A Comparison of ICC-Based Adaptive Mastery Testing and the Waldian Probability Ratio Method

G. Gage Kingsbury and David J. Weiss
University of Minnesota

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 an absolute standard 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, a student is declared a master if his/her score on a conventional classroom achievement test is as high or higher than a prespecified cutoff point or is declared a nonmaster if his/her score on the test is lower than the cutoff point. 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 the amount of time spent on testing to a minimum. 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.

Approaches to Adaptive Mastery Testing

Two attempts that have been made to combine mastery and adaptive testing technologies have been Ferguson's (1969, 1970) application of Wald's Sequential

Probability Ratio Test (SPRT) to mastery testing and Kingsbury and Weiss's (1979a) formulation of an item characteristic curve (ICC) approach to adaptive mastery testing (AMT). 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.

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 prime objective of this paper, then, 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. The first level of comparison presented here is a descriptive comparison based on the theories underlying each of the procedures. This is followed by an empirical comparison of the two testing procedures within the context of a monte carlo simulation of test responses designed to fit a number of theoretical contingencies.

Wald's SPRT Applied to Mastery Testing

The SPRT procedure. 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 products (e.g., light bulbs) 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

$\underline{p}$ = the proportion of defective elements (light bulbs) in the population (consignment);

$\underline{p}_0$ = the proportion of defective elements in the population below which it is always desired to accept the quality of the population; and

$\underline{p}_1$ = 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 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 function. Consequently, the probability of observing W defective elements in a sample of $\underline{m}$ elements (W_m), under H_0: $\underline{p} = \underline{p}_0$ is

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

Under H_1: $\underline{p} = \underline{p}_1$, the probability becomes

$$p_{1_m} = p_1^{(m-W_m)} \ (1 - p_1)^{W_m} \ . \qquad [4]$$

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 willing to be tolerated in rejecting the quality of the consignments that actually do have the quality desired and (2) on the proportion (β) of errors willing to be tolerated 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 specifices that if

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

the consignment is rejected; if

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

the consignment is accepted; and if

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

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 test item responses 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 concentrate on the application of SPRT to a single instructional unit.

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

1. Two criteria of performance (p_0 and p_1), which serve as the lowest level at which a mastery decision will be made and the highest level at which a nonmastery decision will be made and which bound the uncertainty 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 that shows 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 p_0, p_1, α, and β, the statistical work is completed. This puts the power of the SPRT procedure into the hands of the educator quite readily. The procedure is not fully adaptive, however. Items are selected at random or in a fixed sequence; it is only the test length that varies for individuals.

ICC-Based Adaptive Mastery Testing (AMT)

The paradigm for AMT that Kingsbury and Weiss (1979) have proposed makes use of ICC theory and Bayesian statistical theory to adapt the mastery test to the individual's level of skill during the testing process. ICC theory is used to estimate the parameters that most efficiently describe each of the items in the item pool. Given these parameter estimates, it is possible to prescribe a type of adaptive procedure that may 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. 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, an 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, since that

item should provide the most efficient use of testing time. 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 part of the AMT procedure.

The information that an item provides at each point along the achievement continuum may be determined using the ICC model that is assumed to underly individuals' responses to test items. The AMT procedure assumes the 3-parameter logistic ICC model (Birnbaum, 1968). Using this model, the information available in any item is (Birnbaum, 1968, Equation 20.4.16)

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

where

$I_i(\theta)$ = the information available from item i at any achievement level, θ;
c_i = the lower asymptote of the ICC for the item;
D = 1.7, a scaling factor used to allow the logistic ICC to closely approximate a normal ogive;
a_i = the discriminatory power of the item at the inflection point of the ICC;
Ψ = the logistic probability density function;
$L_i(\theta)$ = $a_i(\theta - b_i)$ where b_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. 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. This estimation procedure has been shown to yield biased estimates of trait levels (Kingsbury & Weiss, 1979; McBride & Weiss, 1976). This bias may be attributed to the assumption of a normal distribution of θ in the population made by Owen's procedure or due to inappropriate prior information concerning θ on the individual level (Kingsbury & Weiss, 1979b). The bias inherent in this scoring strategy may render the MISS technique less efficient than it would be under optimal conditions, thereby reducing the efficiency of the AMT technique as a whole.

To use MISS under optimal conditions, trait level estimates should be obtained by maximum likelihood estimation, which yields asymptotically efficient estimates (Birnbaum, 1968). Maximum likelihood estimation techniques are not able, however, to obtain trait level estimates for consistent item response patterns (either all correct or all incorrect) or for item response patterns for which the likelihood function is extremely flat. The Bayesian technique will yield an estimate for any response vector. This inability to estimate θ for

some response patterns mitigated against the use of a maximum likelihood estimation procedure for AMT. Consequently, the Bayesian estimation procedure was used in the AMT procedure on the assumption that the capability to obtain a θ estimate for each individual at each point during the test would outweigh any efficiency lost due to the bias inherent in the estimation procedure. The use of the Bayesian estimation strategy in this study also allowed the use of easily interpretable Bayesian confidence intervals to make the mastery decision.

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 ICC theory, it is possible to generate an analog to the percentage cutoff 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 be close to the individual's level of θ). The analog is based on the use of the test characteristic curve (TCC; Lord & Novick, 1968). The TCC 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 3-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. This assumption yields a TCC of the following form:

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

where

$E(P|\theta)$ = the expected value of the proportion of correct answers observed on the test given at any achievement level;
n = the number of items on the test;
c_i = the estimate of the lower asymptote for the ICC of item i;
a_i = the estimate of the discriminatory power for the item;
b_i = the estimate of the difficulty of the item; and
θ = any given achievement level.

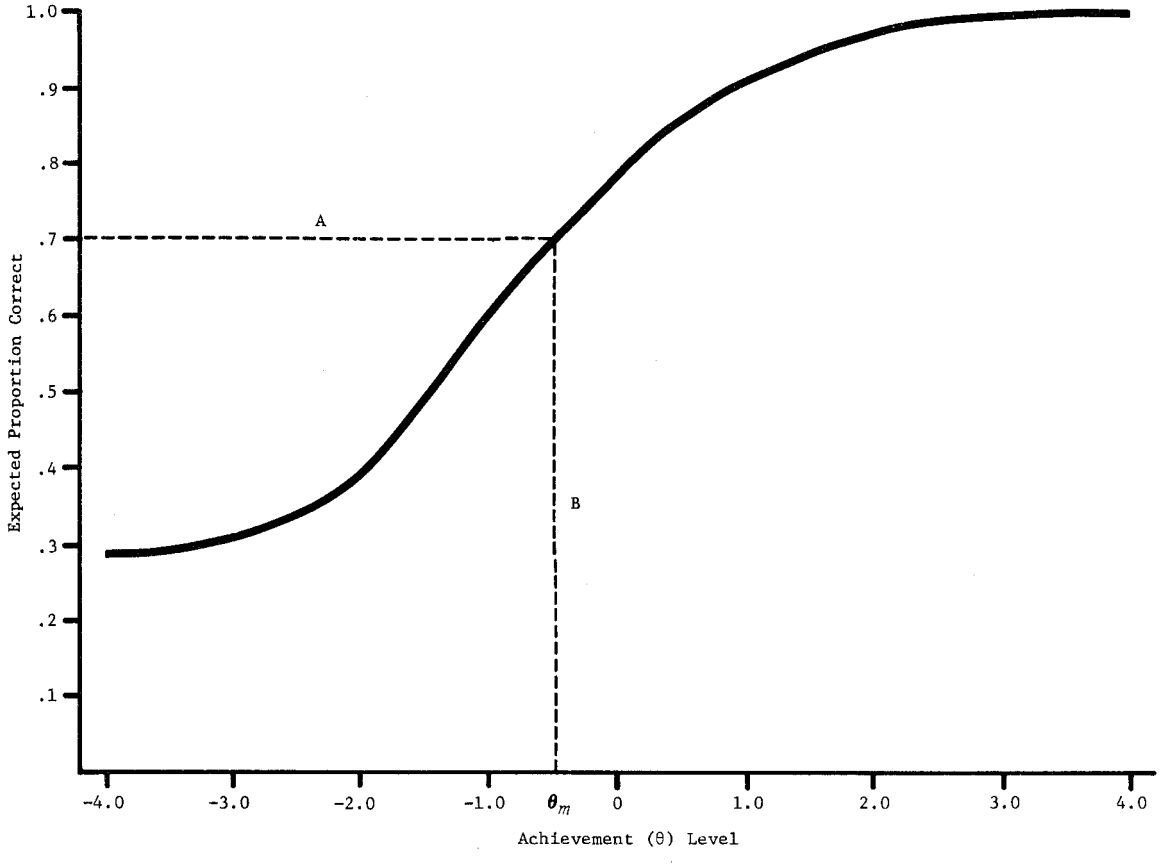
This monotonically increasing function enables the expression of any given level of θ to its most likely proportion correct or, more importantly in this context, to determine the level of θ that will most probably result in any given proportion of correct answers. To exemplify the use of the TCC in determining a level of θ that is comparable to a desired percentage mastery level, a hypothetical TCC is shown in Figure 1. Assuming that some items from the test represented by this TCC are to be administered in some adaptive manner (e.g., MISS) and that a level of θ is to be determined that corresponds to, say, 70% correct performance on the entire test, it may be done using the following steps:

1. Draw a horizontal line (Line A in Figure 1) from the .7 mark on the vertical (expected proportion correct, or P) axis of the TCC figure to the TCC.

2. Drop a vertical line (Line B) from the point of intersection of the TCC and the horizontal line drawn in Step 1 to the horizontal (achievement level, or θ) axis. This point (θ_m) on the achievement level axis is designated the mastery level in terms of the achievement (θ) metric.

3. The mastery level specified in Step 2 above may now be used to make mastery decisions in place of the .7 mastery level originally specified using any subset of items from the original test, provided that individuals' item responses are scored with a method that will put the θ estimate on the same metric as the TCC. Any ICC-based scoring procedure (e.g., Bejar & Weiss, 1979) will result in a θ estimate that will be on the correct metric. This procedure allows the transformation of any desired proportion correct mastery level to the θ metric. Once this transformation is made, ICC theory and its technology may be used to increase the efficiency of present mastery testing techniques.

Figure 1
Hypothetical Test Characteristic Curve Illustrating Conversion from the Proportion Correct Metric to the Achievement Metric



Making the mastery decision using Bayesian confidence intervals. Although any achievement level estimate of any subset of the items from a test obtained using ICC-based scoring will be on the same metric as the TCC for the original

test, two different subsets of items may result in θ 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 ICCs. Thus, ICC-based θ estimates that are on the same metric are comparable except for their differential precision. Comparisons of ICC-based θ estimates should therefore be based on confidence interval estimates instead of the raw achievement level 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--for this application MISS is used to choose the appropriate item at each point in the test-- a point estimator of the individual's achievement level ($\hat{\theta}$) may be determined using Owen's Bayesian scoring algorithm, using 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

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

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 inteval. 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 achievement metric, falls outside the limits of the confidence interval. If it does not, the testing 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 will not include θ_m, the mastery level on the achievement continuum. At this time 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, however, 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.

Wald's SPRT versus ICC-Based AMT Procedure

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 cutoff.
4. Treatment of the achievement metric.

Treatment of items. The SPRT in the simple form 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; 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 his/her 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 mattery (c). This assumption concerning the operating characteristics of the items is less restrictive than the assumption made in the SPRT procedure described above; 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 to be used for mastery testing yields the first point at which it might be decided which of the two models is more appropriate for use in a given situation.

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 one is willing to accept across all the students 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 provides the educator with more flexibility than the AMT procedure, as currently operationalized.

Treatment of mastery level. The SPRT uses an uncertainty region, rather than a single mastery level, to define the mastery and nonmastery regions. The specification of this uncertainty 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 θ 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, which more accurately represents the achievement of the individuals taking the test. This latent continuum 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.

EMPIRICAL COMPARISON OF THE SPRT AND AMT PROCEDURES

To delineate circumstances in which one of the mastery testing procedures might have an advantage over the other, monte carlo simulation was used to compare the two testing procedures under several conditions.

Method

The method used to compare the two variable-length mastery testing procedures to one another, as well as to a conventional (fixed length) testing procedure, consisted of five basic steps:

1. Three 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 three item pools.

3. Conventional tests of three different lengths were drawn from the larg-

er 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 three 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

Three 100-item pools were generated to reflect different types of pools that might be used in a mastery test.

Uniform pool. The uniform pool consisted of 100 items that were perfect replications of one another. Each item had the same discrimination (a = 1.00), difficulty (b = 0.00), and guessing probability (c = .20). 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 -2.50, -2.00, -1.50, -1.00, -0.50, 0.00, 0.50, 1.00, 1.50, 2.00, and 2.50. Nine items at each level of difficulty were used in this pool, along with an additional item with b = 0.00 to bring the pool to 100 items.

a-, b-, and c-variable pool. The a-, b-, and c-variable pool differed from the b-variable pool in that the discriminations and guessing levels of the items were allowed to spread across a range of values. The a values used were .50, 1.00, 1.50, and 2.00. The c values used were .10, .20, and .30. All a and 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).

Item Response Generation

Achievement levels for 500 simulees were drawn from a normal distribution with a mean of zero and a standard deviation of one. Item responses for each of these simulees were then generated for each item in each of the three item pools using the 3-parameter logistic ICC model. That is, knowing the θ level of the simulee and the parameters of the item in question, the probability of a correct response was calculated. A random number was then drawn from a uniform distribution ranging from zero to one. 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 the AMT procedure. The "true" mastery level for each simulee was determined by comparing the θ levels used to generate the item responses with the proportion correct mastery level expressed on the θ metric.

Conventional Tests

Conventional tests of three different lengths (10, 25, and 50 items) were drawn at random from each of the three item pools, with the stipulation that the shortest conventional test served as the first portion of the next longer conventional test and that this test in turn served as the first portion of the longest conventional test. These nine conventional tests 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.

Simulation of the Testing Strategies

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

Conventional test. The conventional test assumed a mastery criterion of 60% correct responses. After all of the items in the conventional test were administered, if the individual answered 60% or more items correctly, the individual was declared a master. If the individual's score was less than 60% correct, the individual was declared a nonmaster.

SPRT procedure. For the SPRT procedure the limits of the uncertainty region were set at proportion-correct values of .50 and .70. Values of α and β were each set to .10. For individuals for whom no decision was made by the Wald procedure before the item pool was exhausted, the mastery decision was made by the conventional procedure, using a mastery proportion of .60.

AMT. For the AMT procedure the mastery levels in each of the 100-item pools corresponding to 60% correct were designed to be equal to $\theta = 0.00$. This mastery level was used with each of the smaller item pools, even though they had not been designed to result in a mastery level of $\theta = 0.00$. This procedure added some sampling error to the AMT procedure, to more appropriately reflect the error that is inherent when using estimated item parameters to determine the mastery level. For the AMT Bayesian scoring procedure, each individual was assumed to have a prior mean of 0.00 and a prior variance of 1.00.

Comparison among Testing Procedures

For each of the three testing procedures (AMT, SPRT, conventional), 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 individuals.

Comparisons within each testing procedure concerning the average test length and the correspondence of decisions with true mastery status were made across all nine combinations of test lengths and item pool types.

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.

Table 1
Mean Number of Items Administered to Each Simulee for Three Mastery Testing Strategies Using Each Type of 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-, 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

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 conventional test 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, again, 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, with larger reductions for the longer MTL levels. At the 10-item, 25-item, and 50-item MTL levels the test length reductions observed were 4%, 33%, and 57%, respectively.

For this pool the AMT procedure resulted in slightly greater reduction in test length 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-, 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 and this reduction was 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% 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 was known: It was the level that was used to generate the item responses. Given this, it was known whether the individual's θ level was actually above or below the prespecified mastery level on the achievement metric ($\theta = 0.00$). Phi correlations between true mastery status and the mastery state determined by each of the three testing procedures for each MTL level and pool type are shown in Table 2.

Uniform pool. For the uniform pool one major trend was observed. For each testing procedure an increase in the MTL level was accompanied by an increase in the correlation between the true and estimated mastery states. (These correlations may be referred to as correspondence coefficients.)

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			
Conventional	.771	.837	.875
AMT	.775	.840	.871
SPRT	.771	.837	.867
b-Variable Pool			
Conventional	.541	.667	.783
AMT	.615	.715	.828
SPRT	.541	.656	.704
a-, b-, and c-Variable Pool			
Conventional	.290	.670	.735
AMT	.470	.733	.787
SPRT	.290	.592	.571

In addition to this major trend, it was observed that for the 10-item and 25-item MTL levels, the AMT procedure produced the highest correspondence coefficient observed (r = .775 and .840, respectively). For the 50-item MTL level the conventional procedure resulted in the highest correspondence (r = .871).

It should be noted that the differences in correspondence between any two MTL levels within any testing procedure (the smallest was .03, between the 25-item and 50-item MTL levels for the SPRT procedure) were much larger than the largest difference noted between any two testing procedures within a single MTL level (.008, for the conventional and SPRT procedures in the 50-item 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. For each MTL level, the AMT procedure resulted in the highest correspondence coefficients. The conventional procedure resulted in the next highest correspondence level for all three MTL levels (tied with the SPRT procedure at the 10-item MTL level).

Differences in correspondence coefficients observed between testing strategies 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. It was also noted that each correspondence level observed was lower for this pool than for the uniform pool across all MTL levels and testing procedures.

a-, b-, and c-variable pool. The same trend of increasing correspondence with increasing MTL level was again noted for the conventional and AMT proce-

dures. For the SPRT procedure the correspondence peaked at $\underline{r}$ = .592 at the 25-item MTL level and dropped to .571 at the 50-item MTL level.

The AMT procedure produced the highest correspondence for all three MTL levels. The conventional procedure resulted in the next highest level of performance at all MTL levels (again tied with the SPRT 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 two testing strategies within a single MTL level. Further, on the average, the correspondence coefficients for this pool were lower than for either of the other pools, with rather large decreases at the 10-item MTL level, particularly for the conventional and SPRT strategies.

Frequency and Type of Errors

To further compare the performance of the three mastery testing strategies the frequency with which each procedure made incorrect decisions (false mastery, 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. This table shows the frequency with which each of the testing procedures made false mastery and false nonmastery decisions in each of the testing conditions. It may be noted that the "Total" column in Table 3 reproduces the information already reported from the correlational analysis, but in a different manner. 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 more false nonmastery decisions than false mastery decisions at all MTL levels. Each procedure also 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 larger MTL levels for all procedures. The differences among the procedures in terms of the types of false decisions made were minimal.

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 conventional and SPRT procedures produced more false mastery than false nonmastery decisions at all MTL levels. The AMT procedure produced more false mastery than false nonmastery decisions at the 10-item MTL level but produced more false nonmastery than false mastery decisions at the two higher MTL levels. For the AMT 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%) at the two higher MTL levels. For the conventional 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 (20% to 80%) was observed at each MTL level.

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 Test	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									
Conventional	3.6	8.0	11.6	2.6	5.6	8.2	2.8	3.4	6.2
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									
Conventional	22.4	2.2	24.6	13.4	3.6	17.0	6.4	4.4	10.8
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-, *b*-, and *c* Variable Pool									
Conventional	0.0	44.6	44.6	2.6	15.2	17.8	7.4	5.8	13.2
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

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

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 levels. For the 50-item MTL level the conventional procedure resulted in a higher frequency of false mastery decisions, but the AMT and SPRT procedures still resulted in higher percentages of false nonmastery decisions. As with the b-variable item pool, the AMT procedure used with this item pool resulted in smaller differences in the frequencies of the two error types than either of the other testing procedures at each MTL level. For the 50-item MTL level the AMT procedure produced a very small difference in the two error rates (.6%). The conventional 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 conventional procedure at the 10-item MTL level).

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 for 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 three testing strategies in the three 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 level in each circumstance. 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. The correspondence fairly systematically decreased from the uniform pool, to the b-variable pool, to the a-, b-, and c-variable pool. The AMT procedure resulted in the highest level of correspondence in all circumstances but one (the conventional test 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 pronounced effects on the correspondence obtained.

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 an 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 a uniform item pool is assumed, the SPRT procedure required the fewest items while resulting in decisions having correspondence coefficients that were quite comparable to the other two procedures. If, however, the item pool includes items with variable a, b, and c parameters, the SPRT procedure may result in the shortest tests, but the AMT procedure will make more accurate classifications. These factors must be considered before any decision is made as to which procedure is "best."

It should also be noted that this simulation was based on the assumption that the latent achievement metric, rather than the propertion-correct metric, was the correct metric; and to the extent that the proportion-correct metric is the correct metric, the findings of this study are less relevant. In addition, several variations on the SPRT procedure and the AMT procedure that were not examined in this study are possible; thus, additional research is necessary before firm conclusions can be drawn concerning the utility of adaptive mastery testing strategies.

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. (NTIS No. AD A047495)

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. (NTIS No. AD A046062)

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. (a)

Kingsbury, G. G., & Weiss, D. J. Relationships among achievement level estimates from three item characteristic curve scoring methods (Research Report 79-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1979. (NTIS No. AD A069815) (b)

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

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

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.

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

ACKNOWLEDGMENTS

This research was supported by funds from Air Force Office of Scientific Research, Army Research Institute, Defense Advanced Research Projects Agency and Office of Naval Research, under Contract N00014-79-C-0172 NR 150-433 with the Personnel and Training Research Programs, Office of Naval Research.