRT models and parameters. Parameter estimation with a large *N* affects performance very little, as does using the 2PL when a 3PL is appropriate. The parameters need not even be directly estimated (Huang, Kalohn, Lin, & Spray, 2000). Multidimensionality also had little effect when examined, but a much more extensive analysis of this topic is necessary. In fact, CCT and multidimensionality is a topic that has received little attention in general, especially while comparing methods of addressing other issues such as item exposure or termination criteria. For example, the numerous problems encountered by Spray, Abdel-fattah, Huang, and Lau (1997) in constructing a multidimensional SPRT procedure might not be an issue with the AMT approach. One misspecification that does have detrimental effects on test efficiency is the specification of the 1PL when the 3PL is appropriate. Considering the proportion of classification testing applications that use a multiple-choice format that enables guessing, especially achievement mastery testing, using 1PL calibrations with the SPRT is not always appropriate.

The SPRT for Multiple Cutscores

Armitage (1950) and Sobel and Wald (1949) independently developed multiple-cutscore extensions of the original SPRT. The developments were purely mathematical, and were not applied to the psychological and educational measurement domain until much later. The bulk of multiple-cutscore SPRT research has focused on the basic questions of parameter specification and item selection. Each study used only one of the two procedures (Sobel & Wald, 1949; Armitage, 1950) with the exception of Jiao, Wang, and Lau (2004), which compared them.

The SPRT was first extended to three categories (Sobel & Wald, 1949), in which case there are two cutscores. Let θ_1 represent the maximally competent examinee in the lowest category, θ_2 represent the midpoint between the two cutscore points, and θ_3 represent the minimally competent examinee in the highest category. The procedure then simultaneously tests two sets of hypotheses: θ_1 vs. θ_2, and θ_2 vs. θ_3. Armitage (1950) suggested another procedure that uses three simultaneous SPRTs to decide between three hypotheses, one for each pair (θ_1 vs. θ_2, θ_2 vs. $\theta_{3,}$ θ_1 vs. θ_3), and defines the evaluative points differently, so that the point above the lower cutscore and the point below the upper cutscore need not be equivalently spaced. The two are then termed θ_2 and θ_3, while the point above the upped cutscore is θ_4. The procedure then tests θ_1 vs. θ_2, θ_3 vs. $\theta_{4,}$ and θ_1 vs. θ_4.

Spray (1993) supported the use of the Armitage procedure as a method for multiple cutscores because it can be extended to *k* categories, as needed by the test user. However, the introduction of more cutscores strains the procedure, causing an increase in ATL and classification error. Spray (1993) noted that because more items are usually required when the examinee's ability is near the decision point, adding more decision points will increase the average test length across the sample.

Spray (1993) was the first study to examine the multiple-cutscore SPRT, simply to determine if it performed within nominal boundaries. Spray demonstrated with a monte carlo simulation study that a CCT with two, three, or four cutscores using Reckase's (1983) SPRT parameter specifications accurately recovered examinee classifications. Error rates stayed low across the ability scale, only peaking at the cutscore points. They still stayed near nominal levels, however, and the stress of two, three, or four cutscores was more prominent in terms of ATL than in misclassification. With four cutscores (five examinee classifications) and a maximum test length of 10, ATL was nearly 10 for examinees with θ near the cutscores. This means that a decision was rarely made using the SPRT alone; the maximum number of items would be reached, and the nearest classification taken. Admittedly, a maximum test length of 10 is quite restrictive, but this effect followed, to a lesser extent, with maximums of 20 and 50 items.

After the feasibility of a multiple-cutscore SPRT was established, Eggen (1999) investigated the application of alternative item selection strategies. However, Eggen chose to use Sobel and Wald's (1949) three-category SPRT. Eggen (1999) was primarily research on item selection, but half of the study involved the three-category case. Because of this focus, no fewer than nine selection strategies were investigated, including six formulations of Kullback-Liebler information (KLI). Eggen also found the three-classification SPRT to work efficiently, with little difference in efficiency between item selection methods that evaluate information with respect to the cutscore. Eggen later compared the three-classification SPRT with three-classification AMT (Eggen & Straetmans, 2000).

Rudner (2002) also investigated item selection in multiple-cutscore classification. Rudner used the *k*-category procedure developed by Armitage (1950) and applied to CCT by Spray (1993). The item pool of 139 items was applied from a state educational assessment test for assigning students to one of four categories. A calibration sample of 1,000 examinees was used to provide parameters for both IRT and Bayesian decision theory. Cut scores were fixed on the θ scale at -0.23, 0.97, and 1.65. Next, trial samples of 10,000 drawn from both a N(1,0) and U(-2.5,2.5) distribution had classification tests simulated under four conditions described in the Item Selection Methods section below. Maximum test length constraints ranged from 5 to 30 items.

Three of the item selection conditions specified the SPRT parameters with CTT methods and utilized cutscore-based item selection, which is appropriate for the SPRT. The fourth condition specified the SPRT parameters with IRT, but utilized estimate-based item selection, which is not appropriate for the SPRT. The three cutscore-based methods were relative entropy, maximum discrimination, and minimum expected cost. Relative entropy is a variant of KLI. Rudner (2002) quantified maximum discrimination as a function of the difference between p_1 and p_2,

$$M_i = \left| \log \frac{P(z_i = 1 \mid m_k)}{P(z_i = 1 \mid m_{k+1})} \right| \tag{14}$$

where $z_i = 1$ is a correct response to item *i*, and the upper mastery group being considered is m_{k+1}. Minimum expected cost is a Bayesian criterion where the test developer is able to define arbitrary cost values for misclassifications and the administration of another item. For instance, if c_{21} is the cost of making a classification decision in Group 2 (d_2) when the examinee is actually in mastery Group 1 (m_1), and c_{12} is vice versa, then the expected cost is

$$B = c_{21} P(d_2 \mid m_1) P(m_1) + c_{12} P(d_1 \mid m_2) P(m_2) \tag{15}$$

Items are selected to minimize expected cost after administration,

$$EC = B(X = 1)P(X = 1) + B(X = 0)P(X = 0) \tag{16}$$

where the probability of each response is multiplied by the expected cost *B* if the examinee were to respond with the associated response.

This study provided a comparison of methods under relatively harsh conditions. Not only were there three cutscore points to put stress on the efficiency of the method, but the fourth category was reserved for only a small percentage of very elite students (θ > 1.65), and the item pool contained only 139 items. Additionally, it is very difficult to make accurate classifications with a maximum test length of only 5 items.

Rudner (2002) found that KLI and minimum expected cost had more accurate classifications into the four categories than maximum discrimination and the IRT formulation. However, the difference in proportion of correct classifications between the IRT formulation and KLI or minimum expected cost ranged only from .004 to .023. The greatest differences were for the 5-item maximum condition, where the proportion correctly classified was far below nominal levels (0.630 to 0.836). Moreover, these results were incomplete, as no data were given on ATL. It is possible that the IRT method used fewer items than the others. Therefore, the only conclusion that can be drawn from this research regarding multiple-cutscore CCT is that it is feasible as long as maximum test length is not too short.

Weissman (2004) also compared the efficiency of item selection methods in a three-cutscore classification test. A pool of 367 items calibrated with the 3PL was employed, with the peaked bank information function reaching its maximum at $\theta = 1.0$. The three cutscores defining the four categories were -0.3, 1.0, and 2, and were determined by using the mode and inflection points of the bank information function. The SPRT was Armitage's (1950) combination, with a fixed IR width at a single value (0.6), but it was unlike other studies in that it was specifically chosen. The justification for this width was that this was approximately twice the standard error of measurement for a test with a classical reliability of 0.91.

Weissman's (2004) classification procedure performed efficiently. The percentage of correct decisions (PCC) reached its minimum at about 92%, both within and across the four categories, with maximum test length of 40 items. However, the results of this study were unlike past research because of the exclusion of examinees for which the SPRT could not make a decision within the short number of items allotted. Since this was the case, PCC was actually *higher* when the maximum test length was only 10 items, and few decisions were being made at all, than with a maximum of 40 items. All of those who reached the maximum were excluded from the results, resulting in an unrealistic portrayal of the efficiency in the results.

All of the multiple cutscore studies discussed up to this point utilized Sobel and Wald's (1949) or Armitage's (1950) method. Jiao, Wang, & Lau (2004) is the only attempt to date to compare the two multiple cutscore SPRT procedures with psychometric simulations, though Govindarajulu (1987) claimed that the two perform comparably from a statistical perspective. If the IRF is nondecreasing and the θ evaluative points are ordered, the comparison of θ_1 and θ_4 is superfluous. Unfortunately, the design of the Jiao et al. study was hampered by two confounds, disabling a direct comparison. First, the three secondary independent variables of maximum test length, item exposure control, and item selection were not completely crossed, but were instead nested. Instead of crossing the 2 x 2 x 2 design to form 8 cells, there were only four cells. Test length and item exposure were confounded so that a test length constraint was imposed only when there was an exposure constraint. This is partially due to the fact that the imposition of item exposure constraints restricted maximum test length to 60 by dividing the 300-item pool into 60 strata of 5 items, but this was unrelated to the minimum test length constraint. Secondly, and much more importantly, the two methods were fixed with different IR widths, transforming it to a study on IR width and ATL, as IR width has such a strong effect on results (Reckase, 1983, Eggen, 1999). Moreover, only two item selection methods were examined: random and FI at the midpoint between the two cutscores, neither of which are the most efficient. Item exposure also had only two levels. These were "no constraints" and selecting items randomly from the items with the five highest levels of information.

The difference in IR width was due to a difference in IR formulation between the two methods. Armitage's method defines $\delta = IR/2 = (\theta_{c1} - \theta_{c2})/2$, where θ_{c1} is the lower cutscore and θ_{c2} is the upper cutscore. The total width 2δ of the IR around each respective cutscore is then fixed at the distance between the decision points, with the upper limit of the lower IR being equal to the lower limit of the upper IR. Sobel and Wald's (1949) method, on the other hand, allows δ

to be chosen. Since the IRs for each cutscore cannot overlap with each other, the maximum value that can be chosen is the value fixed by Armitage's method, where the two indifference regions share a boundary halfway between the decision points. The minimum value that can be chosen is 0, and the experimenter is free to choose any value between these. Since the cutscores were chosen to be -0.5 and 0.5, Armitage's method defines $\delta = 0.5$, with the shared IR boundary at 0.0. Instead of choosing $\delta = 0.5$ for the Sobel and Wald method to build a comparison on equal ground, the authors chose $\delta = 0.1$. When the IR width differs by a factor of 5, this is certain to skew the results, as Reckase (1983) found that, in the simple two-category case, lowering δ from 0.8 to 0.3 increased ATL from 4 to 16. Given that the addition of another cutscore point adds stress onto the decision-making procedure, this difference could be greater in the three-category case.

The results of the study demonstrated this effect. Out of the four conditions, the two with test constraints had nearly equal error rates, as constraints tend to decrease any differences between methods. However, with no constraints, the Sobel and Wald (1949) procedure had half the error of the Armitage (1950) procedure because the Sobel and Wald procedure uses large numbers of items, which the SPRT needs to do to make a decision when the IR width is specified as very small (Reckase, 1983). The Armitage procedure, with its large IR width, made decisions very quickly, and used only a small fraction of the items taken by the Sobel and Wald procedure, but was lacking in accuracy. When the maximum test length constraint was imposed (the minimum test length did not impact this study), this kept the Sobel and Wald procedure from using the number of items that it required, which inflated the observed error rate to that found with the Armitage procedure.

Item selection methods use by Jiao et al. were also very primitive. It has been known for some time in the relevant literature that random item selection does not perform relatively efficiently and some form of intelligent selection is better (e.g., Reckase, 1983; Kingsbury & Weiss, 1983). Random selection should be present as a baseline for comparisons between other methods, but this study used only one intelligent method rather than several and therefore had no need of a baseline. Moreover, if FI is to be used, a better method is needed than simply maximizing FI at the midpoint o the two cutscores. This does not even maximize information in a relevant region of θ. FI should be maximized at the current ability estimate or at the nearest cutscore point (Eggen, 1999; Eggen & Straetmans, 2000). Jiao et al. also employed item exposure controls with random item selection, which is entirely unnecessary.

Moreover, the comparison between the Sobel and Wald (1949) method and the Armitage (1950) method can be made conceptually. If the IR width is set to be the same for both methods, they differ only in the fact that the Armitage method calculates an additional hypothesis test (θ_1 vs. θ_3). Because the other two hypothesis tests (θ_1 vs. θ_2, θ_2 vs. θ_3) are subsumed in the third, the likelihood ratio will always be larger for the third. Therefore, if one of the two hypothesis tests around the two cutscores is able to make a decision, the third hypothesis test is guaranteed to already have made a decision, and therefore does not come into play. The test is only terminated when the two hypothesis tests around the cutscores have significant likelihood ratios.

The primary target of research on the three-classification SPRT has been the effect that additional cutscores have on the results, and how this effect can be addressed. There are two factors that might increase the efficiency of a test where accuracy is reduced below nominal levels by the addition of more cutscores. First, more information is needed across the θ distribution. Since the SPRT requires information only at the decision point (Spray & Reckase, 1994), a test with a single cutscore requires information only at that point. A test with three cutscores, with the target of four categories, requires sufficient information to make decisions at all three cutscores, tripling the requirements of the item bank. Another factor that would help is more information in each item, across a greater range. If each item provides a moderate amount of information at more than one cutscore, the required increase in the size of the item bank will not be as large. Polytomous IRT models address both of these issues.

The SPRT with Polytomous IRT

Another extension of the SPRT is to CCT with polytomously scored items. Lau and Wang (1998; 1999; 2000) suggested that instead of calculating the p_1 and p_2 parameters with a dichotomous IRT model, this could be done with a polytomous model. In a series of monte carlo simulation studies, they investigated the performance of a CCT that employed the GPCM under varying practical conditions, such as item exposure and test length constraints. It was concluded that a polytomous SPRT performed efficiently and that CCTs could be conducted using mixed item types.

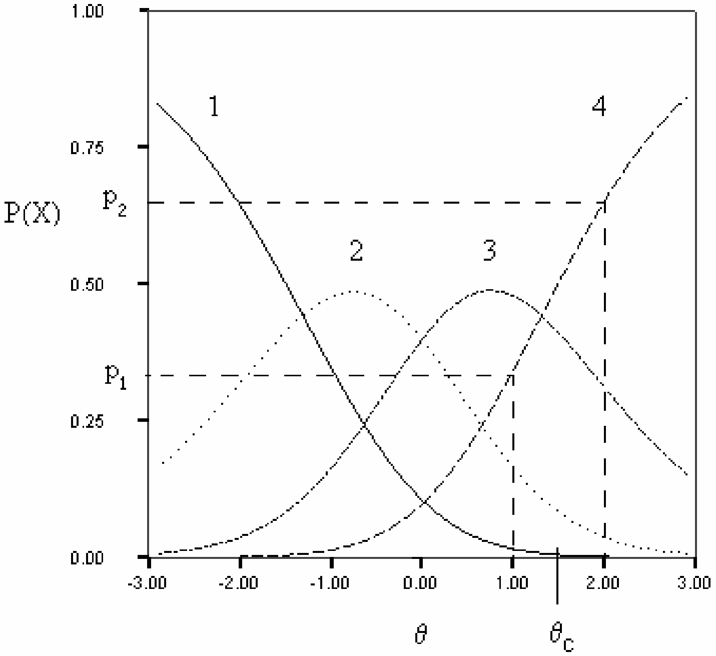
CCT frequently involves multiple-choice achievement items, which can sometimes be modeled more accurately by polytomous IRT models. Dichotomous models score each item as merely correct or incorrect, with the incorrect options assumed to have equal levels of incorrectness. Yet many items are constructed so that the distractor options reflect partial completion of the task or full completion with small errors. In this case, a partial-credit ordering of the responses might fit the data better, theoretically and empirically. Two models that do this are the partial credit model (Masters, 1982) and its generalization, the GPCM (Muraki, 1992).

The important advantage of polytomous IRT, as discussed previously, is information. FI is a function of the slope of the IRF, and with dichotomous IRT that slope is maximized at the point where the item most efficiently discriminates between the possible responses, of which there are only two (correct/keyed and incorrect/non-keyed). With polytomous IRT, there are multiple responses, or categories. If the item has four possible responses, then there are three boundaries between them, with each of those boundaries contributing information for the item. Each item will provide more information when scored polytomously (Dodd, De Ayala, & Koch, 1995) as well as across a greater range. More information generally means greater testing efficiency.

The effect of the GPCM is illustrated in Figure 6. In this example, the same cutscore and IR are used as in Figure 5, and the item discrimination parameter is equal for both items. Note the much greater disparity in the $p_1 = 0.35$ and $p_2 = 0.65$ values, all else being equal. The SPRT will make a decision much more quickly with items such as this.

Moreover, this leads to a more specific advantage. A common practice in CCT is to select items with the greatest information at the cutscore point (Spray & Reckase, 1994). With dichotomously scored items, all other things being equal, items would need to have difficulty locations near the cutscore point. This requires the construction of an item bank with a peaked information function. A large number of items on the same topic with similar difficulty are needed, which might not be feasible. When scored polytomously, items with locations not at the cutscore point might still provide a relatively large amount of information at that point, as compared to the same item scored dichotomously. This allows the item bank to have a greater variance of item difficulty or location parameters. This fact will not only ease the construction of new banks, since less peakedness is required, but also enable the incorporation of existing item banks that are not peaked. Moreover, this might help deal with item exposure issues.

Figure 6: Transforming θ_0 and θ_1 to p_0 and p_1 with a polytomous item and a single cutscore



The adaptation of the SPRT to items scored with ordered polytomous IRT models (Lau & Wang, 1998; 1999; 2000) is quite simple. The same procedure of specifying p_1 and p_2 from θ_1 and θ_2 is utilized, but this is calculated using the category response functions of the polytomous item (Figure 6) rather than the single IRF of a dichotomously scored item (Figure 5). The category response function represents the probability of an examinee endorsing a given response.

Lau and Wang (1998) used 246 dichotomous items and 266 polytomous items from the 1996 National Assessment of Educational Progress. The polytomous items, with ordered partial credit categories, were calibrated with the GPCM (Muraki, 1992), while no information was given regarding the dichotomous calibration; nor were aspects of the item bank, examinee distribution, or SPRT specifications reported. But across conditions of test length and item exposure constraints, pools with polytomous items had mean observed error of 0.028 and ATL of 15.127. On the other hand, dichotomous-only pools had an observed error of 0.038 and an ATL of 22.509. Since virtually no information was provided regarding the design, conclusions are limited, but it was found that polytomous CCTs required fewer items.

Lau and Wang (1999) used the same data, but provided more detail. The mean item difficulty was 1.043, 10,000 examinees were generated from a N(0,1) distribution, and IR width was varied, with values of 0.5 and 1.0. Cutscore location was varied also, at -0.8 and 0.8, both below the mean item difficulty. Only the polytomous items from the previous study were used, and the main comparison was between item selection by FI and KLI, where the KLI boundaries were the same as the IR boundaries. Very little difference was found, supporting other research on the topic (Spray, 1993) and the proposition that the item with the highest FI at the cutscore will also be the item with the highest KLI around the cutscore. Lau and Wang (2000) investigated the procedure once more, incorporating mean item response time into an index for item selection, with little effect. The combined item pool from their first (1998) study was again utilized, and the comparison was between FI and KLI. As before, no difference was found for the latter comparison, so that only FI results were used for conclusion and discussion.

Summary

The SPRT performs even more efficiently when the items are polytomously scored. However, the use of polytomous IRT models might be even more valuable for multiple cutscore CCT. This procedure can easily be adapted to multiple cutscores, a methodology which has not yet been investigated. The *p* parameters can be evaluated at each of the θ values that represent the indifferences regions around the cutscores. This is done simultaneously for each set of hypotheses (θ_1 vs. θ_2, θ_2 vs. $\theta_{3,}$ θ_1 vs. θ_3). Once the *p* values have been calculated, they can be entered into the likelihood ratios for the SPRT. If items are being selected to maximize information at a single cutscore, a given item might provide information at that point or at a different cutscore, depending on the response of the examinee.

This characteristic of polytomous IRT should greatly increase the efficiency of multiple-cutscore SPRT. Suppose a test has three categories, with two cutscores. Let examinee A truly be in the lowest group, but suppose that the examinee correctly guesses the answer to the first few questions. The likelihood ratio of the upper cutscore would then be further from a decision, and a cutscore-based item selection would administer items with the greatest information at the upper cutscore (Eggen, 1999). If these are dichotomously scored, they will provide a large amount of information only at that cutscore. Polytomously scored items, on the other hand, might have a boundary between categories located near the lower cutscore, and therefore provide information at that location. Since the two sets of hypotheses are being simultaneously tested, the procedure will perform less efficiently across a population when items are selected only to maximize information at a single cutscore rather than both.

Adaptive Mastery Testing

The other primary termination criterion utilized in CCT is known as adaptive mastery testing (AMT), though it is easily extended beyond the two-classification mastery testing case to multiple cutscores. Weiss and Kingsbury (1984) suggested using confidence intervals on the θ metric to classify examinees into multiple categories. At each stage in the testing process, usually after each item, the examinee's θ level is estimated using the current set of responses. Then a 1 - α confidence interval is constructed around that estimate using either the CSEM from maximum likelihood estimation or the square root of the Bayesian posterior variance, if Bayesian estimation is used. This confidence interval can be expressed as

$$\hat{\theta} - z_\alpha (CSEM) < \theta < \hat{\theta} + z_\alpha (CSEM) \qquad (17)$$

where z_α is the normal deviate for a 1-α confidence interval, such as 1.96 for a 95% interval. In the two-category case for which it was first developed, if the confidence interval falls completely above the cutscore on the θ metric, the examinee can be classified as above the cutscore. If the confidence interval falls completely below the cutscore, the examinee is classified as below the cutscore. If the confidence interval contains the cutscore, another item is administered.

In the multiple-cutscore case, if this interval falls into a category on the θ metric without overlapping any of the cutscores, a classification can be made. If a cutscore is included in the interval, another item is administered. The introduction of polytomous IRT will not directly affect the calculation of Equation 17. But since the greater information of the polytomous model might lead to increased reductions in the CSEM, decisions might possibly be made with fewer items.

AMT has performed well as a termination criterion in comparison to other methods for dichotomous classification based on test data. The earliest research (Kingsbury & Weiss, 1983) that compared it directly to another method (the SPRT) did not support the sequential SPRT procedure used, but no strong conclusion could be drawn about the SPRT or AMT itself for three reasons. First, as Spray and Reckase (1996) pointed out, one of the two outcomes by which the methods were compared, either misclassification rates or average test length, was not held constant to provide a relatively even comparison on the other. Secondly, the research was designed to compare adaptive and randomly sequential item selection, not to perform an even comparison of the termination criteria of IRT-based AMT and the SPRT. Thirdly, the two methods compared are now outdated: the SPRT was used with random selection of parallel items, and the IRT method used confidence intervals based on the square root of the Bayesian posterior variance (Owen, 1975). Since then, several intelligent and much more efficient SPRT procedures have been developed (Reckase, 1983; Frick, 1992; Spray & Reckase, 1994; Rudner, 2002), both adaptive and sequential, and confidence intervals using standard errors from maximum likelihood estimation have become more favored than the Bayesian posterior variance method (Eggen & Straetmans, 2000) because Bayesian estimation artificially shrinks the posterior variance as compared to the CSEM.

AMT was not investigated further until Frick (1989; 1990; 1992) compared AMT to several CTT-based SPRT procedures with small sample sizes. In several real-data simulations, Frick compared AMT to the SPRT with (1) with random item selection and the original method of SPRT parameter specification; (2) random item selection but subgroup parameter specification; and (3) an adaptive item selection algorithm employing classical difficulty statistics (described below), and subgroup parameter specification.

The SPRT with adaptive item selection was found to have classification accuracy comparable to AMT and the random SPRT, while requiring fewer items. In the one study with 25 examinees, the mean number of items required for the AMT procedure was 14.83, compared to 9.68 for the original SPRT, 10.28 for the random SPRT, and 5.55 for the adaptive SPRT. With 50 examinees in the standardization sample, these were 13.57, 10.23, 8.94, and 6.36 items, respectively. The percent correctly classified ranged from 92.5 to 98.1, which is close to the expected percentage of 95, with the exception of the original SPRT with 25 examinees (88.7). Percentages of accurate classifications were evaluated by a goodness-of-fit χ^2 test; none of these differed significantly from expected values. The same pattern of results was found in a second study, but with even greater differences in ATL. This suggests that CTT-based adaptive testing with the SPRT is a superior method, as compared to AMT, for variable-length testing when standardization samples are very small, which makes IRT parameter estimation difficult.

Nevertheless, there are several deficiencies with this study. First of all, an even comparison was not made. The adaptive SPRT differs from AMT in two ways: the decision procedure and the test theory used for parameter estimation. Next, sample size contributed to the results of the study. IRT parameter estimation requires large numbers of examinees and items. In Frick's (1992) first study, the pool contained 97 real items, and two sample sizes of 25 and 50 were used for parameter estimation. In the second study, there were 85 items, and sample sizes were varied at 25, 50, 75, and 100 examinees. While the comparison of methods with small sample sizes was one of Frick's goals, sample sizes this small make it extremely difficult to estimate IRT parameters, so the practical significance of the study is limited. Additionally, Frick (1992) used Owen's (1975) method of constructing confidence intervals using Bayesian posterior variance, just as Spray and Reckase (1996) did, rather than more modern maximum likelihood estimation.

Spray and Reckase (1996) provided the first matched comparison of the termination criteria. They computed expected classification error rates to match the two procedures before conducting simulations, in an attempt to provide an even comparison on ATL. IRT was used as the basis for both termination criteria, and IRT-based intelligent item selection was also used for both criteria. However, the next item drawn was always the most informative at the decision point (Spray and Reckase, 1994), which might not be optimal for AMT. When expected error rates were held constant, the SPRT procedure required fewer items than the AMT, with the difference becoming greater as the error rates became more constrained. However, they did not use maximum likelihood estimation for the AMT confidence intervals, and with the item selection method held constant, the finding that the SPRT performed noticeably better is still not conclusive.

Eggen and Straetmans (2000) compared the three-classification SPRT with three-classification AMT (Kingsbury and Weiss, 1983), using the same simulation methodology as Eggen (1999), again with the SPRT developed by Sobel and Wald (1949). The main purpose of the research was to compare statistical estimation and statistical testing for CCT under realistic rather than ideal conditions. Both content and exposure controls were introduced as variables, but neither had much of an effect. For the comparison between AMT and SPRT, Eggen and Straetmans approximately matched methods on observed error rates, and compared on ATL across four exposure and content conditions. AMT with 90% confidence intervals produced between 87.7% and 89.9% correct classification , while the SPRT with α = 7.5% produced between 87.4% and 91.1% correct classification . However, when δ was changed from 0.1 to 0.1333 in the SPRT, ATL decreased while errors were unaffected, so the authors chose to only compare the δ = .1333 case to AMT. This decrease was enough to persuade the authors to conclude that statistical testing performed more efficiently, by requiring fewer items. However, when δ = 0.1, the SPRT performed equally with AMT.

Eggen and Straetmans (2000) compared AMT and the SPRT as methods for only three-category classification, yet this is the best comparison between the two procedures for several reasons. First, the square root of the Bayesian posterior variance (Owen, 1975) was replaced by the standard error of the weighted maximum likelihood ability estimate (Warm, 1989). Second, the item selection procedure was crossed. Items were selected to maximize FI at the nearest cutscore, following Spray and Reckase (1994), but this was defined differently since the SPRT does not calculate an ability estimate. The original AMT algorithm of FI at the current estimate was also included. Third, it was conducted under realistic conditions, to make sure that the comparison was not moderated by common constraints such as item exposure and content control. Moreover, the item pool was not generated, but instead used 250 real items calibrated on a sample of 1,198 real examinees from the target population of adult education students, as discussed previously. The only distinct drawback was that that the maximum test length was fixed at a short 25 rather than varied. This might be important regarding the specific practical application of interest to the researchers, but it clouds the results.

Unfortunately, the efforts taken to make a good comparison were superseded by the problems associated with multiple-cutscore classification. As noted before, changing the nominal error rates with the Sobel and Wald (1949) procedure did not change observed error rates. The same was true, to a lesser extent, with AMT. This hampered the comparison between the two methods. It is logical to match methods on observed error or expected error (Spray & Reckase, 1996) and then compare on ATL. But seven out of the eight cases investigated had observed error between 86% and 91%, regardless of the *a priori* error rates. Only the 70% confidence interval with FI at the current estimate had different error rates, and this was still not close to nominal levels, ranging from 83.5 to 85.7. Therefore, all eight of the cases had roughly the same observed error, with varying levels of ATL, which might be due to the ATL being fixed at an unrealistically small value. The AMT procedure performed better with FI at the current ability estimate, which agrees with Thompson and Weiss (2006), but contradicts the findings of Spray and Reckase (1994), who found that cutscore-based and estimate-based selection performed comparably. The authors suggested that this might be due to the item bank structure, the location of the decision points, or to the specifications of the procedures.

Conclusions

Comparisons of termination criteria made to this date are incomplete. With the exception of Eggen and Straetmans (2000), all comparisons were confounded with item selection method, and used the square root of the Bayesian posterior variance rather than the maximum likelihood standard error for AMT. Additionally, no comparisons have been made with polytomous IRT, and only one comparison (Eggen & Straetmans, 2000) involved multiple cutscores.

Item Selection Methods

CCTs, because they are iterative or sequential processes, involve multiple steps, where each step requires the administration of another item to the examinee. The methods of choosing the next item differ in the ways that they evaluate item information and choose the "best" item that should be administered next.

CCT was originally developed with random item selection (Ferguson, 1969); however, Kingsbury and Weiss (1983) demonstrated the relative inefficiency of randomly sequential testing as compared to adaptive testing with intelligent item selection. Since then, research has focused on the application of intelligent IRT-based item selection methods to CCT. Intelligent IRT item selection methods fall into two families: estimate-based (EB) and cutscore-based (CB). With EB selection, the criterion to select the next item involves the current θ estimate of the examinee, as calculated with the examinee's response vector up to that point in the test, regardless of the location of the cutscore(s). Conversely, CB item selection evaluates items with reference to the cutscore, and does not involve the θ estimate. Usually, a CCT with CB item selection does not even estimate θ.

Cutscore-Based Selection

Three approaches have been suggested for CB item selection: information at the cutscore point (Spray & Reckase, 1994), information in a region around the cutscore point (Lin & Spray, 2000), and the maximization of the difference $p_2 - p_1$ in the SPRT termination criterion (Frick, 1992, Lin & Spray, 2000). The latter method was suggested first, as Weitzman (1982) made the important observation that the SPRT will classify examinees with the fewest items when these items maximize the difference between the probabilities p_1 and p_2 that are sequentially entered into the ratio, as described below. However, this was not pursued further until a decade later, when Frick (1992) defined item discrimination as the difference $p_2 - p_1$ as calculated with classical test theory difficulty statistics in the master and nonmaster groups of a calibration sample. Lin and Spray (2000) adapted this approach to IRT, selecting items by maximizing the log of the ratio, which the authors termed a log-odds ratio, between the two parameters when calculated with an IRF,

$$R_i = \frac{\left(p_i(\theta_2)/p_i(\theta_1)\right)^X}{\left(q_i(\theta_2)/q_i(\theta_1)\right)^{1-X}} . \qquad (18)$$

Lin and Spray (2000) also investigated KLI, which evaluates information across a region rather than at a single point. Formally, KLI is an index of the discrepancy between two probability distributions. Within the CCT context, these are conditional distributions at two points on θ. KLI can be described as the expectation over observed responses X_i, from the possible responses x, of the log-likelihood ratio for each item, or

$$K_i(\theta_2 \parallel \theta_1) = E_{\theta_1} \log\left[\frac{L_i(\theta_2; X_i)}{L_i(\theta_1; X_i)}\right] \quad (19)$$

with θ_1 and θ_2 representing two points on θ chosen by the test user and

$$L_i(\theta; x_i) = p_i^{X_i}(\theta)\left[q_i(\theta)\right]^{1-X_i} \quad (20)$$

denoting the likelihood function for the *i*th item. The double vertical bars are standard in this context to emphasize that θ_0 and θ_1 need to be separated, and not viewed as the conditional relationship indicated by a single vertical bar. These are the two endpoints that define the region on which that information is calculated, which can be the same as the indifference region, but does not need to be. With a dichotomous IRT model (Eggen, 1999; Lin & Spray, 2000), this simplifies to

$$K_i(\theta_2 \parallel \theta_1) = p_i(\theta_2)\log\frac{p_i(\theta_2)}{p_i(\theta_1)} + q_i(\theta_1)\log\frac{q_i(\theta_2)}{q_i(\theta_1)} \quad (21)$$

where $p_i(\theta_1)$ is the probability of a correct response at θ_1 and $q_i(\theta_1)$ is the complementary probability of an incorrect response. The θ_1 and θ_2 used to calculate KLI are often the same as with the SPRT, but need not be. As with the SPRT, they are usually determined by adding and subtracting a small constant ε from the cutscore. This is important to note because the literature uses the same notation in the definition of both, defining them both as plus or minus a constant δ around the cutscore point. The possibility of confusion introduced by this is the reason for the use of ε within the context of KLI here.

Lin and Spray (2000) compared all three CB methods that employ IRT--maximum FI, maximum KLI, and log-odds ratio--and found them to be nearly equivalent. This conclusion is not surprising, as all three methods assess the same concept – information at the cutscore – and the item that is the highest in terms of one item selection criterion will also be the highest in terms of the others. As first pointed out by Weitzman (1982), the SPRT will make a decision faster when the two *p* values are most different. This situation occurs with an IRF that has a higher discrimination (slope) parameter. An item that has the most information at the cutscore point (FI) will also have the most information in a small region around that point (KLI), both of which mean that the *p* values will have the greatest difference (log-odds ratio).

A serious drawback to the selection of items at the decision point is that those items which are most informative at that point might become overexposed quite quickly. The same items are given to every examinee, possibly causing the items to become compromised in some testing applications. Their original high level of information might then be eroded. Some form of item exposure control, such as stratum levels (Jiao & Lau, 2003; Kalohn & Spray, 1999) must then be employed. One potentially important question, then, is how adaptive item selection compares to intelligently sequential item selection at the cutscore point under item exposure constraints while comparing the SPRT and AMT. Exposure constraints with adaptive selection are not as important because only examinees at the same ability level will receive the same set of items. This comparison is yet to be made.

Estimate-Based Selection

While intelligently selecting the next item to be administered based solely on item characteristics might be more efficient than random item selection, it does not make use of all available information. Since the broader topic of paramount interest is the interaction between the examinee and an individual item or some set of items – a very general description of the testing process – it is logical to introduce examinee characteristics into the item selection process. This can refer to characteristics of the examinee distribution, as is sometimes used with Bayesian methods (van der Linden, 1990), but to make the test truly adaptive, this must include information concerning the specific examinee being tested. Most notably, this includes the responses to past items in the test – the response vector – and the subsequent scoring of this response vector to obtain the current θ estimate.

Adaptive item selection is defined as the selection of the next item to maximize information at the current θ estimate. This methodology was originally developed for point estimation of ability testing, rather than classification testing, because information varies inversely with the CSEM (Embretson & Reise, 2000), and the provision of the most information at the θ estimation implied the least amount of error present in the estimation of θ at that point in the test. However, this approach is also useful for CCT. The reason for this is that AMT's confidence intervals make use of the CSEM. A greater amount of information at the θ estimate implies a smaller CSEM, which implies a smaller confidence interval, which in turn implies a classification decision that is made with greater accuracy and fewer items. For this reason, adaptive item selection was originally suggested for the AMT termination criterion (Weiss & Kingsbury, 1979), and has been used since (Kingsbury & Weiss, 1983; Thompson & Weiss, 2006).

The first research on FI as an item selection criterion was Weiss and Kingsbury (1979) for AMT and Reckase (1983) for the SPRT termination criterion. Both used adaptive EB item selection, where the next item to be selected was the item from the remaining item bank that had the greatest FI at the most recent θ estimate for the examinee. This was an important development because, while randomly sequential testing with the SPRT might perform much more efficiently than fixed-form conventional tests, it does not compare well with other methods of CCT that make use of intelligent item selection. This is exactly what Kingsbury and Weiss (1983) demonstrated. Reckase's (1983) research demonstrated that using the SPRT with intelligent (i.e, EB) item selection is a viable method, enabling more direct comparisons with other methods in the future (e.g., Spray & Reckase, 1996), as well as exploration among various methods of intelligent item selection (e.g., Lin & Spray, 2000).

Frick (1989; 1990; 1992) developed CCT methods using the SPRT with the classical psychometric model. While not exactly the same, selecting the item with the most information at the current estimate is approximately equivalent to selecting the item which is closest in difficulty to the examinee's current θ estimate and has the highest discrimination, since FI is a function of these two item characteristics. Frick (1992) therefore created an item utility index that attempted to match CTT item difficulty to examinee performance, while simultaneously maximizing discrimination. The first step was computing an incompatibility index, defined as the absolute difference between the proportion of all standardization sample examinees who correctly answered an item and the proportion of items that a given examinee had correctly answered after a specified number of items. An item's discrimination was simply defined as the difference between the proportion of masters who got an item correct *(p_2)* and the proportion of nonmasters who responded correctly *(p_1)*. This is helpful to the SPRT because it maximizes the probability ratio *$(p_2)/(p_1)$*. These two indices were then used to calculate the utility index for each item, which was defined as item discrimination divided by item incompatibility, plus a small arbitrary constant to prevent division by zero. Items were selected to maximize the utility index. The matching of item difficulty to the current estimate of examinee ability on the proportion-correct metric at each stage in the test administration makes this method adaptive under the earlier definitions.

The most important thing missing from this research is a comparison between Frick's adaptive SPRT and Reckase's method. These two procedures both have the SPRT as the termination criterion and adaptive item selection. A comparison would demonstrate how well basing the procedure on CTT approximates the greater complexity and accuracy introduced by IRT. It would be of further interest to investigate how this relationship is moderated by sample size, since the larger the sample used for item calibration, the more IRT is appropriate.

Research Comparing EB and CB Selection

Spray and Reckase (1994) was the first study to compare EB and CB item selection methods. Separate monte carlo simulations were conducted with both AMT and SPRT termination criteria, with the SPRT comparing two item selection methods and Bayesian AMT comparing three. With the SPRT, the two item selection methods were maximum FI at the cutscore and at the true θ, as generated for each examinee. Bayesian AMT compared maximum FI at the cutscore, current θ estimate, and true θ. Each simulation had 1,000 replications, with α and β error rates set at 0.05, and the simulation was repeated with three decision points: -0.5, 0.0, and 1.0. The IR for the SPRT was set to be a symmetric plus or minus 0.5 θ units around the decision point, rather than varied. Methods were evaluated by the ATL function conditional on θ.

Spray and Reckase (1994) concluded that administering the next most informative item at the cutscore point resulted in smaller or equal ATL than using the current θ estimate. This was true for both the SPRT procedure and AMT with Owen's (1975) sequential Bayesian procedure. With AMT, the conditional ATL functions were similar, with the current θ estimate method requiring only slightly more items. With the SPRT, the two methods performed similarly near the cutscore point, but when not near the decision point the conditional ATL function of the true θ method became not only higher but also very erratic. It is likely that there was not enough information in the bank at high ranges of θ to efficiently employ adaptive item selection, which also explains the inability of adaptive item selection to outperform cutscore point selection with AMT. Moreover, although this comparison depended heavily on the information available in the item bank, this was not manipulated or even reported.

It is also important to consider conditional error rates as a dependent variable, but this was not investigated. Considering that error rates will covary strongly with θ level, this might be important in the comparison of item selection strategies. Another issue that might have affected the results is the use of Bayesian priors in the construction of confidence intervals, using Owen's (1975) method. Kingsbury and Weiss (1983) originally suggested this method for AMT, but better methods for estimating conditional standard error of measurement (CSEM) from maximum likelihood estimation have been developed since. The CSEM is better in this application than the posterior Bayesian variance because Owen's method requires the specification of a Bayesian prior distribution. The choice of this distribution can artificially raise or lower the variance, requiring the AMT procedure to then take longer or shorter than it would with CSEM confidence intervals, just as the choice of priors can also introduce bias into the item parameter estimation process (Gifford & Hambleton, 1990). The Bayesian prior N(0,1) in Spray and Reckase (1994) resulted in single-item tests for many simulees. When the decision point was 1.0, a single incorrect response was often enough to produce a nonmastery classification, which the authors suggested might cause high misclassification rates. However, no information regarding this was presented, so the relationship in this case is not known.

The empirical results of Spray and Reckase (1994) do not make logical sense, likely due to deficiencies in the item bank. The SPRT should perform most efficiently when items are selected by maximum information at the decision point, as was found. Doing this provides items with the greatest slope in that area of θ, which transfers into values of p_1 and p_2 that are most distant from each other. This is the same logic that supports the use of KLI and a log-odds ratio (Lin & Spray, 2000). However, the items that should make the AMT procedure more efficient are items that minimize the Bayesian posterior variance or conditional standard error of measurement. Since these two quantities are inversely related to information at the θ estimate, the items that do this most efficiently should be the items that provide the most information at the current estimate. This would cause the confidence interval around the current estimate to shrink more than an item providing information in a different range of θ, which in turn enables a decision to be made more quickly. Thompson and Weiss (2006) found item selection at the current estimate to produce lower ATL than CB selection in 80 out of 80 conditions using real-data simulation with AMT. These results were obtained from data collapsed across the θ dimension, though, rather than being presented as a function of θ as was done in Spray and Reckase (1994).

Eggen and Straetmans (2000) also investigated item selection while comparing AMT and the SPRT, but only partially crossed methods. The AMT procedure used maximum FI at both the current ability estimate and the nearest cut point. No ability estimates are needed when testing with the SPRT as the termination criterion, so the only item selection procedure used with the SPRT was FI at the cut point. Random item selection was also added, merely as a baseline. It was found that CB selection required 3-4 more items than EB selection, but also had slightly higher PCC. Since no matching was performed, it is difficult to draw a conclusion.

Eggen (1999) repeated the comparison of Spray and Reckase (1994) with FI while also introducing the comparison of KLI to the two Fisher methods. Separate simulations were completed for the two and three classification case, with three different information interval sizes (ε = .05, .10, and.15). He also varied the *a priori* error SPRT error rates ($\alpha = \beta$ = .05, .075, and .10) and the size of the indifference region (δ = .10, .11, .12, …, .23). Although other research has also varied the cutscore point and maximum test length, these were fixed at $\theta_c = 0.1$ and a maximum test length of 40.

Several important results were found concerning the item selection methods. First, the size of the KLI region mattered little. Second, although FI at the cutscore produced shorter tests on average than FI at the current θ estimate with comparable accuracy, the difference was very small. Lastly, the same conclusion was drawn concerning the outperformance of the FI methods by KLI. The author suggested that the difference between this research and Spray and Reckase's (1994) might be due to the use of a different IRT model (2PLM) than Spray and Reckase, who used the 3PLM. Thus, while there might be a preference for KLI methods it was not large. The greatest effect of the study on PCC and ATL was of the IR, which provides more evidence that suggests that AMT and SPRT should be matched on PCC by varying the IR before making any comparison.

Lin and Spray (2000) conducted a similar study, with the addition of another item selection method. This log-odds ratio (Equation 18) was compared to the item selection strategies of KLI and FI at the cutscore point under varying conditions, each with 100,000 simulated examinees. There were three cutscore points (-0.32, 0.81, 1.79), eleven sizes of the IR (0.20, 0.21, 0.22, …, 0.30), and two item pools (180, 360). Additionally, item exposure rates were varied: the larger pool of 360 items had stratum depths of *s* = 1 (no constraint), 5, and 10 items, while the smaller pool had stratum depths of *s* = 1 (no constraint) and 5 items. Contrary to Eggen (1999), Lin and Spray (2000) did not impose any constraint on test length because they thought that this would decrease the visibility of any effect.

The authors found only negligible differences between the three item selection methods. For instance, when using the 360-item pool and a stratum depth of $s = 1$ (no exposure constraint), ATL never differed by more than one item among the three, and proportion of correct decisions never differed by more than .002. This was across 33 conditions created by crossing three cutscore points with 11 values of the IR. Other conditions defined by pool size and stratum depth netted similar results, and the authors concluded that there was no difference of importance among the three, and that the common practice of selecting by FI at the cutscore point is justified with the SPRT.

Lau and Wang (1999; 2000) compared FI at the cutscore point to KLI around the cutscore, with endpoints fixed at the endpoints of the IR. Real item parameters were used from both polytomous and dichotomous items in a monte carlo simulation. Like Lin and Spray (2000) and Eggen (1999), it was found that FI and KLI at the cutscore performed nearly identically in terms of PCC and ATL. This occurred across several secondary variables such as test length constraints, cutscore location, item exposure, and item pool size.

As previously discussed, the lack of differences is not surprising. The three item selection methods are highly related conceptually. An item with a large amount of FI at the cutscore point will be an item with high discrimination, or a steep slope. An item with a steep slope will translate to values of p_1 and p_2 that differ much more than an item with little or no discrimination. This difference between p_1 and p_2 is exactly what the log-odds ratio and KLI are assessing directly. So the item with the greatest FI at the decision point is likely to be the item with the greatest values of KLI and the log-odds ratio. The log-odds ratio and KLI are addressing the same concept, but the log-odds ratio is in this case fixed at the indifference region boundaries, while the two points in KLI may vary.

Rudner (2002) also addressed the use of a form of KLI in mastery testing, but it was done from a BDT perspective while using the SPRT as the termination criterion. Rudner's justification for BDT with a classical model was that IRT often requires larger samples and more restrictive assumptions while introducing greater complexity. Yet IRT's major assumptions are of unidimensionality and local independence, which are equivalent; BDT, as well as the SPRT termination criterion, also assumes local independence, so one approach is not more restrictive than the other is in this respect.

Rudner (2002) used classical difficulty parameters from subgroups to specify the SPRT parameters while investigating three item selection criteria from BDT: minimum expected cost, information gain, and maximum classical discrimination. The first was constructed to maximize the BDT termination criterion in an adaptive fashion, while the second is an adaptive variant of KLI, and the third was included as a BDT approximation to maximum FI discrimination at the cutscore point, as supported by Spray and Reckase (1994), and the log-odds ratio of Lin and Spray (2000).

To compare these three item selection criteria, as well as AMT (Kingsbury & Weiss, 1983) with items selected at the true score, Rudner (2002) simulated examinee responses using IRT. The item parameters were sampled from two real item pools, with 139 and 54 items, used in elementary mathematics assessments. Given these small item pools, maximum test length was varied only from 5 to 30, in increments of five, and this constraint of very short tests drawn from small pools likely affected the results. The first assessment classified examinees into two categories, the second into four.

The examinee distributions of 1,000 for the calibration sample (for the BDT item parameters) and 10,000 for the research sample were created by randomly drawing points from two distributions, one normal N(0, 1) and one uniform (-2.5, 2.5), and assigning a true classification based on this true score. However, while the Method section claimed to use two distributions, there was only one table in the Results, and it did not specify which distribution.

There are two important conclusions drawn by Rudner (2002). First, minimum expected cost and information gain noticeably outperformed the other two methods with respect to ATL while producing almost identical proportions of correctly classified examinees. The author notes that, upon further examination, this is due to the two criteria almost always choosing the same item at a given point in the test administration. This conclusion is qualified by the fact that this relatively superior performance by information gain and minimum expected cost is inversely related to maximum test length. The greater accuracy in classification shown by these two procedures has a reduced effect as the maximum number of items is raised. So for longer tests, all four methods should produce roughly equivalent classifications, again supporting the relative equivalence of different item selection approaches.

There were a few noticeable absences in this study, the most obvious being ATL. ATL and PCC should always be examined jointly. Rudner (2002) also imposed a confound by using EB selection with IRT and CB selection with CTT, though CB selection is more appropriate with the SPRT, thereby reducing the effectiveness of the item selection algorithm for the IRT-based condition.

One additional application of KLI was Weissman (2004). Weissman noted that the use of KLI in multiple-cutscore CCT produces a problem. KLI is an index of information across a region, but only one region (θ_0 to θ_1). When there are multiple categories into which examinees can be classified, there are then multiple cutscore points on the θ metric. If the item selection strategy is to maximize information around a cutscore (Spray & Reckase, 1994; 1996), multiple categories necessitate multiple information regions, which KLI is unable to do. Eggen (1999) also noticed this problem. Additionally, there is the issue of defining the KLI region width. Weissman's suggestion was to employ mutual information, a more general form of KLI.

The mutual information between individual examinee responses and examinee trait level, $I(X_i, \theta)$, is the KLI between the joint distribution $f(x_i, \theta)$ and the product of the two marginal distributions $f(x_i)$ and $f(\theta)$. This is expressed as (Weissman, 2004)

$$I(x_i,\theta)=\sum_{x}\sum_{\theta} f(x_i,\theta)\log\left[\frac{f(x_i,\theta)}{f(x_i)\,f(\theta)}\right] \quad (22)$$

Computationally, however, this is evaluated at a discrete number of points. Weissman also suggested another new criterion for item selection in CCT, FI with a posterior weight function.

Weissman compared these two strategies in a four-classification CCT, along with selecting items by FI at the current ability estimate. The comparison used a sample size of 5,000 simulated examinees drawn from a N(0,1) distribution. The item pool had 367 items already calibrated under the 3PL model, centered at θ = 1. The three cutscore points were fixed at -0.3, 1, and 2 with a fixed IR width of 0.6 (δ = 0.3). Maximum test length was introduced as a variable, and the simulation was repeated with maximum test length rising in single increments from 1 to 40. Nominal error was set at 0.10 for each classification group; $\alpha = \beta = 0.05$ at each cutscore except for the decisions for the two end categories (1 and 4), where it was .10. Results showed that FI with a posterior weighted function usually outperformed FI in terms of ATL and observed misclassification rates, and mutual information always outperformed them both. The only exception was for Category 4 with small maximum test length, where the procedures performed more erratically and there was no pattern of superiority. These results suggest that when there are four or more classifications and short tests are needed, mutual information might be more efficient.

Conclusions

Since mutual information's superiority comes with four or more classifications, which is a small proportion of classification testing, its greater complexity does not offer an increase in efficiency for most classification tests. Similarly, KLI is more complex than FI, but does not always offer a substantial increase in efficiency. In the interests of parsimony, therefore, most CCTs can be constructed with an item selection algorithm based on FI. If CTT is used for the psychometric model, analogous indices are available.

While there are not considerable differences in efficiency among information types within CB or EB selection, the specification of CB vs. EB item selection is important. CB selection is appropriate for the SPRT termination criterion (Spray & Reckase, 1994) because it maximizes the difference of the likelihood ratio, while EB selection is appropriate for AMT because it maximizes the reduction in the CSEM used to construct the confidence interval.

Constraints have a moderating effect on differences between competing methods; consider the extreme constraint of requiring all tests to be 50 items long, where no difference would be found in ATL between any item selection methods or termination criteria. Therefore, while they serve useful purposes in practical applications, they are of minor importance in research.

Practical Constraints

When designing a CCT to be used in a live testing program, there is often pressure from external forces to impose constraints on the algorithms to address practical issues. There are three primary types of practical constraints: item exposure, test length, and content area. All three are very important in some testing applications in that they address significant practical issues in testing. However, their relevance is lessened in research comparing algorithms in computerized variable-length testing because the constraints obviously constrain the algorithms from performing as they are designed to do, thereby confounding some conclusions.

This has been supported empirically. Eggen (1999) applied the Kingsbury and Zara (1991) approach to distributing items among three content categories. After an item was administered and no decision had been made, the percentage of items administered in each category was compared to the target percentages, and the next item was selected from the content category with the greatest discrepancy. It was found that, while the introduction of constraints had little absolute impact, it lessened the differences in efficiency between the three item selection methods that were part of this portion of the study. This result was also present, though less strongly, in Eggen and Straetmans (2000), in the comparison between two item selection methods with AMT. Thompson and Weiss (2006) found that test length constraints only served to reduce differences between methods. Spray, Abdel-fattah, Huang, and Lau (1997), Lau (1998), Jiao, Wang, & Lau (2004), also found that constraints reduced differences. Lin and Spray (2000) specifically did not include constraints because they reduce the visibility of comparisons among other variables.

Conclusions Regarding CCT Design

Several important choices are presented to the test user when designing a CCT. The first step is to select a psychometric model that can be used to calibrate the item bank. For several reasons, IRT presents a more viable option, though efficient CCTs can be designed with classical test theory if small sample size necessitates that path. The second step is to select a termination criterion. The SPRT has been shown mathematically to be the uniformly most powerful test of two competing hypotheses (Wald, 1947), but AMT is highly efficient also. Both methods classify examinees with far fewer items than are needed for a fixed-form test, while maintaining high levels of classification accuracy. Both have also been shown to work efficiently in the situation of multiple cutscores, and the SPRT has additionally been shown to be efficient with polytomous IRT. Both of these additional variables were investigated in the current study.

An item selection method must be chosen to work in conjunction with the termination criterion, with several options available. Although conflicting results have been found in the CB/EB comparison, it is evident that the additional complexity introduced by KLI is not necessary for either approach. Therefore, the current study was designed to compare the two paradigms, with FI as the specific operationalization.

Lastly, there is the issue of practical constraints. It is evident from past research that the imposition of constraints is not necessary, and is even detrimental to, simulation research on CCT methods. Therefore, they were not included in this study, but CCT test users are obviously free to impose any constraints necessary for their particular application.

Chapter 3: Method

Purpose

The objective of this study was threefold:

1. To investigate the relative efficiency of multiple-cutscore CCTs scored with polytomous IRT as compared to CCTs scored with dichotomous IRT models;
2. To compare the two termination criteria of AMT and the SPRT in a polytomous CCT, as the two have been compared in previous research but not in the polytomous case;
3. Similarly, to compare the two item selection methods of EB and CB. Although there have been some studies comparing these item selection methods, they have not been compared in the polytomous case.

The primary hypothesis was that there would be an interaction between the number of cutscores and the type of IRT model. The advantages of polytomous IRT were hypothesized to be greater for the two-cutscore case than for a single cutscore test. Operationally, this would be manifested as a greater difference in ATL while maintaining a similar level of PCC.

The method used to compare these conditions was a monte carlo simulation where each simulated examinee was assigned a classification by the CCT, which was compared to the examinee's true classification as determined by the generated θ value. PCC and ATL were the dependent variables.

Two additional independent variables were also investigated for possible interaction with the three primary independent variables: the number of cutscores (one or two) and item bank information structure (uniform or bell-shaped). Thirteen levels of θ, ranging from –3.0 to 3.0 in increments of 0.5, were also investigated, but only to evaluate the effects of the other independent variables conditional on θ. The combination of these independent variables resulted in a 2 (termination criterion) × 2 (item selection method) × 2 (IRT model) × 2 (number of cutscores) × 2 (item bank shape) design with 32 conditions. The independent variables and their levels were:

1. IRT Model
   - a. Dichotomous (3PL)
   - b. Polytomous (GPCM

2. Termination Criterion
   - a. SPRT
   - b. AMT

3. Item Selection Method
   - a. Current θ estimate (EB)
   - b. Cutscore (two-category case) or nearest cutscore (three-category case) (CB)

4. Number of Cutscores
   - a. One cutscore, i.e., two classifications
   - b. Two cutscores, i.e., three classifications

5. Item Bank Shape
   - a. Information function high and flat (uniform item location distribution)
   - b. Information function peaked at cutscore(s) (unimodal bell-shaped item location distribution for one cutscore; bimodal for two cutscores)

Cutscores

The uniform distribution of examinees on θ used in the monte carlo simulation was generated only to allow analysis of the independent variables conditional on θ. However, the cutscores investigated in this study were chosen with the assumption of a N(0,1) distribution, as this is more commonly encountered in practice than a perfectly uniform distribution of examinees. Cutscores were selected for a norm-referenced interpretation since the lack of a real item pool and testing situation precludes a criterion-referenced interpretation. In the two-category condition, with only one cutscore, the cutscore was set at 0.675 on an N(0,1) θ scale. This represents the case where the purpose of the test is to identify the upper 25% of examinees. This corresponds to the situation where the goal of the test is to identify the examinees with the highest level of ability or achievement, such as in educational assessment or employee selection. Similarly, the three-category case was designated with cutscores at –0.675 and 0.675, classifying the top and bottom 25% of the sample. This is an extension of the same situation where the test is designed to not only identify the highest performing examinees, but also the lowest, perhaps for some remedial training.

Item Banks

Six separate banks were constructed by varying the three factors of IRT model, number of classifications, and item bank shape.

1. Dichotomous uniform
2. Polytomous uniform
3. Dichotomous peaked unimodal
4. Polytomous peaked unimodal
5. Dichotomous peaked bimodal
6. Polytomous peaked bimodal

A unimodal item bank was constructed only for the case of a single cutscore, with the information function peaking at that cutscore. A bimodal item bank was constructed only for the two-cutscore case, with information at both cutscores, rather than peaking at a single point. The same uniform bank was used for both the single-cutscore and the two-cutscore cases; hence, there were only six banks, rather than the eight that would be obtained by fully crossing the three aforementioned variables.

The use of different sets of items to compare a polytomous and dichotomous IRT model follows Lau and Wang (1998, 1999, 2000). They used two banks of 266 polytomous items and 246 dichotomously scored items when comparing a polytomous and dichotomous CCT with a single cutscore. It would also be useful to use only one set of items and calibrate them with both IRT models. This would represent the case where one model is appropriate, but an incorrect model is used. For example, item responses could be generated with the GPCM and then calibrated with both the GPCM and 3PL to investigate the loss in efficiency caused by the misspecification of the model, as was shown to have an important effect among dichotomous IRT models by Spray and Reckase (1987). However, this introduces model fit as another independent variable, so two separate sets of items were used in the current study. Additionally, if one model were used to generate data for calibration by both models, the model that was used for generation would have an undue and unfair advantage. That model would fit the calibration data better, and might then lead to better performance in simulations.

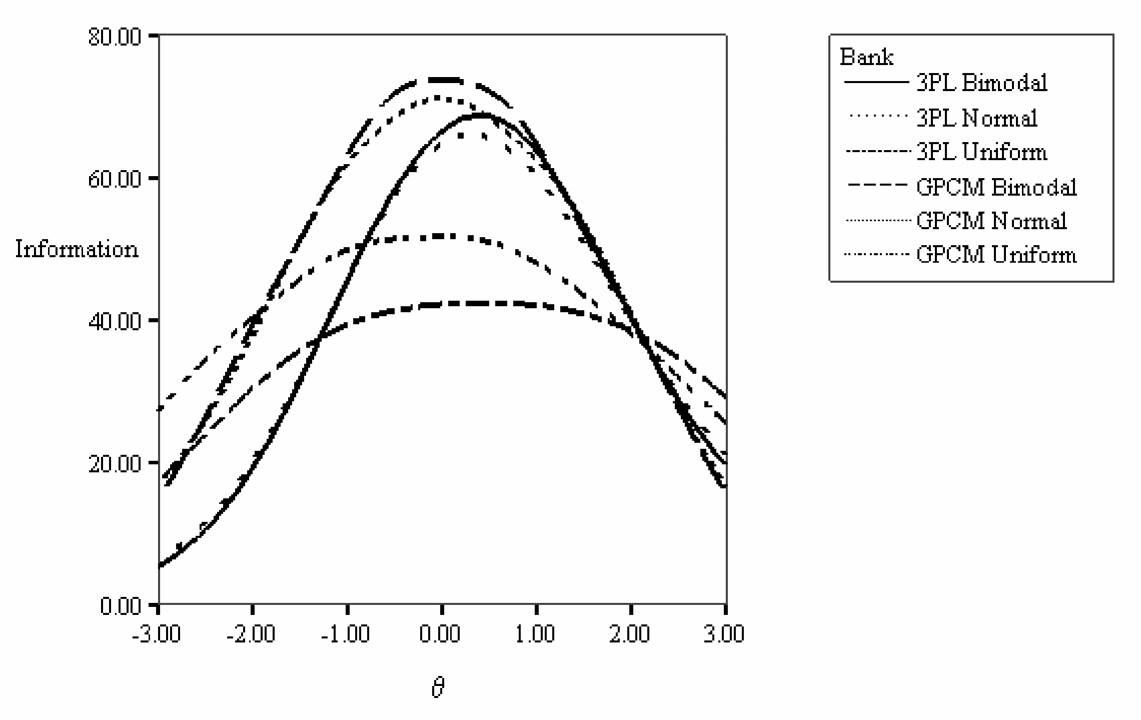
Moreover, in a practical application of a CCT, a bank of items will likely be constructed for the purpose of applying polytomous or dichotomous IRT but not both, and an inappropriate model would likewise not be employed in the real world. For instance, a typical multiple-choice item bank is intended for the 3PL, while a set of constructed-response or performance assessment items is intended for the GPCM or similar scoring method. Items typically are not written for both models simultaneously, and both models are then not equivalently appropriate. Many testing programs utilize both approaches separately because many of the characteristics that are assessed by one type of item cannot be assessed by the other. For example, a multiple-choice examination serves to assess knowledge of facts, while a performance assessment serves to assess the ability to perform tasks correctly. This can be done within the same test, as is done with the NAEP, or in separate tests.

For the polytomous bank in this study, GPCM item parameters were generated. The GPCM is conceptually appropriate for many polytomous achievement items because it involves an ordering of the options to reflect steps involved in determining the correct answer. The GPCM has been used to model polytomous achievement items on exams from the NAEP (Loomis & Borque, 2001; Lau & Wang, 1998, 1999, 2000). It was assumed for the present study that all GPCM items had four responses, which represents many GPCM items on NAEP exams (Lau & Wang, 1998). The GPCM can be used to model multiple-choice items with an explicit ordering of alternatives, but is also used for open-ended or performance assessment items. For the dichotomous banks, 3PL item parameters were generated. The 3PL is appropriate for dichotomously scored multiple-choice achievement items.

Item parameters were generated in an attempt to provide two target shapes of the bank information function, representing the optimal bank information structure for each situation: a uniform bank and a peaked bank. The resultant bank information functions are presented in Figure 7. While the “uniform” and “bimodal” information functions were not strictly so, the uniform function was much flatter than the peaked functions, and the bimodal function provides information across a wider range than the unimodal function. If a difference is found between these banks, it likely would be found between banks that strictly fit the definitions. Moreover, banks of real testing programs are not likely to be strictly uniform or bimodally peaked.

The optimal bank information function for adaptive testing with EB item selection is high and flat, or the item bank contains a large number of highly discriminating items across a range of θ (Weiss, 1974; Thissen, 2000). This implies an approximately uniform distribution of item difficulty, so the item location parameters for the dichotomous IRT bank were uniformly distributed from -3.0 to 3.0 in intervals of $0.03\overline{3}$, producing a bank of 181 items. The item parameters for the dichotomous banks are presented in Appendix A.

Figure 7: Bank information functions



For the polytomous bank with relatively uniform information, the boundary location parameters were generated by first generating the middle of the three parameters for each item, and then adding/subtracting a constant from this value for the other two boundary parameters. The middle parameters were specified at 61 values of θ, from -3.0 to 3.0 in intervals of 0.1, for a bank of 61 items. The constant distance of 1.35 between the category boundaries of the GPCM items was determined by the location of the cutscores; as these were -0.675 and 0.675, a distance of 1.35 (0.675 + 0.675). The rationale for this choice was that it represented the optimal case for SPRT-based multiple-cutscore CCT with polytomous items. If an item is selected because its highest point of information is at one cutscore, the location parameter for that item will likely be near that point. One of the other category boundaries of the item will then be located near the other cutscore, providing a substantial amount of information at that point also. The fact that these items provided information at more than one cutscore when polytomously scored, contributed to the relatively normal peakedness of the bimodal bank information functions even though the location parameters were a bimodal distribution (Figure 7). The item parameters for the polytomous banks are presented in Appendix B.

Since polytomous items have the advantage of more information, the polytomous and dichotomous banks were matched to provide an approximately equivalent bank information function. The polytomous items had three category boundaries, as opposed to the single boundary in dichotomous models, so they provided three times as much information, assuming that item parameters were held constant. Therefore, the polytomous bank contained one-third as many items, for both the uniform and peaked banks. Although 61 items might seem like a small number, the average test length found by Lau and Wang (1998) for polytomous CCT was only 15 items.

Because a bank information function that is peaked at the cutscore is ideally different for a single-cutscore test (unimodal) than a two-cutscore test (bimodal), two separate banks were generated to represent the optimal information case for CB selection, one for a single cutscore and one for two cutscores. In the single cutscore case, the bank was peaked near the cutscore of θ = 0.675, which represents the condition where the purpose of the test was to identify the top 25% of examinees of a normal distribution. This was accomplished by randomly generating item location parameters (b_i) from a N(0.675, 0.25) distribution. The three-classification peaked banks required a bimodal information function, which peaked at both of the two cutscores. These were constructed by generating two pools of items with the location parameters normally distributed around each cutscore, N(-0.675, 0.25) and N(0.675, 0.25), and these were combined to form the bimodal pool. For the polytomous bank, there were 30 items distributed around the lower cutscore, and 31 around the higher cutscore. Similarly, the dichotomous bank had 90 items distributed around the lower cutscore and 91 at the upper cutscore. The relatively small standard deviation of the b_i parameters, only a quarter of the examinee standard deviation, was selected to construct peaked information functions for comparison to the uniform information functions constructed by the uniform item banks.

Bank information functions for the six item banks are shown in Figure 7. Though the 3PL banks had three times as many items as the GPCM banks, the GPCM banks still had more information across most of θ. Note that the distribution of item location or difficulty parameters did not necessarily mean that the information functions were of strictly the same type, due to the fact that each item provided a varying amount of information across θ. However, it is still noteworthy that the bank information functions for the uniform banks were much flatter than the other bank information functions.

Item discrimination also has an effect on information, but since discrimination parameters were randomly generated and therefore uncorrelated with difficulty parameters, they had little effect on the bank information function. The discrimination parameters for each bank were randomly drawn from a N(1.0,0.2) distribution in an effort to produce a high information function found with banks of highly discriminating items. The distribution parameters of 1.0 and 0.2 were arbitrarily chosen, as Lau and Wang (1998, 1999, 2000) did not provide any information concerning the discrimination of the NAEP items used in their polytomous CCTs. Moreover, since the primary purpose of this study was comparison of termination criteria, it was of primary importance that the different banks had similar discrimination parameter distributions, regardless of the values. The same distribution of discrimination parameters was used for generating all banks, so that one bank did not have an advantage over another simply because it had higher discriminations and therefore more information. If one bank had better discriminating items, it would provide more information, predisposing that bank to perform better regardless of shape. Also held constant across the dichotomous banks was the pseudoguessing parameter for the 3PL, which was c = 0.25 for each item, representing the case of a typical four-alternative multiple-choice item with plausible distractors.

Termination Criteria Specifications

Both AMT and the SPRT require certain specifications to be set by the test user. AMT requires specification of the 1-γ confidence interval, while SPRT requires the nominal error rates α and β as well as the width of the IR. (The traditional notation of confidence intervals is that of 1-α, but 1-γ was used here so as not to be confused with the nominal Type I error rate α.) The total nominal error rate γ was here equal to the sum of the nominal Type I (false positive) and Type II (false negative) rates α and β. This study used a 95% AMT confidence interval, which had a nominal error rate of 2.5% on either side.

Though it is not necessary to set the SPRT error rates to be equal, or equivalently for the AMT confidence interval to be symmetrical, this study examined only that case. The SPRT does this by explicitly stating equivalent values of α and β. The specification of the confidence interval for AMT implicitly set $\alpha = \beta = \gamma/2$ at each cutscore. It is possible to specify a different interval width on either side at the cutscore, but this has not yet been investigated. For instance, a 99% interval could be used for all $\theta > \theta_c$, and a 95% interval used for all $\theta < \theta_c$. This would translate to a 0.005 probability of a false positive classification α, and a 0.025 probability of a false negative classification β. This study examined only the simpler scenario where the nominal SPRT rates were each set at .025 and the corresponding AMT confidence intervals were 95%.

Because of the fact that nominal error rates do not always control observed error rates, nominal rates were not varied, even though it is conventional to vary the nominal rates in CCT research. An example of this lack of control is Eggen and Straetmans (2000), who investigated the AMT criterion in four multiple-cutscore situations: a 70% confidence interval with CB selection, a 70% confidence interval with EB selection, a 90% confidence interval with CB selection, and a 90% confidence interval with EB selection. These produced observed PCC of 88.4, 85.4, 89.9, and 89.1 percent, respectively. A similar effect was observed with the multiple-cutscore SPRT. These results might have been due to the strain that multiple cutscores place on the decision procedure, but this observation was also true for the two-classification case in a different study (Eggen, 1999). Eggen found that observed error rates remained relatively the same, regardless of the nominal rates or the item selection method employed. The only effect of multiple cutscores was to increase the overall amount of misclassification. Therefore, only the nominal PCC of 95% was applied in this study.

The SPRT requires an additional arbitrary specification, the IR width. IR size has a direct effect on the results (Reckase, 1983; Eggen & Straetmans, 2000), though some research has not accounted for this fact (Jiao, Wang, & Lau, 2004). Previous research has indicated that a wider IR tends to classify examinees with fewer items, in turn causing a slight increase in classification error as opposed to a smaller IR. Therefore, choosing an SPRT IR that is too large might lead to results with much lower ATL and PCC than the AMT condition that it is intended to be compared to, and the comparison will then not be as even as possible.

Spray and Reckase (1996) suggested that the AMT and SPRT simulations be equated on expected classification error and then compared on test length. They conducted AMT simulations, and then chose nominal SPRT error rates for a fixed IR that produced similar levels of classification error. A similar approach was employed in this study. However, instead of using different nominal error rates for each method, these were kept constant. AMT simulations were run first with a 95% interval, or nominal errors of 2.5% on either side of the cutscore. SPRT simulations were then run with nominal error rates of 2.5%, with the IR systematically varied until *observed* PCC was equivalent to the AMT condition. The two were then compared on average test length. This was done for each of the 16 conditions created by the four variables other than the termination criterion. Eggen and Straetmans (2000) employed this approach in their interpretation of results, but did not vary the IR to make the observed error rates extremely close, as was done in the present study. All observed PCC rates were matched so that the greatest difference in terms of PCC between equivalent conditions of AMT and SPRT was 0.15% (see Table 1).

The SPRT approach that was used for the three-category case was Armitage's (1950) method. This method can be applied to CCT for any number of ordered categories, as opposed to Sobel and Wald's (1949) formulation. Sobel and Wald's is only applicable to the three-category situation, but makes for fewer likelihood ratio comparisons than the Armitage method. Govindarajulu (1987) reported that the two performed comparably in a three-category situation, and since the Armitage method is more generalizable, that method was employed in this study.

Item Selection Methods

The EB item selection method has been shown to be the more efficient of the two item selection methods both in the two-category case (Thompson & Weiss, 2006) and the multiple-cutscore case (Eggen & Straetmans, 2000) with the AMT termination criterion. With the two-category SPRT termination method, prior research indicated that it is more efficient to utilize CB item selection (Spray & Reckase, 1994), but with the multiple-cutscore SPRT termination method, EB selection and CB selection performed with equivalent efficiency (Eggen, 1999). Therefore, both EB and CB selection were applied to each condition in this study, crossed with termination method to investigate the four possible combinations. The current study used only the FI versions of CB and EB selection, as past research has found minimal difference between FI and KLI. Items were selected to maximize FI at the current estimate for EB selection, and the cutscore, or nearest cutscore in the three-classification case, for CB selection

The terms *estimate-based* and *cutscore-based* item selection were used here because the introduction of a second cutscore changes the way that an item selection method is operationalized. This does not occur for EB selection, but does for CB selection. When there is one cutscore, items are always selected based on information at that point. A second cutscore requires the application of a new algorithm.

In the simple two-category case, CB item selection is defined equivalently with both termination criteria. There is only one cutscore, so items are ranked by information at that cutscore. With two cutscores, items are selected to maximize information at the nearest cutscore. However, because AMT operates on the θ scale during every step of the testing process and the SPRT does not, “nearest” must be defined differently for the two termination criteria. Eggen and Straetmans (2000) operationalized CB selection for the AMT termination criterion as the selection of the next most informative item at the cutscore nearest to the current estimate on the θ scale. With the SPRT, the items are selected at the cutscore that is further from a decision, because this reflects the examinee’s ability level being closer to that cutscore. For example, if the examinee’s θ level was near the lower of two cutscores, the decision would be much more difficult to make at the lower cutscore than the higher cutscore. In all probability, the algorithm would quickly recognize that the examinee should not be classified at the higher ability level. This examinee should incorrectly answer the majority of items with difficulty values near the upper cutscore. Therefore, items would be selected to maximize information at the lower cutscore to distinguish between the first and second categories, since the third category would be ruled out for that examinee.

Regardless of the number of categories or the termination criterion, EB selection always operates the same way; items are selected to maximize information at the θ estimate based on item responses up to that point in the test. The EB selection entails more computation than an SPRT-based CTT would normally require, because the SPRT does not use θ estimates in computing the termination criterion, as does AMT.

No practical constraints were applied for two reasons. First, this was an exploratory study of a new methodology, and practical constraints present an opportunity for future research if the methodology is found efficient, accurate, and potentially useful. Second, past research has found that practical constraints tend to reduce differences between competing methodologies (Thompson & Weiss, 2006; Spray, Abdel-fattah, Huang, & Lau 1997; Lau 1998; Jiao, Wang, & Lau, 2004), and often have simple and obvious effects, such as a minimum test length increasing ATL for each condition.

Monte Carlo Simulation

A sample of 13,000 examinees was generated from a U(-3, 3) distribution at thirteen points on θ from –3 to 3 in intervals of 0.5, each with 1,000 examinees. This allowed for the performance of the CCT procedures to be compared conditional on θ, as well as for the entire aggregated distribution, with each θ level contributing equally to the aggregate.

Item responses were simulated by generating a rectangular random number between 0 and 1 for each item for each person, which was then compared to the probability of responding in each of the four item alternatives for the GPCM or two alternatives for the dichotomous case at the true θ level for each examinee (using the appropriate item parameters for that item) to determine the "response." Responses were generated independently for each item for each examinee. The probabilities of responding in a category are defined in Chapter 2 in Equations 1 and 2. For example, suppose for a GPCM item at a given level of θ there was a 0.10 probability of responding with option 1, 0.60 probability of option 2, 0.25 probability of option 3, and a 0.05 probability of option 4. These probabilities were progressively summed to produce boundaries that the random number was compared to. In this example, the first boundary, between categories 1 and 2, was 0.10. The second boundary was 0.70, and the third boundary was 0.95. If the random number for that item was between 0 and 0.1, the examinee "responded" with the first option. If the random number was between 0.10 and 0.70, the examinee responded to the second option, and so on.

The same method was used to simulate responses with the 3PL, but only two options were present. If the random number was above the probability of correctly answering the item for an examinee at the given θ, the simulation scored the item as incorrect. If the random number was below the probability, the simulation scored the item as correct. For example, suppose the probability of a correct response was 0.30. If the random number was below 0.30, the item is "scored" as correct. If the random number was above 0.30, the item response was incorrect.

Once the item was scored, the result was used to update the calculation of the termination criterion. In the case of AMT, the likelihood function was updated, producing a modified estimate of the examinee's ability and SEM. As discussed in Chapters 1 and 2, using the AMT termination criterion if any cutscore was contained within the resulting confidence interval around the θ estimate, another item was administered. If not, the test was terminated for the examinee and a classification assigned. The responses from each item were likewise used to update the likelihood ratio, or ratios in the two-cutscore case, for the SPRT cutscore criterion. Another item was administered if the SPRT termination criterion was not able to make a decision at that point in the test.

Spray (1993) used maximum test lengths of 10, 20, and 50 for simulated CCTs, while Eggen (1999) fixed the maximum at 40 items. However, the imposition of test length constraints tends to reduce the visibility of effects in the results (Thompson & Weiss, 2006). Therefore, no length constraints were used in this study, but should be evaluated in future research on polytomous multiple-cutscore CCT. Examinees that were administered every item in the bank were classified in whichever classification was closest. For example, if the SPRT was greater than 1.0 but less than the upper boundary, examinees were classified into group 2 ("pass").

Before analyzing the simulations, a quality control analysis was performed on the part of the computer program responsible for generation of responses. The data from this substudy are presented in Appendices C and D. Observed proportions of each response were compared to the expected value for a given level of θ, as calculated by the IRF or polytomous response functions. The difference between the two is presented in the column labeled "bias." Values in the bias columns are very near zero; the mean bias for the four polytomous responses are each within 0.001 of zero, with standard deviations of approximately 0.01.

Data Analysis

The simulation was replicated 25 times for each of the 32 conditions of the study to obtain an estimate of sampling variability, which also enabled the use of an ANOVA model to summarize the results. As the item parameters and person parameters did not change, only item responses were generated anew in each replication.

The number of items administered to each of the 13,000 examinees in each of the eight simulated CCTs was used to calculate the ATL for each replication at each value of θ. The generated true θ value for each examinee was compared to the cutscore or cutscores, and a true classification into one of the two or three categories assigned for each group of 1,000 simulees at a given level of θ. This was then compared to the classifications made with the simulated CCTs to determine PCC for each replication, at each level of θ. The relative efficiency of polytomous multiple-cutscore CCT was then gauged by whether it made use of fewer items than dichotomous multiple-cutscore CCT while still maintaining similar rates of misclassification.

The majority of research concerning CCT has subjectively compared procedures on ATL and PCC simultaneously. The notable exception to this was Spray and Reckase (1996), who matched procedures on expected PCC to provide a more objective comparison on ATL alone. Since the current study matched AMT and the SPRT on observed PCC, within the 16 conditions created by the remaining independent variables, the primary method of analysis was an ANOVA summarization of ATL results. PCC results were also summarized with ANOVA, but are of secondary importance since it was a controlled factor. ATL and PCC data were analyzed separately with a 2 (termination criterion) x 2 (item selection method) x 2 (IRT model) x 2 (number of cutscores) x 2 (item bank shape) ANOVA model to obtain relative estimates of effect size. The relative strength of each effect including interactions was gauged by η^2 (Yoes, 1993; Jacobs-Cassuto, 2005), which represents the proportion of variance in the study due to a variable, and was calculated for each effect as

$$SS_{effect}/SS_{total}\ , \tag{23}$$

Where

$$SS_{effect} = \sum (\bar{x}_l - \bar{x}_{grand})^2 \tag{24}$$

with $\bar{x}_l$ as the means for each level of a given effect and $\bar{x}_{grand}$ as the grand mean, and

$$SS_{total} = \sum (x - \bar{x}_{grand})^2\ . \tag{25}$$

It is important to note that ANOVA is only appropriate for *summarization* of the results. Hypothesis tests of PCC and ATL are inappropriate because statistical significance is partially a function of sample size, and sample size (replications) is entirely controlled by the experimenter in monte carlo simulation. It is also important to note that, in this situation, PCC is able to be summarized by ANOVA. In most situations, it is inappropriate to analyze proportions or percentages with ANOVA because they are not normally distributed. However, in this situation PCC was approximately normally distributed around a “true” value of PCC rather than being uniformly distributed across the range 0-100.

Chapter 4: Results

Overview

Table 1 presents the mean ATL and PCC for each of the 32 conditions, averaged across all 25 replications at each condition. ATL ranged from 2.71 to 37.06, and PCC ranged from 94.09 to 98.51. The shortest ATL was for the CCT that utilized CB selection and the GPCM and had only one cutscore. The termination criteria performed equivalently in this condition. The longest ATL was for the CCT that utilized CB selection and the 3PL model with the AMT termination criterion and two cutscores.

Table 1: Mean ATL and PCC for Each Condition

	ATL		PCC	
Condition	Mean	SD	Mean	SD
Uniform Bank				
3PL Model				
AMT Termination				
EB Selection				
1 Cutscore	4.90	0.02	94.51	0.20
2 Cutscores	34.27	0.45	94.09	0.14
CB Selection				
1 Cutscore	5.78	0.03	95.47	0.14
2 Cutscores	37.06	0.42	95.30	0.21
SPRT Termination				
EB Selection				
1 Cutscore	3.76	0.03	94.61	0.12
2 Cutscores	16.69	0.18	94.23	0.19
CB Selection				
1 Cutscore	4.88	0.03	95.61	0.16
2 Cutscores	17.83	0.20	95.16	0.18
GPCM				
AMT Termination				
EB Selection				
1 Cutscore	2.85	0.02	97.59	0.11
2 Cutscores	12.95	0.11	97.91	0.11
CB Selection				
1 Cutscore	3.00	0.02	97.76	0.09
2 Cutscores	13.38	0.07	98.01	0.11
SPRT Termination				
EB Selection				
1 Cutscore	2.85	0.02	97.67	0.16
2 Cutscores	10.32	0.07	98.00	0.11
CB Selection				
1 Cutscore	2.88	0.02	97.62	0.16
2 Cutscores	10.39	0.09	97.97	0.12

	ATL		PCC	
Condition	Mean	SD	Mean	SD
Peaked Bank				
3PL Model				
AMT Termination				
EB Selection				
1 Cutscore	5.31	0.03	95.34	0.14
2 Cutscores	33.43	0.31	96.68	0.14
CB Selection				
1 Cutscore	5.82	0.03	95.91	0.15
2 Cutscores	33.79	0.30	96.38	0.15
SPRT Termination				
EB Selection				
1 Cutscore	4.68	0.02	95.49	0.16
2 Cutscores	19.55	0.17	96.57	0.15
CB Selection				
1 Cutscore	4.82	0.03	95.76	0.14
2 Cutscores	17.77	0.19	96.27	0.17
GPCM				
AMT Termination				
EB Selection				
1 Cutscore	2.89	0.02	97.64	0.11
2 Cutscores	12.32	0.11	98.50	0.11
CB Selection				
1 Cutscore	2.71	0.02	97.82	0.14
2 Cutscores	12.55	0.11	98.46	0.10
SPRT Termination				
EB Selection				
1 Cutscore	2.76	0.02	97.65	0.13
2 Cutscores	9.84	0.07	98.51	0.11
CB Selection				
1 Cutscore	2.71	0.03	97.72	0.13
2 Cutscores	9.59	0.07	98.41	0.10

The result of matching of conditions on PCC in order to provide a more even comparison on ATL, as described in Chapter 3, is seen in Table 1. Conditions that are equivalent for AMT and the SPRT have an approximately equal PCC. As mentioned previously, the greatest difference in PCC is only 0.15%.

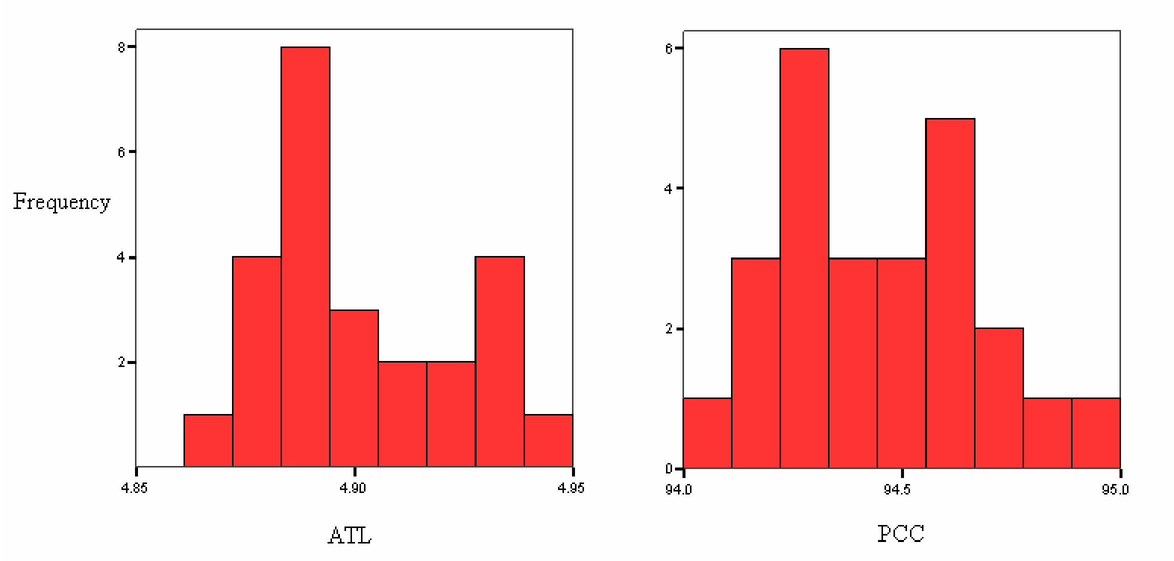
Results were summarized with an ANOVA model, with all replications included at each condition to assess sampling variability in the simulation. Before this summary, the ANOVA assumption of normality within conditions was evaluated by plotting the 25 replications of the first condition in Table 1. This was done because percentage and proportion data are commonly assumed to be uniformly distributed and therefore not appropriate for ANOVA-type analysis, especially with significance tests, although no significance testing was done in these analyses. The descriptive statistics are presented in Table 2, and frequency plots are depicted in Figures 8a and 8b.

Table 2: Descriptive Statistics of Replications in Condition 1

Variable	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
ATL	25	4.87	4.94	4.901	0.020	0.514	-0.950
PCC	25	94.19	94.98	94.500	0.198	0.458	-0.229

Obviously, PCC and ATL were not uniformly distributed, but had an approximately normal distribution, as indicated by the skewness and kurtosis values in Table 2. Moreover, since no significance testing was performed, due to the experimenter control of *N*, the normality assumption was of less importance.

Figure 8: ATL and PCC Distribution in Condition 1



a. ATL

b. PCC

The ANOVA summary for ATL is presented in Table 3, and PCC in Table 4. Several general results are of note. First, the amount of sampling error was very small. The error term η^2 for ATL was less than 0.001, and the error term η^2 for PCC was 0.010, implying that the simulation was stable across replications. If the assumptions employed in the administration of the monte carlo simulations are considered acceptable, such as the responses being generated by comparing U(0,1) random numbers to 3PL and GPCM probabilities, then the PCC and ATL of the CCTs was highly stable within each condition. Each of the five independent variables had a main effect η^2 of at least 0.050 with at least one of the dependent variables, with the exception of item selection.

Table 3: ANOVA Summary for ATL

Source of Variation	SS	df	MS	η^2
Main Effects				
Bank	7.182	1	7.182	<0.001
Termination Criterion	5242.880	1	5242.880	0.064
Item Selection Method	22.842	1	22.842	<0.001
IRT Model	14591.786	1	14591.786	0.177
Number of Cutscores	44778.573	1	44778.573	0.544
Two-way Interactions				
Bank × Termination	41.078	1	41.078	<0.001
Bank × Selection	47.678	1	47.678	0.001
Bank × Model	9.768	1	9.768	<0.001
Bank × Cutscores	16.411	1	16.411	<0.001
Model × Cutscores	8120.859	1	8120.859	0.099
Selection × Model	14.694	1	14.694	<0.001
Selection × Cutscores	0.028	1	0.028	<0.001
Termination × Selection	16.687	1	16.687	<0.001
Termination × Model	2753.710	1	2753.710	0.033
Termination × Cutscores	4290.436	1	4290.436	0.052
Three-way Interactions				
Bank × Termination × Selection	0.511	1	0.511	<0.001
Bank × Termination × Model	36.312	1	36.312	<0.001
Bank × Selection × Model	29.101	1	29.101	<0.001
Bank × Termination × Cutscores	32.120	1	32.120	<0.001
Bank × Model × Cutscores	0.022	1	0.022	<0.001
Bank × Selection × Cutscores	14.376	1	14.376	<0.001
Selection × Model × Cutscores	0.949	1	0.949	<0.001
Termination × Selection × Cutscores	14.537	1	14.537	<0.001
Termination × Model × Cutscores	2148.729	1	2148.729	0.026
Termination × Selection × Model	5.763	1	5.763	<0.001
Four-way Interactions				
Bank × Termination × Model × Cutscores	29.025	1	29.025	<0.001
Bank × Termination × Selection × Cutscores	0.003	1	0.003	<0.001
Bank × Termination × Selection × Model	0.565	1	0.565	<0.001
Bank × Selection × Model × Cutscores	14.537	1	14.537	<0.001
Termination × Selection × Model × Cutscores	4.440	1	4.440	<0.001
Five-way Interaction				
Bank × Termination × Selection × Model × Cutscores	0.677	1	0.677	<0.001
Error	18.727	768	0.024	<0.001
Total	82305.007	799		

Table 4: ANOVA Summary for PCC

Source of Variation	SS	df	MS	η^2
Main Effects				
Bank	111.288	1	111.288	0.073
Termination Criterion	0.164	1	0.164	<0.001
Item Selection Method	13.760	1	13.760	0.009
IRT Model	1212.978	1	1212.978	0.793
Number of Cutscores	33.293	1	33.293	0.022
Two-way Interactions				
Bank × Termination	0.666	1	0.666	<0.001
Bank × Selection	15.015	1	15.015	0.010
Bank × Cutscores	36.159	1	36.159	0.024
Bank × Model	47.103	1	47.103	0.031
Model × Cutscores	3.058	1	3.058	0.002
Selection × Cutscores	0.704	1	0.704	<0.001
Selection × Model	12.575	1	12.575	0.008
Termination × Selection	0.285	1	0.285	<0.001
Termination × Cutscores	0.249	1	0.249	<0.001
Termination × Model	0.073	1	0.073	<0.001
Three-way Interactions				
Bank × Termination × Model	0.525	1	0.525	<0.001
Bank × Termination × Selection	0.270	1	0.270	<0.001
Bank × Model × Cutscores	8.976	1	8.976	0.006
Bank × Selection × Model	11.410	1	11.410	0.007
Bank × Termination × Cutscores	0.032	1	0.032	<0.001
Bank × Selection × Cutscores	1.148	1	1.148	0.001
Selection × Model × Cutscores	0.779	1	0.779	0.001
Termination × Selection × Model	0.003	1	0.003	<0.001
Termination × Model × Cutscores	0.710	1	0.710	<0.001
Termination × Selection × Cutscores	0.307	1	0.307	<0.001
Four-way Interactions				
Bank × Termination × Selection × Cutscores	0.018	1	0.018	<0.001
Bank × Termination × Selection × Model	0.022	1	0.022	<0.001
Bank × Termination × Model × Cutscores	0.063	1	0.063	<0.001
Bank × Selection × Model × Cutscores	1.483	1	1.483	0.001
Termination × Selection × Model × Cutscores	0.000	1	0.000	<0.001
Five-way Interaction				
Bank × Termination × Selection × Model × Cutscores	0.433	1	0.433	<0.001
Error	15.258	768	0.020	0.010
Total	1528.807	799	1513.55	

Additionally, for both dependent variables, more than half the variance was due to a single independent variable. For ATL, 54.4% of the variance was attributable to the number of cutscores with the two-cutscore situation understandably requiring a higher average test length; the ATL for two cutscores (18.874) was higher than for one cutscore (3.911) (Table 5). For PCC, Table 4 shows that 79.3% of the variance was due to IRT Model, where the GPCM classified examinees more accurately (97.949) than the 3PL (95.486) (Table 5).

The means and standard deviations for each main effect are presented in Table 5. The ATL standard deviations were relatively high due to the high ATL for a few conditions (Table 2). The most notable result in this table is that the GPCM required less than half as many items as the 3PL (7.122), while still classifying examinees more accurately (15.664).

Table 5: Main Effects Means and SDs for PCC and ATL

	PCC		ATL	
Variable	Mean	SD	Mean	SD
Bank				
Peaked	97.091	1.09	11.298	9.91
Uniform	96.345	1.54	11.488	10.39
Termination Criterion				
AMT	96.732	1.39	13.953	12.56
SPRT	96.703	1.37	8.833	5.95
Item Selection				
CB	96.849	1.18	11.562	10.30
EB	96.587	1.55	11.224	10.01
IRT Model				
3PL	95.486	0.82	15.664	12.25
GPCM	97.949	0.34	7.122	4.42
Number of Cutscores				
1 Cutscore	96.514	1.23	3.911	1.17
2 Cutscores	96.922	1.50	18.874	9.63

Item Banks

The shape of the item bank – peaked vs. uniform – had virtually no effect on ATL ($\eta^2 < 0.001$) and a moderate effect on PCC ($\eta^2 = 0.073$). The uniform bank required an average of 0.19 more items, while producing a slightly lower PCC (96.345 vs. 97.091; Table 5). The lack of difference is not surprising, given that there was a substantial amount of information available in both banks for the CCT algorithm to use.

The primary reason that item bank was included as a variable was to examine its interaction with the item selection criteria. Theoretically, CB item selection should be more efficient with a bank that has an information function peaked at the cutscore(s), while a more uniform bank is more appropriate for EB item selection. However, the interaction of Item Selection and Bank had little effect relative to other effects in the ANOVA (ATL $\eta^2 = 0.001$, PCC $\eta^2 = 0.010$). This might be due to the fact that not enough information was available for all examinees, as discussed below.

Item bank did have a slight interaction with the number of cutscores (PCC $\eta^2 = 0.024$). The uniform bank produced equivalent PCC for both situations (Table 6), but the peaked bank resulted in higher PCC for the two-cutscore situation (97.507) as opposed one cutscore (96.674). This is likely due to the fact that the three-classification peaked bank was different than the two-classification peaked bank, with a bimodal distribution rather than limiting available information to a narrow region around a single cutscore. This is especially interesting since the addition of one more cutscore increased the burden on the CCT; the bimodal distribution offset this effect to the point that the PCC was higher in the three-classification case.

Table 6: Interaction of Item Bank and
Number of Cutscores for ATL and PCC

	ATL		PCC	
Item Bank	Mean	SD	Mean	SD
Peaked				
1 Cutscore	3.960	2.61	96.674	7.98
2 Cutscores	18.636	22.03	97.507	5.01
Uniform				
1 Cutscore	3.863	2.50	96.353	8.12
2 Cutscores	19.112	23.45	96.336	7.01

Item bank also had a moderate interaction with IRT model (PCC $\eta^2 = 0.031$). For the uniform bank, the GPCM had PCC that was nearly 3% higher than the 3PL (7). For the peaked bank, this difference was reduced to less than 2%, as the 3PL PCC increased by about 1% while the GPCM PCC remained essentially the same. This implies that the GPCM provided enough information across θ with both banks, while the 3PL performed better with the concentration of information in a single region near the cutscore.

Table 7: Interaction of Item Bank and
IRT Model for ATL and PCC

	ATL		PCC	
Item Bank	Mean	SD	Mean	SD
Peaked				
3PL Model	15.679	21.83	96.102	8.01
GPCM	6.917	9.21	98.079	4.79
Uniform				
3PL Model	15.648	23.22	94.871	9.15
GPCM	7.327	9.92	97.819	5.20

Termination Criterion

Termination criterion had the opposite outcome of item bank, with no effect on PCC ($\eta^2 < 0.001$) but a moderate effect on ATL ($\eta^2 = 0.064$). This was due to AMT requiring 5.12 more items on average (Table 5). The lack of difference in PCC was due to the fact that AMT and SPRT were matched on PCC to provide a comparison of ATL on even ground, which was achieved in that respect.

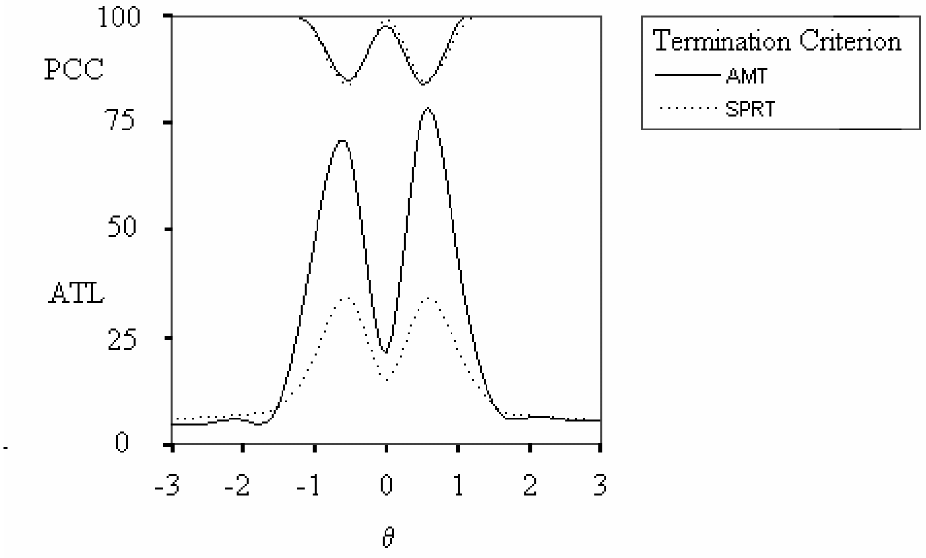
Upon closer examination of the data, it was found that the reason AMT required more items on average was that the procedure had a difficult time classifying examinees that were very close to the cutscore in the three-classification situation (Table 8). Examinees at θ = 0.50 or -0.50 were only 0.175 units from the cutscores of 0.675 or -0.675. For AMT to make a 95% confidence interval decision on an examinee with an estimated θ of 0.50, the CSEM had to be less than 0.175/1.96 = 0.089. The moderately sized item banks used did not have enough information available to reduce CSEM to this level, so these examinees were administered every item in the bank, increasing the average ATL for AMT (Figure 9). Due to this specific situation, if the bank had contained many more highly discriminating items, this difference between AMT and the

SPRT would most likely have been reduced. In any case, AMT still performed extremely efficiently, classifying examinees with an average of less than 14 items and a PCC of 96.732, and had a lower ATL than the SPRT at the extremes of θ with three classifications (Figure 9).

Table 8: Conditional Means and Standard Deviations of ATL for Values of θ From -3.00 to +3.00

	AMT				SPRT			
	1.00		2.00		1.00		2.00	
θ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
-3.00	2.10	1.17	4.72	2.59	1.63	0.73	6.13	4.12
-2.50	2.12	1.16	4.92	2.69	1.64	0.71	6.39	3.92
-2.00	2.18	1.14	5.69	3.01	1.69	0.69	7.00	3.69
-1.50	2.32	1.09	9.52	4.66	1.83	0.64	9.01	3.42
-1.00	2.61	1.04	48.10	23.04	2.12	0.59	20.73	4.42
-0.50	3.25	1.12	67.32	27.77	2.72	0.68	33.95	7.68
0.00	4.96	1.39	21.15	10.65	4.29	0.99	15.14	5.19
0.50	9.59	0.77	76.68	34.95	8.73	0.91	33.47	6.94
1.00	8.00	1.37	43.58	20.66	7.78	1.29	22.41	6.04
1.50	4.92	1.79	9.32	4.47	4.68	1.53	9.12	3.07
2.00	4.13	1.88	6.31	3.66	3.72	1.46	6.81	3.04
2.50	3.94	1.91	5.78	3.62	3.47	1.45	6.05	3.22
3.00	3.89	1.91	5.66	3.60	3.40	1.43	5.77	3.34

Figure 9: ATL and PCC as a Function of θ and Termination Criterion for Two Cutscores



The interaction of termination criterion and item selection method was of primary interest. This effect was also very small (ATL $\eta^2 < 0.001$, PCC $\eta^2 < 0.001$). EB selection required slightly fewer items for both termination criteria (Table 9).

Table 9: Interaction of Termination Criterion
and Item Selection for ATL and PCC

	ATL		PCC	
Item Selection	Mean	SD	Mean	SD
CB				
AMT	14.266	23.10	96.882	7.02
SPRT	8.857	9.66	96.816	6.91
EB				
AMT	13.639	23.05	96.582	7.48
SPRT	8.808	9.37	96.591	7.20

This result was also likely related to the above situation regarding examinees near the cutscores; under ideal conditions, EB item selection should have performed better with AMT. However, in this case, AMT was able to easily classify examinees with the exception of the single group of examinees near the cutscore, regardless of item selection method. Examinees at that point were administered the entire item bank with both selection methods, substantially increasing the average ATL for both and likely clouding any comparison.

Item Selection

The main effect of item selection was minimal with both ATL ($\eta^2 < 0.001$) and PCC ($\eta^2 = 0.009$). CB selection used 0.338 items more on average, but also classified with slightly more accuracy, with 0.262 higher PCC (Table 5). As previously discussed, this might be related to the fact that more items were needed to classify examinees very near the cutscore than were in the item bank. Item selection method also had very little interaction with the other independent variables, the largest being a minimal interaction with IRT model (PCC $\eta^2 = 0.008$). The two item selection methods performed comparably with the GPCM in terms of ATL and PCC, but for the 3 PL, the PCC for EB selection was 0.513 lower than for CB selection (Table 10).

Table 10: Interaction of IRT Model and
Item Selection for ATL and PCC

	ATL		PCC	
Item Selection	Mean	SD	Mean	SD
CB				
3PL Model	15.968	22.61	95.743	8.32
GPCM	7.155	9.56	97.955	5.05
EB				
3PL Model	15.968	22.46	95.230	8.91
GPCM	7.155	9.59	97.943	4.95

IRT Model

IRT model had a large effect on ATL ($\eta^2 = 0.177$) and a very large effect on PCC ($\eta^2 = 0.793$). The large effect on ATL was due to the GPCM requiring less than half as many items (7.122) as the 3PL (15.664) on average (Table 5). While substantial, this was actually smaller than expected; it was expected that the GPCM would need approximately one-third as many items, since it contained three times as much information per item. However, this shortfall was due to the sizable difference in PCC; the PCC with the GPCM (97.949) was much higher than the PCC with the 3PL (95.486). Note that the PCC with the GPCM was above the nominal rate of 95%; the procedure was so efficient that it actually classified examinees more accurately than it was designed for. If the nominal PCC rate was relaxed to bring the observed PCC down to 95%, it would require even fewer items, and if it used only 2 fewer items on average, GPCM CCTs would indeed use one-third as many items as 3PL CCTs. So, when this allowance is made, the expectation was supported.

IRT model also had modest two-way ATL interactions with termination criterion (η^2 = 0.033) and the number of classifications (η^2 = 0.099). When the 3PL was the psychometric model, the SPRT had an ATL of 11.248 and AMT had an ATL of 20.079, a difference of 8.831 items (Table 11).

Table 11: Interaction of IRT Model and Termination Criterion for ATL and PCC

	ATL		PCC	
Termination Criterion	Mean	SD	Mean	SD
AMT				
3PL Model	20.079	29.35	95.510	8.78
GPCM	7.827	11.33	97.954	5.02
SPRT				
3PL Model	11.248	10.74	95.463	8.47
GPCM	6.417	7.36	97.944	4.98

However, for the GPCM, these were 6.417 and 7.827, respectively, a difference of only 1.41 items. This substantial reduction in difference was again likely due to the fact that not enough information was available in the item bank for the AMT at examinee θ values near the cutscore, causing most CCTs for those examinees to administer every item in the bank. Because the GPCM bank had one-third as many items, the overall average was lower.

The interaction between IRT model and proportion of correct classifications was also of primary interest in this study, and the hypothesis was supported that polytomous IRT’s advantage in efficiency would be greater for the two-cutscore case than the one-cutscore case (Table 12).

Table 12: Interaction of IRT Model and Number of Cutscores for ATL and PCC

	ATL		PCC	
Number of Cutscores	Mean	SD	Mean	SD
1 Cutscore				
3PL Model	4.996	2.37	95.344	9.58
GPCM	2.827	2.26	97.683	5.94
2 Cutscores				
3PL Model	26.331	27.97	95.629	7.56
GPCM	11.417	11.89	98.215	3.82

For the single-cutscore case, the 3PL ATL was 4.996 and the GPCM ATL was 2.827, for a reduction of approximately 44%. In the two-cutscore case the 3PL ATL was 26.331 and the

GPCM ATL was 11.417, for a reduction of approximately 57%. This interaction accounted for nearly 10% of the variance in ATL, and is seen in Figures 10a and 10b, where the disparity in ATL was much greater for the three-classification case. Additionally, there was a moderate three-way interaction of IRT model, number of classifications, and termination criterion (ATL η^2 = 0.026), where the advantage of the GPCM was even greater for AMT (Figure 11).

Figure 10: ATL and PCC as a Function of θ and IRT Model for

a. One Cutscore b. Two Cutscores

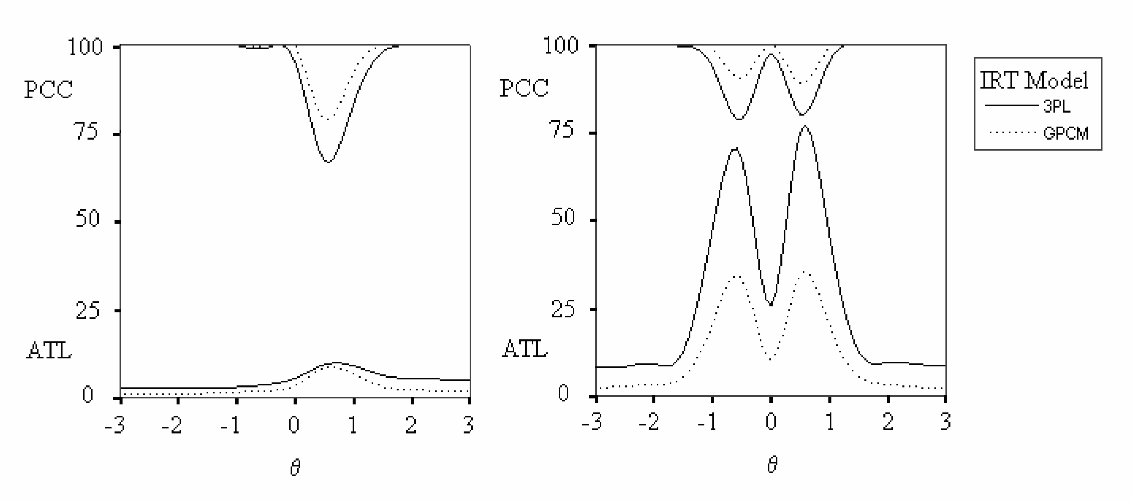
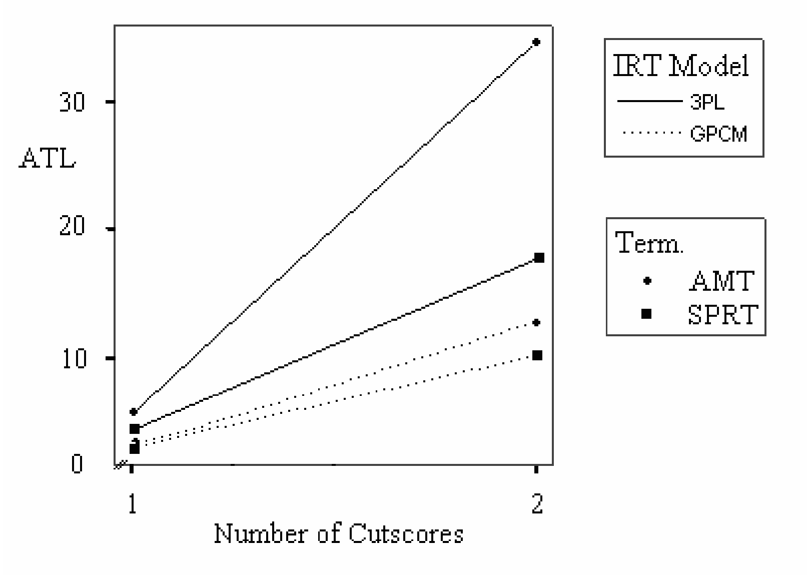


Figure 11: Interaction of IRT Model, Termination Criterion, and Number of Cutscores



Number of Cutscores

The number of cutscores had a moderate effect on PCC (η^2 = 0.022) and a substantial effect on ATL (η^2 = 0.544). The PCC for two cutscores (96.922) was slightly higher than for one (96.514), and the ATL for two (18.874) was much higher than for one (3.911) (Table 5). The number of cutscores also had an interaction with termination criterion (ATL $\eta^2 = 0.052$). There was very little difference between the termination criteria in terms of ATL for the single-cutscore case, with an AMT ATL of 4.155 and an SPRT ATL of 3.667 (Table 13). However, the difference was large for the two-cutscore case, with an AMT ATL of 23.750 and a SPRT ATL of 13.999. Again, this might be due to the issue of examinees close to a cutscore in the two-cutscore case.

Table 13: Interaction of Termination Criterion and Number of Cutscores for ATL and PCC

	ATL		PCC	
Termination Criterion	Mean	SD	Mean	SD
AMT				
1 Cutscore	4.155	2.64	96.510	8.25
2 Cutscores	23.750	29.43	96.954	6.09
SPRT				
1 Cutscore	3.667	2.44	96.517	7.85
2 Cutscores	13.998	11.03	96.890	6.16

Chapter 5: Discussion and Conclusions

The 32 conditions in the study resulted in 5 main effects and 26 interactions. The majority of these accounted for very small proportions of the total variance, as apparent in the ANOVA summary tables. However, the primary expectation of the study was supported – that the advantage in efficiency offered by the GPCM would be even greater for the two-cutscore case than for the one-cutscore case. This was true for both termination criteria, the SPRT and AMT, because the GPCM offers more information across θ, which is advantageous to AMT, and at any given cutscore on θ, which is advantageous to the SPRT.

Main Effects

Several main effects had substantial strength. The most obvious effect was that ATL was much higher for the two-cutscore case than for one cutscore, as found in previous multiple-cutscore research (Spray, 1993; Eggen & Straetmans, 2000). The presence of a second cutoff required the administration of more than four times as many items (18.874), on average, than the single cutscore case (3.911) (Table 5). Even with more than four times as many items, the PCC for the two-cutscore case was only 0.412 higher.

Unlike the other independent variables, however, the number of cutscores is not often a choice that is presented to a test developer, but is usually necessitated by the intended application. Once the number of cutoffs required for a particular application is established, several choices are presented such as termination criterion and item selection algorithm. Therefore, this effect is of little practical importance since it often cannot be avoided. In any case, test developers that are aware of the greater test length of multiple-cutscore tests can ensure that adequate time is allotted, and the item bank has a sufficient amount of information to accurately classify examinees.

The GPCM, because of the greater amount of information per item, required less than half as many items (7.122) as the 3PL model (15.664) (Table 5). This supports previous research (Lau & Wang, 1998; 1999) and also the use of GPCM-scored items that can be administered in a reasonable amount of time. If GPCM-scored items for a certain test require three or four times as much time for an examinee to respond, this advantage in ATL is of little practical use.

The SPRT termination criterion required fewer items (8.833) to make a decision than AMT (13.953) with only a 0.029 difference in PCC. However, this effect is increased by the issue noted in Chapter 4, where examinees with θ values of 0.5 and -0.5 were administered all the items in the bank because the bank did not have enough information at that point to make a decision with AMT. If cutscores were more distant from these examinees, such as 0.75, the results might have been different, as they might have if more highly discriminating items were used.

CB item selection resulted in slightly higher ATL (11.562) than EB item selection 11.224, although it also led to slightly more accurate classification of examinees (Table 5). There was also a relatively small effect of item bank, with the uniform bank resulting in slightly higher ATL (11.488) than the peaked bank (11.298), even though the peaked bank resulted in higher PCC (97.091) than the uniform bank (96.345) (Table 5). This supports the use of a bank with more information near the cutscores, especially since the “uniform” bank was only uniform in its distribution of item location parameters and not the bank information function; the information function was still bell-shaped (Figure 7). The small main effects for these two variables are not notable, as they were included in the study for possible interaction with the remaining dependent variables.

Interactions

The primary expectation of the current study was demonstrated, that polytomous CCTs are very efficient in their classification of examinees, and this was especially true for the multiple cutscore case. With two cutscores, examinees were classified by the polytomous CCT with an ATL of only 11.418 items and a PCC of 98.215, compared with dichotomous ATL of 26.331 and PCC of 95.629 (Table 12). The advantages were even greater when AMT was the termination criterion of the CCT, again possibly due to the inability of AMT to make a decision for the two-cutscore case with dichotomous IRT.

The interaction of IRT model and the number of cutscores is important for two types of practical applications, both with one cutscore and multiple cutscores: testing programs that currently use CCT but not polytomous IRT, and testing programs that currently use polytomous IRT but not CCT. The former is the common situation of testing programs that use variable-length CCT methods to administer tests, but only make use of the ubiquitous dichotomously-scored multiple-choice item. But while polytomous IRT might offer substantial reductions in testing time with increased accuracy, not all testing applications lend themselves to the GPCM as it requires an explicit ordering of the item responses. The nominal response IRT model (Bock, 1972) represents an alternative option, where the item responses are not explicitly ordered but still scored polytomously. The application of the nominal response model to CCT offers an opportunity for future research.

The latter situation is not as common, but represents the more appropriate opportunity for polytomous CCT. In this situation, a polytomous IRT model such as the GPCM is already assumed as the psychometric model. However, the examinations are being administered as conventional tests in which the examinee receives the same items in the same order, often with substantial numbers of items, especially if the test is primarily dichotomously scored and only a few polytomous items are present. This is the case in testing programs that principally rely on dichotomous multiple-choice items, but also require a few higher-fidelity polytomous items such as an open-response item or a performance simulation.

The second most important result was the interaction of termination criterion and number of cutscores. In the one-cutscore case, the ATL for AMT (4.155) was only a fraction of an item more than the ATL for the SPRT (3.667), though this might be related to ATL being so low that it prevented any difference from being more noticeable. The PCC for these two cases was also similar, with AMT PCC of 96.510 and SPRT PCC of 96.517 (Table 13). Yet for the two-cutscore case, the difference between AMT and the SPRT was substantial: the ATL for AMT was 23.750 and the ATL for the SPRT was 13.999, a difference of nearly 10 items. PCC remained similar at 96.954 for AMT and 96.890 for the SPRT. This suggests that while AMT might be highly efficient and commonly used for one-cutscore CCT, it would be more advantageous for the test user to use the SPRT as the termination criterion for multiple-cutscore CCT. However, it may be due to the aforementioned situation of there not being enough information in the bank at the two cutscores to classify those examinees whose θ value falls near the cutscore.

The next largest interaction with ATL as the dependent variable was termination criterion and IRT model, with results similar to the interaction with number of cutscores. As found in previous CCT research with dichotomous IRT, the SPRT used fewer items (11.248) than AMT (20.079) but also had lower PCC (95.463) than AMT (95.510) (Table 11), though it should be noted that this might again be due to the lack of information in the dichotomous bank near the cutscores. With the GPCM, the difference in ATL was much less (6.417 vs. 7.827), but the SPRT was more accurate (97.944) than AMT (95.510). When considered in conjunction with the interaction of termination criterion and number of cutscores, it appears that termination criterion is of little consequence for single-cutscore tests with the GPCM.

An important expected effect that was not found was an interaction between termination criterion and item selection method. It was expected that AMT would perform better with EB selection and SPRT with CB selection. Since this does not agree with the results of previous research (Spray & Reckase, 1994; Eggen, 1999; Eggen & Straetmans, 2000), it also presents an appropriate target for future research. Lack of effect might be due to the previously described situation with examinees near the cutoff. The bank information functions might have also played a part.

As expected, ATL with the GPCM (7.122) was much lower than ATL with the 3PL (15.664), as the GPCM offered a considerable advantage in available information. This was previously demonstrated in the one-cutscore case by Lau and Wang (1998; 1999; 2000). The difference between GPCM and the 3PL in the current study was accentuated by the fact that GPCM conditions had noticeably higher PCC even though they required less than half as many items. IRT model does present a choice to the test developer, though this is occasionally restricted by the content of the test. If items cannot easily be written so that the responses represent ordered steps, then an alternative IRT model must be chosen. Frequently, this is the 3PL, as it is appropriate for many types of multiple-choice items where examinee guessing is a factor, but other polytomous IRT models exist that can be applied, such as the nominal response model.

Additionally, the observed difference in nominal and observed PCC with the GPCM is of practical significance. When designing a CCT with the GPCM, it is important for the test designer to remember that the observed error might result in less than nominal error. In this case, it was less than half the nominal error of 5%. In the case where the purpose of the test is to screen a large number of examinees, and a modest amount of misclassification is acceptable, the nominal PCC might be relaxed so that classifications can be made with fewer items. This effect differs from all others in that it is a comparison of nominal and observed error *within* each condition, rather than the comparison of observed error across conditions.

Future Research

Although this study had a discrete uniform θ distribution so that results could be examined conditional on θ, future research could consider a continuous distribution of examinees. This would likely affect observed PCC, as large numbers of examinees would not be clustered in close proximity to the cutscores used in this study, though other studies would likely use different cutscores. However, the researcher would need to justify the use of a specific distribution, such as if it is previously established as in Eggen and Straetmans (2000).

Additionally, an interesting variable would be polytomous/dichotomous model misspecification. IRT model misspecification in terms of assumed dimensions and number of parameters has been previously investigated with the SPRT termination criterion. Although IRT-based CCT with the SPRT is fairly robust in terms of the assumption of unidimensionality (Spray, Abdel-fattah, Huang, and Lau, 1997; Lau, 1998), it has been shown not to be robust to the choice of the number of IRT model parameters (Reckase, 1983; Kalohn & Spray, 1999; Jiao & Lau, 2003). The current study used separately generated sets of items that were explicitly 3PL or GPCM, but it is possible to take a data set of items that is able to be calibrated with a polytomous model and also calibrate it with a dichotomous model.

The lack of involvement of real data is itself a limitation of the current study. Monte carlo simulations are useful and appropriate for preliminary research on a topic, but the absence of real data hampers generalizability. A post hoc (real-data) simulation methodology would be quite useful for extending the line of research on polytomous multiple-cutscore CCT. This would obviously provide greater fidelity to a specific application, but it would also enable the even comparison of polytomous IRT and dichotomous IRT, because the sample could be used to calibrate the items with competing models. The use of a subset of the sample in a cross-validation approach might not be necessary, as that approach has little effect on post hoc simulation in one study (Thompson & Weiss, 2007). The limitation with post hoc simulation is the requirement of a large enough data set to obtain accurate item parameter estimates.

Three other suggestions for future research are related to items and their IRT parameters. Future research along these lines should use larger item banks. A smaller bank size was chosen because past research (Lau & Wang, 1998; 1999; 2000) indicated that very few items are necessary, on average, to obtain accurate decisions. However, in an effort to make more information available at regions of θ where it was most needed, another option would be to generate a larger item bank, but place a maximum constraint on test length so that the entire large bank is not administered to some examinees.

To provide a clearer comparison between bank characteristics, future research could generate item banks to reflect specific intended characteristics. Uniform bank information functions could be higher and flatter, to provide more opportunity for adaptive item selection to select highly discriminating items across θ. Likewise, the banks with bimodal distributions of item location/difficulty parameters could have bimodal information functions. This can be accomplished by having more items in a narrower region around each cutoff.

Additionally, this study provided an optimal situation for polytomous multiple-cutscore CCT because the distance between item boundaries was specifically chosen so that items would provide a large amount of information at both cutoffs simultaneously. In an actual testing program, it is unrealistic to expect this to be the case with every item. A more realistic item pool, especially with real items, would provide a more accurate assessment of polytomous multiple-cutscore CCT in realistic situations. However, the use of a real item pool would preclude any manipulation of the bank information function.

Future research should also explore the gains in efficiency offered by polytomous CCT in the four-classification case. This is the scoring situation of the NAEP examinations (Loomis & Bourque, 2001), which classifies examinees as *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The NAEP is a nationwide testing program where the cost/benefit ratio of developing a CCT program would be favorable, as it becomes more beneficial for larger-scale programs. It can also be applied to testing for five categories, which represents the use of CCT to assign academic grades (Kingsbury & Weiss, 1984). However, it must be noted that the efficiency of the test will likely decrease with an increase in the number of cutscores (Spray, 1993).

In addition, the relationship between nominal and observed error rates should be explored further. Previous research has found a difference (e.g., Spray, 1993; Lin and Spray, 2000), but has yet to be explored as the primary focus of a study to determine the factors that play a major role in affecting this difference. A higher amount of information available in the item bank would reduce observed error. So would a cutscore that falls nearer to the extremes of the examinee distribution, which would make it easier to classify most examinees. Maximum test lengths, on the other hand, end the test before the termination criterion can be satisfied, increasing error (Spray, 1993; Rudner, 2002).

Future research should also consider the application of composite hypotheses with the SPRT (Weitzman, 1982), where the statistical test evaluates whether the examinee's θ is in the region of θ above or below the cutscore, rather than equal to the arbitrarily defined points θ_1 and θ_2. This more accurately reflects the conceptual goal of CCT. Additionally, this approach is equivalent to that of AMT under certain conditions (Thompson, 2007b), leading to a more unified theory of CCT.

Conclusions

This study demonstrated that the advantages offered by CCT with a polytomous IRT model (Lau & Wang, 1998; 1999; 2000) are present to an even greater extent with CCT with multiple cutscores. Future research is necessary to further explore the relationship of secondary independent variables to this effect, but it is important to note just how efficient multiple-cutscore CCTs were with polytomous IRT: only 11.418 items were needed, on average, and 98.215% of the simulated examinees were correctly classified. The savings of testing time and expense across large numbers of examinees would be considerable if this CCT model was applied, and it presents a significant opportunity to testing programs where the model is feasible.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, 12*, 137-144.

Braun, H., Bejar, I.I., and Williamson, D.M. (2006). *Rule-based methods for automated scoring: Application in a licensing context*. In Williamson, D.M., Mislevy, R.J., and Bejar, I.I. (Eds.) Automated scoring of complex tasks in computer-based testing. Mahwah, NJ: Erlbaum.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Buchanan, P.G., Vucinic, T.J., Rigos, J.J., & Gleim, I.N. (2004). Preparing for success on the revised CPA exam. *Journal of Accountancy Online Issues, January 2004*. Retrieved in Feb. 5, 2007, from http://www.aicpa.org/PUBS/JOFA/jan2004/cpaexam.htm.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.

Eggen, T. J. H. M, & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*, 713-734.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Epstein, K. I., & Knerr, C. S. (1977). Applications of sequential testing procedures to performance testing. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1977.

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

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

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

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

Gifford, J. A., & Hambleton, R. K. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement*, *14*, 33-43.

Govindarajulu, Z. (1987). *The sequential statistical analysis of hypothesis testing, point and interval estimation, and decision theory*. Columbus, OH: American Sciences Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Huang, C.-Y., Kalohn, J.C., Lin, C.-J., and Spray, J. (2000). *Estimating item parameters from classical indices for item pool development with a computerized classification test*. (Research Report 2000-4). Iowa City, IA: ACT, Inc.

Jacobs-Cassuto, M.S. (2005). *A comparison of adaptive mastery testing using testlets with the 3-parameter logistic model*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Jiao, H., & Lau, A. C. (2003). *The effects of model misfit in a computerized classification test.* Paper presented at the annual meeting of the National Council of Educational Measurement, Chicago, IL, April 2003.

Jiao, H., Wang, S., & Lau, C. A. (2004). *An investigation of two combination procedures of SPRT for three-category classification decisions in a computerized classification test*. Paper presented at the annual meeting of the American Educational Research Association, San Antonio, April 2004.

Kalohn, J. C., & Spray, J. A. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement, 36*, 47-59.

Kingsbury, G.G., & Weiss, D.J. (1979). An adaptive testing strategy for mastery decisions. Research report 79-05. Minneapolis: University of Minnesota, Psychometric Methods Laboratory.

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 theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

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

Lau, C. A., & Wang, T. (1998). *Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Lau, C. A., & Wang, T. (1999). *Computerized classification testing under practical constraints with a polytomous model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Lau, C. A., & Wang, T. (2000). *A new item selection procedure for mixed item type in computerized classification testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

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

Lin, C.-J. & Spray, J.A. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. (Research Report 2000-8). Iowa City, IA: ACT, Inc.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1972). Sequential testing for dichotomous decisions. *Educational and Psychological Measurement, 32,* 85-95.

Loomis, S. C., & Borque, M. L. (Eds.) (2001). *National Assessment of Education Progress Achievement Levels 1992-1998 for Science*. U.S. Department of Education.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

O'Neill, T. R., Marks, C. & Liu, W. (2006). *Assessing the Impact of English as a Second Language Status on Licensure Examinations*. CLEAR Exam Review, 17(1), p. 19-23.

Ostini, R., & Nering, M. L. (2005). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

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

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2006). *Practical considerations in computer-based testing*. New York: Springer.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Rudner, L. M. (2002). *An examination of decision-theory adaptive testing procedures*. Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

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

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

Sobel, M. & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics*, *20*, 502-522.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Research Report 93-7). Iowa City, Iowa: ACT, Inc.

Spray, J. A., Abdel-fattah, A. A., Huang, C., and Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional* (Research Report 97-5). Iowa City, Iowa: ACT, Inc.

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

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computerized adaptive test*. Paper presented at the Annual Meeting of the National Council for Measurement in Education (New Orleans, LA, April 5-7, 1994).

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

Sullum, J. (2004). Flower power: Free the florists. *Reason Magazine Online*, Retrieved February 5, 2007, from http://www.reason.com/news/show/35591.html.

Thissen, D. (2000). Reliability and measurement precision. In Wainer, H. (Ed.), *Computerized Adaptive Testing: A primer*. Mahwah, NJ: Erlbaum.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1). Available online: http://pareonline.net/getvn.asp?v=12&n=1

Thompson, N. A. (2007b). *Computerized classification testing with composite hypotheses*. Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Thompson, N. A., & Weiss, D.J. (2007). *Item selection with adaptive mastery testing*. Under revision.

van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological measurement* (pp. 129-156). Boston, MA: Kluwer-Nijhof.

Vos, H. J. (1998). Optimal sequential rules for computer-based instruction. *Journal of Educational Computing Research, 19*, 133-154.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24*, 271-292.

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

Warm, T. A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Weiss, D. J. (1974). Strategies of adaptive ability measurement. Technical Report 74-05. Minneapolis, MN: University of Minnesota Psychometric Methods Program.

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

Weissman, A. (2004). *Mutual information item selection in multiple-category classification CAT*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego, CA.

Weitzman, R. A. (1982a). Sequential testing for selection. *Applied Psychological Measurement, 6*, 337-351.

Weitzman, R. A. (1982b). Use of sequential testing to prescreen prospective entrants into military service. In D. J. Weiss (Ed.), *Proceedings of the 1982 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1982.

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

Appendix A: Dichotomous IRT Item Parameters

	Uniform	Normal	Bimodal	
Item	*b*	*b*	*b*	*a*
1	-3.00	0.28	-0.47	1.24
2	-2.97	1.30	-0.64	0.88
3	-2.93	-0.39	-0.89	1.13
4	-2.90	0.78	-0.65	1.02
5	-2.87	0.58	-0.64	0.87
6	-2.83	-0.24	-1.04	0.84
7	-2.80	0.94	-0.57	0.89
8	-2.77	-0.54	-0.81	1.08
9	-2.73	0.90	-0.84	0.87
10	-2.70	0.50	-0.51	1.22
11	-2.67	-0.14	-1.02	1.21
12	-2.63	-1.00	-0.88	1.12
13	-2.60	-0.14	-0.60	0.54
14	-2.57	0.04	-0.93	1.08
15	-2.53	-0.72	-0.71	0.70
16	-2.50	0.77	-0.28	0.97
17	-2.47	-1.08	-0.88	0.96
18	-2.43	-0.80	-0.58	1.14
19	-2.40	-2.30	-0.81	1.01
20	-2.37	-1.04	-0.85	1.20
21	-2.33	0.27	-0.94	0.82
22	-2.30	-0.69	-0.68	1.08
23	-2.27	0.23	-0.89	1.22
24	-2.23	1.25	-0.57	0.88
25	-2.20	-0.85	-0.26	1.15
26	-2.17	-0.71	-0.80	1.08
27	-2.13	-1.47	-0.89	0.98
28	-2.10	-0.50	-0.33	0.85
29	-2.07	2.03	-0.61	0.94
30	-2.03	0.02	-1.41	0.84
31	-2.00	0.23	-0.54	0.88
32	-1.97	1.07	-0.80	1.03
33	-1.93	0.59	-0.73	1.02
34	-1.90	-0.15	-0.92	1.19
35	-1.87	-0.78	-0.64	0.88
36	-1.83	-0.24	-0.04	0.95
37	-1.80	-0.09	-0.70	0.73
38	-1.77	-1.95	-0.54	1.04
39	-1.73	-0.60	-0.81	0.98
40	-1.70	-0.04	-0.51	1.01
41	-1.67	-1.41	-0.45	1.01
42	-1.63	-0.47	-0.47	0.94
43	-1.60	1.29	-0.46	1.28
44	-1.57	0.84	-0.77	0.71
45	-1.53	-0.18	-0.35	0.72
46	-1.50	0.08	-0.26	1.23
47	-1.47	-1.03	-0.98	0.89
48	-1.43	0.22	-1.64	0.99
49	-1.40	-0.27	-0.83	1.29
50	-1.37	-0.97	-0.59	1.10
51	-1.33	0.16	-0.25	1.09
52	-1.30	1.70	-0.45	1.11
53	-1.27	0.36	-0.63	0.98
54	-1.23	-0.28	-0.54	1.28
55	-1.20	0.38	-0.61	0.95
56	-1.17	0.17	-0.76	0.97
57	-1.13	0.53	-0.48	1.32
58	-1.10	-0.20	-0.69	1.26
59	-1.07	-0.12	-0.98	1.14
60	-1.03	-1.02	-0.82	1.02
61	-1.00	-0.93	-0.52	1.33
62	-0.97	-1.41	-0.69	0.66
63	-0.93	1.27	-0.43	1.04
64	-0.90	0.42	-0.84	0.81
65	-0.87	1.56	-0.63	1.04
66	-0.83	-1.08	-0.81	1.21
67	-0.80	-0.31	-0.63	0.94
68	-0.77	1.67	-0.26	0.73
69	-0.73	-0.52	-0.68	1.13
70	-0.70	1.41	-0.73	1.05
71	-0.67	-1.24	-0.35	1.02
72	-0.63	0.23	-0.93	1.20
73	-0.60	0.26	-0.78	1.05
74	-0.57	0.89	-0.48	0.81
75	-0.53	1.61	-0.75	1.03
76	-0.50	-0.09	-0.45	0.91
77	-0.47	0.14	-0.40	0.67
78	-0.43	-1.02	-1.10	0.94
79	-0.40	0.24	-1.18	1.13
80	-0.37	-0.41	-0.51	0.46
81	-0.33	-0.87	-0.81	1.19
82	-0.30	-0.33	-0.85	0.80
83	-0.27	-0.20	-0.38	1.09
84	-0.23	-1.23	-0.82	0.98
85	-0.20	-0.49	-0.71	0.62
86	-0.17	-0.97	-0.28	0.76
87	-0.13	-0.94	-0.48	0.62
88	-0.10	1.89	-0.63	0.90
89	-0.07	0.35	-0.72	0.92
90	-0.03	1.83	-0.82	0.86

Appendix A (cont.): Dichotomous IRT Item Parameters

	Uniform	Normal	Bimodal	
Item	*b*	*b*	*b*	*a*
91	0.00	-0.79	0.70	1.34
92	0.03	1.30	0.76	1.03
93	0.07	1.15	1.16	0.96
94	0.10	-0.32	0.77	0.56
95	0.13	-0.84	0.68	0.94
96	0.17	-0.30	0.64	1.10
97	0.20	0.56	0.55	0.96
98	0.23	2.51	0.67	0.81
99	0.27	1.13	0.44	1.17
100	0.30	1.18	0.53	0.71
101	0.33	-0.13	0.61	0.94
102	0.37	0.18	0.71	1.30
103	0.40	0.46	0.60	1.26
104	0.43	-0.97	1.01	1.27
105	0.47	0.55	0.46	0.90
106	0.50	-1.60	0.94	0.90
107	0.53	-0.72	0.83	1.24
108	0.57	0.07	0.23	1.05
109	0.60	-0.89	0.28	0.90
110	0.63	-0.61	0.83	1.05
111	0.67	0.33	0.71	1.06
112	0.70	-0.90	0.90	0.94
113	0.73	-0.60	0.78	0.94
114	0.77	0.82	0.59	0.84
115	0.80	0.68	0.92	0.75
116	0.83	0.35	0.90	1.00
117	0.87	-1.07	0.98	0.96
118	0.90	1.25	0.43	1.24
119	0.93	0.59	1.05	0.69
120	0.97	-0.26	0.61	0.86
121	1.00	-0.15	0.07	1.01
122	1.03	0.83	0.93	1.32
123	1.07	-0.82	0.93	0.85
124	1.10	1.26	0.47	0.80
125	1.13	0.06	0.61	1.17
126	1.17	1.50	0.67	1.06
127	1.20	-2.01	0.89	0.75
128	1.23	-0.37	0.54	0.76
129	1.27	0.04	0.85	1.11
130	1.30	-0.34	0.28	1.20
131	1.33	-1.41	0.70	0.53
132	1.37	0.81	0.51	0.95
133	1.40	0.03	0.83	1.18
134	1.43	0.91	0.94	1.02
135	1.47	0.05	0.60	1.36
136	1.50	0.35	0.42	0.93
137	1.53	-0.02	0.63	0.60
138	1.57	1.02	0.18	1.08
139	1.60	-0.19	0.64	0.68
140	1.63	0.00	0.99	0.90
141	1.67	-0.58	0.68	0.94
142	1.70	-0.29	0.77	0.93
143	1.73	-1.47	0.95	0.60
144	1.77	0.87	0.59	1.31
145	1.80	1.64	0.62	0.98
146	1.83	-1.31	0.27	0.87
147	1.87	1.00	0.75	0.98
148	1.90	0.35	0.65	0.65
149	1.93	0.33	1.31	1.12
150	1.97	0.15	0.40	0.63
151	2.00	-1.09	0.54	1.21
152	2.03	-0.49	0.17	0.76
153	2.07	-0.59	0.50	1.26
154	2.10	2.31	0.30	1.05
155	2.13	0.52	0.64	1.11
156	2.17	0.13	0.57	1.15
157	2.20	1.22	0.67	0.89
158	2.23	0.88	0.68	0.99
159	2.27	-0.58	0.52	1.10
160	2.30	-0.70	0.51	0.98
161	2.33	-0.01	0.53	0.78
162	2.37	-0.15	0.74	0.69
163	2.40	-1.03	0.74	1.05
164	2.43	1.49	0.42	0.86
165	2.47	0.69	0.40	1.28
166	2.50	0.19	0.74	0.88
167	2.53	-1.46	0.35	0.70
168	2.57	-1.35	1.16	1.29
169	2.60	-0.89	0.67	0.79
170	2.63	0.74	1.15	1.02
171	2.67	0.39	0.78	1.28
172	2.70	-0.16	0.89	1.21
173	2.73	-0.92	0.51	1.25
174	2.77	-0.94	0.70	0.86
175	2.80	-0.63	0.48	0.81
176	2.83	-0.10	0.92	0.92
177	2.87	0.72	0.79	0.52
178	2.90	-0.46	0.63	1.35
179	2.93	-0.93	0.96	1.19
180	2.97	0.97	0.64	0.69
181	3.00	0.33	0.53	1.04

Appendix B: Polytomous IRT Item Parameters

			Uniform			Normal			Bimodal	
Item	*a*	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
1	1.04	-4.35	-3.00	-1.65	0.14	1.49	2.84	-1.89	-0.54	0.81
2	0.83	-4.25	-2.90	-1.55	-1.04	0.31	1.66	-2.63	-1.28	0.07
3	1.14	-4.15	-2.80	-1.45	-1.40	-0.05	1.30	-2.28	-0.93	0.42
4	0.96	-4.05	-2.70	-1.35	-3.25	-1.90	-0.55	-2.08	-0.73	0.62
5	1.34	-3.95	-2.60	-1.25	-0.22	1.13	2.48	-2.17	-0.82	0.53
6	1.17	-3.85	-2.50	-1.15	-1.21	0.14	1.49	-1.87	-0.52	0.83
7	1.00	-3.75	-2.40	-1.05	-0.37	0.98	2.33	-2.24	-0.89	0.46
8	0.61	-3.65	-2.30	-0.95	-1.57	-0.22	1.13	-2.27	-0.92	0.43
9	1.04	-3.55	-2.20	-0.85	-0.11	1.24	2.59	-1.67	-0.32	1.03
10	0.88	-3.45	-2.10	-0.75	-1.44	-0.09	1.26	-1.58	-0.23	1.12
11	1.17	-3.35	-2.00	-0.65	-1.11	0.24	1.59	-1.74	-0.39	0.96
12	0.97	-3.25	-1.90	-0.55	-2.29	-0.94	0.41	-1.79	-0.44	0.91
13	0.97	-3.15	-1.80	-0.45	-3.08	-1.73	-0.38	-1.71	-0.36	0.99
14	0.97	-3.05	-1.70	-0.35	-1.30	0.05	1.40	-2.26	-0.91	0.44
15	0.91	-2.95	-1.60	-0.25	-0.43	0.92	2.27	-1.92	-0.57	0.78
16	1.38	-2.85	-1.50	-0.15	-0.86	0.49	1.84	-1.60	-0.25	1.10
17	1.03	-2.75	-1.40	-0.05	-0.57	0.78	2.13	-2.05	-0.70	0.65
18	1.02	-2.65	-1.30	0.05	-0.82	0.53	1.88	-1.55	-0.20	1.15
19	1.25	-2.55	-1.20	0.15	-1.65	-0.30	1.05	-1.99	-0.64	0.71
20	1.07	-2.45	-1.10	0.25	0.37	1.72	3.07	-2.01	-0.66	0.69
21	1.31	-2.35	-1.00	0.35	-1.38	-0.03	1.32	-2.81	-1.46	-0.11
22	0.96	-2.25	-0.90	0.45	-1.38	-0.03	1.32	-1.69	-0.34	1.01
23	1.37	-2.15	-0.80	0.55	-1.05	0.30	1.65	-1.93	-0.58	0.77
24	1.24	-2.05	-0.70	0.65	-0.86	0.49	1.84	-2.46	-1.11	0.24
25	1.30	-1.95	-0.60	0.75	-1.32	0.03	1.38	-2.06	-0.71	0.64
26	0.82	-1.85	-0.50	0.85	-1.33	0.02	1.37	-2.18	-0.83	0.52
27	0.73	-1.75	-0.40	0.95	-1.35	0.00	1.35	-1.95	-0.60	0.75
28	0.98	-1.65	-0.30	1.05	-0.24	1.11	2.46	-2.02	-0.67	0.68
29	1.08	-1.55	-0.20	1.15	-0.42	0.93	2.28	-2.18	-0.83	0.52
30	0.95	-1.45	-0.10	1.25	-2.28	-0.93	0.42	-1.78	-0.43	0.92
31	0.99	-1.35	0.00	1.35	0.38	1.73	3.08	-0.72	0.63	1.98
32	1.01	-1.25	0.10	1.45	-1.58	-0.23	1.12	-0.82	0.53	1.88
33	1.24	-1.15	0.20	1.55	-1.09	0.26	1.61	-0.88	0.47	1.82
34	1.40	-1.05	0.30	1.65	-0.11	1.24	2.59	-0.69	0.66	2.01
35	0.85	-0.95	0.40	1.75	-2.17	-0.82	0.53	-0.53	0.82	2.17
36	1.06	-0.85	0.50	1.85	-2.72	-1.37	-0.02	-0.59	0.76	2.11
37	1.12	-0.75	0.60	1.95	-0.04	1.31	2.66	-0.49	0.86	2.21
38	1.18	-0.65	0.70	2.05	-1.52	-0.17	1.18	-0.68	0.67	2.02
39	1.25	-0.55	0.80	2.15	-2.74	-1.39	-0.04	-0.61	0.74	2.09
40	0.68	-0.45	0.90	2.25	-3.41	-2.06	-0.71	-0.40	0.95	2.30

Appendix B (cont.): Polytomous IRT Item Parameters

		Uniform			Normal			Bimodal		
Item	*a*	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
41	0.83	-0.35	1.00	2.35	-1.47	-0.12	1.23	-0.50	0.85	2.20
42	1.33	-0.25	1.10	2.45	-1.46	-0.11	1.24	-0.19	1.16	2.51
43	1.00	-0.15	1.20	2.55	-1.74	-0.39	0.96	-0.44	0.91	2.26
44	0.66	-0.05	1.30	2.65	-3.24	-1.89	-0.54	-0.62	0.73	2.08
45	0.69	0.05	1.40	2.75	-1.14	0.21	1.56	-0.76	0.59	1.94
46	1.20	0.15	1.50	2.85	-1.25	0.10	1.45	-0.91	0.44	1.79
47	1.14	0.25	1.60	2.95	-3.09	-1.74	-0.39	-0.67	0.68	2.03
48	0.94	0.35	1.70	3.05	-0.91	0.44	1.79	-0.67	0.68	2.03
49	1.01	0.45	1.80	3.15	-1.47	-0.12	1.23	-0.21	1.14	2.49
50	0.76	0.55	1.90	3.25	-3.60	-2.25	-0.90	-0.76	0.59	1.94
51	1.04	0.65	2.00	3.35	-2.22	-0.87	0.48	-0.31	1.04	2.39
52	1.16	0.75	2.10	3.45	-0.57	0.78	2.13	-0.27	1.08	2.43
53	0.96	0.85	2.20	3.55	-1.26	0.09	1.44	-0.32	1.03	2.38
54	0.94	0.95	2.30	3.65	-1.54	-0.19	1.16	-0.74	0.61	1.96
55	0.89	1.05	2.40	3.75	-3.09	-1.74	-0.39	-1.03	0.32	1.67
56	1.31	1.15	2.50	3.85	-2.02	-0.67	0.68	-0.53	0.82	2.17
57	1.07	1.25	2.60	3.95	-1.07	0.28	1.63	-1.37	-0.02	1.33
58	0.89	1.35	2.70	4.05	-0.73	0.62	1.97	-1.14	0.21	1.56
59	1.11	1.45	2.80	4.15	-2.33	-0.98	0.37	-0.55	0.80	2.15
60	0.71	1.55	2.90	4.25	-1.56	-0.21	1.14	-0.74	0.61	1.96
61	0.85	1.65	3.00	4.35	-2.17	-0.82	0.53	-0.73	0.62	1.97

Appendix C: Dichotomous IRT Observed and Predicted Responses

	θ = -3			θ = -2			θ = -1			θ = 0		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
1	0.26	0.25	0.00	0.30	0.28	0.02	0.42	0.44	-0.01	0.80	0.80	0.00
2	0.26	0.27	-0.01	0.33	0.34	-0.01	0.50	0.53	-0.02	0.78	0.79	-0.01
3	0.27	0.26	0.00	0.33	0.33	0.00	0.59	0.59	0.00	0.89	0.89	0.00
4	0.26	0.26	0.00	0.31	0.32	0.00	0.51	0.51	-0.01	0.82	0.82	0.01
5	0.29	0.27	0.02	0.34	0.34	0.01	0.53	0.53	0.01	0.80	0.79	0.01
6	0.29	0.29	0.00	0.43	0.40	0.03	0.63	0.64	-0.01	0.85	0.86	-0.01
7	0.28	0.27	0.01	0.30	0.33	-0.03	0.52	0.51	0.01	0.77	0.78	-0.01
8	0.23	0.26	-0.03	0.37	0.33	0.04	0.57	0.56	0.01	0.85	0.86	-0.01
9	0.29	0.28	0.01	0.36	0.36	-0.01	0.59	0.58	0.01	0.84	0.83	0.01
10	0.26	0.25	0.00	0.26	0.28	-0.02	0.47	0.45	0.02	0.83	0.81	0.02
11	0.25	0.26	-0.01	0.35	0.34	0.02	0.62	0.63	-0.02	0.91	0.92	-0.01
12	0.27	0.26	0.01	0.34	0.33	0.02	0.58	0.58	0.00	0.88	0.88	-0.01
13	0.30	0.32	-0.02	0.39	0.41	-0.02	0.58	0.56	0.02	0.75	0.73	0.02
14	0.26	0.27	-0.01	0.32	0.34	-0.03	0.62	0.60	0.02	0.86	0.89	-0.03
15	0.31	0.30	0.01	0.37	0.38	-0.01	0.56	0.56	0.00	0.75	0.78	-0.03
16	0.24	0.26	-0.02	0.24	0.29	-0.05	0.45	0.43	0.02	0.70	0.71	-0.01
17	0.28	0.27	0.01	0.32	0.35	-0.04	0.59	0.59	0.00	0.87	0.86	0.01
18	0.25	0.26	-0.01	0.32	0.30	0.02	0.47	0.48	-0.01	0.81	0.82	-0.01
19	0.25	0.27	-0.02	0.33	0.34	0.00	0.55	0.56	-0.01	0.85	0.85	0.00
20	0.26	0.26	0.00	0.31	0.32	-0.01	0.58	0.57	0.02	0.88	0.89	-0.01
21	0.26	0.29	-0.03	0.38	0.39	-0.01	0.61	0.61	0.00	0.82	0.84	-0.02
22	0.26	0.26	0.00	0.30	0.31	-0.01	0.52	0.52	0.00	0.83	0.83	0.00
23	0.26	0.26	0.01	0.32	0.32	0.00	0.59	0.58	0.01	0.91	0.90	0.01
24	0.26	0.27	-0.01	0.32	0.33	-0.01	0.50	0.51	0.00	0.79	0.78	0.02
25	0.24	0.25	-0.01	0.27	0.27	0.00	0.38	0.39	-0.01	0.71	0.72	-0.01
26	0.26	0.26	0.00	0.32	0.32	-0.01	0.55	0.56	-0.01	0.85	0.86	-0.01
27	0.28	0.27	0.01	0.36	0.35	0.01	0.60	0.59	0.01	0.87	0.86	0.01
28	0.25	0.27	-0.01	0.28	0.31	-0.03	0.46	0.46	0.01	0.70	0.71	-0.01
29	0.27	0.27	0.00	0.32	0.32	0.00	0.52	0.51	0.00	0.81	0.80	0.01
30	0.32	0.32	0.00	0.50	0.48	0.03	0.74	0.73	0.01	0.90	0.91	-0.01
31	0.27	0.27	0.00	0.33	0.33	0.00	0.50	0.50	0.00	0.78	0.77	0.01
32	0.27	0.27	0.00	0.31	0.33	-0.03	0.55	0.56	-0.02	0.86	0.85	0.01
33	0.27	0.26	0.01	0.34	0.32	0.01	0.53	0.54	-0.01	0.84	0.84	0.00
34	0.27	0.26	0.00	0.33	0.33	0.00	0.60	0.60	0.01	0.91	0.90	0.01
35	0.30	0.27	0.03	0.33	0.34	0.00	0.53	0.53	0.00	0.79	0.79	0.00
36	0.26	0.26	0.01	0.29	0.28	0.01	0.42	0.38	0.04	0.64	0.64	0.00
37	0.27	0.29	-0.02	0.41	0.37	0.04	0.55	0.56	-0.01	0.77	0.78	-0.01
38	0.28	0.26	0.02	0.29	0.30	-0.01	0.47	0.48	-0.01	0.78	0.79	-0.01
39	0.27	0.27	0.00	0.34	0.34	0.00	0.57	0.57	0.00	0.85	0.85	0.00
40	0.27	0.26	0.01	0.31	0.30	0.01	0.48	0.48	0.01	0.76	0.78	-0.02
41	0.26	0.26	0.00	0.30	0.30	0.00	0.46	0.46	0.00	0.75	0.76	-0.01
42	0.24	0.26	-0.02	0.29	0.31	-0.02	0.49	0.48	0.01	0.75	0.76	-0.01
43	0.25	0.25	0.00	0.28	0.28	0.00	0.41	0.43	-0.02	0.80	0.80	0.00
44	0.30	0.30	0.00	0.39	0.39	0.00	0.54	0.57	-0.03	0.78	0.79	-0.01
45	0.26	0.28	-0.02	0.36	0.34	0.03	0.48	0.48	0.00	0.72	0.70	0.02

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = 1			θ = 2			θ = 3		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
1	0.97	0.97	0.00	1.00	1.00	0.00	1.00	1.00	0.00
2	0.94	0.94	0.00	0.98	0.99	0.00	1.00	1.00	0.00
3	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
4	0.95	0.96	-0.01	0.99	0.99	0.00	1.00	1.00	0.00
5	0.94	0.94	0.00	0.99	0.99	0.00	1.00	1.00	0.00
6	0.96	0.96	0.00	0.99	0.99	0.00	1.00	1.00	0.00
7	0.93	0.94	-0.01	0.99	0.99	0.00	0.99	1.00	0.00
8	0.98	0.97	0.01	1.00	1.00	0.00	1.00	1.00	0.00
9	0.95	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00
10	0.97	0.97	0.00	1.00	1.00	0.00	1.00	1.00	0.00
11	0.99	0.99	0.00	1.00	1.00	0.00	1.00	1.00	0.00
12	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
13	0.85	0.86	-0.01	0.94	0.94	0.00	0.99	0.97	0.01
14	0.98	0.98	0.01	1.00	1.00	0.00	1.00	1.00	0.00
15	0.92	0.91	0.01	0.97	0.97	-0.01	0.99	0.99	0.00
16	0.90	0.92	-0.02	0.98	0.98	0.00	1.00	1.00	0.00
17	0.97	0.97	0.00	0.99	0.99	0.00	1.00	1.00	0.00
18	0.97	0.97	0.00	0.99	1.00	0.00	1.00	1.00	0.00
19	0.97	0.97	0.00	0.99	0.99	0.00	1.00	1.00	0.00
20	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
21	0.96	0.95	0.01	0.99	0.99	0.00	1.00	1.00	0.00
22	0.97	0.97	0.00	0.99	1.00	0.00	1.00	1.00	0.00
23	0.98	0.99	-0.01	1.00	1.00	0.00	1.00	1.00	0.00
24	0.92	0.94	-0.01	0.98	0.98	0.00	1.00	1.00	0.00
25	0.93	0.94	-0.01	0.99	0.99	0.00	1.00	1.00	0.00
26	0.96	0.97	-0.01	1.00	1.00	0.00	1.00	1.00	0.00
27	0.97	0.97	0.00	0.99	0.99	0.00	1.00	1.00	0.00
28	0.93	0.90	0.02	0.97	0.98	0.00	0.99	0.99	0.00
29	0.95	0.95	0.01	0.99	0.99	0.00	1.00	1.00	0.00
30	0.98	0.98	0.01	1.00	0.99	0.00	1.00	1.00	0.00
31	0.93	0.93	0.00	0.98	0.98	0.00	1.00	1.00	0.00
32	0.96	0.97	-0.01	0.99	1.00	-0.01	1.00	1.00	0.00
33	0.95	0.97	-0.01	0.99	0.99	0.00	1.00	1.00	0.00
34	0.98	0.99	-0.01	1.00	1.00	0.00	1.00	1.00	0.00
35	0.94	0.94	0.00	0.99	0.99	0.00	1.00	1.00	0.00
36	0.89	0.88	0.00	0.98	0.97	0.00	0.99	1.00	-0.01
37	0.90	0.92	-0.02	0.97	0.98	0.00	0.99	0.99	0.00
38	0.95	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00
39	0.96	0.97	0.00	0.99	0.99	0.00	1.00	1.00	0.00
40	0.94	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00
41	0.95	0.94	0.01	0.99	0.99	0.00	1.00	1.00	0.00
42	0.95	0.94	0.01	0.99	0.99	0.00	1.00	1.00	0.00
43	0.96	0.97	-0.01	1.00	1.00	0.00	1.00	1.00	0.00
44	0.93	0.92	0.01	0.98	0.98	0.00	0.99	0.99	0.00
45	0.88	0.88	0.00	0.96	0.96	0.00	0.99	0.99	0.00

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = -3			θ = -2			θ = -1			θ = 0		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
46	0.25	0.25	-0.01	0.27	0.27	0.00	0.38	0.38	0.00	0.73	0.73	0.01
47	0.28	0.28	-0.01	0.36	0.38	-0.03	0.62	0.62	0.00	0.85	0.86	-0.01
48	0.34	0.32	0.02	0.54	0.52	0.02	0.81	0.81	0.00	0.95	0.96	-0.01
49	0.26	0.26	0.00	0.30	0.30	-0.01	0.57	0.56	0.02	0.87	0.90	-0.03
50	0.25	0.26	-0.01	0.33	0.30	0.03	0.50	0.49	0.01	0.84	0.81	0.03
51	0.25	0.26	-0.01	0.30	0.28	0.02	0.40	0.40	0.00	0.71	0.71	0.00
52	0.25	0.26	0.00	0.29	0.29	0.00	0.42	0.45	-0.03	0.77	0.78	0.00
53	0.27	0.26	0.00	0.34	0.32	0.02	0.47	0.51	-0.04	0.79	0.81	-0.02
54	0.26	0.25	0.00	0.29	0.28	0.00	0.46	0.45	0.01	0.82	0.82	-0.01
55	0.26	0.27	-0.01	0.30	0.32	-0.02	0.51	0.51	-0.01	0.79	0.80	0.00
56	0.27	0.27	0.00	0.30	0.34	-0.04	0.55	0.55	0.00	0.83	0.83	-0.01
57	0.22	0.25	-0.03	0.27	0.27	0.00	0.43	0.43	0.00	0.82	0.81	0.00
58	0.26	0.26	0.01	0.29	0.29	0.00	0.50	0.51	0.00	0.87	0.86	0.01
59	0.25	0.27	-0.01	0.32	0.34	-0.02	0.63	0.62	0.01	0.90	0.90	0.00
60	0.27	0.27	0.01	0.33	0.34	-0.01	0.55	0.57	-0.01	0.85	0.85	-0.01
61	0.26	0.25	0.01	0.29	0.28	0.02	0.46	0.44	0.02	0.81	0.82	-0.02
62	0.33	0.30	0.03	0.39	0.39	0.00	0.56	0.56	0.00	0.73	0.76	-0.03
63	0.25	0.26	-0.01	0.29	0.29	-0.01	0.45	0.45	0.00	0.75	0.76	-0.01
64	0.28	0.29	-0.01	0.40	0.38	0.02	0.59	0.58	0.01	0.82	0.82	0.00
65	0.28	0.26	0.02	0.32	0.31	0.00	0.51	0.51	0.00	0.81	0.82	-0.01
66	0.25	0.26	-0.01	0.31	0.31	0.00	0.54	0.55	-0.01	0.88	0.88	0.00
67	0.26	0.27	0.00	0.33	0.33	0.00	0.49	0.52	-0.02	0.81	0.80	0.01
68	0.28	0.27	0.01	0.29	0.33	-0.04	0.46	0.46	-0.01	0.69	0.69	0.00
69	0.28	0.26	0.02	0.31	0.31	0.00	0.52	0.51	0.01	0.85	0.84	0.01
70	0.27	0.26	0.00	0.33	0.32	0.01	0.56	0.54	0.03	0.84	0.84	0.00
71	0.24	0.26	-0.02	0.29	0.29	0.00	0.44	0.43	0.00	0.71	0.74	-0.03
72	0.26	0.26	0.00	0.33	0.33	0.00	0.62	0.60	0.02	0.90	0.90	0.00
73	0.24	0.26	-0.03	0.35	0.33	0.03	0.54	0.55	-0.01	0.83	0.85	-0.02
74	0.29	0.27	0.01	0.32	0.33	-0.02	0.51	0.50	0.02	0.78	0.75	0.04
75	0.26	0.26	0.00	0.33	0.33	0.00	0.55	0.54	0.01	0.85	0.84	0.01
76	0.24	0.26	-0.02	0.30	0.31	-0.01	0.49	0.47	0.02	0.76	0.75	0.01
77	0.26	0.29	-0.03	0.35	0.35	-0.01	0.50	0.50	0.00	0.71	0.71	0.00
78	0.30	0.28	0.01	0.41	0.39	0.02	0.64	0.66	-0.01	0.87	0.89	-0.02
79	0.28	0.27	0.00	0.36	0.38	-0.02	0.70	0.69	0.01	0.93	0.93	-0.01
80	0.35	0.34	0.01	0.43	0.43	0.00	0.53	0.55	-0.02	0.71	0.70	0.01
81	0.27	0.26	0.01	0.32	0.31	0.01	0.54	0.55	-0.01	0.88	0.88	0.00
82	0.30	0.29	0.01	0.36	0.38	-0.02	0.59	0.59	0.00	0.80	0.82	-0.02
83	0.23	0.26	-0.03	0.26	0.29	-0.02	0.42	0.43	-0.01	0.75	0.75	0.00
84	0.26	0.27	-0.01	0.34	0.34	0.00	0.56	0.57	-0.01	0.84	0.85	0.00
85	0.31	0.31	0.00	0.39	0.40	-0.02	0.57	0.57	0.01	0.75	0.76	-0.01
86	0.29	0.27	0.02	0.33	0.32	0.01	0.47	0.46	0.01	0.68	0.69	-0.01
87	0.29	0.30	-0.01	0.36	0.38	-0.02	0.51	0.53	-0.01	0.74	0.72	0.03
88	0.29	0.27	0.02	0.33	0.33	0.00	0.50	0.52	-0.02	0.78	0.79	-0.01
89	0.25	0.27	-0.02	0.36	0.34	0.02	0.55	0.54	0.01	0.83	0.82	0.01
90	0.27	0.28	-0.01	0.37	0.36	0.01	0.58	0.58	0.01	0.84	0.83	0.02

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = 1			θ = 2			θ = 3		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
46	0.95	0.95	0.00	1.00	0.99	0.00	1.00	1.00	0.00
47	0.97	0.96	0.01	0.99	0.99	0.00	1.00	1.00	0.00
48	0.99	0.99	0.00	1.00	1.00	0.00	1.00	1.00	0.00
49	0.99	0.99	0.00	1.00	1.00	0.00	1.00	1.00	0.00
50	0.96	0.96	0.00	1.00	0.99	0.00	1.00	1.00	0.00
51	0.94	0.93	0.00	0.99	0.99	0.00	1.00	1.00	0.00
52	0.96	0.95	0.00	1.00	0.99	0.00	1.00	1.00	0.00
53	0.96	0.95	0.01	0.99	0.99	-0.01	1.00	1.00	0.00
54	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
55	0.95	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00
56	0.97	0.96	0.01	0.99	0.99	0.00	1.00	1.00	0.00
57	0.97	0.97	-0.01	1.00	1.00	0.00	1.00	1.00	0.00
58	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
59	0.99	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
60	0.97	0.97	0.00	1.00	0.99	0.00	1.00	1.00	0.00
61	0.98	0.98	0.01	1.00	1.00	0.00	1.00	1.00	0.00
62	0.90	0.90	0.00	0.98	0.97	0.01	0.99	0.99	0.00
63	0.95	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00
64	0.95	0.95	0.01	0.98	0.99	-0.01	1.00	1.00	0.00
65	0.96	0.96	0.00	0.99	0.99	0.00	1.00	1.00	0.00
66	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
67	0.95	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00
68	0.87	0.87	0.00	0.97	0.96	0.01	0.99	0.99	0.00
69	0.97	0.97	0.00	1.00	1.00	0.00	1.00	1.00	0.00
70	0.97	0.97	0.00	0.99	0.99	0.00	1.00	1.00	0.00
71	0.93	0.93	-0.01	0.99	0.99	0.00	1.00	1.00	0.00
72	0.98	0.99	0.00	1.00	1.00	0.00	1.00	1.00	0.00
73	0.96	0.97	-0.01	1.00	1.00	0.00	1.00	1.00	0.00
74	0.91	0.91	-0.01	0.98	0.98	0.00	0.99	0.99	0.00
75	0.96	0.97	-0.01	1.00	0.99	0.00	1.00	1.00	0.00
76	0.93	0.93	0.00	0.98	0.98	0.00	1.00	1.00	0.00
77	0.88	0.87	0.01	0.96	0.95	0.00	0.99	0.99	0.00
78	0.98	0.98	0.00	1.00	1.00	0.00	1.00	1.00	0.00
79	0.99	0.99	0.00	1.00	1.00	0.00	1.00	1.00	0.00
80	0.83	0.82	0.01	0.91	0.91	0.00	0.95	0.96	-0.01
81	0.99	0.98	0.01	1.00	1.00	0.00	1.00	1.00	0.00
82	0.95	0.94	0.00	0.99	0.99	0.01	1.00	1.00	0.00
83	0.96	0.95	0.01	0.99	0.99	0.00	1.00	1.00	0.00
84	0.97	0.97	0.01	1.00	0.99	0.00	1.00	1.00	0.00
85	0.90	0.89	0.01	0.95	0.96	-0.01	0.99	0.99	0.00
86	0.87	0.88	-0.01	0.97	0.96	0.00	0.99	0.99	0.00
87	0.88	0.87	0.01	0.95	0.95	0.00	0.99	0.98	0.01
88	0.95	0.94	0.01	0.99	0.99	0.00	1.00	1.00	0.00
89	0.94	0.95	-0.01	0.99	0.99	0.00	0.99	1.00	0.00
90	0.95	0.95	0.00	0.99	0.99	0.00	1.00	1.00	0.00

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = -3			θ = -2			θ = -1			θ = 0		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
91	0.24	0.25	-0.01	0.26	0.25	0.01	0.28	0.27	0.01	0.39	0.38	0.02
92	0.25	0.25	0.00	0.24	0.26	-0.01	0.27	0.28	-0.01	0.44	0.41	0.04
93	0.25	0.25	0.00	0.27	0.25	0.02	0.26	0.27	-0.01	0.34	0.35	-0.01
94	0.27	0.27	0.00	0.30	0.30	0.00	0.38	0.37	0.01	0.49	0.49	-0.01
95	0.25	0.25	-0.01	0.27	0.26	0.01	0.28	0.30	-0.02	0.45	0.44	0.01
96	0.24	0.25	-0.01	0.26	0.26	0.01	0.28	0.28	-0.01	0.41	0.42	-0.02
97	0.24	0.25	-0.01	0.27	0.26	0.01	0.33	0.31	0.02	0.47	0.47	0.00
98	0.25	0.26	-0.01	0.28	0.27	0.01	0.31	0.32	-0.01	0.46	0.46	-0.01
99	0.24	0.25	-0.02	0.27	0.26	0.01	0.28	0.29	-0.01	0.46	0.47	-0.01
100	0.29	0.26	0.03	0.27	0.28	-0.01	0.37	0.35	0.02	0.49	0.51	-0.02
101	0.26	0.25	0.00	0.28	0.26	0.02	0.31	0.30	0.00	0.44	0.46	-0.01
102	0.21	0.25	-0.04	0.23	0.25	-0.02	0.27	0.27	0.01	0.35	0.38	-0.03
103	0.24	0.25	-0.01	0.25	0.25	0.00	0.27	0.27	0.00	0.42	0.41	0.01
104	0.26	0.25	0.01	0.28	0.25	0.03	0.28	0.26	0.02	0.31	0.33	-0.02
105	0.27	0.25	0.02	0.29	0.27	0.02	0.33	0.32	0.01	0.51	0.50	0.01
106	0.26	0.25	0.01	0.27	0.26	0.01	0.28	0.29	0.00	0.38	0.39	-0.01
107	0.26	0.25	0.01	0.24	0.25	-0.02	0.26	0.27	-0.01	0.38	0.36	0.01
108	0.25	0.25	0.00	0.27	0.26	0.01	0.33	0.33	0.01	0.58	0.55	0.03
109	0.26	0.26	0.01	0.27	0.27	0.00	0.36	0.34	0.02	0.54	0.55	0.00
110	0.25	0.25	-0.01	0.28	0.26	0.03	0.28	0.28	0.00	0.40	0.39	0.01
111	0.24	0.25	-0.01	0.28	0.26	0.02	0.26	0.28	-0.02	0.42	0.41	0.01
112	0.27	0.25	0.02	0.23	0.26	-0.02	0.30	0.28	0.02	0.42	0.39	0.02
113	0.24	0.25	-0.01	0.27	0.26	0.01	0.31	0.29	0.02	0.40	0.42	-0.02
114	0.27	0.25	0.01	0.28	0.27	0.01	0.32	0.32	0.00	0.48	0.48	0.00
115	0.28	0.26	0.02	0.28	0.27	0.01	0.31	0.31	0.00	0.43	0.43	0.01
116	0.26	0.25	0.01	0.26	0.26	0.01	0.27	0.28	-0.01	0.40	0.38	0.02
117	0.29	0.25	0.04	0.26	0.26	0.01	0.28	0.28	0.00	0.35	0.38	-0.03
118	0.26	0.25	0.01	0.25	0.25	-0.01	0.25	0.29	-0.03	0.46	0.47	-0.01
119	0.27	0.26	0.01	0.29	0.27	0.02	0.32	0.31	0.01	0.42	0.42	0.00
120	0.25	0.25	-0.01	0.28	0.27	0.01	0.33	0.32	0.02	0.48	0.47	0.01
121	0.26	0.25	0.00	0.27	0.27	0.00	0.35	0.35	0.00	0.61	0.60	0.01
122	0.25	0.25	0.00	0.24	0.25	-0.01	0.25	0.26	-0.01	0.32	0.33	-0.01
123	0.23	0.25	-0.03	0.26	0.26	0.00	0.32	0.29	0.02	0.40	0.41	-0.01
124	0.26	0.26	0.01	0.29	0.28	0.02	0.34	0.34	0.00	0.52	0.51	0.01
125	0.22	0.25	-0.04	0.26	0.25	0.01	0.28	0.28	0.00	0.41	0.42	-0.01
126	0.27	0.25	0.02	0.25	0.26	-0.01	0.29	0.29	0.01	0.44	0.42	0.02
127	0.25	0.26	-0.01	0.24	0.27	-0.03	0.32	0.31	0.01	0.41	0.43	-0.02
128	0.27	0.26	0.01	0.30	0.28	0.02	0.34	0.34	0.00	0.50	0.50	0.00
129	0.25	0.25	-0.01	0.26	0.25	0.01	0.28	0.27	0.01	0.38	0.38	0.00
130	0.28	0.25	0.03	0.24	0.26	-0.02	0.30	0.30	0.00	0.53	0.52	0.01
131	0.30	0.28	0.03	0.32	0.31	0.01	0.41	0.38	0.03	0.52	0.51	0.01
132	0.25	0.25	-0.01	0.26	0.26	-0.01	0.30	0.31	-0.01	0.50	0.48	0.02
133	0.24	0.25	-0.01	0.23	0.25	-0.02	0.24	0.27	-0.03	0.42	0.37	0.05
134	0.25	0.25	0.00	0.26	0.26	0.00	0.26	0.28	-0.02	0.39	0.37	0.02
135	0.25	0.25	0.00	0.27	0.25	0.02	0.26	0.27	-0.01	0.39	0.40	-0.01

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = 1			θ = 2			θ = 3		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
91	0.75	0.75	0.00	0.97	0.96	0.00	1.00	1.00	0.00
92	0.69	0.70	-0.01	0.93	0.92	0.00	0.99	0.99	0.01
93	0.56	0.58	-0.01	0.88	0.85	0.03	0.97	0.97	0.00
94	0.68	0.67	0.01	0.82	0.82	0.00	0.93	0.92	0.01
95	0.76	0.72	0.04	0.92	0.92	0.00	0.98	0.98	0.00
96	0.74	0.75	0.00	0.95	0.95	0.00	0.99	0.99	0.00
97	0.76	0.76	0.00	0.94	0.94	0.00	0.99	0.99	0.00
98	0.72	0.71	0.01	0.91	0.90	0.01	0.97	0.97	-0.01
99	0.80	0.82	-0.02	0.97	0.97	0.00	1.00	1.00	0.00
100	0.75	0.73	0.02	0.89	0.89	0.00	0.97	0.96	0.00
101	0.70	0.74	-0.03	0.94	0.93	0.01	0.99	0.98	0.00
102	0.77	0.74	0.03	0.96	0.96	0.00	1.00	1.00	0.00
103	0.77	0.78	-0.01	0.96	0.97	-0.01	0.99	1.00	0.00
104	0.64	0.62	0.02	0.90	0.92	-0.02	0.99	0.99	0.00
105	0.79	0.77	0.01	0.94	0.94	0.00	0.99	0.99	0.00
106	0.67	0.64	0.03	0.90	0.88	0.03	0.97	0.97	0.00
107	0.68	0.69	-0.01	0.94	0.94	-0.01	0.99	0.99	0.00
108	0.84	0.85	-0.01	0.96	0.97	-0.02	1.00	1.00	0.00
109	0.82	0.81	0.00	0.93	0.95	-0.02	0.98	0.99	-0.01
110	0.69	0.68	0.01	0.91	0.92	-0.01	0.99	0.99	0.01
111	0.71	0.72	-0.01	0.93	0.93	0.00	0.99	0.99	0.00
112	0.63	0.66	-0.02	0.90	0.89	0.01	0.97	0.98	0.00
113	0.70	0.69	0.01	0.90	0.91	-0.01	0.98	0.98	0.00
114	0.73	0.73	0.00	0.92	0.91	0.01	0.98	0.98	0.00
115	0.62	0.64	-0.02	0.85	0.85	0.00	0.94	0.95	-0.01
116	0.63	0.66	-0.02	0.90	0.90	-0.01	0.98	0.98	0.00
117	0.65	0.63	0.02	0.88	0.88	0.00	0.99	0.97	0.01
118	0.83	0.83	0.00	0.97	0.97	0.00	1.00	1.00	0.00
119	0.63	0.61	0.01	0.82	0.82	0.00	0.93	0.93	-0.01
120	0.74	0.73	0.01	0.93	0.91	0.02	0.98	0.98	0.00
121	0.87	0.87	0.00	0.97	0.97	0.00	0.99	1.00	0.00
122	0.64	0.65	-0.01	0.94	0.94	0.00	0.99	0.99	0.00
123	0.66	0.64	0.01	0.87	0.87	0.00	0.96	0.96	0.00
124	0.77	0.76	0.01	0.93	0.92	0.01	0.96	0.98	-0.01
125	0.75	0.76	-0.02	0.96	0.96	0.01	0.99	0.99	0.00
126	0.74	0.73	0.01	0.94	0.94	0.00	0.99	0.99	0.00
127	0.67	0.65	0.02	0.85	0.85	0.00	0.95	0.95	0.00
128	0.75	0.73	0.02	0.91	0.90	0.00	0.97	0.97	-0.01
129	0.67	0.68	-0.01	0.95	0.92	0.02	0.99	0.99	0.00
130	0.86	0.86	0.00	0.97	0.98	-0.01	1.00	1.00	0.00
131	0.67	0.68	-0.01	0.83	0.82	0.00	0.92	0.92	0.01
132	0.79	0.77	0.03	0.94	0.94	0.00	0.99	0.99	0.00
133	0.69	0.69	0.00	0.93	0.94	0.00	0.99	0.99	0.00
134	0.62	0.65	-0.03	0.88	0.90	-0.02	0.98	0.98	0.00
135	0.80	0.79	0.01	0.97	0.97	0.00	1.00	1.00	0.00

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = -3			θ = -2			θ = -1			θ = 0		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
136	0.25	0.25	-0.01	0.29	0.27	0.02	0.32	0.32	0.00	0.52	0.51	0.01
137	0.26	0.27	-0.01	0.29	0.30	-0.01	0.37	0.37	0.00	0.48	0.51	-0.03
138	0.23	0.25	-0.03	0.26	0.26	0.00	0.33	0.33	0.00	0.58	0.56	0.01
139	0.29	0.26	0.03	0.29	0.28	0.01	0.37	0.35	0.02	0.51	0.49	0.01
140	0.26	0.25	0.01	0.27	0.26	0.01	0.28	0.28	0.00	0.38	0.39	0.00
141	0.25	0.25	0.00	0.25	0.26	-0.01	0.29	0.30	0.00	0.44	0.44	0.00
142	0.28	0.25	0.03	0.27	0.26	0.01	0.29	0.29	0.00	0.44	0.42	0.02
143	0.23	0.26	-0.03	0.28	0.29	-0.01	0.32	0.34	-0.02	0.44	0.46	-0.01
144	0.27	0.25	0.02	0.24	0.25	-0.01	0.25	0.27	-0.02	0.44	0.41	0.03
145	0.24	0.25	-0.01	0.26	0.26	0.00	0.29	0.30	0.00	0.43	0.45	-0.01
146	0.24	0.26	-0.01	0.28	0.28	0.00	0.35	0.35	0.00	0.56	0.55	0.01
147	0.24	0.25	-0.01	0.28	0.26	0.02	0.29	0.29	0.00	0.40	0.42	-0.02
148	0.30	0.26	0.04	0.29	0.29	0.00	0.36	0.35	0.00	0.49	0.50	0.00
149	0.24	0.25	-0.01	0.26	0.25	0.01	0.26	0.26	0.00	0.31	0.31	0.00
150	0.25	0.27	-0.02	0.31	0.30	0.01	0.37	0.39	-0.02	0.54	0.55	-0.01
151	0.26	0.25	0.00	0.26	0.25	0.00	0.28	0.28	0.00	0.44	0.44	0.01
152	0.26	0.26	0.00	0.29	0.29	0.00	0.37	0.39	-0.01	0.62	0.58	0.03
153	0.26	0.25	0.01	0.27	0.25	0.02	0.29	0.28	0.01	0.46	0.44	0.02
154	0.27	0.25	0.01	0.28	0.26	0.02	0.32	0.32	0.00	0.53	0.53	0.01
155	0.25	0.25	0.00	0.26	0.26	0.01	0.30	0.28	0.02	0.42	0.42	0.00
156	0.26	0.25	0.01	0.25	0.26	-0.01	0.25	0.28	-0.04	0.41	0.44	-0.02
157	0.26	0.25	0.01	0.27	0.26	0.01	0.30	0.31	-0.01	0.45	0.45	-0.01
158	0.24	0.25	-0.02	0.26	0.26	0.00	0.29	0.29	0.00	0.43	0.43	0.00
159	0.23	0.25	-0.02	0.25	0.26	-0.01	0.30	0.29	0.01	0.45	0.46	-0.01
160	0.23	0.25	-0.02	0.28	0.26	0.02	0.28	0.31	-0.03	0.49	0.47	0.01
161	0.27	0.26	0.01	0.28	0.28	0.00	0.33	0.34	-0.01	0.50	0.50	0.00
162	0.24	0.26	-0.02	0.28	0.28	0.00	0.35	0.34	0.01	0.49	0.47	0.01
163	0.26	0.25	0.00	0.23	0.26	-0.03	0.27	0.28	-0.02	0.42	0.41	0.01
164	0.28	0.26	0.02	0.29	0.27	0.02	0.36	0.33	0.03	0.52	0.51	0.00
165	0.24	0.25	-0.01	0.27	0.25	0.02	0.30	0.28	0.02	0.46	0.47	-0.01
166	0.28	0.25	0.03	0.25	0.26	-0.02	0.33	0.30	0.03	0.42	0.44	-0.02
167	0.26	0.26	0.00	0.27	0.29	-0.03	0.37	0.38	0.00	0.56	0.55	0.01
168	0.24	0.25	-0.01	0.25	0.25	0.00	0.27	0.26	0.02	0.31	0.30	0.01
169	0.25	0.26	-0.01	0.30	0.27	0.03	0.30	0.32	-0.03	0.48	0.47	0.01
170	0.24	0.25	-0.01	0.25	0.25	0.00	0.26	0.27	-0.01	0.34	0.34	0.00
171	0.26	0.25	0.01	0.26	0.25	0.01	0.26	0.27	-0.01	0.36	0.37	-0.01
172	0.25	0.25	0.00	0.28	0.25	0.03	0.26	0.27	-0.01	0.35	0.35	0.00
173	0.24	0.25	-0.01	0.22	0.25	-0.04	0.28	0.28	0.00	0.43	0.44	-0.01
174	0.25	0.25	0.00	0.27	0.26	0.01	0.34	0.31	0.03	0.46	0.45	0.01
175	0.27	0.26	0.01	0.28	0.27	0.01	0.32	0.34	-0.02	0.49	0.51	-0.01
176	0.25	0.25	0.00	0.27	0.26	0.02	0.29	0.29	0.01	0.39	0.39	0.00
177	0.27	0.28	0.00	0.32	0.31	0.01	0.39	0.38	0.01	0.49	0.50	-0.01
178	0.24	0.25	-0.01	0.28	0.25	0.03	0.26	0.27	-0.01	0.40	0.39	0.01
179	0.24	0.25	-0.01	0.25	0.25	-0.01	0.27	0.26	0.00	0.33	0.34	-0.01
180	0.25	0.26	-0.01	0.27	0.28	-0.01	0.34	0.35	-0.01	0.49	0.49	0.00
181	0.28	0.25	0.03	0.26	0.26	0.00	0.31	0.30	0.02	0.46	0.46	0.00

Appendix C (cont.): Dichotomous IRT Observed and Predicted Responses

	θ = 1			θ = 2			θ = 3		
Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
136	0.79	0.79	0.01	0.94	0.94	-0.01	0.99	0.99	0.00
137	0.69	0.70	-0.01	0.87	0.85	0.01	0.93	0.94	-0.01
138	0.86	0.86	-0.01	0.97	0.97	0.00	0.99	1.00	0.00
139	0.72	0.70	0.01	0.87	0.87	0.00	0.95	0.95	0.00
140	0.62	0.63	-0.01	0.86	0.87	-0.01	0.97	0.97	0.00
141	0.73	0.72	0.01	0.93	0.92	0.01	0.98	0.98	0.00
142	0.68	0.69	-0.01	0.90	0.91	-0.01	0.98	0.98	0.00
143	0.65	0.64	0.01	0.80	0.81	-0.01	0.94	0.92	0.02
144	0.77	0.79	-0.01	0.97	0.97	0.00	1.00	1.00	0.00
145	0.76	0.74	0.02	0.94	0.93	0.01	0.99	0.99	0.01
146	0.82	0.81	0.01	0.94	0.95	0.00	0.99	0.99	0.00
147	0.72	0.70	0.02	0.93	0.92	0.02	0.98	0.98	0.00
148	0.71	0.70	0.01	0.85	0.86	-0.01	0.94	0.95	-0.01
149	0.49	0.52	-0.03	0.85	0.84	0.01	0.97	0.97	0.00
150	0.74	0.74	-0.01	0.89	0.89	0.00	0.94	0.96	-0.01
151	0.78	0.79	-0.01	0.97	0.97	0.01	1.00	1.00	0.00
152	0.79	0.81	-0.02	0.93	0.94	-0.01	0.98	0.98	0.00
153	0.82	0.81	0.01	0.97	0.97	0.00	1.00	1.00	0.00
154	0.83	0.83	0.00	0.96	0.97	-0.01	1.00	0.99	0.00
155	0.74	0.75	-0.01	0.94	0.95	-0.01	0.99	0.99	0.00
156	0.80	0.77	0.02	0.96	0.96	0.01	1.00	0.99	0.00
157	0.74	0.72	0.03	0.91	0.91	0.00	0.99	0.98	0.01
158	0.70	0.72	-0.03	0.94	0.93	0.01	0.99	0.99	0.00
159	0.78	0.78	-0.01	0.95	0.96	0.00	1.00	0.99	0.00
160	0.78	0.77	0.01	0.94	0.94	0.00	0.99	0.99	0.00
161	0.75	0.74	0.01	0.91	0.91	0.00	0.98	0.97	0.00
162	0.67	0.68	-0.01	0.87	0.86	0.01	0.95	0.95	0.00
163	0.73	0.71	0.02	0.94	0.93	0.01	0.98	0.99	-0.01
164	0.78	0.78	0.01	0.93	0.93	0.00	0.99	0.98	0.00
165	0.84	0.84	0.00	0.97	0.98	-0.01	1.00	1.00	0.00
166	0.70	0.70	0.01	0.90	0.90	0.00	0.98	0.98	0.00
167	0.76	0.76	-0.01	0.91	0.91	0.00	0.97	0.97	0.00
168	0.56	0.56	0.00	0.90	0.90	0.00	0.99	0.99	0.00
169	0.71	0.71	0.01	0.88	0.89	-0.02	0.97	0.97	0.01
170	0.57	0.58	0.00	0.85	0.86	-0.01	0.96	0.97	-0.01
171	0.72	0.71	0.01	0.96	0.95	0.01	1.00	0.99	0.00
172	0.67	0.67	0.00	0.93	0.93	0.00	0.99	0.99	0.00
173	0.82	0.80	0.01	0.97	0.97	0.00	1.00	1.00	0.00
174	0.72	0.71	0.01	0.90	0.90	-0.01	0.97	0.98	0.00
175	0.76	0.75	0.01	0.92	0.92	0.01	0.97	0.98	-0.01
176	0.65	0.65	0.00	0.89	0.88	0.01	0.97	0.97	0.00
177	0.67	0.66	0.01	0.83	0.81	0.02	0.90	0.91	-0.01
178	0.78	0.78	0.01	0.98	0.97	0.01	0.99	1.00	0.00
179	0.63	0.64	-0.01	0.92	0.92	0.00	0.99	0.99	0.00
180	0.70	0.70	0.00	0.86	0.87	-0.01	0.96	0.96	0.00
181	0.78	0.77	0.01	0.95	0.95	0.00	0.99	0.99	0.00

Appendix D: Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
-3	1	0.87	0.88	-0.01	0.13	0.12	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	2	0.62	0.61	0.02	0.34	0.36	-0.02	0.03	0.03	0.00	0.00	0.00	0.00
-3	3	0.79	0.80	-0.01	0.21	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	4	0.82	0.81	0.01	0.17	0.18	-0.01	0.01	0.00	0.01	0.00	0.00	0.00
-3	5	0.87	0.87	0.00	0.13	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	6	0.91	0.90	0.00	0.09	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	7	0.80	0.78	0.02	0.20	0.21	-0.01	0.00	0.01	0.00	0.00	0.00	0.00
-3	8	0.65	0.66	-0.01	0.32	0.31	0.01	0.03	0.04	0.00	0.00	0.00	0.00
-3	9	0.92	0.91	0.00	0.08	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	10	0.91	0.89	0.02	0.09	0.11	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
-3	11	0.94	0.92	0.01	0.06	0.08	-0.02	0.00	0.00	0.00	0.00	0.00	0.00
-3	12	0.87	0.88	-0.01	0.13	0.12	0.01	0.01	0.00	0.00	0.00	0.00	0.00
-3	13	0.88	0.89	-0.01	0.12	0.11	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	14	0.78	0.77	0.02	0.21	0.23	-0.01	0.00	0.01	0.00	0.00	0.00	0.00
-3	15	0.85	0.84	0.01	0.15	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	16	0.96	0.96	0.00	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	17	0.81	0.84	-0.03	0.19	0.16	0.03	0.00	0.00	0.00	0.00	0.00	0.00
-3	18	0.93	0.93	0.00	0.07	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	19	0.90	0.90	0.00	0.10	0.10	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	20	0.86	0.86	0.00	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	21	0.59	0.60	-0.01	0.40	0.39	0.01	0.01	0.01	0.00	0.00	0.00	0.00
-3	22	0.88	0.89	-0.01	0.11	0.11	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	23	0.93	0.92	0.01	0.07	0.08	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	24	0.77	0.75	0.01	0.23	0.24	-0.01	0.01	0.00	0.00	0.00	0.00	0.00
-3	25	0.89	0.89	0.00	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	26	0.73	0.75	-0.02	0.25	0.24	0.01	0.02	0.01	0.00	0.00	0.00	0.00
-3	27	0.77	0.78	0.00	0.21	0.21	0.00	0.01	0.01	0.00	0.00	0.00	0.00
-3	28	0.82	0.83	-0.02	0.18	0.16	0.01	0.01	0.00	0.00	0.00	0.00	0.00
-3	29	0.83	0.82	0.01	0.17	0.18	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	30	0.88	0.88	0.01	0.12	0.12	-0.01	0.00	0.00	0.00	0.00	0.00	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
-3	31	0.98	0.98	0.01	0.02	0.02	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	32	0.97	0.98	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	33	0.99	0.99	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	34	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	35	0.97	0.97	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	36	1.00	0.99	0.01	0.01	0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	37	1.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	38	0.98	0.99	-0.01	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	39	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	40	0.96	0.95	0.01	0.04	0.05	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	41	0.98	0.97	0.01	0.02	0.03	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	42	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	43	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	44	0.91	0.94	-0.03	0.09	0.06	0.03	0.00	0.00	0.00	0.00	0.00	0.00
-3	45	0.94	0.93	0.00	0.07	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	46	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	47	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	48	0.97	0.98	-0.01	0.03	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	49	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	50	0.94	0.95	0.00	0.06	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	51	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	52	1.00	1.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	53	0.99	0.99	0.01	0.01	0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	54	0.97	0.97	-0.01	0.03	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	55	0.95	0.95	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	56	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	57	0.96	0.95	0.01	0.04	0.05	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	58	0.93	0.94	-0.01	0.07	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-3	59	0.99	0.99	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	60	0.94	0.94	0.00	0.06	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-3	61	0.96	0.96	-0.01	0.04	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
-2	1	0.55	0.53	0.02	0.42	0.44	-0.01	0.03	0.03	0.00	0.00	0.00	0.00
-2	2	0.20	0.23	-0.03	0.58	0.56	0.02	0.21	0.20	0.01	0.01	0.01	0.00
-2	3	0.34	0.34	0.00	0.58	0.59	-0.01	0.08	0.07	0.01	0.00	0.00	0.00
-2	4	0.43	0.44	0.00	0.50	0.50	0.00	0.07	0.06	0.00	0.00	0.00	0.00
-2	5	0.40	0.39	0.01	0.56	0.57	-0.02	0.05	0.04	0.01	0.00	0.00	0.00
-2	6	0.54	0.55	-0.02	0.44	0.43	0.02	0.02	0.02	0.00	0.00	0.00	0.00
-2	7	0.37	0.37	0.01	0.55	0.55	0.00	0.08	0.08	-0.01	0.00	0.00	0.00
-2	8	0.35	0.36	-0.01	0.49	0.47	0.02	0.15	0.16	0.00	0.01	0.01	0.00
-2	9	0.65	0.63	0.02	0.33	0.35	-0.02	0.02	0.02	0.00	0.00	0.00	0.00
-2	10	0.64	0.64	0.01	0.33	0.34	-0.01	0.02	0.02	0.00	0.00	0.00	0.00
-2	11	0.62	0.62	0.00	0.38	0.37	0.01	0.01	0.02	-0.01	0.00	0.00	0.00
-2	12	0.56	0.57	-0.01	0.40	0.40	0.00	0.04	0.03	0.01	0.00	0.00	0.00
-2	13	0.61	0.60	0.01	0.37	0.37	-0.01	0.02	0.03	0.00	0.00	0.00	0.00
-2	14	0.39	0.36	0.03	0.52	0.55	-0.03	0.10	0.09	0.01	0.00	0.00	0.00
-2	15	0.50	0.51	0.00	0.45	0.45	0.00	0.05	0.05	0.00	0.00	0.00	0.00
-2	16	0.72	0.72	0.01	0.27	0.28	-0.01	0.01	0.01	0.00	0.00	0.00	0.00
-2	17	0.44	0.45	-0.01	0.51	0.50	0.01	0.05	0.05	0.00	0.00	0.00	0.00
-2	18	0.68	0.68	0.01	0.30	0.31	-0.01	0.02	0.01	0.00	0.00	0.00	0.00
-2	19	0.49	0.49	0.00	0.48	0.48	0.00	0.02	0.03	-0.01	0.00	0.00	0.00
-2	20	0.46	0.47	-0.02	0.50	0.48	0.02	0.04	0.04	0.00	0.00	0.00	0.00
-2	21	0.10	0.11	-0.01	0.69	0.68	0.01	0.21	0.20	0.01	0.00	0.00	0.00
-2	22	0.62	0.61	0.01	0.37	0.37	0.00	0.02	0.02	-0.01	0.00	0.00	0.00
-2	23	0.54	0.53	0.01	0.45	0.45	0.00	0.01	0.02	0.00	0.00	0.00	0.00
-2	24	0.24	0.25	-0.01	0.65	0.65	0.00	0.11	0.10	0.01	0.00	0.00	0.00
-2	25	0.43	0.45	-0.02	0.54	0.52	0.02	0.03	0.03	0.00	0.00	0.00	0.00
-2	26	0.41	0.39	0.02	0.49	0.51	-0.02	0.09	0.10	-0.01	0.00	0.00	0.00
-2	27	0.46	0.47	-0.02	0.46	0.45	0.02	0.08	0.08	0.00	0.00	0.00	0.00
-2	28	0.47	0.47	0.01	0.47	0.48	-0.01	0.06	0.05	0.00	0.00	0.00	0.00
-2	29	0.38	0.39	-0.02	0.56	0.55	0.02	0.06	0.06	0.00	0.00	0.00	0.00
-2	30	0.54	0.57	-0.03	0.44	0.40	0.04	0.02	0.03	-0.01	0.00	0.00	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
-2	31	0.89	0.90	0.00	0.11	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	32	0.89	0.88	0.00	0.11	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	33	0.90	0.91	-0.01	0.10	0.09	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	34	0.96	0.96	0.00	0.05	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	35	0.88	0.89	-0.01	0.12	0.11	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	36	0.92	0.93	-0.01	0.08	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	37	0.94	0.95	-0.01	0.06	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	38	0.94	0.93	0.00	0.06	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	39	0.95	0.95	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	40	0.87	0.86	0.01	0.12	0.14	-0.01	0.01	0.00	0.00	0.00	0.00	0.00
-2	41	0.89	0.89	0.00	0.11	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	42	0.99	0.98	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	43	0.93	0.93	-0.01	0.08	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	44	0.83	0.82	0.01	0.16	0.17	-0.01	0.01	0.01	0.00	0.00	0.00	0.00
-2	45	0.79	0.80	-0.01	0.20	0.19	0.01	0.01	0.01	0.00	0.00	0.00	0.00
-2	46	0.90	0.90	0.00	0.10	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	47	0.94	0.93	0.01	0.07	0.07	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	48	0.88	0.89	-0.01	0.12	0.11	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	49	0.96	0.96	0.00	0.05	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	50	0.81	0.83	-0.02	0.19	0.17	0.02	0.01	0.01	0.00	0.00	0.00	0.00
-2	51	0.96	0.95	0.01	0.04	0.05	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	52	0.97	0.97	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	53	0.95	0.94	0.01	0.05	0.06	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-2	54	0.89	0.88	0.00	0.11	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	55	0.82	0.81	0.01	0.17	0.19	-0.02	0.01	0.01	0.00	0.00	0.00	0.00
-2	56	0.97	0.96	0.00	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	57	0.76	0.75	0.01	0.24	0.24	0.00	0.01	0.01	0.00	0.00	0.00	0.00
-2	58	0.81	0.78	0.03	0.18	0.21	-0.03	0.01	0.01	0.00	0.00	0.00	0.00
-2	59	0.94	0.94	0.00	0.06	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-2	60	0.79	0.82	-0.03	0.20	0.18	0.02	0.01	0.01	0.00	0.00	0.00	0.00
-2	61	0.86	0.86	0.00	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
-1	1	0.12	0.12	0.00	0.59	0.60	-0.01	0.27	0.27	0.01	0.02	0.01	0.01
-1	2	0.03	0.03	0.00	0.33	0.34	-0.01	0.54	0.51	0.03	0.10	0.11	-0.02
-1	3	0.03	0.04	-0.01	0.49	0.50	0.00	0.45	0.43	0.01	0.03	0.03	0.00
-1	4	0.10	0.09	0.01	0.53	0.54	-0.01	0.34	0.35	0.00	0.02	0.03	0.00
-1	5	0.03	0.04	-0.01	0.58	0.57	0.01	0.38	0.38	0.00	0.01	0.01	0.00
-1	6	0.10	0.11	-0.01	0.64	0.64	0.00	0.25	0.25	0.01	0.00	0.01	0.00
-1	7	0.05	0.06	-0.02	0.51	0.50	0.02	0.41	0.41	0.00	0.03	0.03	0.00
-1	8	0.11	0.11	0.00	0.44	0.42	0.02	0.37	0.38	-0.01	0.08	0.09	-0.01
-1	9	0.18	0.19	-0.01	0.63	0.62	0.01	0.18	0.19	0.00	0.01	0.01	0.00
-1	10	0.24	0.24	0.00	0.56	0.57	-0.01	0.19	0.18	0.01	0.01	0.01	0.00
-1	11	0.13	0.15	-0.02	0.68	0.65	0.03	0.18	0.19	-0.01	0.00	0.00	0.00
-1	12	0.15	0.16	-0.01	0.60	0.59	0.01	0.24	0.24	0.00	0.01	0.01	0.00
-1	13	0.19	0.19	0.00	0.60	0.60	0.00	0.21	0.21	0.00	0.01	0.01	0.00
-1	14	0.05	0.06	-0.01	0.47	0.48	-0.02	0.44	0.42	0.02	0.04	0.04	0.00
-1	15	0.15	0.14	0.02	0.56	0.56	0.00	0.27	0.29	-0.02	0.02	0.02	0.00
-1	16	0.18	0.17	0.01	0.69	0.71	-0.02	0.13	0.12	0.01	0.00	0.00	0.00
-1	17	0.08	0.09	-0.01	0.58	0.56	0.02	0.32	0.33	-0.01	0.02	0.02	0.00
-1	18	0.24	0.24	0.00	0.61	0.61	0.00	0.14	0.15	-0.01	0.01	0.00	0.00
-1	19	0.07	0.08	0.00	0.63	0.63	0.01	0.28	0.29	-0.01	0.01	0.01	0.01
-1	20	0.08	0.09	-0.02	0.60	0.58	0.02	0.31	0.31	0.00	0.01	0.01	0.00
-1	21	0.00	0.00	0.00	0.24	0.24	0.00	0.68	0.67	0.02	0.08	0.09	-0.02
-1	22	0.19	0.19	-0.01	0.62	0.60	0.03	0.18	0.20	-0.02	0.01	0.01	0.00
-1	23	0.09	0.08	0.01	0.68	0.67	0.01	0.23	0.25	-0.02	0.00	0.00	0.00
-1	24	0.02	0.02	0.00	0.42	0.42	0.00	0.52	0.53	0.00	0.04	0.04	0.00
-1	25	0.06	0.06	0.00	0.63	0.61	0.02	0.30	0.32	-0.02	0.01	0.01	0.00
-1	26	0.08	0.09	-0.01	0.48	0.48	0.00	0.39	0.38	0.01	0.05	0.05	0.00
-1	27	0.14	0.16	-0.02	0.50	0.50	-0.01	0.32	0.31	0.02	0.04	0.04	0.01
-1	28	0.10	0.10	0.00	0.54	0.56	-0.02	0.34	0.32	0.02	0.02	0.02	0.00
-1	29	0.05	0.06	-0.02	0.55	0.53	0.02	0.39	0.39	0.00	0.02	0.02	0.00
-1	30	0.16	0.17	0.00	0.56	0.59	-0.03	0.27	0.23	0.04	0.01	0.01	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
-1	31	0.59	0.60	-0.02	0.39	0.38	0.02	0.02	0.02	0.00	0.00	0.00	0.00
-1	32	0.57	0.56	0.01	0.40	0.41	-0.01	0.03	0.03	0.00	0.00	0.00	0.00
-1	33	0.56	0.55	0.01	0.42	0.43	-0.01	0.02	0.02	0.00	0.00	0.00	0.00
-1	34	0.67	0.67	0.00	0.32	0.32	0.00	0.01	0.01	0.00	0.00	0.00	0.00
-1	35	0.64	0.65	-0.01	0.34	0.33	0.01	0.03	0.02	0.00	0.00	0.00	0.00
-1	36	0.68	0.67	0.01	0.31	0.32	-0.01	0.01	0.01	0.00	0.00	0.00	0.00
-1	37	0.71	0.72	-0.01	0.28	0.27	0.01	0.01	0.01	0.00	0.00	0.00	0.00
-1	38	0.66	0.65	0.01	0.33	0.34	-0.01	0.02	0.01	0.01	0.00	0.00	0.00
-1	39	0.70	0.69	0.01	0.30	0.30	0.00	0.01	0.01	0.00	0.00	0.00	0.00
-1	40	0.63	0.64	-0.01	0.33	0.32	0.00	0.04	0.03	0.01	0.00	0.00	0.00
-1	41	0.65	0.65	0.00	0.32	0.32	0.00	0.03	0.02	0.00	0.00	0.00	0.00
-1	42	0.86	0.86	0.00	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-1	43	0.75	0.71	0.03	0.25	0.28	-0.03	0.01	0.01	0.00	0.00	0.00	0.00
-1	44	0.57	0.57	0.00	0.37	0.37	0.00	0.06	0.05	0.01	0.00	0.00	0.00
-1	45	0.55	0.53	0.01	0.38	0.40	-0.02	0.08	0.06	0.01	0.00	0.00	0.00
-1	46	0.54	0.53	0.01	0.44	0.44	0.00	0.02	0.02	-0.01	0.00	0.00	0.00
-1	47	0.63	0.65	-0.02	0.35	0.34	0.01	0.02	0.01	0.00	0.00	0.00	0.00
-1	48	0.63	0.61	0.01	0.35	0.36	-0.01	0.02	0.03	0.00	0.00	0.00	0.00
-1	49	0.79	0.79	0.00	0.20	0.20	-0.01	0.01	0.01	0.00	0.00	0.00	0.00
-1	50	0.54	0.55	0.00	0.40	0.40	0.00	0.05	0.05	0.00	0.00	0.00	0.00
-1	51	0.78	0.77	0.01	0.22	0.23	-0.01	0.01	0.01	0.00	0.00	0.00	0.00
-1	52	0.82	0.81	0.01	0.18	0.19	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
-1	53	0.76	0.75	0.02	0.23	0.25	-0.02	0.01	0.01	0.00	0.00	0.00	0.00
-1	54	0.59	0.59	0.00	0.38	0.39	0.00	0.03	0.03	0.00	0.00	0.00	0.00
-1	55	0.45	0.46	-0.01	0.49	0.48	0.01	0.06	0.07	-0.01	0.00	0.00	0.00
-1	56	0.73	0.74	-0.01	0.26	0.26	0.01	0.00	0.00	0.00	0.00	0.00	0.00
-1	57	0.30	0.30	0.00	0.60	0.60	0.01	0.10	0.10	0.00	0.00	0.00	0.00
-1	58	0.39	0.41	-0.02	0.53	0.51	0.02	0.08	0.08	0.00	0.00	0.00	0.00
-1	59	0.69	0.69	-0.01	0.30	0.30	0.01	0.01	0.01	0.00	0.00	0.00	0.00
-1	60	0.54	0.54	0.00	0.41	0.40	0.01	0.05	0.06	-0.01	0.00	0.00	0.00
-1	61	0.58	0.57	0.01	0.38	0.39	-0.01	0.04	0.04	0.01	0.00	0.00	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
0	1	0.01	0.01	0.00	0.24	0.24	0.01	0.61	0.61	-0.01	0.15	0.15	0.00
0	2	0.00	0.00	0.00	0.08	0.08	0.00	0.49	0.48	0.01	0.43	0.44	-0.01
0	3	0.00	0.00	0.00	0.11	0.10	0.01	0.62	0.62	0.00	0.27	0.28	-0.01
0	4	0.01	0.01	0.00	0.18	0.18	-0.01	0.58	0.60	-0.02	0.24	0.22	0.02
0	5	0.00	0.00	0.00	0.13	0.11	0.02	0.68	0.69	-0.01	0.19	0.21	-0.01
0	6	0.01	0.01	0.00	0.26	0.23	0.03	0.63	0.64	-0.01	0.10	0.12	-0.02
0	7	0.00	0.00	0.00	0.14	0.13	0.00	0.61	0.60	0.01	0.26	0.27	-0.01
0	8	0.02	0.02	0.00	0.20	0.19	0.01	0.49	0.49	0.01	0.30	0.31	-0.02
0	9	0.01	0.02	0.00	0.32	0.32	-0.01	0.58	0.57	0.01	0.09	0.09	-0.01
0	10	0.03	0.03	0.00	0.35	0.36	-0.01	0.50	0.51	-0.01	0.12	0.10	0.02
0	11	0.01	0.01	0.00	0.29	0.28	0.01	0.62	0.62	0.00	0.08	0.09	-0.01
0	12	0.02	0.02	0.00	0.25	0.28	-0.03	0.61	0.58	0.03	0.13	0.13	0.00
0	13	0.02	0.02	0.00	0.29	0.31	-0.02	0.57	0.56	0.01	0.13	0.11	0.02
0	14	0.00	0.00	0.00	0.13	0.13	0.00	0.58	0.58	-0.01	0.29	0.28	0.01
0	15	0.01	0.01	0.00	0.25	0.24	0.01	0.58	0.58	0.00	0.16	0.17	-0.01
0	16	0.01	0.01	0.00	0.35	0.34	0.02	0.59	0.61	-0.02	0.05	0.05	0.00
0	17	0.00	0.01	0.00	0.19	0.18	0.01	0.62	0.62	0.00	0.20	0.20	0.00
0	18	0.01	0.03	-0.01	0.34	0.37	-0.03	0.57	0.53	0.04	0.08	0.07	0.00
0	19	0.00	0.00	0.00	0.18	0.17	0.00	0.67	0.68	0.00	0.15	0.15	0.00
0	20	0.01	0.01	0.00	0.20	0.19	0.01	0.62	0.63	-0.01	0.18	0.18	0.00
0	21	0.00	0.00	0.00	0.01	0.02	0.00	0.42	0.43	-0.01	0.57	0.55	0.01
0	22	0.02	0.02	0.00	0.34	0.32	0.02	0.53	0.56	-0.03	0.11	0.11	0.01
0	23	0.00	0.00	0.00	0.18	0.18	-0.01	0.70	0.70	0.00	0.12	0.12	0.00
0	24	0.00	0.00	0.00	0.06	0.06	0.00	0.58	0.59	-0.01	0.36	0.36	0.00
0	25	0.00	0.00	0.00	0.17	0.14	0.03	0.66	0.69	-0.02	0.17	0.17	0.00
0	26	0.01	0.01	0.00	0.17	0.17	0.00	0.55	0.55	-0.01	0.27	0.27	0.01
0	27	0.02	0.02	0.00	0.25	0.25	0.00	0.53	0.52	0.01	0.20	0.21	0.00
0	28	0.01	0.01	0.00	0.20	0.20	0.01	0.59	0.60	-0.01	0.20	0.19	0.01
0	29	0.00	0.00	0.00	0.14	0.14	0.00	0.62	0.62	0.00	0.24	0.24	0.01
0	30	0.01	0.02	-0.01	0.27	0.29	-0.02	0.59	0.57	0.02	0.13	0.13	0.01

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
0	31	0.17	0.18	-0.01	0.63	0.60	0.03	0.19	0.21	-0.02	0.01	0.01	0.00
0	32	0.14	0.15	-0.01	0.63	0.60	0.02	0.23	0.24	-0.01	0.01	0.01	0.00
0	33	0.11	0.10	0.01	0.64	0.65	-0.01	0.25	0.24	0.00	0.00	0.01	0.00
0	34	0.15	0.14	0.01	0.72	0.71	0.01	0.13	0.15	-0.01	0.00	0.00	0.00
0	35	0.26	0.26	0.00	0.55	0.56	-0.01	0.18	0.17	0.01	0.01	0.01	0.00
0	36	0.21	0.22	0.00	0.61	0.62	-0.02	0.17	0.16	0.01	0.01	0.00	0.00
0	37	0.23	0.25	-0.02	0.64	0.63	0.01	0.13	0.12	0.01	0.00	0.00	0.00
0	38	0.17	0.17	0.00	0.65	0.66	-0.01	0.18	0.17	0.01	0.00	0.00	0.00
0	39	0.18	0.18	0.00	0.69	0.67	0.01	0.13	0.14	-0.02	0.00	0.00	0.00
0	40	0.32	0.32	0.00	0.50	0.50	-0.01	0.18	0.17	0.01	0.01	0.01	0.00
0	41	0.26	0.27	-0.02	0.55	0.55	0.00	0.18	0.17	0.01	0.01	0.01	0.01
0	42	0.41	0.38	0.03	0.56	0.58	-0.02	0.04	0.04	-0.01	0.00	0.00	0.00
0	43	0.27	0.28	-0.01	0.60	0.59	0.01	0.13	0.13	0.00	0.00	0.00	0.00
0	44	0.28	0.25	0.03	0.48	0.51	-0.03	0.23	0.22	0.01	0.02	0.02	0.00
0	45	0.18	0.21	-0.03	0.51	0.51	0.00	0.27	0.26	0.02	0.03	0.03	0.01
0	46	0.10	0.10	0.00	0.62	0.64	-0.02	0.28	0.26	0.02	0.01	0.01	0.00
0	47	0.18	0.18	0.00	0.65	0.65	0.01	0.17	0.17	-0.01	0.00	0.00	0.00
0	48	0.20	0.20	0.00	0.60	0.59	0.01	0.19	0.20	-0.01	0.01	0.01	0.00
0	49	0.38	0.38	0.01	0.54	0.54	0.00	0.08	0.08	0.00	0.00	0.00	0.00
0	50	0.21	0.20	0.01	0.51	0.53	-0.02	0.26	0.25	0.01	0.02	0.02	0.00
0	51	0.32	0.33	-0.01	0.58	0.58	0.01	0.09	0.09	0.00	0.00	0.00	0.00
0	52	0.33	0.34	-0.01	0.60	0.59	0.01	0.07	0.07	0.00	0.00	0.00	0.00
0	53	0.31	0.33	-0.02	0.58	0.56	0.02	0.11	0.10	0.00	0.00	0.00	0.00
0	54	0.17	0.18	-0.01	0.59	0.59	0.00	0.24	0.22	0.01	0.01	0.01	0.00
0	55	0.12	0.11	0.01	0.52	0.53	-0.02	0.33	0.33	0.00	0.04	0.03	0.01
0	56	0.21	0.21	0.00	0.69	0.68	0.01	0.10	0.11	-0.01	0.00	0.00	0.00
0	57	0.03	0.04	-0.01	0.43	0.45	-0.02	0.51	0.47	0.04	0.03	0.04	-0.01
0	58	0.08	0.09	-0.01	0.51	0.51	0.01	0.37	0.37	0.01	0.03	0.04	0.00
0	59	0.23	0.22	0.00	0.65	0.63	0.02	0.12	0.14	-0.02	0.00	0.00	0.00
0	60	0.21	0.21	0.00	0.54	0.52	0.02	0.22	0.25	-0.02	0.03	0.02	0.00
0	61	0.22	0.20	0.03	0.55	0.56	-0.01	0.22	0.23	-0.01	0.01	0.01	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
1	1	0.00	0.00	0.00	0.03	0.03	0.00	0.40	0.41	-0.01	0.58	0.57	0.01
1	2	0.00	0.00	0.00	0.01	0.01	0.01	0.19	0.21	-0.02	0.79	0.78	0.01
1	3	0.00	0.00	0.00	0.01	0.01	0.00	0.23	0.24	-0.01	0.76	0.75	0.01
1	4	0.00	0.00	0.00	0.02	0.02	0.00	0.35	0.34	0.01	0.64	0.64	0.00
1	5	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.25	-0.02	0.76	0.74	0.02
1	6	0.00	0.00	0.00	0.02	0.02	-0.01	0.39	0.41	-0.02	0.60	0.57	0.03
1	7	0.00	0.00	0.00	0.01	0.01	0.00	0.30	0.28	0.02	0.69	0.71	-0.02
1	8	0.00	0.00	0.00	0.05	0.05	0.01	0.35	0.34	0.01	0.60	0.61	-0.01
1	9	0.00	0.00	0.00	0.04	0.05	0.00	0.49	0.49	0.00	0.47	0.46	0.00
1	10	0.00	0.00	0.00	0.08	0.08	0.00	0.52	0.50	0.02	0.41	0.42	-0.01
1	11	0.00	0.00	0.00	0.04	0.03	0.01	0.45	0.47	-0.02	0.52	0.51	0.01
1	12	0.00	0.00	0.00	0.04	0.04	0.00	0.44	0.44	-0.01	0.52	0.52	0.01
1	13	0.00	0.00	0.00	0.05	0.05	0.00	0.48	0.47	0.01	0.48	0.48	0.00
1	14	0.00	0.00	0.00	0.01	0.01	0.00	0.29	0.28	0.01	0.70	0.71	-0.01
1	15	0.00	0.00	0.00	0.03	0.04	-0.01	0.42	0.40	0.02	0.56	0.56	-0.01
1	16	0.00	0.00	0.00	0.03	0.03	0.00	0.57	0.54	0.02	0.41	0.43	-0.02
1	17	0.00	0.00	0.00	0.02	0.02	0.00	0.33	0.35	-0.01	0.65	0.64	0.01
1	18	0.00	0.00	0.00	0.07	0.07	0.00	0.55	0.53	0.02	0.39	0.41	-0.02
1	19	0.00	0.00	0.00	0.02	0.01	0.01	0.37	0.35	0.02	0.62	0.64	-0.03
1	20	0.00	0.00	0.00	0.03	0.02	0.01	0.34	0.36	-0.02	0.64	0.63	0.01
1	21	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.08	-0.01	0.93	0.92	0.01
1	22	0.00	0.00	0.00	0.05	0.05	-0.01	0.48	0.48	0.00	0.48	0.47	0.01
1	23	0.00	0.00	0.00	0.01	0.01	0.00	0.36	0.37	-0.01	0.64	0.63	0.01
1	24	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.17	-0.01	0.84	0.83	0.01
1	25	0.00	0.00	0.00	0.00	0.01	0.00	0.30	0.31	-0.01	0.70	0.68	0.01
1	26	0.00	0.00	0.00	0.02	0.03	0.00	0.29	0.33	-0.04	0.68	0.64	0.04
1	27	0.00	0.00	0.00	0.06	0.06	0.01	0.39	0.40	-0.01	0.55	0.55	0.01
1	28	0.00	0.00	0.00	0.02	0.02	0.00	0.35	0.36	-0.01	0.63	0.62	0.02
1	29	0.00	0.00	0.00	0.01	0.01	0.00	0.27	0.29	-0.02	0.72	0.70	0.02
1	30	0.00	0.00	0.00	0.05	0.04	0.01	0.45	0.45	0.00	0.50	0.51	-0.01

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
1	31	0.02	0.02	0.00	0.31	0.31	0.00	0.56	0.57	-0.01	0.11	0.11	0.00
1	32	0.01	0.01	0.00	0.26	0.26	-0.01	0.60	0.59	0.01	0.13	0.13	0.00
1	33	0.00	0.00	0.00	0.22	0.22	0.00	0.67	0.66	0.01	0.11	0.12	-0.01
1	34	0.01	0.01	0.00	0.31	0.29	0.02	0.63	0.65	-0.02	0.05	0.06	-0.01
1	35	0.04	0.04	0.00	0.38	0.38	0.00	0.49	0.49	0.00	0.09	0.09	0.00
1	36	0.02	0.02	0.00	0.35	0.36	0.00	0.55	0.55	0.00	0.08	0.07	0.00
1	37	0.01	0.02	-0.01	0.38	0.40	-0.02	0.55	0.52	0.02	0.06	0.05	0.01
1	38	0.01	0.01	0.00	0.30	0.31	-0.01	0.61	0.60	0.01	0.08	0.08	0.00
1	39	0.01	0.01	0.00	0.32	0.34	-0.02	0.62	0.59	0.03	0.05	0.06	-0.01
1	40	0.08	0.08	0.00	0.39	0.40	-0.01	0.44	0.43	0.01	0.10	0.09	0.00
1	41	0.05	0.05	0.00	0.40	0.39	0.01	0.47	0.48	-0.01	0.09	0.09	0.00
1	42	0.04	0.04	0.00	0.55	0.56	-0.01	0.40	0.39	0.01	0.01	0.01	-0.01
1	43	0.04	0.04	0.01	0.42	0.42	0.00	0.48	0.49	-0.01	0.06	0.06	0.00
1	44	0.07	0.06	0.01	0.33	0.34	-0.01	0.45	0.46	-0.01	0.15	0.14	0.01
1	45	0.03	0.04	-0.01	0.29	0.31	-0.01	0.49	0.49	0.00	0.19	0.16	0.02
1	46	0.00	0.00	0.00	0.21	0.21	0.00	0.65	0.66	0.00	0.14	0.13	0.01
1	47	0.01	0.01	0.00	0.32	0.32	0.00	0.60	0.59	0.01	0.07	0.08	-0.01
1	48	0.02	0.02	-0.01	0.33	0.33	0.01	0.55	0.55	0.00	0.11	0.11	0.00
1	49	0.06	0.06	0.00	0.51	0.51	0.00	0.41	0.40	0.01	0.02	0.03	-0.01
1	50	0.03	0.03	0.00	0.31	0.30	0.01	0.50	0.51	-0.01	0.15	0.15	0.00
1	51	0.04	0.05	-0.01	0.49	0.47	0.01	0.43	0.44	-0.01	0.04	0.04	0.01
1	52	0.05	0.04	0.01	0.48	0.50	-0.02	0.44	0.43	0.01	0.03	0.03	0.00
1	53	0.05	0.05	0.00	0.48	0.46	0.02	0.42	0.44	-0.02	0.05	0.05	0.00
1	54	0.02	0.02	0.00	0.30	0.30	-0.01	0.56	0.56	0.00	0.13	0.12	0.01
1	55	0.01	0.01	0.00	0.20	0.21	0.00	0.59	0.58	0.01	0.20	0.21	-0.01
1	56	0.01	0.01	0.00	0.39	0.38	0.01	0.57	0.57	0.00	0.04	0.04	-0.01
1	57	0.00	0.00	0.00	0.08	0.09	-0.01	0.61	0.59	0.02	0.31	0.32	-0.01
1	58	0.00	0.01	0.00	0.18	0.17	0.00	0.58	0.57	0.01	0.24	0.25	-0.01
1	59	0.02	0.02	0.00	0.36	0.37	-0.02	0.55	0.55	0.01	0.07	0.06	0.01
1	60	0.05	0.04	0.01	0.29	0.31	-0.02	0.51	0.50	0.01	0.16	0.16	0.00
1	61	0.02	0.03	0.00	0.31	0.31	0.00	0.56	0.54	0.02	0.12	0.13	-0.02

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.11	0.01	0.88	0.89	-0.01
2	2	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.06	-0.01	0.95	0.94	0.01
2	3	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.96	0.96	0.00
2	4	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.10	0.00	0.90	0.90	0.00
2	5	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.00	0.97	0.97	0.00
2	6	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.09	0.00	0.91	0.91	0.00
2	7	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.07	-0.01	0.94	0.93	0.01
2	8	0.00	0.00	0.00	0.01	0.01	0.00	0.18	0.16	0.02	0.81	0.83	-0.02
2	9	0.00	0.00	0.00	0.01	0.00	0.00	0.15	0.15	0.00	0.85	0.85	0.00
2	10	0.00	0.00	0.00	0.01	0.01	0.00	0.24	0.21	0.03	0.75	0.78	-0.03
2	11	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.11	0.00	0.89	0.89	0.00
2	12	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.14	0.02	0.84	0.86	-0.02
2	13	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.16	0.01	0.84	0.84	0.00
2	14	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.07	-0.01	0.94	0.93	0.01
2	15	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.13	0.00	0.87	0.87	0.00
2	16	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.00	0.89	0.89	0.00
2	17	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.09	0.02	0.90	0.91	-0.02
2	18	0.00	0.00	0.00	0.01	0.00	0.00	0.18	0.19	-0.01	0.82	0.81	0.00
2	19	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.06	0.01	0.93	0.94	-0.01
2	20	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08	-0.01	0.92	0.92	0.01
2	21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	-0.01	1.00	0.99	0.01
2	22	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.17	0.00	0.84	0.83	0.00
2	23	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.05	-0.01	0.96	0.95	0.01
2	24	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	-0.01	0.98	0.98	0.01
2	25	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.95	0.95	0.00
2	26	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.00	0.89	0.89	0.00
2	27	0.00	0.00	0.00	0.01	0.01	0.01	0.15	0.17	-0.03	0.84	0.82	0.02
2	28	0.00	0.00	0.00	0.01	0.00	0.00	0.10	0.10	0.00	0.89	0.90	-0.01
2	29	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.06	0.01	0.93	0.94	-0.01
2	30	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.15	0.01	0.85	0.85	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
2	31	0.00	0.00	0.00	0.05	0.05	0.01	0.46	0.47	-0.01	0.49	0.49	0.01
2	32	0.00	0.00	0.00	0.03	0.04	0.00	0.43	0.43	0.00	0.54	0.53	0.01
2	33	0.00	0.00	0.00	0.01	0.02	-0.01	0.40	0.40	0.00	0.59	0.58	0.01
2	34	0.00	0.00	0.00	0.02	0.02	0.00	0.50	0.50	0.00	0.48	0.48	0.00
2	35	0.00	0.00	0.00	0.08	0.09	-0.01	0.49	0.51	-0.01	0.42	0.40	0.03
2	36	0.00	0.00	0.00	0.04	0.06	-0.01	0.53	0.52	0.01	0.43	0.43	0.01
2	37	0.00	0.00	0.00	0.05	0.06	-0.01	0.57	0.56	0.01	0.38	0.38	0.00
2	38	0.00	0.00	0.00	0.03	0.03	0.00	0.51	0.49	0.02	0.46	0.47	-0.02
2	39	0.00	0.00	0.00	0.04	0.04	0.00	0.52	0.53	-0.01	0.44	0.44	0.01
2	40	0.01	0.01	0.00	0.18	0.15	0.03	0.48	0.50	-0.02	0.34	0.35	-0.01
2	41	0.00	0.00	0.00	0.11	0.10	0.01	0.51	0.51	0.00	0.37	0.39	-0.01
2	42	0.00	0.00	0.00	0.10	0.10	0.00	0.67	0.68	-0.01	0.23	0.22	0.02
2	43	0.00	0.00	0.00	0.08	0.09	0.00	0.57	0.56	0.01	0.35	0.36	-0.01
2	44	0.01	0.01	0.00	0.10	0.11	-0.01	0.48	0.46	0.02	0.41	0.42	-0.01
2	45	0.01	0.00	0.00	0.07	0.08	-0.02	0.43	0.44	-0.01	0.50	0.47	0.02
2	46	0.00	0.00	0.00	0.02	0.02	0.00	0.41	0.39	0.02	0.58	0.60	-0.02
2	47	0.00	0.00	0.00	0.04	0.04	0.00	0.48	0.50	-0.02	0.48	0.47	0.01
2	48	0.00	0.00	0.00	0.07	0.06	0.01	0.47	0.48	-0.01	0.46	0.46	0.00
2	49	0.00	0.00	0.00	0.14	0.14	0.00	0.57	0.60	-0.04	0.29	0.26	0.04
2	50	0.00	0.00	0.00	0.08	0.07	0.01	0.43	0.45	-0.01	0.49	0.48	0.00
2	51	0.00	0.00	0.00	0.11	0.11	0.00	0.59	0.59	0.00	0.30	0.30	0.01
2	52	0.00	0.00	0.00	0.10	0.10	0.00	0.64	0.63	0.01	0.26	0.27	-0.01
2	53	0.00	0.00	0.00	0.12	0.12	0.00	0.58	0.57	0.01	0.30	0.31	-0.01
2	54	0.00	0.00	0.00	0.04	0.05	-0.01	0.48	0.46	0.02	0.48	0.49	-0.01
2	55	0.00	0.00	0.00	0.03	0.03	0.00	0.37	0.37	0.00	0.61	0.60	0.00
2	56	0.00	0.00	0.00	0.05	0.04	0.01	0.54	0.57	-0.03	0.41	0.39	0.02
2	57	0.00	0.00	0.00	0.01	0.01	0.00	0.23	0.23	0.00	0.77	0.77	0.00
2	58	0.00	0.00	0.00	0.02	0.02	0.00	0.34	0.33	0.01	0.64	0.65	-0.01
2	59	0.00	0.00	0.00	0.05	0.06	-0.01	0.55	0.54	0.01	0.40	0.41	-0.01
2	60	0.00	0.00	0.00	0.07	0.08	-0.01	0.45	0.45	0.01	0.47	0.47	0.00
2	61	0.00	0.00	0.00	0.06	0.06	0.00	0.46	0.46	0.00	0.48	0.48	0.00

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
3	1	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.98	0.98	0.00
3	2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.99	0.98	0.00
3	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1.00	0.99	0.00
3	4	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.98	0.98	0.00
3	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
3	6	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.99	0.99	0.00
3	7	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.99	0.99	0.00
3	8	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.07	0.00	0.94	0.93	0.00
3	9	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.97	0.97	0.00
3	10	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.94	0.94	0.00
3	11	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.98	0.98	0.00
3	12	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.97	0.97	0.00
3	13	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.01	0.96	0.97	-0.01
3	14	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.98	0.99	0.00
3	15	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.97	0.97	0.00
3	16	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.98	0.99	-0.01
3	17	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.98	0.98	0.00
3	18	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.96	0.96	0.00
3	19	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.99	0.99	0.00
3	20	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.01	0.98	0.99	-0.01
3	21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
3	22	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.97	0.96	0.00
3	23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1.00	1.00	0.00
3	24	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	1.00	1.00	0.00
3	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1.00	1.00	0.00
3	26	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	-0.01	0.98	0.97	0.01
3	27	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.06	-0.01	0.95	0.94	0.01
3	28	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.98	0.98	0.00
3	29	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.99	0.99	0.00
3	30	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.97	0.97	0.01

Appendix D (cont.): Polytomous IRT Observed and Predicted Responses

		Response = 1			Response = 2			Response = 3			Response = 4		
θ	Item	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias	Obs.	Pred.	Bias
3	31	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.13	-0.01	0.88	0.87	0.01
3	32	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.08	0.00	0.93	0.92	0.00
3	33	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.09	0.00	0.91	0.91	0.00
3	34	0.00	0.00	0.00	0.01	0.01	0.00	0.24	0.23	0.01	0.76	0.76	0.00
3	35	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.17	-0.01	0.84	0.83	0.01
3	36	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.18	0.01	0.80	0.82	-0.01
3	37	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.12	-0.01	0.89	0.88	0.01
3	38	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.13	0.00	0.88	0.87	0.00
3	39	0.00	0.00	0.00	0.03	0.03	0.00	0.31	0.30	0.01	0.66	0.67	-0.01
3	40	0.00	0.00	0.00	0.01	0.01	0.00	0.26	0.24	0.02	0.73	0.75	-0.02
3	41	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.75	0.75	0.00
3	42	0.00	0.00	0.00	0.01	0.01	0.00	0.22	0.22	-0.01	0.78	0.77	0.00
3	43	0.00	0.00	0.00	0.02	0.02	-0.01	0.28	0.26	0.02	0.70	0.72	-0.02
3	44	0.00	0.00	0.00	0.02	0.01	0.01	0.23	0.22	0.01	0.75	0.77	-0.01
3	45	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08	0.00	0.92	0.92	0.00
3	46	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.13	0.00	0.87	0.87	0.00
3	47	0.00	0.00	0.00	0.01	0.00	0.00	0.20	0.17	0.03	0.80	0.82	-0.03
3	48	0.00	0.00	0.00	0.01	0.01	0.00	0.30	0.29	0.01	0.69	0.70	-0.01
3	49	0.00	0.00	0.00	0.00	0.01	-0.01	0.21	0.20	0.01	0.79	0.79	0.00
3	50	0.00	0.00	0.00	0.01	0.01	0.00	0.26	0.25	0.01	0.73	0.74	-0.01
3	51	0.00	0.00	0.00	0.00	0.01	0.00	0.25	0.24	0.00	0.75	0.75	0.00
3	52	0.00	0.00	0.00	0.01	0.01	0.00	0.25	0.26	-0.01	0.74	0.73	0.01
3	53	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.16	0.00	0.84	0.84	0.00
3	54	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.12	-0.02	0.90	0.88	0.02
3	55	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.14	0.00	0.86	0.86	0.00
3	56	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.05	-0.01	0.96	0.95	0.01
3	57	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.90	0.90	0.00
3	58	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.17	0.01	0.83	0.83	-0.01
3	59	0.00	0.00	0.00	0.01	0.01	0.00	0.21	0.22	-0.01	0.78	0.77	0.01
3	60	0.00	0.00	0.00	0.01	0.01	0.00	0.17	0.18	-0.01	0.82	0.81	0.01
3	61	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.13	-0.01	0.88	0.87	0.01