A COMPARISON OF ADAPTIVE, SEQUENTIAL, AND CONVENTIONAL TESTING STRATEGIES FOR MASTERY DECISIONS

G. Gage Kingsbury
and
David J. Weiss

RESEARCH REPORT 80-4
NOVEMBER 1980

COMPUTERIZED ADAPTIVE TESTING LABORATORY
PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

This research was supported by funds from the Air Force Human Resources Laboratory, the Air Force Office of Scientific Research, the Army Research Institute, and the Office of Naval Research, and monitored by the Office of Naval Research.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Research Report 80-4

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A Comparison of Adaptive, Sequential, and Conventional Testing Strategies for Mastery Decisions

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
G. Gage Kingsbury and David J. Weiss

8. CONTRACT OR GRANT NUMBER(s)
N00014-79-C-0172

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P.E.: 6115N Proj.: RR042-04
T.A.: RR042-04-01
W.U.: NR 150-433

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
November 1980

13. NUMBER OF PAGES
28

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
This research was supported by funds from the Air Force Human Resources Laboratory, the Air Force Office of Scientific Research, the Army Research Institute, and the Office of Naval Research, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

achievement testing
mastery testing
adaptive testing
tailored testing
computerized testing
criterion-referenced testing
sequential testing
latent trait theory
item characteristic curve theory
item response theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two procedures for making mastery decisions with variable length tests and a conventional mastery testing procedure were compared in monte carlo simulation. The simulation varied the characteristics of the item pool used for testing and the maximum test length allowed. The procedures were compared in terms of the mean test length needed to make a decision, the validity of the decisions made by each procedure, and the types of classification errors made by each procedure. Both of the variable test length procedures were found to result in important reductions in mean test length from the conventional test length. The Sequential Probability Ratio Test (SPRT) procedure resulted in greater test length reductions, on the average, than the Adaptive Mastery Testing (AMT) procedure. However, the AMT procedure resulted both in more valid mastery decisions and in more balanced error rates than the SPRT procedure under all conditions. In addition, the AMT procedure produced the best combination of test length and validity.

CONTENTS

A Comparison of Adaptive, Sequential, and Conventional Testing Strategies for Mastery Decisions

The use of criterion-referenced achievement test interpretation has gained great support within the educational measurement community since its introduction less than two decades ago (Glaser & Klaus, 1962). It is intuitively appealing to educators to be able to measure students' performances against absolute standards of behavior on prespecified learning objectives, and the use of criterion-referenced test interpretation gives educators this capability. One of the most basic forms of criterion-referenced test interpretation involves classifying students into two categories--one containing students who have achieved a sufficient command of the subject matter (mastery) and the other containing students who have not achieved a sufficient command of the subject matter (nonmastery). Traditionally, students are declared masters when their scores on a conventional classroom achievement test are as high or higher than a prespecified cutoff level, or are declared nonmasters if their scores on the test are lower than the cutoff level. This form of classroom testing has been called mastery testing and can be useful (1) in determining the degree of student proficiency within a classroom and (2) as a diagnostic tool to identify individuals who need further training in specific instructional areas (Nitko & Hsu, 1974).

As traditional mastery testing has been developing its own technology, adaptive testing technology has also developed to allow educators to make maximum use of classroom testing time while reducing to a minimum the amount of time spent on testing. The use of adaptive testing techniques has recently been shown to be effective in reducing test length while obtaining high-fidelity achievement level estimates in several instructional settings (e.g., Bejar, Weiss, & Gialluca, 1977; Brown & Weiss, 1977).

Mastery and adaptive testing technologies have each shown their usefulness in the academic setting for different, but compatible, reasons. It is therefore not surprising that a fusion of the two techniques should occur in order to allow mastery testing to be accomplished in the shortest possible class time while maintaining the accurate decisions necessary for correct diagnoses of student instructional problems.

Alternatives to Conventional Mastery Tests

Two alternatives have been proposed to conventional mastery tests in which all test items are administered to every student: (1) Ferguson's (1969, 1970) application of Wald's (1947) Sequential Probability Ratio Test (SPRT) and (2) Kingsbury and Weiss's (1979) formulation of an item response theory (IRT; Lord & Novick, 1968) adaptive mastery testing method. Both of these testing procedures attempt to accomplish two common ends. First, the procedures seek to shorten the length of the test. Second, the procedures use statistical techniques designed to hold the number of misclassifications (i.e., individuals for whom the wrong decision is made) to some acceptable minimum. The methods by which these two procedures attempt to accomplish these ends are quite different, however.

The very fact that two procedures exist that attempt to accomplish the same basic ends through different techniques renders a comparison of the two methods desirable. The primary objective of this study was a comparison of the efficiency with which these two procedures for mastery testing achieved their goals of reducing test length while obtaining a high percentage of correct decisions.

SPRT Applied to Mastery Testing

The SPRT. Wald's (1947) SPRT was originally designed as a quality control test for use in a manufacturing setting. It was designed to determine whether a large consignment of light bulbs (or any other product) contained a small enough proportion of defective bulbs to pass some prespecified quality criterion while only testing a small sample of the light bulbs in the consignment. Wald's solution to this problem was to draw light bulbs sequentially from the consignment, to test the light bulb drawn at each stage, and to determine at each stage the relative probabilities of the following two hypotheses:

$$H_0 : p = p_0 \quad [1]$$

$$H_1 : p = p_1 \quad [2]$$

where

p = the proportion of defective elements (light bulbs) in the population (consignment);
p_0 = the proportion of defective elements in the population below which it is always desired to accept the quality of the population; and
p_1 = the proportion of defective elements in the population above which it is always desired to reject the quality of the population.

Since each stage of the sampling procedure may be viewed as a Bernoulli trial (given that each element is sampled at random without replacement from the population of equivalent elements and is assigned either nondefective or defective status), the probability of observing a certain number of defective elements in a sample of a certain size, given that either H_0 or H_1 is true, may be described with the binomial probability density function. Consequently, the probability of observing W defective elements in a sample of m elements (W_m), under H_0: $p = p_0$ is

$$P_{0_m} = p^{(m-W_m)} \; (1 - p_0)^{W_m}. \quad [3]$$

Under H_1: $p = p_1$, the probability becomes

$$P_{1_m} = p_1^{(m-W_m)} \; (1 - p_1)^{W_m}. \quad [4]$$

The ratio of these two probabilities yields an index of the relative strengths of the two hypotheses such that at each stage in the sampling procedure the quality of the consignment may be either rejected or accepted, or sampling of elements may be continued. The stringency of the test is based (1) on the proportion (α) of errors one is willing to tolerate in rejecting the quality of the consignments that actually do have the quality desired, and (2) on the proportion (β) of errors one is willing to tolerate in accepting the quality of consignments that do not actually have the minimum acceptable quality.

In its final log form the test used by the SPRT at each stage of sampling specifies that if

$$\text{Log}\ \frac{p_{1m}}{p_{0m}} \geq \text{Log}\ \frac{1-\beta}{\alpha}\ , \qquad [5]$$

the consignment is rejected; if

$$\text{Log}\ \frac{p_{1m}}{p_{0m}} \leq \text{Log}\ \frac{\beta}{1-\alpha} \qquad [6]$$

the consignment is accepted; and if

$$\text{Log}\ \frac{\beta}{1-\alpha} < \text{Log}\ \frac{p_{1m}}{p_{0m}} < \text{Log}\ \frac{1-\beta}{\alpha}\ , \qquad [7]$$

sampling continues.

Wald (1947) has shown that this testing procedure results in error levels approximating α and β across consignments. Further, it has been shown that the probability of not obtaining a decision for a consignment approaches zero as the sample size increases.

Ferguson's application to mastery testing. Ferguson (1969) has applied the SPRT within a mastery testing situation using students' responses to test items in place of light bulbs and a domain of items that represents an instructional objective instead of a consignment. The quality that Ferguson evaluated was students' command of the content area being tested. Ferguson also branched through an instructional hierarchy, applying the SPRT to various objectives of instruction. The present study, however, will consider only the application of the SPRT to a single instructional unit.

To employ the SPRT in a mastery testing situation, the educator must specify the following:

1. Two performance criterion levels--$\underline{p}_0$, serving as the lowest level at which a mastery decision will be made and $\underline{p}_1$, serving as the highest level at which a nonmastery decision will be made . These two values bound the neutral region in which testing will continue.

2. Two levels of error acceptance (α and β), which determine the strictness of the decision test and should reflect the relative costs of the two error types.

3. A maximum test length to constrain the testing time for individuals who are very difficult to classify.

One characteristic of this form of adaptive mastery testing is that it is fairly simple to implement within a classroom situation. The decision rule is easily incorporated into a chart (e.g., Wald, 1947, Sec. 5.3.2) showing the teacher or the student how many questions need to be answered correctly or incorrectly for each test length in order to terminate the test. Once the charts are made for various values of $\underline{p}_0$, $\underline{p}_1$, α, and β, the statistical work is completed and the SPRT procedure can be readily used by the educator. Since the procedure is sequential, however, it is not fully adaptive. Items are selected at random or in a fixed sequence; it is only the test length that varies for individuals.

IRT-Based Adaptive Mastery Testing

The paradigm for adaptive mastery testing (AMT) that Kingsbury and Weiss (1979) have proposed makes use of IRT and Bayesian statistical theory to adapt the mastery test to the individual's level of skill during the testing process. IRT is used (1) to estimate the parameters that most efficiently describe each of the items in the item pool and (2) to select the best items for each testee. Given the IRT parameter estimates, it is possible to apply an adaptive item selection and test termination procedure that will allow mastery decisions that are quite accurate to be made while shortening the length of the test needed for most individuals.

The AMT procedure is based on three integrated procedures (see Kingsbury & Weiss, 1979). These are

1. A procedure for individualizing the administration of test items,

2. A method for converting a traditional (proportion correct) mastery level to the latent achievement metric, and

3. A procedure for making mastery decisions using Bayesian confidence intervals.

Individualized item selection. To make mastery testing a more efficient process, it is desirable to reduce the length of each individual's test (1) by eliminating test items that provide little information concerning an individual's achievement level and (2) by terminating the AMT procedure after enough information has been gathered so that the mastery decision can be made with a high degree of confidence. To operationalize this goal, the item to be administered to an individual at any point during the testing procedure is selected on the basis of the amount of information that the item provides concerning the individual's achievement level estimate at that point in the test; that item is the item that most reduces the uncertainty in the person's achievement estimate. A procedure that selects and administers the most informative item at each point in an adaptive test--the maximum information search and selection (MISS) technique--has been described by Brown and Weiss (1977) and is used in the AMT procedure.

The information that an item provides at each point along the achievement continuum may be determined using the IRT model that is assumed to underly the individual's responses to the test items. The AMT procedure assumes the three-parameter logistic item response function (IRF), or item characteristic curve, model (Birnbaum, 1968). Using this model, the information available in any test. item is (Birnbaum, 1968, Equation 20.4.16)

$$I_i(\theta) = (1 - c_i)\, D^2 a_i^2 \psi^2 \; [DL_i(\theta)] \;/\; \{\psi[DL_i(\theta)] + c_i \psi^2 [-DL_i(\theta)], \qquad [8]$$

where

$\text{I}_\text{i}(\theta)$ = the information available from item $\underline{\text{i}}$ at any achievement level, θ;

$\underline{\text{c}}_\text{i}$ = the lower asymptote of the IRF for the item;

D = 1.7, a scaling factor used to allow the logistic IRF to closely approximate a normal ogive;

$\underline{\text{a}}_\text{i}$ = the discriminatory power of the item at the inflection point of the IRF;

= the logistic probability density function;

$\text{L}_\text{i}(\theta)$, = $\underline{\text{a}}_\text{i}(\theta - \underline{\text{b}}_\text{i})$, where $\underline{\text{b}}_\text{i}$ is the difficulty of the item; and

ψ = the cumulative logistic function.

If it is assumed that the achievement level estimate ($\hat{\theta}$) is the best estimate of the actual achievement level (θ), the item information of each of the items not yet administered may be evaluated at $\hat{\theta}$ at any point during the test by substituting $\hat{\theta}$ for θ in Equation 8. The item that has the highest information value at the individual's current level of $\hat{\theta}$ is thus chosen to be administered next. For this study a Bayesian estimator of the individual's achievement level, developed by Owen (1969), was used.

Mastery level. The classical mastery testing procedure specifies a percentage of the items on a test that must be correctly answered by an individual in order for him/her to be declared a master. Using IRT, it is possible to generate an analogue to the percentage correct mastery level of classical theory for use in adaptive testing, even though the use of MISS will tend to result in each person answering about 50% of the items correctly given a large enough item pool (because items administered will most probably have difficulty levels very close to the individual's level of $\hat{\theta}$). The analogue is based on the use of the test response function (TRF), or test characteristic curve (Lord & Novick, 1968). The TRF is the function that relates the achievement continuum to the expected proportion of correct answers that a person at any level of θ may be expected to obtain if all of the items on the test are administered.

For this procedure the assumption was made that a three-parameter logistic ogive described the functional relationship between the latent trait (achievement) and the probability of observing a correct response to any of the items on the test (i.e., the IRF). This assumption yields a TRF of the following form:

$$E(P|\theta) = \left(\sum_{i=1}^{n} (1 - c_i) + c_i \, \frac{1 + \exp[1.7a_i(b_i - \theta)]}{\exp[1.7a_i(b_i - \theta)]} \right) \bigg/ n \qquad [9]$$

where

$E(P|\theta)$ = the expected value of the proportion of correct answers observed on the test, given any achievement level;

$\underline{n}$ = the number of items on the test;

$\underline{c}_{\underline{i}}$ = the estimate of the lower asymptote for the IRF of item $\underline{i}$;

$\underline{a}_{\underline{i}}$ = the estimate of the item discrimination;

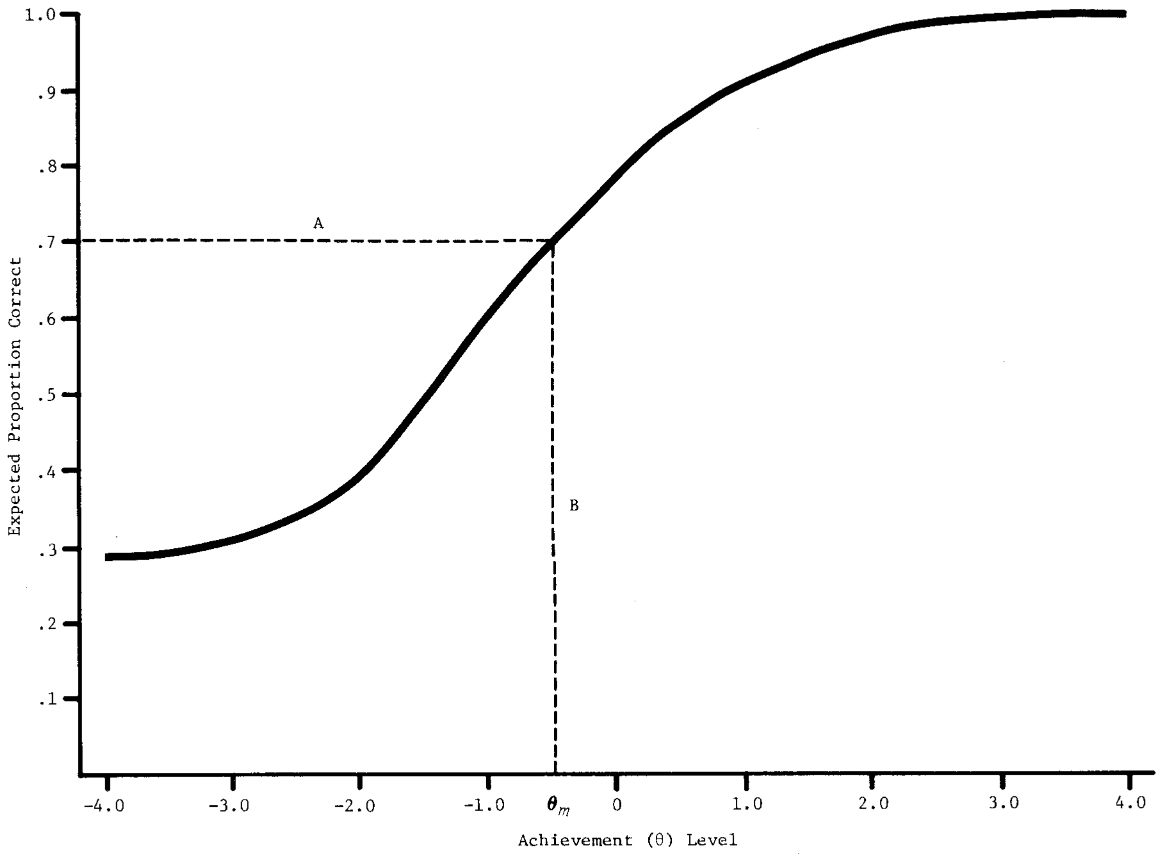
$\underline{b}_{\underline{i}}$ = the estimate of the difficulty of the item; and

θ = any given achievement level.

This monotonically increasing function enables expressing any given level of θ as its most likely proportion correct or, more importantly in this context, determining the level of θ that will most probably result in any given proportion of correct answers (Kingsbury & Weiss, 1979, pp. 3-4). Once this determination is made, IRT and its technology may be used to increase the efficiency of present mastery testing techniques.

Figure 1 shows the TRF for a hypothetical test, which may be used to illustrate how the transformation of a specified cutoff score from the proportion-correct metric to the latent achievement (θ) metric can be made. In Figure 1 a horizontal line (A) has been drawn from the prespecified cutoff score (70% cor-

Figure 1
Hypothetical Test Response Function Illustrated Conversion from the Proportion-Correct Metric to the Achievement Metric



rect) to the TRF. From this point of intersection, a vertical line (B) has been drawn to the θ metric. The point of intersection on the achievement metric (θ_m) is then the point on the θ metric to be used as the cutoff value, since a score higher than θ_m is most likely to yield a proportion-correct score above the cutoff score and a score lower than θ_m is most likely to yield a proportion-correct score below the cutoff score.

Making the mastery decision using Bayesian confidence intervals. Although any achievement level estimate obtained using IRT-based scoring of any subset of the items from a test will be on the same metric as the TRF for the original test, two different subsets of items may result in achievement level estimates that are not equally informative. For example, if one test consisted of many items and the other used only a few items, the longer test would probably yield a more precise θ estimate, provided that the items in the two tests had similar IRFs. Thus, IRT-based θ estimates that are on the same metric are comparable except for their differential precision. Comparisons of IRT-based θ estimates should therefore be based on confidence interval estimates instead of the point estimates.

For this reason, the AMT strategy makes mastery decisions with the use of Bayesian confidence intervals. Specifically, after each item is selected and administered to an individual (using MISS to choose the appropriate item at each point in the test), a point estimate of the individual's achievement level (θ) may be determined using Owen's Bayesian scoring algorithm, based on information gained from all items administered previously. Given this point estimate and the corresponding variance estimate, also obtained using Owen's procedure, a Bayesian confidence interval may be defined such that

$$\hat{\theta}_i - 1.96(\sigma_i^2)^{\frac{1}{2}} \leq \theta \leq \hat{\theta}_i + 1.96(\sigma_i^2)^{\frac{1}{2}}, \quad \text{with } p = .95 \qquad [10]$$

where

$\hat{\theta}_i$ = the Bayesian point estimate of achievement level, calculated following item i;

σ_i^2 = the Bayesian posterior variance following item i; and

θ = the true achievement level.

Equation 10 may be interpreted as meaning that the probability is .95 that the true value of the achievement level parameter, θ, is within the bounds of the confidence interval. It might also be said that there was 95% confidence that the true parameter value lies within the confidence interval.

After this confidence interval has been generated, it is a simple matter to determine whether or not θ_m, the achievement level earlier designated as the mastery level on the θ metric, falls outside the limits of the confidence interval. If it does not, the AMT procedure administers another item to the individual and recalculates the confidence interval. This procedure continues until, after some item has been administered, the confidence interval calculated does not include θ_m, the mastery level on the achievement continuum. At this point the testing procedure terminates and a mastery decision is made. If the lower limit of the confidence interval falls above the specified mastery level, θ_m, the individual is declared a master; if the upper limit of the confidence interval falls below θ_m, the individual is declared a nonmaster.

Given a finite item pool size, the testing procedure may exhaust the pool before a decision can be made in this manner. It is possible to make a decision concerning mastery for any of these individuals based on whether the Bayesian point estimate of their achievement level ($\hat{\theta}$) is above or below the specified mastery level, θ_m. These decisions, however, cannot be made with the same degree of confidence as those made with confidence intervals that do not contain the mastery level.

At any stage during the testing procedure, the individual's θ estimate will fall to one side or the other of the specified mastery cutoff, so for any one person only half of the dual inequality in Equation 10 needs to be evaluated at any time during the test. Thus, the decision rule implied by Equation 10 is functionally equivalent to using a one-sided Bayesian confidence interval with $p = .975$. The error rate that the procedure actually produces will be examined further below.

SPRT versus AMT

The two mastery testing strategies described above differ in a number of characteristics. The most salient of these differences are as follows:

1. Treatment of the items in the domain.
2. Treatment of the uncertainty of decisions.
3. Treatment of the mastery level.
4. Treatment of the achievement metric.

Treatment of items. The SPRT, as outlined above, treats all of the items in the mastery test as if they were perfect replicates of each other. Thus, an individual's response to a particular item is viewed solely as a probabilistic function of the individual's true mastery status. This assumption is most appropriate in the production setting in which Wald originally designed his procedure--that is, each light bulb can be expected to be like every other light bulb. This assumption may be less tenable in the mastery testing situation, where an individual's responses to test items may vary as a function of differential characteristics of the items themselves, as well as the individual's mastery status.

The AMT procedure assumes that if items differ, their individual characteristics may be described by a logistic ogive that varies as a function of the item's power to discriminate among individuals with different achievement levels (a), the item's difficulty (b), and the ease with which an individual may answer the item correctly with no knowledge of the subject matter (c). This assumption concerning the operating characteristics of the items is less restrictive than the assumption made in the SPRT; but to the extent that the items do not conform to the logistic form specified, the assumption might still restrict the efficiency of the AMT procedure.

Both mastery testing procedures, therefore, postulate some systematic similarities among the test items. To the extent that one of the postulations is closer to the actual state of the world than the other, it might be expected that the corresponding procedure would perform more efficiently. Thus, the characteristics of the item pool itself is the first point at which the two testing strategies diverge.

Treatment of uncertainty. The SPRT makes use of traditional hypothesis testing methods to determine the point at which an individual's item responses are sufficient evidence for making a decision concerning his/her mastery status. Here "sufficient" is defined in terms of the α and β error rates that are acceptable for the group of students being tested. α and β may be set independently to reflect the educator's concerns over the relative costs of the two error types.

The AMT procedure uses a symmetric Bayesian confidence interval to make the mastery decision. This functionally sets α equal to β and, by doing so, implies equal costs for the two error types. To the extent that the costs of the two error types are not equal, the SPRT procedure provides the educator with more flexibility than the AMT procedure as currently operationalized.

Treatment of mastery level. The SPRT uses a neutral region, rather than a single mastery level, to define the mastery and nonmastery regions. The specification of this neutral region is based on a decision by the educator concerning the range that appropriately reflects uncertainty as to whether the student's performance is actually the performance of a master or a nonmaster. By contrast, the AMT procedure defines a single mastery level and determines whether an individual is significantly above or below the mastery level using a Bayesian confidence interval.

This difference between the two testing procedures renders tentative any comparison that might be made. The performance of the SPRT procedure will vary widely as a function of the uncertainty band chosen. For the AMT technique this uncertainty is not directly taken into account. Any comparison between the two techniques is conditional upon the width and absolute bounds of the uncertainty region.

Treatment of the achievement metric. The decisions made by the SPRT are dependent on the percentage of items that are correctly answered for any specific test length. Thus, the metric of achievement assumed in this procedure is the proportion-correct metric. The AMT procedure assumes, due to the differential properties of the items in the item pool, that there is a nonlinear transformation of the proportion-correct metric that more accurately represents the achievement of the individuals taking the test. This latent continuum (the θ metric) serves as the achievement metric for the AMT procedure.

This difference in the achievement metric again renders comparisons between the two procedures somewhat difficult, since the "true" achievement levels of individuals must be postulated to fit one of these metrics. Any differences noted in the performance of the two procedures may be due to this difference in the achievement metrics assumed.

Method

Monte Carlo simulation was used to delineate circumstances in which one of the mastery testing procedures might have an advantage ove the other. The method used to compare the two variable-length mastery testing procedures (AMT and SPRT) to one another, as well as to a conventional (fixed length) testing procedure, consisted of five basic steps:

1. Four item pools were generated in which the items differed from one another to different degrees.

2. Item responses were generated for 500 simulated subjects (simulees) for each of the items in the four item pools.

3. Conventional tests of three different lengths were drawn at random from the larger item pools; these conventional tests served as item pools from which the SPRT and AMT procedures drew items.

4. The AMT and SPRT procedures were simulated for each of the four different item pool types and the three conventional test lengths.

5. Comparisons were drawn among the three types of tests (AMT, SPRT, conventional) concerning the degree of correspondence between the decisions made by the three test types and the true mastery status. Further comparisons were made based on the average test length that each test type required to reach its decisions.

Item Pool Generation

Four 100-item pools were generated to reflect different types of pools that might be used in a mastery test. Items were randomly ordered during the generation of each item pool.

Uniform pool. The uniform pool consisted of 100 items that were perfect replications of one another. Each item had discrimination (*a*) of 1.0, difficulty (*b*) of 0.0, and lower asymptote (guessing level, or *c*) of .2. This pool was designed to correspond to the SPRT procedure's assumption that all items in the test are similar.

b-variable pool. The *b*-variable pool varied from the uniform pool only in that the items had a range of difficulty levels. Eleven values of *b* were assigned to an approximately equal number of items in the pool. The values of *b* chosen were -2.5, -2.0, -1.5, -1.0 -.5, 0.0, .5, 1.0, 1.5, 2.0, and 2.5. Nine items at each level of difficulty were used in this pool, along with an additional item with *b* = 0.0 to bring the pool to 100 items. This item pool structure is shown in Appendix Table A.

a- and b- variable pool. The *a*- and *b*- variable pool differed from the *b*-variable pool in that the discriminations (*a*) of the items differed across a range of values. The *a* values used were .5, 1.0, 1.5, and 2.0. Each level of discrimination was equally represented in the item pool, and each level of item discrimination occurred with approximately the same frequency at each level of difficulty. This item pool structure is shown in Appendix Table B.

a-, b-, and c-variable pool. The *a*-, *b*-, and *c*-variable pool differed from the *a*- and *b*-variable pool in that the guessing levels of the items were allowed to spread across a range of values. The *c* values used were .1, .2, and .3. All *c* values were approximately equally represented. The parameter estimates were arranged such that each level of difficulty was represented by items that had approximately the same average _a_ level and the same average _c_ level (i.e., the pool was approximately rectangular). This item pool structure is shown in Appendix Table C.

Mastery Level Conversion

Each of the four item pools described above was designed so that a proportion-correct mastery level of .60 would correspond to an achievement level of 0.0 when conversion via the TRF was done. Figures 2a through 2d show the TRFs for the uniform; _b_-variable; _a_- and _b_-variable; and _a_-, _b_-, and _c_-variable item pools, respectively. These figures were developed by plotting the values of $E(P|\theta)$ given by Equation 9 as a function of θ, using each of the 100 items that made up each item pool to evaluate Equation 9. In each of these figures, the .6 proportion-correct mastery cutoff has been converted to the θ metric. In each item pool the converted value observed was $\theta_m=0.0$, as expected.

It should also be noted from these four figures that the TRF for the uniform item pool was quite different from the TRFs for the other three item pools (which were quite similar to one another). The TRF for the uniform item pool was much steeper for θ levels close to 0.0. This was to be expected, since each item in the uniform item pool had a difficulty level of 0.0, whereas in the other item pools the difficulty of items differed from one another, and each difficulty level was approximately equally represented. The uniform item pool thus has all of the characteristics of a peaked test concentrating its ability to differentiate contiguous θ levels immediately around $\theta=0.0$. The other item pools have the characteristics of rectangular tests, with more widespread measurement capacity. This spreading of the measurement capacity is represented directly in the shallow, steadily increasing TRFs for these three item pools.

Item Response Generation

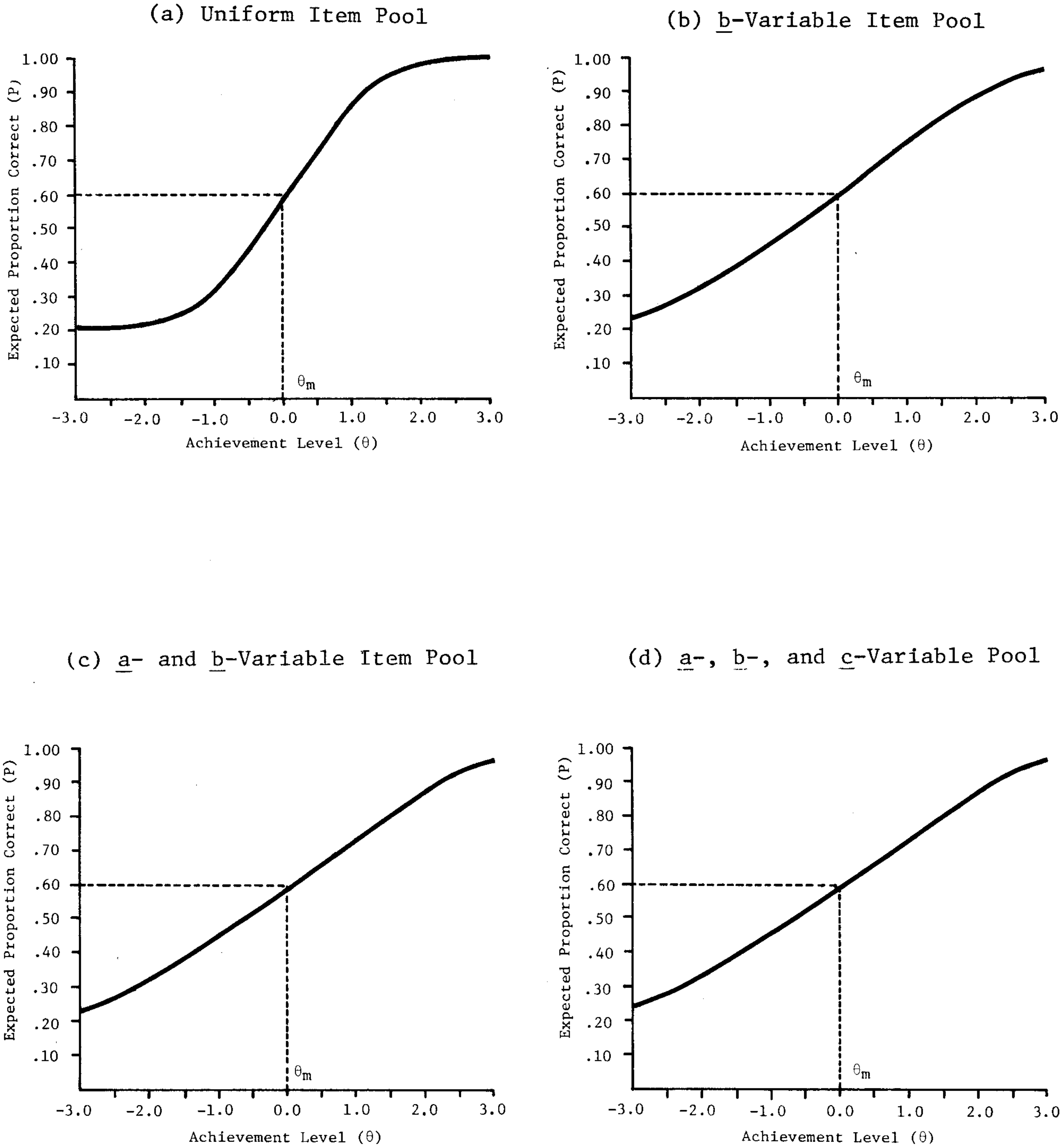
Achievement levels for 500 simulees were drawn from a normal distribution with a mean of 0.0 and a standard deviation of 1.0. Item responses for each of these simulees were then generated for each item in each of the four item pools using the three-parameter logistic IRT model. That is, knowing the θ level of the simulee and the parameters of the item in question, the expected probability of a correct response was calculated applying Equation 9, above, to the item. A random number was then drawn from a uniform distribution ranging from 0 to 1. If this number was lower than the probability of a correct response, the simulee was given a correct response to the item. If the number was higher than the correct response probability, the simulee was given an incorrect response.

Thus, in this study the achievement metric and the item response generator correspond closely to the model assumed by IRT. The "true" mastery level for each simulee was determined by comparing the θ level used to generate the item responses with the proportion correct mastery level expressed on the θ metric.

Conventional Tests

Conventional tests (CTs) of three different lengths (10, 25, and 50 items) were drawn at random from each of the four item pools, with the stipulation that

Figure 2
Test Characteristic Curves for Four Item Pools



(a) Uniform Item Pool

(b) *b*-Variable Item Pool

(c) *a*- and *b*-Variable Item Pool

(d) *a*-, *b*-, and *c*-Variable Pool

the shortest test served as the first portion of the next longer test and that this test in turn served as the first portion of the longest test. For each of the three variable pools (Appendix Tables A, B, and C), Items 1 to 10 comprised the 10-item CT, Items 1 to 25 comprised the 25-item CT, and Items 1 to 50 comprised the 50-item CT. (This procedure was not necessary for the uniform pool, since all items were identical.) These 12 CTs served as subpools from which the AMT and SPRT procedures drew items during the simulations. This random sampling from a larger domain of items was designed to correspond to the traditional mastery testing paradigm and to the random sampling model underlying the SPRT procedure.

Simulation of the Testing Strategies

Using the item response data for the 500 simulees and the item parameters available for each of the items (for the AMT procedure), the three testing strategies (AMT, SPRT, CT) were employed to make mastery decisions for each individual. Each testing procedure was used with each of the 12 subpools.

CT procedure. Two different decision processes were used with the CTs to examine the effect of choosing an achievement metric. The first decision process used proportion-correct scores (CT/PC) and a mastery criterion of 60% correct responses. After all of the items in the CT were administered, if the simulee answered 60% or more items correctly, the simulee was declared a master. If the simulee's score was less than 60% correct, the simulee was declared a nonmaster. This decision rule corresponded to the normal classroom use of conventional mastery tests with the proportion-correct achievement metric.

The second decision process used with the CT tests employed the latent achievement continuum and Bayesian scoring (CT/B) to make mastery decisions. Using the TRF the proportion-correct mastery level was converted to a mastery level on the θ metric. Then, after all of the items in the CT were administered to a simulee, the obtained response pattern was scored using the known IRF parameters for the items and Owen's Bayesian scoring procedure, assuming a prior distribution having a mean of 0.0 and a variance of 1.0. If the final θ estimate was higher than the mastery level on the θ metric, the simulee was declared a master. If the final estimate was lower than the mastery level, the simulee was declared a nonmaster.

SPRT procedure. For the SPRT procedure the limits of the neutral region were set at proportion-correct values of .5 and .7. Values of α and β were each set to .10. For simulees for whom no decision was made by the SPRT procedure before the item pool was exhausted, the mastery decision was made in the same way as it was for the conventional testing procedure, using a mastery proportion-correct value of .6.

AMT procedure. For the AMT procedure the mastery levels in each of the 100-item pools corresponding to 60% correct were determined from the TRF. This mastery level was used with each of the smaller item pools, even though they had not been designed to result in the same mastery level on the θ metric. This procedure added some sampling error to the AMT procedure in order to more appropriately reflect the error inherent when using estimated item parameters to determine the mastery level. As with the Bayesian scoring of the CT, each individual was given a prior distribution with a mean of 0.0 and a variance of 1.0 for the Bayesian scoring of the adaptive test.

Comparison Among the Testing Procedures

For each of the three testing procedures (AMT, SPRT, and CT), the value of the procedure may be judged by the average length of the test required to make the mastery decision and by how well the decisions that are made reflect the true state of nature. Specifically, the AMT and SPRT procedures were compared in terms of the average reduction in the length of the test required to make mastery decisions across the entire group of individuals. Further, all three procedures were compared in terms of how well the decisions they made corresponded with the true mastery status of the simulees.

Comparisons within each testing procedure concerning the average test length and the correspondence of decisions with true mastery status were made across all 12 combinations of test lengths and item pool types. For the CT procedure, the correspondence of decisions with true mastery status was examined using both the CT/PC and CT/B scores.

Results

Test Length

Table 1 shows the mean test length required by each of the testing procedures to make a decision concerning the mastery status of the simulees in the test group.

Uniform pool. As can be seen from Table 1, the AMT procedure resulted in some test length reduction for each maximum test length (MTL), with the reduction in test length increasing as the MTL increased. For the 10-item MTL, the percentage by which the CT length was reduced was 9.7%; for the 25-item MTL the reduction was 36%; and for the 50-item MTL the observed reduction was 54%.

For the SPRT procedure, also, increasing test length reduction was noted as MTL increased; and some reduction was noted at each level of MTL. For the 10-item MTL, the reduction observed was 12%. The 25-item MTL resulted in a 48% reduction. For the 50-item MTL the reduction was 69%. At all MTL levels the SPRT procedure resulted in a greater reduction of test length than the AMT procedure.

b-variable pool. For the pool in which the difficulty levels of the items differed, the data in Table 1 show the same trends that were noted for the uniform pool. The AMT procedure reduced the test length at each MTL, and the reduction increased with the MTL level. For the 10-item, 25-item, and 50-item MTL levels, the AMT procedure reduced test length by 6%, 28%, and 46%, respectively.

The SPRT procedure also reduced test length at each MTL level; however, in this case, the reductions were larger for the longer MTL levels, relative to those obtained with AMT. At the 10-item, 25-item, and 50-item MTL levels the test length reductions observed were 4%, 33%, and 57%, respectively.

Table 1
Mean Number of Items Administered to Each Simulee for Three Mastery Testing Strategies Using Each Item Pool, at Three Maximum Test Lengths

Item Pool and Testing Strategy	Maximum Test Length 10	25	50
Uniform Pool			
Conventional	10.00	25.00	50.00
AMT	9.03	15.99	23.00
SPRT	8.75	13.12	15.39
b-Variable Pool			
Conventional	10.00	25.00	50.00
AMT	9.43	18.09	27.17
SPRT	9.62	16.79	21.41
a- and b-Variable Pool			
Conventional	10.00	25.00	50.00
AMT	8.55	15.78	24.07
SPRT	9.41	15.78	18.55
a-, b-, and c-Variable Pool			
Conventional	10.00	25.00	50.00
AMT	8.73	16.35	23.39
SPRT	8.62	13.42	15.70

For this pool the AMT procedure resulted in slightly greater reduction in test length (average of 9.43 items) than the SPRT procedure (average of 9.62 items) at the 10-item MTL level, whereas the SPRT procedure resulted in greater test length reductions for the longer MTL levels. Across all MTL levels, both procedures reduced test length somewhat less for this item pool than for the uniform item pool.

a- and b-variable pool. The data in Table 1 indicate that the AMT procedure resulted in test length reduction at each MTL level with this item pool. Test length reduction was greater for the larger MTL levels. For the 10-item, 25-item, and 50-item MTL levels, the reductions in test lengths were 14%, 37%, and 52%, respectively.

For the SPRT procedure, test length reduction was again observed, increasing with MTL. The reductions in test length noted were 6%, 37%, and 63%, respectively, for the 10-item, 25-item, and 50-item MTL levels.

With this item pool, the AMT procedure produced a greater decrease in test length (mean of 8.55 items) than the SPRT procedure (mean of 9.41 items) for the 10-item MTL. For the 25-item MTL level, both procedures resulted in the same mean reductions in test length (15.78 items), but the SPRT procedure resulted in the greater test length reduction for the 50-item MTL. The two testing procedures resulted in mean test lengths which were, on the average, slightly longer than those observed for the uniform pool but shorter than those observed for the b-variable pool.

a-, b-, and c-variable pool. Table 1 shows that when the AMT procedure was used with this item pool, test length was again reduced at each MTL, with this reduction greater for the longer MTL levels. For the 10-item, 25-item, and 50-item MTL levels, the observed reductions in test length were 13%, 35%, and 53%, respectively.

For the SPRT procedure with this item pool, test length reduction was once more observed, with an increasing reduction as the MTL increased. The reductions noted were 14%, 46%, and 69%, respectively, for the 10-item, 25-item, and 50-item MTL levels.

For this item pool the SPRT procedure terminated using a smaller average number of items for each MTL. Further, the degree of test length reduction in this pool for both procedures at all MTL levels was quite similar to that observed for the uniform item pool.

Correspondence with True Mastery Status

For each of the simulees in the sample, the true θ level (the θ level that was used to generate the item responses) was known. Given this, it was known whether the individual's θ level was actually above or below the mastery level as determined on the achievement metric ($\theta = 0.0$). Phi correlations between true mastery status and the mastery state (correspondence coefficients) determined by each of the testing procedures for each MTL level and pool type are shown in Table 2.

Uniform pool. The major trend observed for the uniform pool was that for each testing procedure an increase in the MTL level was accompanied by an increase in correspondence coefficients. In addition, it was observed that for the 10-item and 25-item MTL levels, the AMT procedure produced the highest correspondence coefficients observed (r = .775 and .840, respectively). For the 50-item MTL level the CT/PC procedure resulted in the highest correspondence (r = .875). The CT/B procedure resulted in the lowest correspondence coefficient observed at each MTL level. However, the differences in correspondence between MTL levels within any testing procedure were generally larger than the differences noted between testing procedures within a single MTL level.

b-variable pool. The same major trend that was found for the uniform pool was again observed in the b-variable pool: Each testing strategy resulted in higher correspondence as the MTL level increased. However, for this item pool the AMT procedure resulted in the highest correspondence coefficient at each MTL level. The CT/B procedure resulted in the next highest correspondence coefficient for the two highest MTL levels, and the CT/PC procedure tied with the SPRT procedure for the next highest correspondence at the 10-item MTL level.

Differences in correspondence coefficients observed between testing procedures within an MTL level were larger in this pool than in the uniform pool but were still somewhat smaller than the differences noted between MTL levels, on the average. For this pool each correspondence level observed was lower than for the uniform pool across all MTL levels and testing procedures.

Both the AMT procedure and the CT/B resulted in higher correspondence coef-

Table 2
Phi Correlations Between Observed Mastery State and True Mastery State for Each Mastery Testing Strategy, Using Each Type of Item Pool, at Three Maximum Test Lengths

Item Pool and Testing Strategy	Maximum Test Length		
	10	25	50
Uniform Pool			
CT/PC	.771	.837	.875
CT/B	.706	.803	.863
AMT	.775	.840	.871
SPRT	.771	.837	.867
b-Variable Pool			
CT/PC	.541	.667	.783
CT/B	.533	.714	.791
AMT	.615	.715	.828
SPRT	.541	.656	.704
a- and b-Variable Pool			
CT/PC	.626	.719	.771
CT/B	.638	.763	.788
AMT	.638	.756	.778
SPRT	.626	.698	.720
a-, b-, and c-Variable Pool			
CT/PC	.290	.670	.735
CT/B	.485	.741	.804
AMT	.470	.733	.787
SPRT	.290	.592	.571

ficients at the 25-item MTL level (.715 and .714) than the SPRT procedure did at the 50-item MTL level (.704).

a- and b- variable pool. Again, each testing strategy resulted in higher correspondence coefficients as MTL increased. In this pool, however, CT/B resulted in the highest correspondence coefficients at each MTL level (tied with the AMT procedure at the 10-item MTL level). The AMT procedure resulted in the next highest correspondence coefficients at the 25-item and 50-item MTL levels.

Differences in correspondence coefficients among the various testing procedures within any MTL level still showed a tendency to be smaller than differences within any particular testing procedure across MTL levels. The correspondence coefficients observed for the CT/B and the AMT procedure at the 25-item MTL level (.763 and .756) were each higher than the correspondence coefficient observed for the SPRT procedure at the 50-item MTL level.

a-, b-, and c-variable pool. The same pattern of increasing correspondence with increasing MTL was again noted for the CT/PC, CT/B, and AMT procedures. For the SPRT procedure the correspondence peaked at r = .592 at the 25-item MTL and dropped to .571 at the 50-item MTL. The CT/B procedure resulted in the highest correspondence coefficients at all three MTL levels. The AMT procedure produced the next highest correspondence for all three MTL levels. The SPRT procedure resulted in the lowest level of correspondence at all MTL levels (tied with the CT/PC procedure at the 10-item MTL level).

Once again, the average difference in correspondence was much greater between MTL levels within testing strategies than between testing strategies within a single MTL level. Further, on the average, the correspondence coefficients for this pool were lower than for the other pools, with rather large decreases at the 10-item MTL level, particularly for the CT/PC and SPRT strategies.

With this item pool, the SPRT testing procedure produced a lower correspondence coefficient at the 50-item MTL level (.571) than any of the other three procedures produced at the 25-item MTL level (.670 for the CT/PC procedure, .741 for the CT/B procedure, and .733 for the AMT procedure).

Correspondence as a Function of Test Length

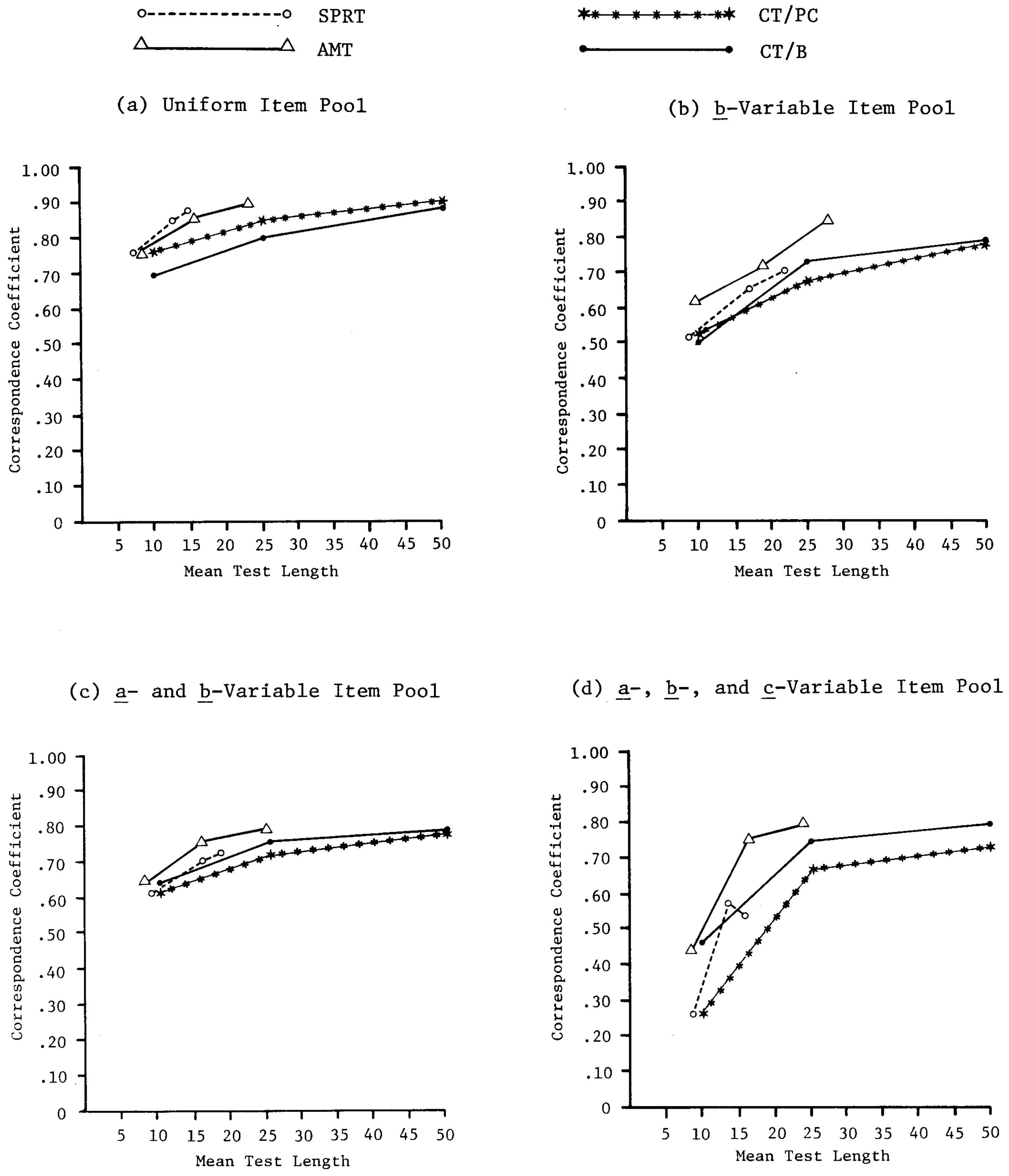
Figures 3a through 3d combine the observations made above concerning test length and correspondence. These figures show the correspondence coefficients (as reported in Table 2) as a function of the mean number of items administered by each testing strategy at each MTL level (as reported in Table 1) separately for each item pool. A testing strategy can be said to be most efficient to the extent that it yields the highest correspondence coefficient for any given mean test length, or combination of highest correspondence level and shortest test length.

For the uniform item pool, Figure 3a indicates that the SPRT procedure was the most efficient, as defined above. At all mean test lengths for which data were available, the SPRT procedure resulted in the combination of highest correspondence coefficient and shortest test length. The correspondence coefficient observed for the SPRT at a mean test length of 15.39 items (the longest mean test length observed for the SPRT procedure) was .867. To achieve this correspondence level (interpolating from the data in Figure 3a), the AMT procedure would need to administer approximately 21 items, the CT/PC procedure would need about 45 items, and the CT/B procedure would need more than 50 items.

For the _b_-variable item pool, Figure 3b indicates that the AMT procedure was the most efficient testing procedure. The correspondence coefficient observed for the AMT procedure using an average test length of 27.17 items (the longest mean test length observed for the AMT procedure) was .828. As Figure 3b shows, this was the highest coefficient observed for any test strategy for any mean test length using this item pool. Each of the CT procedures would have required more than 50 items to achieve this level of correspondence. For the SPRT procedure, the longest mean test length observed was 21.41 items. At this mean test length the SPRT procedure resulted in a correspondence level of .704, whereas for the same test length the AMT procedure would have resulted in a correspondence level (again through linear interpolation) of approximately .745.

For the _a_- and _b_- variable item pool, Figure 3c shows that the AMT procedure was again the most efficient test procedure at all observed mean test lengths, achieving the combination of highest correspondence levels and lowest

Figure 3
Correspondence Coefficients as a Function of the Mean Number of Items Administered by Each Testing Strategy for Each Item Pool



(a) Uniform Item Pool

(b) b-Variable Item Pool

(c) a- and b-Variable Item Pool

(d) a-, b-, and c-Variable Item Pool

test lengths for all three MTLs. At a mean test length of 24.07 items (the longest mean test length observed for the AMT procedure), the AMT procedure resulted in a correspondence coefficient of .778. The CT/PC procedure would require more than 50 items to produce as high a correspondence level, and the CT/B procedure would require about 39 items. For the SPRT procedure, the longest mean test length observed was 18.55 items. At this test length the SPRT procedure resulted in a correspondence coefficient of .720, whereas interpolation shows that the AMT procedure at the same test length would yield a correspondence coefficient of about .765.

Figure 3d shows that the AMT procedure was also the most efficient testing procedure at all observed mean test lengths with the a-, b-, and c-variable item pool. At its longest observed mean test length, 23.39 items, the AMT procedure resulted in a correspondence coefficient of .787. The CT/PC procedure would need to administer more than 50 items to achieve this correspondence level, and the CT/B procedure would need about 46 items. The longest mean test length observed with the SPRT was 15.70 items, which resulted in a correspondence level of .571. Interpolation of the data in Figure 3d shows that the AMT would probably result in a correspondence coefficient of about .710 at this test length.

The data in Figure 3 also show that the differences in the efficiencies of the testing strategies were much more pronounced in the a-, b-, and c-variable item pool than in any of the other three item pools. For instance, in the a-, b-, and c-variable item pool the scoring system used with the conventional tests made a consistent difference in the correspondence coefficient observed at each test length. The magnitude of this difference was much greater with this item pool than with any other item pool.

Frequency and Type of Errors

To further compare the performance of the mastery testing strategies the frequency with which each procedure made incorrect decisions of false mastery and false nonmastery was examined; the percentage of decision errors made by each of the testing strategies with each of the item pools at each MTL is shown in Table 3. The "Total" column in Table 3 reproduces in a different manner the information already reported from the correlational analysis. For each situation in which a high correlation was noted, a correspondingly low total error rate is noted in Table 3, as expected.

Uniform pool. For the uniform pool each of the testing strategies resulted in the same general pattern of errors across MTL levels. Each procedure resulted in fewer errors of each type with increased MTL. The difference in the frequencies of false mastery and false nonmastery decisions was smaller with higher MTL levels for all procedures except the CT/B procedure. The differences among the procedures in terms of the types of false decisions made were minimal, again except for the CT/B procedure, which resulted in higher total error rates at each MTL level than any other testing strategy.

b-variable pool. For this item pool the patterns of errors made by the different testing strategies were less regular than in the uniform pool. The CT/PC and SPRT procedures produced more false mastery than false nonmastery decisions at all MTL levels. The AMT procedure produced more false mastery than

Table 3
Percentage of Incorrect Decisions by Type of Error Made by Each Testing Strategy Using Each Type of Item Pool, at Three Maximum Test Lengths

Item Pool and Testing Strategy	Maximum Test Length								
	10			25			50		
	False Mastery	False Non-Mastery	Total	False Mastery	False Non-Mastery	Total	False Mastery	False Non-Mastery	Total
Uniform Pool									
CT/PC	3.6	8.0	11.6	2.6	5.6	8.2	2.8	3.4	6.2
CT/B	5.4	9.4	14.8	5.0	4.8	9.8	3.8	3.0	6.8
AMT	3.6	7.8	11.4	3.0	5.0	8.0	3.0	3.4	6.4
SPRT	3.6	8.0	11.6	2.6	5.6	8.2	3.2	3.4	6.6
b-Variable Pool									
CT/PC	22.4	2.2	24.6	13.4	3.6	17.0	6.4	4.4	10.8
CT/B	14.0	9.2	23.2	7.8	6.4	14.2	4.8	5.6	10.4
AMT	12.2	7.0	19.2	6.6	7.6	14.2	3.4	5.2	8.6
SPRT	22.4	2.2	24.6	14.2	3.4	17.6	11.4	3.6	15.0
a- and b-Variable Pool									
CT/PC	13.4	5.4	18.8	6.8	7.2	14.0	7.2	4.2	11.4
CT/B	9.8	8.2	18.0	6.0	5.8	11.8	4.4	6.2	10.6
AMT	9.8	8.2	18.0	5.4	6.8	12.2	3.8	7.4	11.2
SPRT	13.4	5.4	18.8	8.4	6.6	15.0	9.2	4.8	14.0
a-, b-, and c-Variable Pool									
CT/PC	0.0	44.6	44.6	2.6	15.2	17.8	7.4	5.8	13.2
CT/B	8.4	18.0	26.4	5.0	8.0	13.0	3.8	6.0	9.8
AMT	8.0	19.4	27.4	5.2	8.2	13.4	5.0	5.6	10.6
SPRT	0.0	44.6	44.6	2.0	21.0	23.0	3.8	19.4	23.2

false nonmastery decisions at the 10-item MTL level, but produced more false nonmastery than false mastery decisions at the two higher MTL levels. The CT/B procedure produced more false mastery than false nonmastery decisions for the two lowest MTL levels, but the trend was reversed for the highest MTL level. For the AMT procedure and the CT/B procedure the discrepancy in the frequencies of the two types of errors was smaller than for the other two procedures at all three MTL levels and was quite small (less than 2%) for the two higher MTL levels. For the CT/PC procedure the difference in the frequencies of the two types of errors was quite small at the highest MTL level; but for the SPRT procedure, a fairly large discrepancy between the two error rates (8% to 20%) was observed at each MTL.

In all testing conditions but one (AMT with a 25-item MTL), the use of the b-variable item pool resulted in higher discrepancies between the two observed error rates (as well as higher absolute error rates) than when the uniform pool was used.

a- and b-variable pool. Using this item pool, the SPRT procedure resulted in more false mastery decisions than false nonmastery decisions for all three MTL levels. The CT/B procedure resulted in a predominance of false mastery decisions for the two lowest MTL levels and a predominance of false nonmastery decisions at the 50-item MTL level. The CT/PC procedure produced a greater percentage of false mastery decisions at the 10-item and 50-item MTL levels, but a greater percentage of false nonmastery decisions at the 25-item MTL level. The AMT procedure resulted in a greater percentage of false mastery decisions at the 10-item MTL level, but a greater percentage of false nonmastery decisions at the 25-item and 50-item MTL levels.

Small discrepancies in error rates (less than 2%) were observed for the CT/B procedure at all MTL levels for the AMT procedure at the 10-item and 25-item MTL levels, and for the SPRT procedure and the CT/PC procedure at the 25-item MTL level only. The SPRT procedure resulted in the largest discrepancy in error rates at all MTL levels (tied with the CT/PC procedure at the 10-item MTL).

a-, b-, and c-variable pool. For this item pool, each of the testing procedures resulted in higher frequencies of false nonmastery decisions than false mastery decisions for the 10-item and 25-item MTL. For the 50-item MTL the CT/PC procedure resulted in a higher frequency of false mastery decisions, but the CT/B, AMT, and SPRT procedures still resulted in higher percentages of false nonmastery decisions.

The AMT procedure used with this item pool resulted in smaller differences in the frequencies of the two error types than any of the other testing procedures at the 50-item MTL level. The CT/B and AMT procedures tied for the lowest discrepancy in error rates (3%) at the 25-item MTL level. The CT/B procedure resulted in the lowest discrepancy in error rates (9.6%) at the 10-item MTL level. For the 50-item MTL level the AMT procedure produced a very small difference in the two error rates (.6%). The CT/PC procedure also produced a small difference in the two error rates for the 50-item MTL level (1.6%). The SPRT procedure resulted in the highest difference between the two error rates at all MTL levels (tied with the CT/PC procedure at the 10-item MTL.)

One interesting result was observed when the errors made with the *b*-variable item pool were compared with those made using the *a*-, *b*-, and *c*-variable item pool. For the *b*-variable pool, each of the testing procedures was more likely to make false mastery decisions than false nonmastery decisions. This tendency was reversed for the *a*-, *b*-, and *c*-variable item pool, where each of the procedures made more false nonmastery decisions than false mastery decisions. These trends were most noticeable for each of the testing procedures at the 10-item MTL level and most noticeable for the SPRT procedure across all MTL levels.

It is probable that these trends were artifacts of the random sampling of items used to create the conventional tests, since the shorter conventional tests would be less representative of the item domain due to the small sample of items taken. The results obtained here would be explained by a very easy 10-item conventional test being drawn from the *b*-variable pool and a very difficult 10-item test being drawn from the *a*-, *b*-, and *c*-variable pool. In fact, the mean *b*-value for the 10-item conventional test drawn from the *b*-variable pool was -.80; for the *a*-, *b*-, and *c*-variable pool, it was 1.25. This would also explain the observation that the SPRT procedure most clearly showed these trends, since the SPRT procedure used shorter test lengths, on the average, than the other two procedures to make its final decisions and therefore was most prone to small-sample artifacts.

Discussion and Conclusions

Several trends were noted in the data concerning the performance of the different testing strategies with the four different item pools. In every instance the AMT and SPRT procedures produced reductions in the mean test length required to make mastery decisions. This reduction increased with the MTL in each item pool. The AMT procedure resulted in reductions of 6% to 54% from the length of the conventional test. The SPRT procedure resulted in reductions of 4% to 69%. On the average, the SPRT procedure required fewer items to make the mastery decision.

The correspondence between the estimated mastery status and the true mastery status systematically increased with MTL for all testing procedures in each item pool (with the exception of the SPRT procedure used with the *a*-, *b*-, and *c*-variable item pool). Either the AMT procedure or the CT/B procedure resulted in the highest level of correspondence in all circumstances but one (the conventional tests performed best for the 50-item MTL with the uniform pool). On the average, though, the differences between different MTL levels were more pronounced than differences between testing procedures. Further, the type of item pool used had important effects on the correspondence obtained.

In the *b*-variable and the *a*- and *b*-variable item pools, the AMT and CT/B procedures resulted in higher levels of correspondence at the 25-item MTL level then did the SPRT procedure at the 50-item MTL level. For the *a*-, *b*-, and *c*-variable pool, the AMT, CT/B, and CT/PC procedures each resulted in higher correspondence at the 25-item MTL level than did the SPRT procedure at the 50-item MTL level.

When correspondence coefficients were examined as a function of the mean number of items administered by each of the testing strategies, the results differed depending on the type of item pool used. For the uniform item pool the SPRT procedure was the most efficient testing procedure (i.e., it produced the highest correspondence level for any mean test length) for all observed mean test lengths. Using each of the other item pools, the AMT procedure was the most efficient testing strategy at all observed mean test lengths. Both the AMT and SPRT procedures were more efficient than either conventional testing procedure at all observed mean test lengths using the uniform and *b*-variable item pools. Using the *a*- and *b*-variable and *a*-, *b*-, and *c*-variable item pools, only the AMT procedure was consistently more efficient than the conventional testing procedures.

The AMT procedure resulted in the most even frequencies in the types of decision errors made across most MTL levels and item pools. This was desirable, since both error types were assumed to have the same relative cost. Further, it was noted that the SPRT procedure was most susceptible to small-sample artifacts, resulting in the largest imbalance in the frequencies with which the two types of errors were made.

To prescribe the best testing strategy of those described here requires specification of priorities and conditionals. If testing time is at a premium in a course of instruction, then it might be important to shorten the length of tests in the course without much loss of decision accuracy. Both the AMT and SPRT procedures were designed to accomplish this. Results from this study indicate that the performance of these strategies is affected greatly by the characteristics of the items available for use. If a uniform item pool is available in which all items are equal in difficulty, discrimination, and guessing parameters, then the SPRT procedure will require the fewest items while resulting in decisions having correspondence coefficients that are quite comparable to the other three procedures. If, however, the item pool includes items with variable *a*, *b*, and *c* parameters, the SPRT procedure results in the shortest tests, but each of the other procedures will make more accurate classifications. Using a realistic item pool of this type, the AMT procedure provides the optimal combination of high decision correspondence and short test length. These factors must be considered before any decision is made as to which procedure is "best."

Several variations on the SPRT procedure and the AMT procedure that were not examined in this study are possible. For instance, Reckase (1980) has suggested that the SPRT procedure might be expanded to use the three-parameter logistic IRT model. This would remove the restrictive assumption from the SPRT model that all test items have identical characteristics. At the same time, this more general response model would allow use of the differential characteristics of the items as a basis for item selection and presentation. Comparison of the performance of the AMT procedure to this expanded SPRT testing procedure should be considered as a subject for future research.

References

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. *An information comparison of conventional and adaptive tests in the measurement of classroom achievement* (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Brown, J. M., & Weiss, D. J. *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977.

Ferguson, R. L. *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). University of Pittsburgh, Learning and Research Development Center, 1969. (ERIC Document Reproduction No. ED 037 089)

Ferguson, R. L. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. (Doctoral dissertation, University of Pittsburgh, 1969) *Dissertation Abstracts International*, 1970, *30*, 3856A. (University Microfilms No. 70-4530).

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in system development*. Chicago: Holt, Rinehart, & Winston, 1962.

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

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

Nitko, A., & Hsu, T. C. Using domain referenced tests for student placement, diagnosis, and attainment in a system of adaptive individualized instruction. *Educational Technology*, 1974, *14*, 48-53.

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

Reckase, M. D. Some decision procedures for use with tailored testing. In D. J. Weiss (Ed)., *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980.

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

APPENDIX: SUPPLEMENTARY TABLES

Table A
Item Parameter Values for Each of the Items
in the b-Variable Item Pool

Item				Item				Item			
Number	a	b	c	Number	a	b	c	Number	a	b	c
1	1.0	-2.5	.2	35	1.0	-1.5	.2	68	1.0	1.5	.2
2	1.0	-1.0	.2	36	1.0	-1.5	.2	69	1.0	-0.5	.2
3	1.0	-1.5	.2	37	1.0	0.5	.2	70	1.0	-0.5	.2
4	1.0	2.5	.2	38	1.0	-1.0	.2	71	1.0	-0.5	.2
5	1.0	-2.5	.2	39	1.0	0.5	.2	72	1.0	2.0	.2
6	1.0	0.5	.2	40	1.0	1.5	.2	73	1.0	2.0	.2
7	1.0	-1.5	.2	41	1.0	2.5	.2	74	1.0	-1.0	.2
8	1.0	-0.5	.2	42	1.0	-0.0	.2	75	1.0	2.0	.2
9	1.0	-2.5	.2	43	1.0	0.5	.2	76	1.0	2.0	.2
10	1.0	1.0	.2	44	1.0	2.5	.2	77	1.0	-2.5	.2
11	1.0	2.0	.2	45	1.0	0.0	.2	78	1.0	1.5	.2
12	1.0	-1.0	.2	46	1.0	0.0	.2	79	1.0	-0.5	.2
13	1.0	-1.0	.2	47	1.0	0.5	.2	80	1.0	2.0	.2
14	1.0	0.5	.2	48	1.0	-2.0	.2	81	1.0	-2.0	.2
15	1.0	0.0	.2	49	1.0	2.5	.2	82	1.0	1.0	.2
16	1.0	0.0	.2	50	1.0	-2.5	.2	83	1.0	-1.5	.2
17	1.0	1.0	.2	51	1.0	-0.5	.2	84	1.0	-2.0	.2
18	1.0	1.5	.2	52	1.0	0.5	.2	85	1.0	-1.5	.2
19	1.0	1.5	.2	53	1.0	-2.5	.2	86	1.0	2.0	.2
20	1.0	-1.5	.2	54	1.0	1.0	.2	87	1.0	2.0	.2
21	1.0	2.5	.2	55	1.0	2.5	.2	88	1.0	-1.0	.2
22	1.0	-2.5	.2	56	1.0	-2.0	.2	89	1.0	-1.0	.2
23	1.0	0.0	.2	57	1.0	-2.0	.2	90	1.0	1.5	.2
24	1.0	-2.5	.2	58	1.0	0.5	.2	91	1.0	-1.5	.2
25	1.0	-1.0	.2	59	1.0	1.0	.2	92	1.0	0.0	.2
26	1.0	2.5	.2	60	1.0	1.0	.2	93	1.0	0.0	.2
27	1.0	2.5	.2	61	1.0	-0.5	.2	94	1.0	1.5	.2
28	1.0	-1.0	.2	62	1.0	-2.0	.2	95	1.0	-0.5	.2
29	1.0	-2.0	.2	63	1.0	2.0	.2	96	1.0	1.0	.2
30	1.0	1.0	.2	64	1.0	0.0	.2	97	1.0	0.5	.2
31	1.0	-2.0	.2	65	1.0	1.5	.2	98	1.0	-2.0	.2
32	1.0	0.0	.2	66	1.0	-0.5	.2	99	1.0	2.5	.2
33	1.0	1.5	.2	67	1.0	1.0	.2	100	1.0	-2.5	.2
34	1.0	-1.5	.2								

Table B
Item Parameter Values for Each of the Items
in the a- and b-Variable Item Pool

Item Number	a	b	c	Item Number	a	b	c	Item Number	a	b	c
1	1.5	-0.5	.2	35	1.0	0.5	.2	68	1.0	2.0	.2
2	1.0	0.0	.2	36	2.0	2.5	.2	69	1.5	-0.5	.2
3	0.5	-1.0	.2	37	1.0	0.5	.2	70	0.5	0.5	.2
4	0.5	-1.5	.2	38	1.5	-2.5	.2	71	0.5	1.5	.2
5	1.5	-0.5	.2	39	0.5	-2.5	.2	72	1.0	-1.5	.2
6	1.5	2.0	.2	40	1.0	-2.0	.2	73	1.0	-2.5	.2
7	1.5	-2.0	.2	41	1.5	-2.0	.2	74	1.0	-1.0	.2
8	1.5	-1.0	.2	42	0.5	0.0	.2	75	0.5	2.0	.2
9	0.5	-1.0	.2	43	0.5	-2.5	.2	76	0.5	1.0	.2
10	1.5	1.0	.2	44	2.0	0.5	.2	77	1.5	1.0	.2
11	0.5	1.0	.2	45	2.0	-2.0	.2	78	1.0	2.5	.2
12	1.5	-2.0	.2	46	0.5	-1.5	.2	79	0.5	0.0	.2
13	2.0	0.0	.2	47	1.0	-1.5	.2	80	1.5	0.5	.2
14	0.5	-0.5	.2	48	1.5	0.5	.2	81	1.5	1.5	.2
15	1.0	1.5	.2	49	2.0	-2.0	.2	82	1.5	-0.5	.2
16	2.0	-2.5	.2	50	2.0	-1.5	.2	83	2.0	1.5	.2
17	0.5	-1.5	.2	51	1.0	0.5	.2	84	1.0	1.0	.2
18	1.0	2.5	.2	52	1.0	-1.0	.2	85	2.0	-2.0	.2
19	2.0	2.0	.2	53	1.0	-2.0	.2	86	2.0	2.0	.2
20	2.0	0.0	.2	54	2.0	1.0	.2	87	2.0	-1.5	.2
21	1.5	2.0	.2	55	0.5	-2.5	.2	88	1.0	1.5	.2
22	0.5	2.5	.2	56	1.0	-2.5	.2	89	2.0	1.0	.2
23	1.0	0.0	.2	57	1.5	-1.0	.2	90	2.0	2.0	.2
24	2.0	2.0	.2	58	1.5	1.0	.2	91	1.5	1.5	.2
25	0.5	-2.5	.2	59	2.0	-1.0	.2	92	0.5	2.5	.2
26	1.5	2.5	.2	60	1.0	0.5	.2	93	2.0	1.5	.2
27	1.0	0.0	.2	61	1.5	0.0	.2	94	2.0	-0.5	.2
28	1.5	2.0	.2	62	1.0	-1.5	.2	95	0.5	-2.5	.2
29	1.0	2.5	.2	63	2.0	-1.0	.2	96	0.5	0.5	.2
30	1.0	2.5	.2	64	2.0	-1.5	.2	97	1.0	0.0	.2
31	0.5	-1.0	.2	65	1.5	2.0	.2	98	2.0	1.5	.2
32	0.5	-1.0	.2	66	2.0	-0.5	.2	99	0.5	0.0	.2
33	2.0	-2.5	.2	67	1.5	-0.5	.2	100	1.5	-0.5	.2
34	0.5	1.5	.2								

Table C
Item Parameter Values for Each of the Items
in the a-, b-, and c-Variable Item Pool

Item Number	a	b	c	Item Number	a	b	c	Item Number	a	b	c
1	1.5	2.5	.1	35	0.5	1.0	.3	68	1.0	-1.5	.2
2	1.0	2.5	.3	36	1.0	2.5	.1	69	0.5	0.0	.3
3	2.0	0.5	.3	37	1.0	-2.0	.1	70	1.0	0.0	.1
4	1.0	1.5	.1	38	0.5	-2.5	.2	71	1.5	1.0	.2
5	1.5	1.5	.2	39	2.0	2.0	.1	72	1.5	1.0	.1
6	2.0	-1.5	.1	40	1.5	2.0	.3	73	2.0	-1.5	.2
7	2.0	2.0	.2	41	2.0	-0.5	.1	74	1.5	-2.0	.1
8	1.5	1.0	.2	42	1.0	-2.5	.3	75	1.5	-0.5	.3
9	2.0	1.0	.1	43	2.0	-1.0	.2	76	0.5	1.5	.3
10	1.5	1.5	.5	44	0.5	-1.0	.2	77	0.5	2.0	.2
11	1.5	-1.0	.3	45	0.5	-1.0	.1	78	1.5	-0.5	.2
12	1.0	-1.5	.1	46	0.5	2.5	.3	79	2.0	2.0	.2
13	0.5	-0.5	.1	47	1.5	-2.0	.2	80	1.5	0.5	.1
14	1.0	0.5	.3	48	1.0	2.5	.2	81	1.0	-1.0	.2
15	1.5	-2.5	.1	49	1.0	0.5	.3	82	1.0	2.5	.1
16	0.5	-2.5	.3	50	2.0	-2.5	.1	83	1.0	2.5	.3
17	0.5	-1.5	.3	51	1.0	0.0	.3	84	1.0	-2.0	.1
18	1.0	-1.0	.1	52	1.5	2.0	.3	85	1.0	0.5	.2
19	2.0	-2.5	.3	53	1.5	2.0	.3	86	0.5	0.5	.1
20	2.0	0.0	.2	54	2.0	2.5	.3	87	2.0	2.0	.1
21	1.5	2.0	.2	55	2.0	-1.5	.2	88	0.5	-2.5	.1
22	2.0	-2.0	.2	56	0.5	0.0	.2	89	1.5	-0.5	.3
23	0.5	0.5	.2	57	1.0	0.5	.2	90	2.0	1.5	.3
24	0.5	1.0	.2	58	1.0	0.0	.2	91	0.5	-2.5	.3
25	1.5	-0.5	.3	59	1.0	0.0	.3	92	2.0	0.0	.3
26	2.0	-2.0	.3	60	1.5	-2.0	.1	93	1.0	-2.0	.3
27	2.0	-0.5	.3	61	1.5	0.0	.2	94	0.5	2.5	.1
28	2.0	1.0	.3	62	1.0	1.5	.2	95	0.5	-2.5	.1
29	2.0	-1.0	.2	63	1.5	-0.5	.1	96	0.5	0.0	.2
30	1.5	1.0	.2	64	0.5	-1.0	.2	97	0.5	-1.5	.1
31	2.0	1.5	.1	65	2.0	1.5	.2	98	0.5	-1.5	.1
32	0.5	-1.0	.3	66	1.5	0.5	.1	99	1.5	-0.5	.2
33	1.0	-1.5	.2	67	1.0	1.0	.1	100	0.5	1.5	.3
34	2.0	-2.0	.2								