AN ADAPTIVE TESTING STRATEGY FOR MASTERY DECISIONS

G. Gage Kingsbury

and

David J. Weiss

Research Report 79-5
September 1979

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

This research was supported by funds from the Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, 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 79-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) An Adaptive Testing Strategy 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) N00014-76-C-0627
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, MN 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.: 61153N PROJ.:RR042-04 T.A.: RR042-04-01 W.U.: NR150-389
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, VA 22217		12. REPORT DATE September 1979
		13. NUMBER OF PAGES 36
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 Army Research Institute, Air Force Human Resources Laboratory, Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, 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)

latent trait test theory
item response theory
response-contingent testing
individualized testing
achievement testing
computerized testing
adaptive testing
sequential testing
branched testing
testing
tailored testing
programmed testing
automated testing

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

In an attempt to increase the efficiency of mastery testing while maintaining a high level of confidence for each mastery decision, the theory and technology of item characteristic curve (ICC) response theory (Lord & Novick, 1968) and adaptive testing were applied to the problem of judging individuals' competencies against a prespecified mastery level to determine whether each individual is a "master" or a "nonmaster" of a specified content domain. Items from two conventionally administered classroom mastery tests administered in a

CONTENTS

CONTENTS

An Adaptive Testing Strategy for Mastery Decisions

During the past 15 years, considerable interest in the psychological and educational measurement community has been directed toward the evaluation of student competency in various fields of study. In the simplest case, competency in a field has been operationalized as some minimum skill level above which a student is declared a "master" and below which a student is declared a "nonmaster." Mastery testing has been developed as an implementation of the more general criterion-referenced test interpretation model formulated by Glaser and Klaus (1962) and expanded upon by many since then (e.g., Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham, 1971; Popham & Husek, 1969).

"Mastery" has typically been defined by subject matter experts as the minimum percentage of items that a student should be able to answer from a given set of test items in order to be classified as proficient. Therefore, a student who correctly answered only the minimum acceptable percentage of items on a test of this type would be declared a master, and a student who correctly answered one item less would be declared a nonmaster in the subject matter area. So that all of the mastery decisions made would be comparable, mastery testing has traditionally required all students to answer the same set of test questions.

This approach to mastery testing has several problems. First, a student whose test score is far above the specified cutoff score would be said to be a master of the subject matter; similarly, a student whose score was just barely above the cutoff score would also be declared a master, but presumably that decision would be made with less confidence. Thus, classical mastery testing results in different levels of intuitive confidence for students whose raw scores fall at different distances above or below the cutoff, which results in decisions with different dependabilities for students with different raw scores.

This problem has been discussed on the group level by Livingston (1972) in a study discussing the reliability of criterion-referenced tests as a function of the mean score level of the testee group. Hambleton and Novick (1973) and Davis and Diamond (1974) have specified methods to develop cutoff rules designed to yield certain desired ratios of false positive and false negative decisions through the use of the differential accuracy of decisions made at different raw score levels, but little research has been directed toward equalizing the confidence levels in decisions made by a mastery test across all levels of performance. Hambleton and Novick (1973) have suggested that the use of Bayesian point estimation of students' mastery scores might improve the accuracy of mastery decisions; it will be shown in this report that the use of Bayesian confidence interval estimates may be useful in equalizing the confidence in decisions made across all levels of observed performance.

A second problem with the classical mastery testing paradigm is that each student tested is given the same set of test questions, even though the set of questions may be inappropriate for any reasonably precise measurement at some achievement levels. In the mastery testing area, attempts have been made to adapt the test to each student (e.g., Ferguson, 1970); but these attempts have almost universally assumed that all items administered were of equal quality. It is possible,through the use of item characteristic curve (ICC) response theory (Lord & Novick, 1968), to distinguish between items which yield different amounts of information concerning different trait levels.

Several authors (e.g., Bejar, Weiss , & Gialluca, 1977; McBride & Weiss, 1976; Urry, 1977) have demonstrated that adaptive testing procedures using ICC response theory can reduce test length with no reduction in measurement precision. These testing procedures adapt the difficulty and information characteristics of each individual's test by drawing from large item pools items that are matched to the individual's estimated trait level. These results indicate that by making use of all of the information available about the test items and the individual's estimated achievement levels, the application of adaptive testing procedures using ICC response theory to a traditional mastery testing situation might result in a decrease in the test length needed to make confident decisions concerning each individual's mastery status.

Objectives

This report describes the design and application of an adaptive mastery testing strategy that eliminates these problems of the traditional mastery testing approach. The adaptive mastery testing strategy is designed to reduce the average test length for each student, while equalizing the level of confidence in decisions made across the entire range of the achievement continuum. This report compares the performance of the conventional and adaptive mastery testing procedures within the context of one course of instruction in terms of efficiency, information characteristics, and level of correspondence between masery decisions.

The Adaptive Mastery Testing Procedure

The adaptive mastery testing (AMT) procedure is designed to administer achievement test items selected from a classical mastery test, but not all items are administered to each student. The test items administered to a given student are selected to provide the most information concerning the achievement level of that student. Mastery decisions are made with a specified degree of confidence for each student,using a cutoff point prespecified on the achievement continuum.

There are three important components of the AMT procedure. The first involves converting the mastery level to the achievement metric. The second component is the item-selection technique used to determine which items should be administered to a specific student. The final component of the AMT strategy involves the manner in which the mastery decision is made and the degree of confidence that can be placed in the decision once it has been made.

Mastery and the Achievement Metric

The classical mastery testing procedure specifies a percentage of the items on a test that must be correctly answered by a student in order to be declared a master. Using ICC theory, it is possible to generate an analogue to the "percentage" cutoff of classical theory for use in adaptive testing. This is necessary, since in an adaptive test each individual will tend to answer about 50% of the items correctly, given a large enough item pool, because the items administered will be selected to be close to the individual's achievement level (Vale & Weiss, 1975; Weiss, 1973). The ICC analogue of proportion correct is based on the use of the test characteristic curve (TCC). The TCC is the function that relates the ICC achievement continuum to the expected proportion of correct answers that an individual at any achievement level may be expected to obtain if all of the items on the test were administered.

For this study the assumption was made that a three-parameter logistic ogive would describe 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:

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

where

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

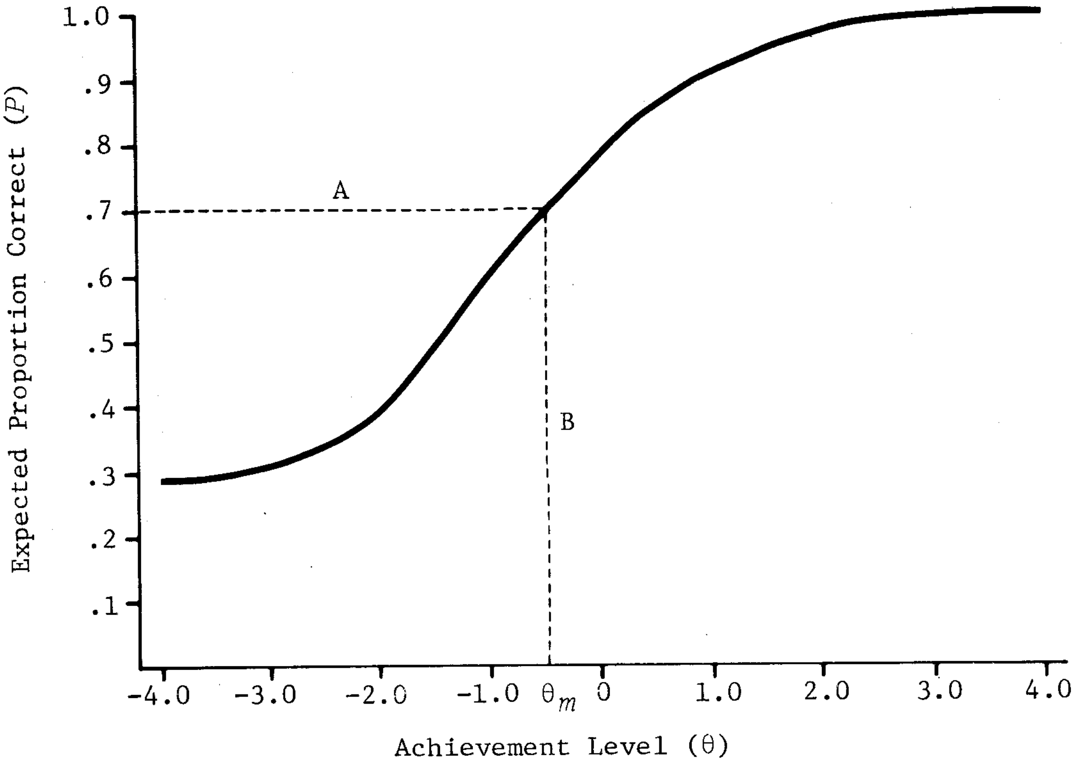
Thus, as Equation 1 indicates, the expected proportion correct at a given level of achievement (θ) is the average, over all items in the test, of the probability of a correct response for each item, given the three ICC item parameters for each item and assuming a logistic ICC.

This monotonically increasing function permits relating any achievement level to its most likely proportion correct or, more importantly in this context, determining the achievement level (θ) which will most probably result in any given proportion of correct answers. An example of the use of the TCC in determining an achievement level that is comparable to a desired "percentage" cutoff is shown in Figure 1 using a hypothetical TCC. To determine a level of achievement that corresponds to, for example, a 70% mastery level on the test items which comprise the TCC, these steps would be followed:

1. Draw a horizontal line (line A in Figure 1) from the P=.7 mark on the vertical (expected proportion correct, or P) axis of the TCC plot 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 using the achievement metric.

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



3. The cutoff point specified in Step 2 may now be used to make mastery decisions in place of the *P*=.7 mastery level originally specified. Once the mastery level is expressed in the achievement metric (θ), rather than in terms of proportion correct, it is no longer necessary to administer all the items in the test to obtain an achievement level estimate for an individual--and a corresponding mastery decision. An achievement level estimate can then be obtained using any subset of items from the original test, provided that the individual's item responses are scored with a method that will put the achievement level estimate on the same metric as the TCC. Any ICC-based scoring procedure (Bejar & Weiss, 1979), in conjunction with the original item parameter estimates, will result in an achievement level estimate which will be on the θ metric.

This procedure allows conversion of any desired proportion correct mastery level to the θ metric. Once this transfer is made, ICC theory and adaptive testing strategies may be used to increase the efficiency of mastery testing techniques.

Adaptive Item Selection and Scoring

To make mastery testing a more efficient process, the objectives of the AMT strategy were (1) to reduce the length of each student's test by eliminating test items which provided little information concerning the student's achievement level and (2) to terminate the AMT procedure after enough information had been obtained so that the mastery decision could be made with a high degree of confidence.

To operationalize the first objective, items were selected to be administered to student at each point during the testing procedure on the basis of the amount of information that the item provided concerning the student's achievement level estimate at that point in testing. The administration of the test item which provides the most information concerning the student's present achievement level estimate 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 testing procedure was described by Brown and Weiss (1977), and this procedure was used in the present study. This procedure uses an adaptive maximum information search and selection (MISS) technique for the sequential selection of test items to be administered to each individual.

Item selection. The information that an item provides at each point along the achievement continuum can be determined from the ICC parameters of the item. Using the unidimensional three-parameter logistic ICC model (Birnbaum, 1968) to describe responses to the five-alternative multiple-choice items used in this study, the information available in any item is (Birnbaum, 1968, Equation 20.4.16)

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

where

$I_i(\theta)$ = the information available from item i at any achievement level θ;

a_i = the ICC discrimination parameter of the item;

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;

$L_i(\theta)$ = $a_i(\theta - b_i)$, where b_i is the ICC difficulty parameter of the item;

ψ = the logistic probability density function; and

Ψ = the cumulative logistic function.

If it assumed that the achievement level estimate ($\hat{\theta}$) is the best estimate of the true achievement level (θ), item information levels of each of the items not yet administered can be evaluated using $\hat{\theta}$ at any point during the test. The item which has the highest information value at the individual's current level of $\hat{\theta}$ is thus chosen to be administered next. Appendix A (adapted from Brown & Weiss, 1977) gives an example of the use of the MISS procedure to select items.

Estimation of θ. For this study a Bayesian estimator (Owen, 1969) of the student's achievement level ($\hat{\theta}$) was used. Details of the scoring procedure have been provided by Brown and Weiss (1977, pp. 4-5); Bejar and Weiss (1979) have provided an explanation and scoring programs for Owen's method.

Owen's θ estimation procedure has been shown to yield biased estimates of trait levels (Kingsbury & Weiss, 1979; Lord, 1976; 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 (Lord, 1976) and/or to inappropriate prior information concerning θ on the individual level (Kingsbury & Weiss, 1979). The bias inherent in this scoring method may render the MISS technique less efficient than it would be under optimal conditions, and thereby may reduce the efficiency of the AMT technique as a whole.

To use MISS under optimal conditions, θ estimates should be obtained through the use of a maximum likelihood estimation technique, which yields asymptotically efficient estimates (Birnbaum, 1968). Maximum likelihood θ estimation techniques are not able, however, to obtain trait level estimates for consistent item response patterns (either all correct or all incorrect responses) or for item response patterns for which the likelihood function is extremely flat. Owen's Bayesian scoring method will yield an estimate for any response pattern. The inability of the maximum likelihood procedures to estimate θ for some response patterns mitigates against the use of a maximum likelihood estimation procedure in this situation, since it would be necessary to assign arbitrary θ estimates during the early stages of item selection and scoring. Thus, the Bayesian scoring procedure was used in order to obtain θ estimates for each student after each item administered by the adaptive testing procedure, even though some efficiency might have been lost in the AMT due to the bias inherent in the estimation procedure. Use of the Bayesian θ estimation procedure in this study also allowed the use of easily interpretable Bayesian confidence intervals to make the mastery decision.

Bayesian Confidence Intervals: Making the Mastery Decision

Any achievement level estimate ($\hat{\theta}$) obtained using ICC-based scoring of any subset of the items from the original test and their ICC item parameters will be on the same metric as the TCC for the original test. This allows immediate comparison between any achievement level estimate ($\hat{\theta}$) and any point on the achievement metric (e.g., θ_m). However, 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 that were too easy for a given individual and the other used the same number of equally discriminating items at about the appropriate difficulty level for that individual, the second test would yield a much more accurate achievement level estimate for that individual. Achievement level estimates that are on the same metric are comparable if their differential precision is taken into account. To do this, confidence interval estimates for the $\hat{\theta}$'s should be compared instead of the point estimates ($\hat{\theta}$). For this reason, the AMT strategy makes mastery decisions with the use of Bayesian confidence intervals.

After each item was selected using MISS and administered to a student, a point estimate of the student's achievement level ($\hat{\theta}$) was determined using Owen's Bayesian scoring algorithm and the responses obtained from all items previously administered. Given this point estimate and the corresponding variance estimate for the $\hat{\theta}$, also obtained using Owens' procedure (see Brown & Weiss, 1977, Equations 3 and 5, pp. 4-5), a Bayesian confidence interval may be defined such that:

$$\hat{\theta}_i - 1.96(\hat{\sigma}_i^2)^{\frac{1}{2}} \leq \theta \leq \hat{\theta}_i + 1.96(\hat{\sigma}_i^2)^{\frac{1}{2}}, \text{ with } P = .95, \qquad [3]$$

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

$\hat{\sigma}_i^2$ = the Bayesian posterior variance estimate following item i,

and θ = the true achievement level.

This statement may be interpreted as meaning that the probability that the true value of the achievement level parameter, θ, is within the bounds of the confidence interval is .95. Alternatively, it might also be concluded with 95% confidence that the true parameter value (θ) lies within the confidence interval. Confidence intervals at differing confidence levels can be constructed using appropriate z-values from a normal distribution in place of the 1.96 in Equation 3.

After this confidence interval has been generated, it can be determined whether or not θ_m, the achievement level earlier designated as the mastery level using the TCC (see Figure 1), falls outside the limits of the confidence interval. If it does not, another item is administered to the student, and the confidence interval is recalculated using the updated $\hat{\theta}$ and its updated variance. 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 testing is terminated, and a mastery decision is made. If the lower limit of the confidence interval falls above the specified mastery level, θ_m, the student is declared a master. If, on the other hand, the upper limit of the confidence interval falls below θ_m, the student is declared a nonmaster. Given a finite size item pool, the testing procedure may, in some cases, exhaust the item pool before a decision can be made. This will occur for students with $\hat{\theta}$ values close to θ_m. It is possible to make a mastery decision for these students based simply on whether the Bayesian point estimate of their achievement level ($\hat{\theta}$) is above or below θ_m. However, for these students, mastery decisions will not be made with the same confidence levels as those made for students for whom the confidence interval falls completely above or below θ_m.

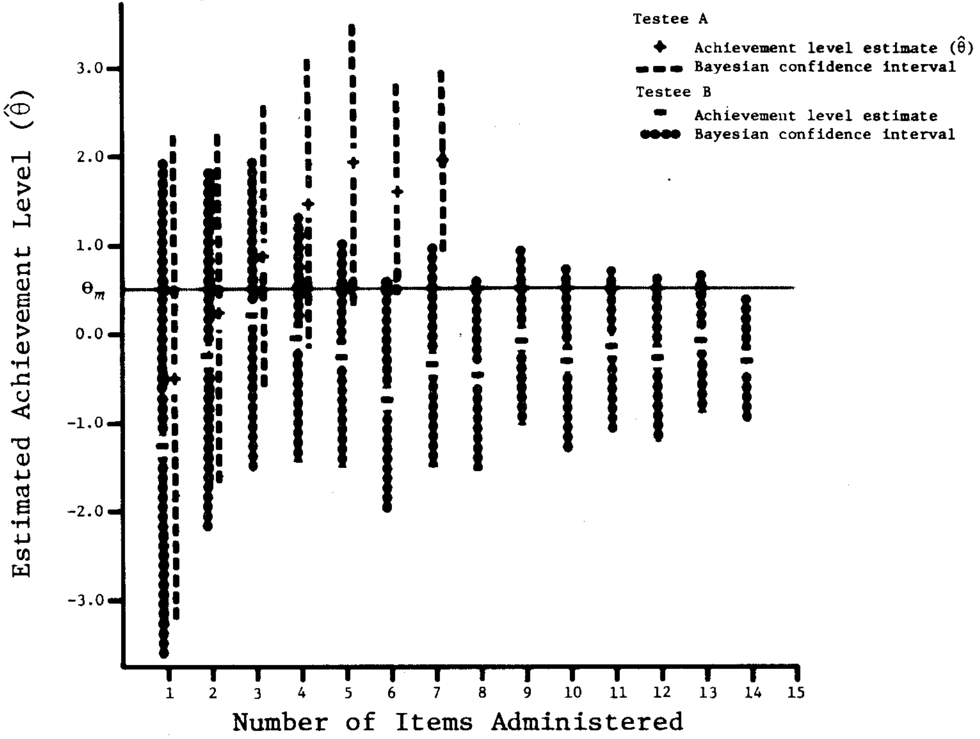
Illustration

Figure 2 shows the result of the AMT procedure for two hypothetical testees, A and B. Achievement level point estimates ($\hat{\theta}$) and error bands, which indicate the appropriate Bayesian confidence intervals, are shown for each testee after each item was administered. An arbitrary mastery level, $\theta_m = .50$, was chosen for this example; normally, however, the mastery level would be determined by the TCC transformation of an existent proportion correct mastery criterion.

For Testee A, the first θ estimate was below θ_m, but the confidence interval around this estimate contained θ_m. Thus, the θ estimate was not precise enough to make a confident decision; consequently, testing continued for Testee A. After each item was administered, a new θ estimate and a corresponding confidence interval were calculated. For the first 6 items administered to

Testee A, the confidence interval around the θ estimate contained θ_m, and testing continued. After the administration of the *7th* item, the entire confidence interval around the θ estimate for Testee A was above θ_m. This implied that the θ estimate was precise enough to allow a confident decision to be made for Testee A. Testee A was declared a master at this point, and testing was terminated.

Figure 2
Example of the AMT Procedure: Achievement Level Point Estimates and Bayesian Confidence Intervals after Each Item Administered to Two Hypothetical Testees, Testee A and B



For Testee B, the same type of procedure was followed. For the first 13 items administered to Testee B, the confidence interval around $\hat{\theta}$ contained θ_m. The *14th* item administered to Testee B resulted in a $\hat{\theta}$ and confidence interval which fell completely below θ_m. At that point, testing was terminated and Testee B was declared a nonmaster of the subject area.

It should be noticed that Testee A had a final θ estimate ($\hat{\theta} \simeq 1.9$) that was much closer to the mastery level than the final θ estimate for Testee B ($\hat{\theta} \simeq -.30$). Therefore, much more precise measurement was needed for Testee B than for Testee A to make mastery decisions with comparable confidence levels, and several more items were administered to Testee B than to Testee A, to obtain the additional precision needed in order to make the mastery decision.

Method

The AMT strategy was evaluated using real-data simulation (Weiss, 1973). In this approach, test item response data obtained from the administration of a conventional paper-and-pencil multiple-choice achievement test were used to simulate the administration of the AMT strategy. That is, items were selected by the AMT strategy for each student from the conventional test already administered. Item responses obtained in the conventional test were used by the AMT strategy and scored as described above. If a mastery decision could not be made after a given item was used, another item from the conventional test was selected by the MISS approach, and the previously obtained item response was used by the AMT strategy. This procedure was continued until the AMT strategy could make a mastery decision or until all items in the conventional test pool had been administered.

Subjects and Tests

Item response data were obtained from trainees undergoing the Weapon Mechanics course at the Lowry Air Force Base Technical Training Center during 1977 and 1978. This course is computer-managed, and trainees proceed at their own pace through 13 well-specified blocks of instruction. During each block, several tests are given from which mastery decisions are made. Trainees are given several attempts to pass each test in each block.

For this study two block tests of different lengths were arbitrarily chosen to investigate the properties of the AMT procedure. Specifically, data used were the item responses of 200 trainees to the first test in the first block of instruction (Test 11) and the item responses of 200 trainees to the first test in the third block of instruction (Test 31). These tests consisted of 30 and 50 conventionally administered 5-alternative multiple-choice items, respectively. Only the trainees' performances in their first attempt to pass the tests were used for this study.

Fitting the ICC Response Model

Estimation of item parameters. The procedure used for the estimation of the three item parameters of the logistic ICC response model was developed by Urry (1976). This procedure obtains initial estimates for the discrimination (a) and the difficulty (b) parameters for an item through the use of a direct conversion of the classical item parameters and the individuals' raw scores (number correct). A value of the lower asymptote parameter (c) is found which minimizes a χ^2 goodness-of-fit statistic for the item. These initial values are made more precise through the use of an ancillary correction procedure (Fisher, 1950). To obtain more precise estimates of the parameters, the entire procedure is repeated replacing the individuals' raw scores with Bayesian modal estimates (Samejima, 1969) of their achievement levels.

Urry's item parameterization method excludes items which meet any of the following rejection criteria during the first stage of the procedure:

1. a less than .80,
2. b less than -4.00 or greater than 4.00, and
3. c greater than .30.

If an item is excluded on the basis of one of these criteria during the initial stage of the parameterization procedure, it receives no parameter estimates in either stage of the procedure. These restrictive criteria are removed after the first phase of the calibration, and no further culling of the items is done. Thus, the final values of the parameter estimates for those items which survive the first phase are not constrained by the rejection criteria.

Evaluating the fit of the model. To examine the usefulness and appropriateness of the unidimensional three-parameter logistic ICC model with data of the type provided by the Weapon Mechanics course, two questions were investigated:

1. Does factor analysis of the intercorrelations between item responses result in only a single common factor? That is, is the use of a unidimensional model justified by the presence of only a single nonrandom dimension?
2. Do parameter estimates obtained from these data correspond to the range of parameter estimates obtained in previous studies that have shown this type of model to be useful in increasing testing efficiency?

To answer the first question, principal axis factor analyses were performed separately on the data from Test 11 and Test 31. Matrices of item intercorrelations (phi coefficients) were calculated from the raw item-response data for the 200 trainees on each of the tests using the PEARSON CORR computer subroutine from the *Statistical Package for the Social Sciences* (*SPSS*; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1970).

The resultant 30 × 30 (Test 11) and 50 × 50 (Test 31) item intercorrelation matrices were each factor analyzed by the iterative principal axis factor analysis subroutine from *SPSS*. The initial communality estimate for each of the items was the squared multiple correlation of the item with all other items in the test. The analysis iterated until successive communality estimates differed by a negligible amount.

To determine the amount of random variation in the final factor-analytic solutions, parallel analyses were conducted following the suggestion of Horn (1965). This entailed factor analyses of sets of random data that were generated to parallel the original data, using the same number of "items" and "subjects." Eigenvalues obtained for factors in the random data were used to determine whether factors obtained from the analysis of the real data were "true" factors or residual factors. If the eigenvalue of a factor obtained from the real data was larger than that for the corresponding random-data factor, the real-data factor was considered to be a true factor; but if the eigenvalue was similar to that obtained from the random-data factor, then the real-data factor was considered to be a residual factor of no real importance.

To answer the second question posed above, the parameter estimates obtained for these two tests were compared to the estimates obtained in two other studies (Bejar, Weiss, & Kingsbury, 1977; Brown & Weiss, 1977) that used a unidimensional three-parameter logistic ICC model to attempt to improve testing accuracy in achievement testing situations. Further comparisons were made between the parameter estimates obtained from the present data and the guidelines expressed by Urry (1977) to indicate whether the use of an adaptive testing item pool will improve the quality or efficiency of trait measurement. Urry's guidelines are as follows:

1. The a parameter estimates of the items in the pool should exceed .80.
2. The b parameter estimates should be widely and evenly distributed between -2.00 and +2.00.
3. The c parameter estimates should be less than .30.

To the extent that parameter estimates obtained from Tests 11 and 31 followed Urry's guidelines and showed close correspondence to other item pools that have proven to be useful in adaptive testing, it could be concluded that the items used in this study would show some usefulness with the unidimensional three-parameter ICC model.

Simulation of AMT

In order to simulate the AMT strategy, a computer program was designed to "administer" the one item in the item pool (which included all of the items from the conventional test not rejected by the calibration procedure) providing the most information at a trainee's current level of $\hat{\theta}$. Each trainee began the test with $\hat{\theta}$ of 0.0 and a prior variance of 1.0. The trainee's response taken from his/her original responses to the conventional test was used by the Bayesian scoring routine to produce a new θ estimate. Then the item with the most information at this new $\hat{\theta}$ was chosen to be administered next. (No item was administered more than once to a trainee.) A new θ estimate was found using the trainee's response to this item, and then another item was chosen based on the new θ estimate.

The program continued to choose items to be administered until the trainee's $\hat{\theta}$ was shown to be either above or below a given mastery level, θ_m, with a prespecified degree of confidence. A 95% Bayesian symmetric confidence interval was calculated around the trainee's $\hat{\theta}$ after each item was administered. The AMT strategy continued until this confidence interval failed to include the prespecified mastery level; when this occurred, the AMT procedure was terminated. A lower limit of three items was set for the length of the AMT to avoid anomalous results that might occur from making mastery decisions based on a small number of item responses. For trainees for whom a mastery decision could not be made with the AMT procedure before all items were administered, mastery was determined by whether the final $\hat{\theta}$ was above or below θ_m.

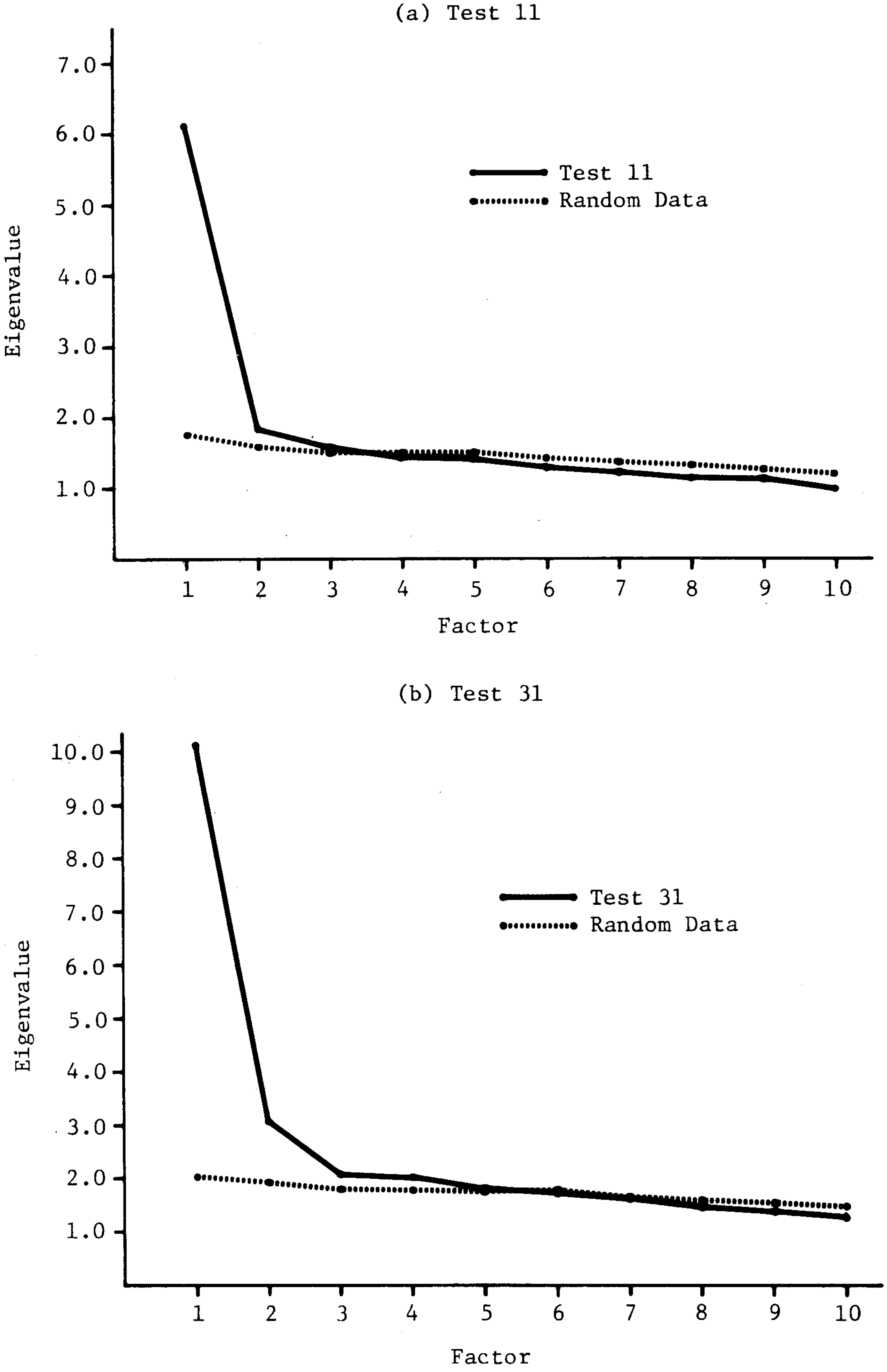
During the simulation, three different mastery levels were used corresponding to proportion correct mastery levels of P=.7, .8, and .9. These mastery levels were calculated from the TCC for each test, as described above. To maximize the comparability between the conventional and adaptive mastery testing strategies, the conventional test was truncated to include only the items which were not rejected by the calibration procedure. In addition, the conventional test was scored by Owen's Bayesian scoring method, and the same mastery levels were used for both testing strategies.

Comparison of Efficiency: AMT versus Conventional Testing

If the AMT strategy were a more efficient testing procedure than the conventional mastery testing procedure, it would reduce test length while administering items with high enough information to maintain a very high correlation between decisions made by the AMT and the conventional approach. Consequently, to determine whether the AMT procedure reduced the number of items given to trainees without reducing the quality of the mastery decisions made for those trainees, three criteria were evaluated separately for Test 11 and Test 31 for the AMT and conventional testing procedures at each of the mastery levels:

1. The mean number of items administered to trainees,
2. The mean information obtained after all items were administered, and
3. Relationships between mastery decisions made at the termination of the testing by the AMT and conventional procedures.

Figure 3
Eigenvalues of the First 10 Common Factors Extracted From Item Intercorrelations for Test 11 and Test 31 and for Parallel Random-Data Factors



(a) Test 11

(b) Test 31

In addition, to examine the characteristics of the two testing procedures more closely, the mean information obtained from each procedure was plotted for each testing strategy as a function of the achievement level estimate for each mastery level.

Results

Applicability of the ICC Model

Factor analysis. Eigenvalues of the first 10 factors extracted from item intercorrelations for Test 11 and Test 31 and the random data parallel analysis for each test are shown in Appendix Table B-1; these values are plotted in Figure 3. For Test 11 (Figure 3a) the first three factors had higher eigenvalues than their corresponding random-data factors. However, only the first factor differed substantially from the corresponding random-data factor. Thus, for Test 11 it was not unreasonable to infer that only the first factor was a "true" factor underlying trainees' responses, since the eigenvalues of the other factors resembled those of the random factors and the first factor accounted for more than three times the amount of common variance than any other factor.

For Test 31 (Figure 3b) the eigenvalues of the first five factors extracted each exceeded the eigenvalues of their corresponding random factor, but only the first two factors exceeded the random-data values by a substantial amount. The first factor accounted for 20.5% of the common variance extracted by the 10-factor solution, and the second factor accounted for 6.2% of the common variance. No other factor accounted for more than 5% of the variance. These data indicate that there were probably two real factors underlying trainees' responses to Test 31. This two-factor solution might indicate that a multidimensional latent trait model should be postulated to explain trainees' responses to Test 31. However, because the first factor accounted for over three times as much variance as the second factor, the unidimensional model could still be used; data presented by Reckase (1978) indicate that if a dominant first factor exists, items calibrated using a unidimensional model will adequately measure that first factor.

Estimation of the ICC parameters. Tables 1 and 2 show the ICC parameter estimates obtained for each of the items in Test 11 and Test 31, respectively. Of the items in the conventional test, 17% (5 items) from Test 11 were rejected by the parameterization procedure, while 24% (12 items) were rejected for Test 31. These losses are comparable to losses observed during other investigations of achievement tests using this parameterization procedure; Bejar, Weiss, and Kingsbury (1977) lost 22% of their total pool during item parameterization, and Brown and Weiss (1977) lost 13% of their total pool.

For Test 11, values of the a parameter estimates ranged from .63 to 4.69, with a mean of 1.48 and a standard deviation of .98. Values of the b parameter estimates ranged from -2.35 to 1.32, with a mean of -.98 and a standard deviation of 1.01. Values of the c parameter estimates ranged from .00 to .49, with a mean of .27 and a standard deviation of .138.

For Test 31, values of estimates of the a parameter ranged from .63 to 3.42, with a mean of 1.16 and a standard deviation of .65. Values of the b parameter estimates were from -1.86 to 3.18, with a mean of -.58 and a standard deviation of 1.08. The c parameter estimates ranged from .00 to .77, with a

Table 1
ICC Item Parameter Estimates for the Items in Test 11

Item Number	*a* Discrimination	*b* Difficulty	*c* Lower Asymptote
1	--[a]	--	--
2	.81	-.88	.22
3	--	--	--
4	.92	-1.58	.18
5	.66	-1.06	.37
6	.70	-1.18	.36
7	2.75	-1.98	.12
8	1.77	.81	.49
9	1.52	.26	.48
10	--	--	--
11	--	--	--
12	.63	-1.89	.29
13	1.38	-1.64	.31
14	1.70	-1.01	.37
15	1.17	-1.61	.25
16	.67	-1.90	.29
17	1.46	-.74	.27
18	.75	-.93	.12
19	.65	-1.24	.20
20	1.08	-1.71	.36
21	4.69	.98	0
22	2.16	-1.51	.16
23	2.16	-1.55	.19
24	1.32	.56	.30
25	1.21	-1.54	.36
26	--	--	--
27	3.58	-2.35	--
28	1.04	1.32	.46
29	.83	-1.69	.09
30	1.31	-.46	.43

[a] Missing values indicate that the item was rejected by the parameter estimation procedure.

mean of .28 and a standard deviation of .16. For both of these tests the parameter estimates obtained were well within the range established by two earlier studies that examined achievement tests using the same item parameterization method (Bejar, Weiss, & Kingsbury, 1977; Brown & Weiss, 1977).

Examination of the item parameter estimates obtained from Test 11 and Test 31, using Urry's guidelines for a good adaptive testing item pool, indicated the following:

1. For both Test 11 and Test 31, 76% of the items had *a* values exceeding .80, while the average value for both tests exceeded 1.00.
2. The *b* values were fairly widely and evenly distributed between -2.0 and 1.0, but the distribution was rather sparse above 1.0. Considering the small numbers of items in the two item pools, the distribution of the *b* values seems appropriate, though the pools might have been

Table 2
ICC Item Parameter Estimates for the Items in Test 31

	a	*b*	*c*
Item Number	Discrimination	Difficulty	Lower Asymptote
1	--[a]	--	--
2	.70	-1.40	.33
3	3.39	-1.86	--
4	1.95	3.18	.77
5	.88	-1.78	.37
6	.65	- .82	.14
7	.71	- .68	.39
8	.81	-1.85	.38
9	.66	-1.84	.35
10	--	--	--
11	1.18	-.74	.37
12	--	--	--
13	.95	-.90	.36
14	2.55	-1.39	.01
15	--	--	--
16	.94	-.44	.13
17	1.13	-1.43	.23
18	.92	-.46	.38
19	1.03	-.49	.13
20	.79	.26	.16
21	.80	-1.04	.35
22	1.01	-.65	.15
23	.80	-1.11	.19
24	.79	.98	.27
25	--	--	--
26	1.05	.09	.41
27	.95	-.23	.39
28	1.11	-1.64	.20
29	1.54	-1.56	.14
30	.73	-.44	.11
31	.63	-1.54	.06
32	--	--	--
33	.95	.40	.17
34	1.20	1.13	.45
35	1.07	.45	.27
36	3.42	-1.74	--
37	--	--	--
38	--	--	--
39	--	--	--
40	--	--	--
41	--	--	--
42	--	--	--
43	1.04	-.77	.37
44	1.18	-.49	.39
45	1.03	-.97	.36
46	.74	-1.83	.21
47	1.08	-.56	.38
48	1.02	.80	.37
49	.83	.29	.42
50	1.70	1.06	.33

[a] Missing values indicate that the item was rejected by the parameter estimation procedure.

slightly too easy to meet Urry's second guideline. However, Urry's guidelines were proposed for ability tests for which it is desired to measure precisely across a wide range of ability, whereas the data of this study were from a mastery achievement test for which it was desired to classify students on either side of a mastery level. Thus, the distribution of *b* values would not be expected to conform with Urry's second recommendation.

3. Fifty-six percent of the items in Test 11 and 47% of the items in Test 31 obtained *c* estimates below .30. The average *c* estimate for each test was less than .30.

Thus, in light of Urry's guidelines and the earlier studies, examination of the item parameters obtained indicated that the parameter estimates obtained from Test 11 and Test 31 were similar to those obtained for items which had previously been used to improve achievement measurement; consequently, the items were appropriate for investigating the AMT strategy.

Conversion of the Mastery Level to the ICC Metric

The ICC item parameter estimates for each test were used in Equation 1 to obtain the TCC for each test. Figure 4 shows the resulting TCC for Test 11 (Figure 4a), using item parameters for the 25 items that survived the calibration procedure, and for Test 31 (Figure 4b), based on the 38 items for which parameter estimates were available on that test. Conversion of the proportion correct mastery levels (P=.7, .8, and .9) to the achievement metric (θ) are also shown.

Test 11 had a slightly steeper TCC than did Test 31, reflecting the higher average discrimination of its items. The lower average *b* level of the Test 11 items (i.e., easier items) is reflected in the fact that the TCC for Test 11 is shifted to the left along the achievement level, or θ, axis in comparison to Test 31. The relatively equal average *c* parameters for the two tests are reflected in the values of the TCC at θ=-4.0.

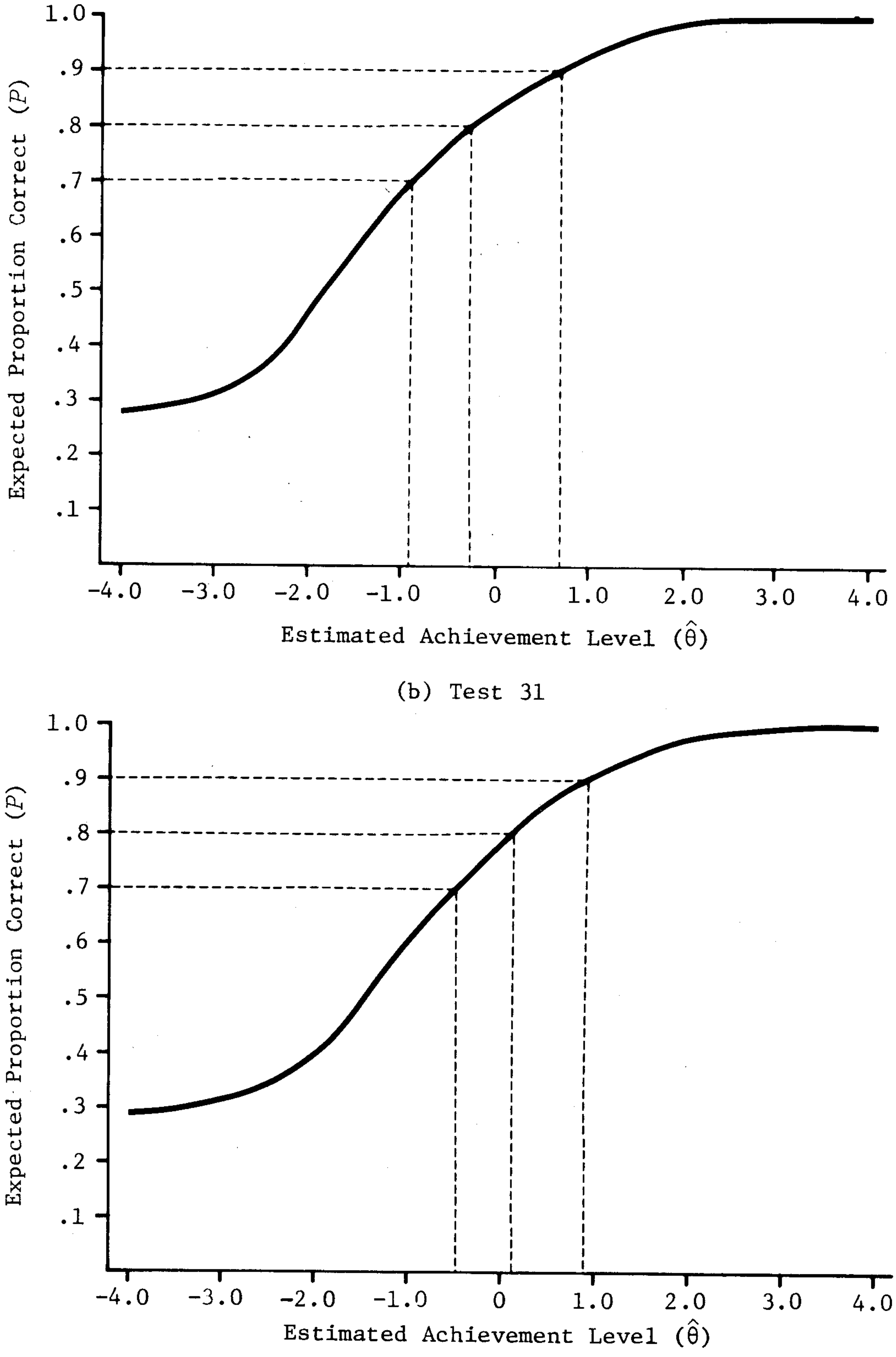
For Test 11 the P=.7 mastery level was converted to θ=-.90 on the achievement metric, the P=.8 mastery level was converted to θ=-.23, and the P=.9 mastery level was converted to θ=.75. For Test 31 the P=.7 mastery level was converted to θ=-.48; the P=.8 level, to θ=.12; and the P=.9 level, to θ=.91 on the achievement metric. It can be seen that for both tests the conversion was nonlinear, reflecting the gain in potential discriminability resulting from consideration of the unique operating characteristics of each item.

Test Length

Table 3 shows the mean number of items, the average amount of information obtained from each item administered, and the number of individuals from various subsamples under the AMT and conventional strategies at each of the three different mastery levels. The four subgroups for which these data are presented are (1) the total group of trainees, (2) the groups of trainees declared masters by the relevant testing procedure, (3) the groups of trainees declared nonmasters by the relevant testing procedure, and (4) the groups of trainees for which the AMT procedure made decisions with full confidence (i.e., trainees for whom the mastery level, θ_m, fell outside the 95% confidence interval at some point during the test and terminated the AMT procedure). Frequency distributions of

Figure 4
Test Characteristic Curves for Test 11 and Test 31, with Conversion of Three Mastery Levels (P=.7, .8, and .9) from the Proportion-Correct Metric to the Achievement Metric

(a) Test 11



(b) Test 31

numbers of items administered for each of these subgroups are in Appendix Table B-2 for Test 11 and Appendix Table B-3 for Test 31.

Table 3
Sample Size (N), Mean Test Length ($\bar{L}$), and Mean Information Per Item ($\bar{I}$) for AMT and Conventional (Conv) Test for Tests 11 and 31 at Three Mastery Levels for Total Group and Three Subgroups

Test, Mastery Level, and Testing Strategy	Group											
	Total			Mastery			Nonmastery			High Confidence		
	N	$\bar{L}$	$\bar{I}$	N	$\bar{L}$	$\bar{I}$	N	$\bar{L}$	$\bar{I}$	N	$\bar{L}$	$\bar{I}$
Test 11												
P=.7												
Conv	199	25	.29	172	25	.28	27	25	.39	154	25	.30
AMT	199	12.8	.32	174	12.3	.28	25	16.4	.52	154	9.2	.36
P=.8												
Conv	199	25	.29	135	25	.29	64	25	.30	100	25	.40
AMT	199	17.4	.29	126	17.4	.28	73	17.4	.32	100	9.9	.55
P=.9												
Conv	199	25	.29	43	25	.50	156	25	.23	132	25	.27
AMT	199	13.1	.34	34	20.8	.50	165	11.5	.27	132	7.0	.38
Test 31												
P=.7												
Conv	200	38	.22	127	38	.18	73	38	.28	122	38	.21
AMT	200	21.8	.30	134	19.4	.26	68	26.4	.36	122	11.4	.44
P=.8												
Conv	200	38	.22	73	38	.15	127	38	.26	117	38	.24
AMT	200	23.4	.26	74	27.7	.20	126	20.9	.32	117	13.1	.41
P=.9												
Conv	200	38	.22	27	38	.12	173	38	.23	151	38	.24
AMT	200	14.7	.23	28	38	.12	172	10.9	.30	151	7.2	.40

Total group. For the total group of trainees responding to Test 11, the AMT procedure reduced the average number of items administered ($\bar{L}$) substantially at every mastery level. The minimum reduction in number of items administered that was noted was for the P=.8 mastery level, where test length for the conventional test was 25 items, compared to a mean test length for the AMT procedure of $\bar{L}$=17.4 items; this reduction of 7.6 items represents a minimum test length reduction of 30.4% of the conventional test length. The maximum test length reduction was 48.8% of the conventional test (12.2 items) when a mastery level of P=.7 was used. For the same group of trainees, a gain in the average amount of information ($\bar{I}$) obtained from each item administered was noted for the AMT procedure at the P=.7 and P=.9 mastery levels. The gains in information per item administered were .03 information units (IU), or a 10% increase at the P=.7 mastery level, and .05 IU, or a 17% increase, at the P=.9 mastery level.

For the total group of trainees responding to Test 31, the same two trends were noted. First, test length was reduced with the use of the AMT procedure at each mastery level. The minimum reduction of test length was noted with the use of the *P*=.8 mastery level, for which the conventional test length of 38 items was reduced to a mean AMT length of $\overline{L}$=23.4 items--a reduction of 38.4% in mean test length. The greatest reduction in test length was noted for the .9 mastery level at which the mean AMT length was 14.7--a reduction in test length of 61.3%.

The second trend was that the AMT procedure provided more information with each item administered than the conventional test for all mastery levels. The smallest increase in information was .01 IU per item (a 5% increase), for the *P*=.9 mastery level. The largest gain in the mean information per item was .08 IU (a 36% increase), for the *P*=.7 mastery level. For mastery levels *P*=.8 and *P*=.9, the percent reduction in test length under AMT was greater for Test 31 than that noted for Test 11. The increase in information per item noted for AMT was greater for Test 31 than for Test 11 at all three mastery levels.

Appendix Tables B-2 and B-3 show that test lengths for the AMT procedure for different trainees were quite variable. For most of the trainees, either a very long test (as long as the conventional test) was needed, or a very short test (8 items or less) was sufficient. This U-shaped distribution of test lengths was obtained for both Test 11 and Test 31 across all mastery levels.

Mastery groups. When only those trainees were considered who were judged to be masters for Test 11 at one of the mastery levels by the AMT or the conventional testing procedure, test length reduction was again noted for the AMT procedure at all three mastery levels. For mastery levels *P*=.7 and *P*=.8, adaptive tests for those in the mastery group were approximately the same mean length as those for the total group; but for mastery level *P*=.9 adaptive tests for the mastery group were much longer (20.8 versus 13.1 items on the average). In comparison with the conventional test, for the AMT procedure in the mastery group alone the minimum test length reduction was 4.2 items, or 16.8% of the conventional test length of 25 items, at the *P*=.9 mastery level; and the maximum test length reduction was 12.7 items, or 50.8% of the conventional test length, at the *P*=.7 mastery level.

The AMT procedure and the conventional testing procedure provided almost identical mean amounts of information ($\overline{I}$) for items administered to the mastery groups, even though the AMT procedure administered fewer items at each mastery level. However, for these groups interpretation of the differences in mean information ($\overline{I}$) is obscured by the fact that the two different testing procedures gave trainees with different achievement levels mastery status. A clearer comparison of information provided by the two testing procedures is shown below.

For the groups of trainees labeled as masters for Test 31, test-length reduction was observed with the use of AMT for only two of the three mastery levels examined. At the *P*=.7 mastery level, mean test length was reduced by 18.6 items, or a reduction of 48.9% of the conventional test length, by use of AMT. For the *P*=.8 mastery level the mean test length was reduced by 10.3 items, or a reduction of 27.1% of the conventional test length. For the *P*=.9 mastery level the AMT procedure never reached a decision of mastery in less than 38 items, the length of the conventional test.

For Test 31, the AMT procedure resulted in higher mean information per item than the conventional test for the *P*=.7 mastery level (a difference of

.08 IU per item, or a 44% increase over the conventional test) and the *P*=.8 mastery level (.05 IU per item higher, a 33% increase). At the *P*=.9 mastery level the conventional test and the adaptive test administered items with equal average information.

As the mastery level became higher, for both Test 11 and Test 31 there was a trend for greater numbers of items to be administered before a decision of mastery could be made. This resulted from the fact that the higher mastery levels fell above the steepest portion of the TCCs, as is shown in Figure 4. This would imply that the entire conventional test would have more difficulty discriminating among trainees at these mastery levels; consequently, the AMT procedure would have to use more of the items from the conventional test in order to determine whether a trainee was above or below the higher mastery levels. This trend may be clearly seen in Appendix Tables B-2 and B-3. For each test, trainees were placed in the mastery group for mastery level *P*=.7 with a wide range of test lengths. As the mastery level was raised, trainees were more likely to be declared masters only after a larger number of items were administered, until for Test 31 at the *P*=.9 mastery level, all those who were declared masters took all of the items in the item pool before the mastery decision was made.

Nonmastery groups. For the trainees who were declared nonmasters for Test 11, using either the adaptive or conventional testing procedures, reductions in test length were observed at every mastery level with the AMT procedure. The smallest reduction in test length, 7.6 items, was observed for the *P*=.8 mastery level and accounted for 30.4% of the conventional test length. The largest reduction is test length was 13.5 items at the *P*=.9 mastery level, or 54% of the conventional test length. At each mastery level for Test 11, more mean information was obtained from each item administered to the nonmasters by the AMT procedure than by the conventional procedure. The smallest increase in information per item was .02 IU (a 6.7% increase), for the *P*=.8 mastery level. The largest increase in mean information was .13 IU (a 33.3% increase) per item, for the *P*=.7 mastery level.

For the trainees declared nonmasters for Test 31, reductions in mean test length were again noted with the AMT procedure at each mastery level. The minimum mean decrease in test length was 11.6 items, or 30.5% of the conventional test length of 38 items, at the *P*=.7 mastery level. The maximum reduction in average test length was 27.1 items, or 71.3% of the conventional test length, at the *P*=.9 mastery level. As the criterion level increased, the number of items needed by the AMT procedure to make the nonmastery decision steadily decreased.

For the nonmastery groups administered Test 31, the mean information per item was higher at each mastery level for the AMT procedure than for the conventional testing procedure. The minimum increase in information was .06 IU (a 23% increase) per item administered, for the *P*=.8 mastery level; and the maximum increase observed was .8 IU per item (a 28.6% increase), for the *P*=.7 mastery level.

Across both Tests 11 and 31, there was a tendency for the adaptive test to administer fewer items before making a decision of nonmastery as the mastery level increased. The sole exception to this trend was observed for Test 11 at the *P*=.8 mastery level, which showed a slight increase in the number of items administered when compared with the P=.7 mastery level for that test. For both tests a higher mean information was obtained for each item administered by the AMT procedure at each mastery level. No consistent trend was noted in the differences in average information per item across mastery levels for the two tests.

High-confidence groups. The high-confidence groups included only those trainees for whom the AMT procedure terminated with full confidence, i.e., trainees for whom the Bayesian confidence interval failed to include the mastery level at some test length at or before the exhaustion of the items from the conventional test item pool. For Test 11 the AMT procedure terminated with high confidence for a minimum of 50% of the group of trainees, at the P=.8 mastery level. The largest high-confidence group was 77% (N=154) of the total group of trainees, at the P=.7 mastery level.

Test length was reduced considerably by the AMT procedure at all criterion levels for the high-confidence groups. The minimum reduction in mean test length was observed for the P=.8 mastery level and was 15.1 items, or 60.4% of the conventional test length. The largest mean reduction in test length observed was 18 items, or 72% of the conventional test length, at the P=.9 mastery level. Modal test length for the high-confidence groups for Test 11 at all mastery levels was 3 items (see Appendix Table B-2), or only 12% of the length of the conventional test (an 88% reduction). The AMT procedure produced greater mean information per item at each mastery level. The smallest observed increase was .06 IU (a 20% increase) per item administered, for the P=.7 mastery level. The largest mean increase was .15 IU per item (a 37.5% increase), at the P=.8 level. For Test 31 the minimum number of trainees in the high-confidence group was 117, or 58% of the total group, at the P=.8 mastery level. The largest high-confidence group was 151, or 76% of the total trainee group, for the mastery level P=.9.

Test length for the AMT procedure was much shorter than the conventional test at each criterion level. The smallest reduction in mean test length was 24.9 items, or 65.5% of the conventional test length, for the P=.8 mastery level. The largest average reduction in test length was 30.8 items, or 81.1% of the total conventional test length, for the P=.9 mastery level. Similar to Test 11, modal test lengths for Test 31 were quite short: 4 items for the P=.7 mastery level, 5 items for the P=.8 mastery level, and 3 items (for 57% of the high-confidence group) at the P=.9 mastery level.

The AMT procedure produced higher mean information per item than the conventional testing procedure at all mastery levels. The minimum increase in mean information per item was .16 IU (an increase of 66.7% over the mean information provided by the conventional test), for the P=.9 mastery level. The maximum mean information increase that was observed was .23 IU per item (a 112% increase), for the P=.7 mastery level.

For both Test 11 and Test 31 the AMT procedure made confident decisions for between 50% and 77% of the total group at each mastery level. For the trainees in the high-confidence groups, the average adaptive test length ranged from 19% to 39% of the original conventional test length, while modal test lengths were only 8% to 6% of the conventional test length (i.e., over 90% reduction). Also, the adaptive testing procedure resulted in 20% to 119.5% increase in the mean amount of information obtained per item over the conventional test. The increase in mean information per item was greater for Test 31 than for Test 11 at all criterion levels.

Correspondence Between Decisions

Table 4 shows the Pearson product-moment (phi) correlations between the decisions made by the AMT and conventional testing procedures across all three criterion levels for Test 11 and Test 31. The lowest correlation observed was .67, for Test 11 at the *P*=.9 mastery level. The highest correlation was .97, for Test 31 at the *P*=.8 mastery level. The correlations between mastery decisions for Test 31 were higher than for Test 11 at all mastery levels. In addition, the average decision variance in common between the two testing procedures was 79% of the total decision variance.

Table 4
Phi Correlations Between Mastery Decisions Made by AMT and Conventional Testing Procedures for Test 11 and Test 31, at Three Mastery Levels

	Mastery Level		
Test	*P*=.7	*P*=.8	*P*=.9
Test 11	.91	.88	.67
Test 31	.93	.97	.94

To examine more completely the correspondence in decisions made by the AMT and conventional procedures, Table 5 shows joint frequency distributions of decisions for the two testing procedures at each of the three mastery levels for Test 11 and Test 31. The lowest level of agreement between the AMT and conventional testing procedures was noted for Test 11 at the *P*=.9 mastery level, where the two testing procedures agreed for 178, or 89.4% of the 199 trainees tested. The highest level of agreement was 98.5%, for Test 31 at the *P*=.8 and *P*=.9 mastery levels. Across both tests and all criterion levels, the two procedures agreed for 95.9% of the trainees tested. For the longer test (Test 31) the two procedures agreed for 97.9% of the trainees, and for the shorter test (Test 11) the two procedures agreed for 94.0% of the trainees.

Table 5
Joint Distributions of Mastery Decisions Made by AMT and Conventional Tests 11 and 31 at Three Mastery Levels

Mastery Level	Test 11		Test 31	
and AMT Decision	Mastery	Nonmastery	Mastery	Nonmastery
P=.7				
AMT Mastery	171	3	126	6
AMT Nonmastery	1	24	1	67
P=.8				
AMT Mastery	125	1	72	2
AMT Nonmastery	10	63	1	125
P=.9				
AMT Mastery	28	6	26	2
AMT Nonmastery	15	150	1	171

Information Functions

Figures 5 and 6 show the information obtained by Conventional Tests 11 and 31, respectively, and adaptive testing procedures as a function of estimated achievement level ($\hat{\theta}$). (Points plotted in these figures are based on mean information obtained from trainees within a plus or minus .1 range around a given $\hat{\theta}$; numerical values of information are shown in Appendix Table B-4.) Figures 5 and 6 each show three adaptive testing information curves--one for each mastery level examined--and one conventional test curve.

Figure 5 shows that Test 11 was poorly designed to make mastery decisions at middle-range mastery levels ($\hat{\theta}$ between -.5 and +.5, or proportion correct of about *P*=.75 to *P*=.85), since the test's information was predominantly concentrated at low achievement levels ($\hat{\theta}$<-1.0), with an information spike caused by a single highly discriminating item (Item 28; see Table 1) at about 1.0 on the achievement continuum. Information functions for the AMT strategy at each of the three mastery levels closely approximated the conventional information function in the region near each respective mastery level ($\hat{\theta}$=.8, -.2, -9). In addition, as achievement level moved away from the mastery levels, the AMT information functions fell below the information function for the conventional test, particularly at the lower achievement levels. Further, as the difference between the achievement level and the mastery level increased, the difference in amounts of information used by the AMT procedure and the conventional procedure tended to become larger. However, for the *P*=.8 mastery level an upturn in the information function occurred below the -1.3 achievement level, and the difference in information between the conventional and adaptive procedure decreased slightly. The same type of upturn was noted for the *P*=.9 mastery level, for $\hat{\theta}$ levels below -1.1.

Figure 6 shows that for Test 31 the conventional test information function was monotone decreasing within the observed range of trainees' achievement levels. This implies that Test 31 provided its most precise measurement at low achievement levels and that differences between the two testing procedures should be most noticeable at low achievement levels. The AMT information functions for Test 31 in Figure 6 reinforce the trends noted in Test 11 for each of the mastery levels. That is,

1. The AMT information functions each closely approximated the conventional test information function in the region of the achievement continuum near the appropriate mastery level.
2. For achievement levels beyond the region near the mastery level, the AMT information function was lower than the conventional test information function.
3. The difference in information between the AMT and conventional testing procedures was greater for achievement levels further from the specified mastery level, up to a point.
4. At the lower end of the achievement continuum ($\hat{\theta}$<-.5), an increase in the amount of information provided by the AMT procedure was noted for each of the mastery levels examined. The point on the $\hat{\theta}$ continuum at which the upturn was noted was lower for each successively lower criterion level.

For Test 31 one additional result was noted that did not appear in the Test 11 AMT data: For both the *P*=.8 and *P*=.9 criterion levels, a final downturn in the

Figure 5
Mean Obtained Information as a Function of Estimated Achievement Level for AMT and Conventional Test 11 at Three Mastery Levels

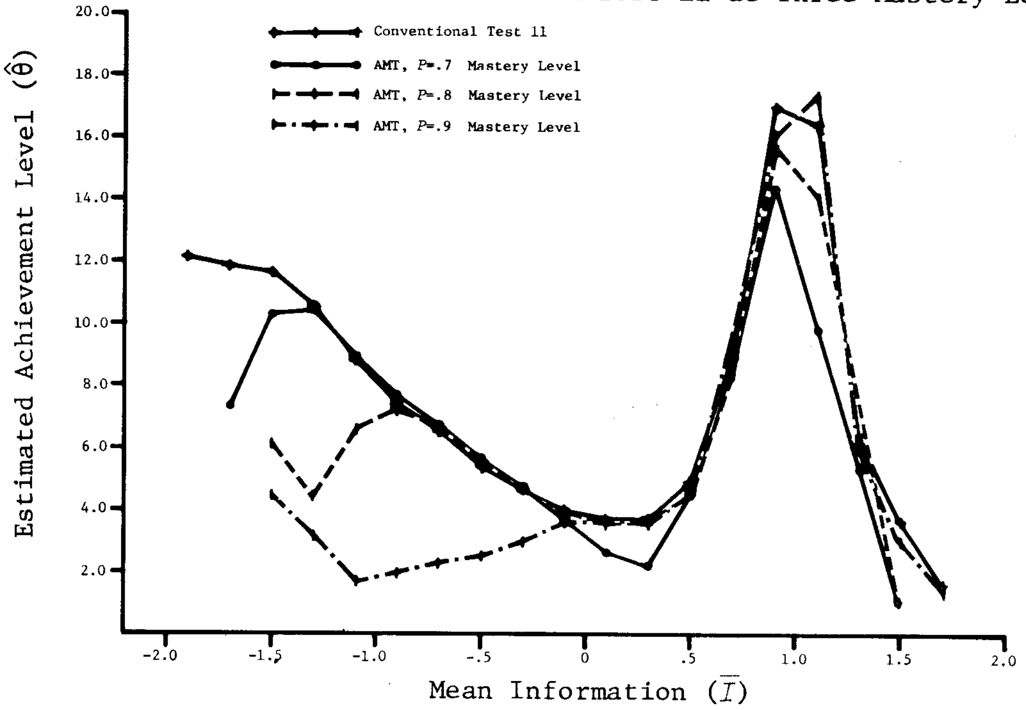
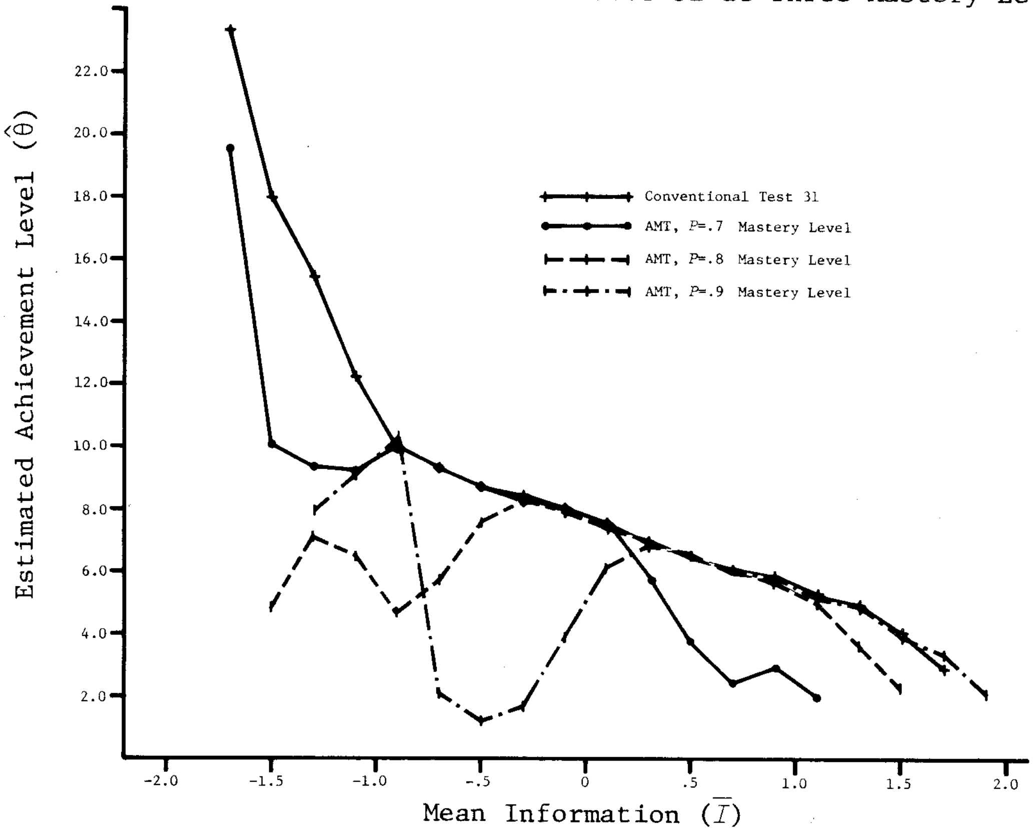


Figure 6
Mean Obtained Information as a Function of Estimated Achievement Level for AMT and Conventional Test 31 at Three Mastery Levels



information functions for the AMT procedure was observed at the lowest obtained $\hat{\theta}$ levels. This implies that the observed upturns in information may have been one side of an information spike, possibly caused by the minimum limit of three items placed on the AMT procedure.

Discussion and Conclusions

The unidimensional three-parameter logistic ICC model was fit to two conventional tests that were previously used to make mastery decisions in a military training course. Data originally gathered during the training course were used to evaluate, in real-data simulation, the efficiency of the proposed adaptive mastery testing (AMT) procedure in terms of the number of items administered, the information obtained, and the degree of agreement between the AMT and conventional testing procedures. The AMT procedure was simulated assuming three different mastery levels, stated in terms of the achievement metric, through the use of the test characteristic curves (TCCs) for the two conventional tests. The results of these simulations indicated that the proposed AMT procedure reduced the number of items administered during the average test, while at the same time making decisions which were very much the same as those made by the conventional testing procedure.

The AMT procedure reduced the average test length for the entire group of trainees by 30% to 61% of the conventional test length. The reductions in test length observed varied across different mastery levels for both of the conventional tests. When specific subgroups of the samples were considered, mean test length reductions of up to 81% of the items in the conventional test were again observed in almost every subgroup examined at each mastery level and for both tests. The only subgroup for which no test length reduction was observed for the AMT strategy was the group passing Test 31 at the highest criterion level (P=.90 correct). For the groups of trainees for which the AMT procedure was able to make high-confidence decisions, AMT mean test lengths were 60% to 81% shorter than the conventional tests across all mastery levels examined. Further, high-confidence decisions were made for 50% to 77% of the trainees at each mastery level.

At each mastery level for each test, agreement was high between the decisions made by the adaptive and conventional testing procedures. The two procedures made the same decision for approximately 96% of the cases across all circumstances. Using the larger item pool (Test 31), the two procedures agreed for about 98% of the cases. The lowest agreement level observed was approximately 89%.

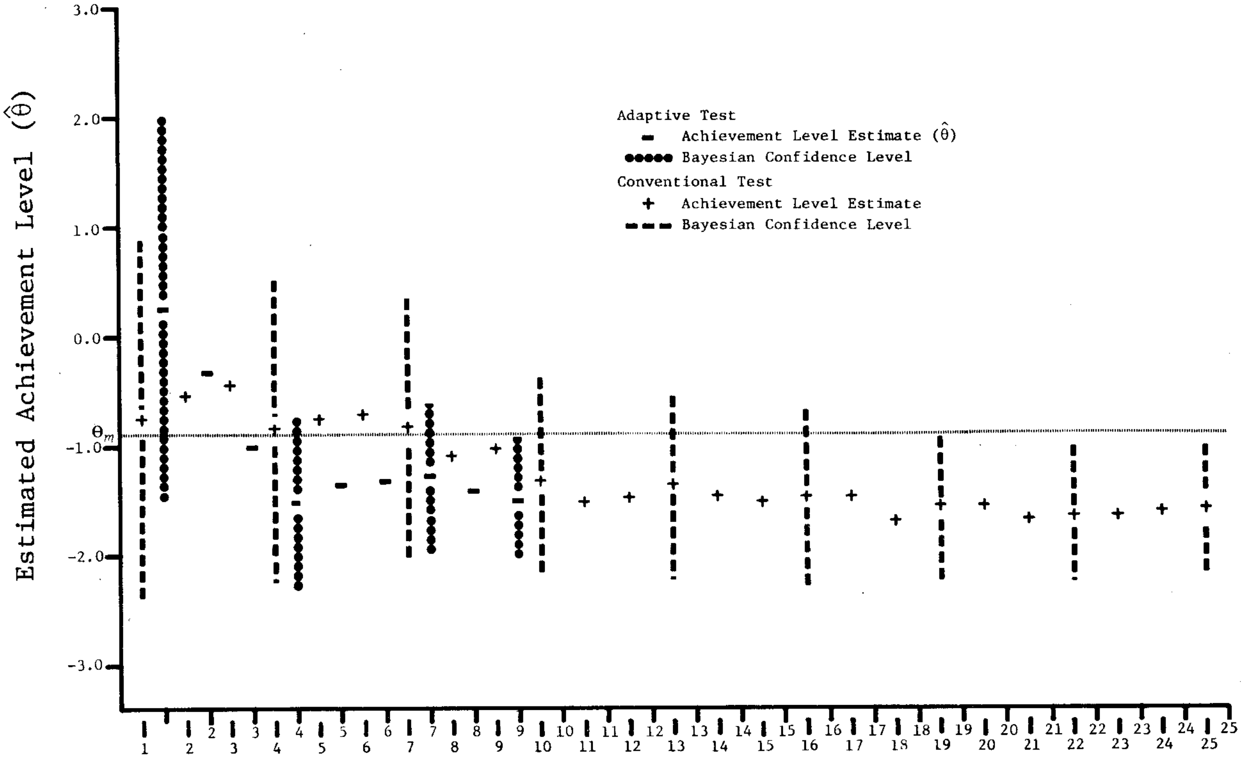
At each mastery level examined, the information functions observed for the adaptive tests closely approximated the information functions obtained for the relevent conventional test at achievement levels close to the mastery level, and fell below the conventional test information functions for more extreme achievement levels. For the achievement levels very different from the mastery level, the difference between the information functions for the two testing procedures reached a maximum; and at the most extreme achievement levels the difference in information decreased slightly.

Thus, the AMT procedure was shown to make mastery decisions very similar to those made by the conventional testing procedure, while administering fewer items, by using the information in the item pool that was available to make high-confidence decisions.

The test-length reduction observed using the AMT procedure may be attributed to two characteristics of the procedure. First, the AMT strategy administered to a trainee only those items which provided the most precise measurement at the trainee's current level of $\hat{\theta}$. Second, the AMT procedure terminated the test as soon as enough information was available to make a decision at a predetermined level of confidence concerning the trainee's mastery level. The termination rule allowed the test to terminate prior to the exhaustion of the item pool, if enough information was available in the items, and the item administration procedure presented the most informative items early in the testing session.

Each of these characteristics of the AMT procedure can be more clearly seen by examination of the Bayesian point estimates and the associated confidence intervals obtained from a trainee's responses after each item administered by the AMT and conventional testing procedures. One such record is shown in Figure 7 for a trainee responding to Test 11. The $\hat{\theta}$ estimates plotted in Figure 7 include 95% Bayesian confidence intervals for the $\hat{\theta}$ estimate after the first item and after every third item administered thereafter for both AMT and conventional procedures (even though the confidence interval was not used for making the mastery decision with the conventional procedure).

Figure 7
Achievement Level Estimates for Trainee 14 after Each Item Administered by AMT and Conventional Testing Procedures for Test 11, with 95% Bayesian Confidence Intervals Indicated after Every Third Item (*P*=.7 Mastery Level)



It may be seen from Figure 7 that both testing procedures made a nonmastery decision for the trainee (i.e., determined that the trainee's true achievement level fell below the specified mastery level), even though both procedures estimated the trainee's achievement level as being above the mastery level for the first few items. The conventional test θ estimates were above the mastery level for the first 7 items; the adaptive test θ estimates dropped below the mastery level after only 2 items. The AMT procedure made the mastery decision after administering 9 items, compared with the conventional test length of 25 items. At each test length greater than a single item, the Bayesian confidence interval around the conventional test θ estimate was larger than the confidence interval around the AMT θ estimate. This indicates the greater measurement precision available to the AMT procedure due to the adaptive item administration procedure.

Further, it may be noted in Figure 7 that the conventional test strategy finally resulted in a Bayesian confidence interval that fell completely below the mastery level after 19 items were administered (still over twice the test length of the adaptive test); but since the conventional testing procedure does not terminate even after this high-confidence level is reached, 6 more items were administered before the test ended. This illustrative example showed that the AMT procedure was far more economical than the conventional procedure in terms of test length,due to the adaptive item selection procedure and the use of the Bayesian confidence interval as a termination mechanism.

Additional Advantages of the AMT Strategy

The ICC-based adaptive mastery testing strategy described in this report has several other advantages over conventional testing procedures used to make mastery decisions. As has been demonstrated with these data, use of the ICC metric and related achievement estimation procedures can result in mastery decisions for most trainees (50% to 77%) with known and predetermined levels of confidence. Coupled with appropriate design of mastery testing item pools using ICC concepts, the percentage of high-confidence decisions could be substantially increased until mastery decisions could be made for virtually all students at the same high and predetermined level of confidence. Design of such mastery testing item pools would include a concentration of highly discriminating items around the mastery level, plus sufficient numbers of highly discriminating items elsewhere along the achievement continuum to permit high-confidence decisions to be made for all students. Actual numbers of items required at various discrimination levels could be estimated using Owen's Bayesian scoring procedure and information on the difficulties and discriminations of items to estimate in advance the values of the Bayesian posterior variance (which is used to construct the Bayesian confidence intervals used in the AMT procedure) at the expected levels of θ.

If the mastery testing item pool is not designed in advance to permit high-confidence decisions for each student, the AMT procedure still permits the tester to determine the confidence level of each mastery decision made, even if it is not a high-confidence decision. This can be determined by locating the distance of the mastery level, θ_m, from the student's estimated achievement level, $\hat{\theta}$. This distance can then be treated as a standardized deviation from the mean of a normal distribution, with a variance equal to the estimated posterior variance; and .50 plus the area of the portion of the normal distribution included in that deviation will then give the confidence level for a given mastery decision for that student. In this way, a confidence level for the mastery decision can be attached to each such decision. As a result, instructional decisions based on lower confidence level mastery decisions can be made more tentatively.

A further advantage of the ICC-based AMT strategy is that it can be extended to the multiple-content area mastery testing problem with further savings in test administration time. In many training environments, it is desirable to measure mastery on a number of learning objectives at the same point in time. Using conventional testing procedures to measure mastery on 6 objectives, for example, the student would have to take 6 different tests with a fixed number of items, for a potential total of over 100 items. However, since the AMT strategy utilizes the same item selection and scoring procedures that Brown and Weiss (1977) used in their intercontent branching adaptive testing strategy, the AMT strategy can operate in the same fashion; all that differs is the intrasubtest termination rule. Thus, in the multicontent branching AMT strategy, the achievement level estimates used to make the mastery decisions in each of a number of content-based mastery tests would be used to serve as entry points for beginning testing (using appropriate multiple regression equations) in subsequent mastery tests in the battery. If there is any correlation between mastery decisions made on the separate subtests, the use of an intercontent branching AMT should result in substantial additional savings in testing time over that obtained by use of the AMT strategy in each subtest separately.

The AMT procedure described above, or an improved version, should thus be extremely useful in a training sequence in which many subject areas are taught and tested within a short time, thus putting a premium on testing time. A self-paced instructional setting in which a student is given more than one attempt to demonstrate mastery of a content area with a single test may also benefit from an AMT procedure that would allow students to take different items on each attempt, thus avoiding the problem of students merely "learning" the test, without learning the subject matter.

The AMT procedure should be tested in an actual classroom situation. Further research should also be conducted to determine whether conventional mastery testing or the AMT procedure result in mastery decisions which more accurately predict external performance criteria.

References

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, January 1979. (NTIS No. AD A067752)

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)

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977. (NTIS No. AD A044828)

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)

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles, CA: UCLA Graduate School of Education, Center for the Study of Evaluation, 1974.

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, 1970). Dissertion Abstracts International, 1970, 30, 3856A. (University Microfilms No. 70-4530)

Fisher, R. A. Contributions to mathematical statistics. New York, NY: John Wiley & Sons, 1950.

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

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

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

Horn, J. L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30, 179-185.

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.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F. M. Discussion. In W. A. Gorman (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, September 1976. (NTIS No. PB-261-694)

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)

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical package for the social sciences. New York, NY: McGraw-Hill, 1970.

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

Popham, W. J. (Ed.), Criterion-referenced measurement--an introduction. Englewood Cliffs, NJ: Educational Technology Publications, 1971.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34 (4, Pt. 2, Monograph No. 17).

Urry, V. W. A five year quest: Is computerized adaptive testing feasible? In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (U.S. Civil Service Commission, Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-000-00940-9)

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

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (NTIS No. AD A018758)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

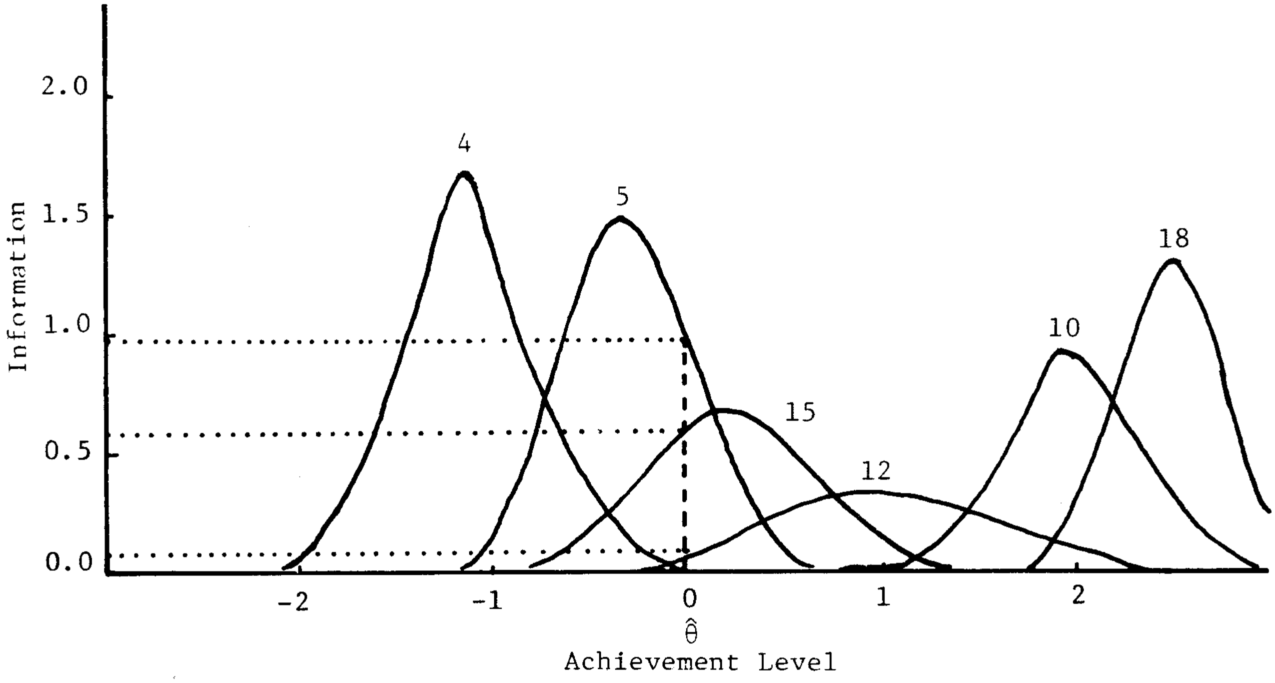
Appendix A

Illustration of MISS Procedure for Choosing Items for AMT

The essential characteristics of the adaptive testing strategy employed in this study have been described in previous sections. However, to understand the method more completely, it is helpful to see the results of its application with an actual testee.

Figure A-1 shows estimated item information curves for six items from Test 1. (There would probably be many items in the test, but only six were chosen to simplify the illustration.) The height of the information curve at a given achievement level ($\hat{\theta}$) indicates the amount of information provided by the item. Most of the items are fairly "peaked"; that is, they provide information over a relatively narrow range of the achievement continuum. While the information curves overlap to some degree, different items provide different amounts of information at a given point on the achievement continuum. The guiding principle for the adaptive procedure is to administer the item which provides the most information at the current achievement estimate ($\hat{\theta}$).

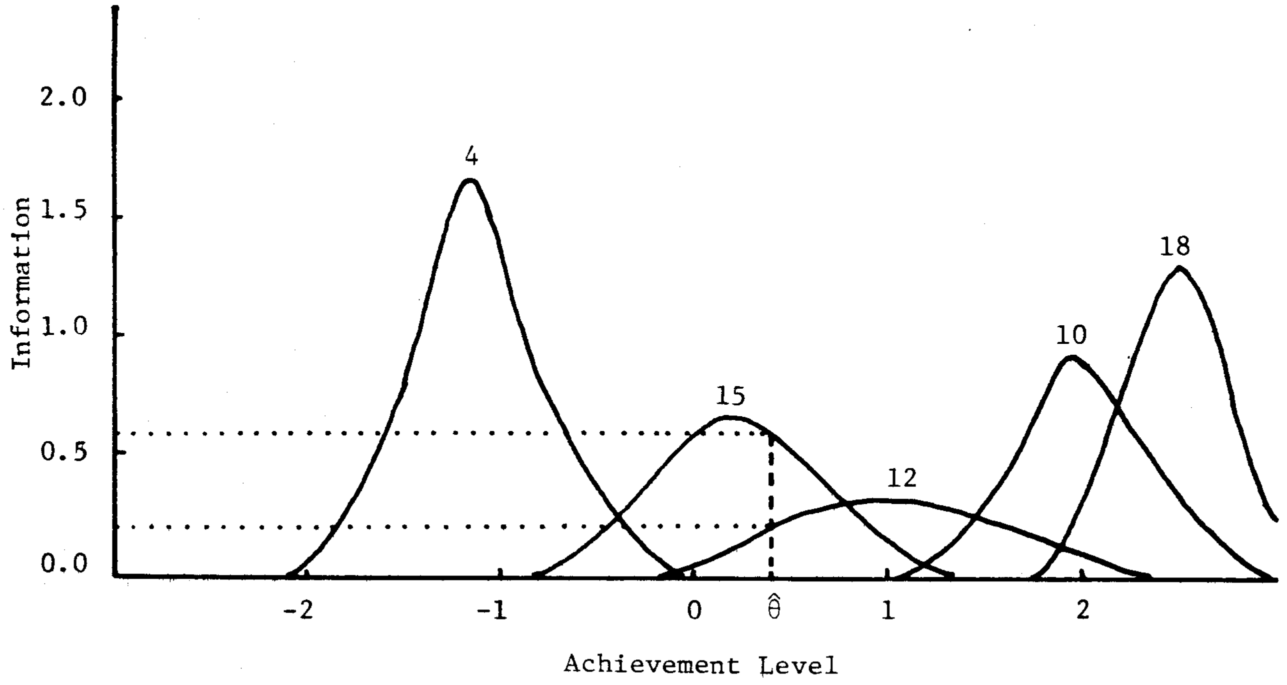
Figure A-1
Estimated Item Information Curves for Six Items from Test 1



For a testee beginning Test 1, the initial achievement estimate was $\hat{\theta}=0$; this is shown by the vertical dashed line in Figure A-1. Of the six items in the example, only three items had essentially nonzero information values at $\hat{\theta}=0$; these values, shown by the horizontal dotted lines in Figure A-1, were .95 for Item 5, .60 for Item 15, and .10 for Item 12. Applying the rule that the item selected is the one which provides the most information at the current $\hat{\theta}$, Item 5 would be selected for administration.

Figure A-2 shows the revised value of $\hat{\theta}$=.46 derived from the Bayesian scoring routine, assuming that a correct answer was given to Item 5. The confidence interval surrounding this $\hat{\theta}$ is assumed to contain the mastery level, so testing would continue. The information curve for Item 5, which was already administered, is not shown in Figure A-2. At the new value of $\hat{\theta}$, only Items 15 and 12 provide significant values of information. Since Item 15 has an information value of .60 and Item 12 has a value of .20, Item 15 would be selected as the second item to be administered to this testee.

Figure A-2
Estimated Item Information Curves for Five Items from Test 1



Assuming that the testee had correctly answered Item 15, the value of $\hat{\theta}$ would increase to .92. The confidence interval around this new $\hat{\theta}$ still contains the arbitrary mastery level, so testing would continue. At $\hat{\theta}$=.92, only Item 12 would provide significant amounts of information, and it would be administered next. Thus, at each step during the testing procedure, the item which provides the most information concerning the testee's current level of $\hat{\theta}$ is administered. In a larger item pool, testing would continue in this fashion until it was possible to make a mastery decision with a prespecified level of confidence, at which point the test would terminate.

Appendix B

Supplementary Tables

Table B-1
Eigenvalues of the First 10 Common Factors Extracted from Item Intercorrelations for Test 11 and Test 31, and for Parallel Random-Data Factors

	Test 11		Test 31	
Factor	Real Data	Random Data	Real Data	Random Data
1	6.14	1.75	10.23	2.04
2	1.85	1.61	3.08	1.90
3	1.60	1.52	2.10	1.84
4	1.41	1.51	2.06	1.82
5	1.38	1.50	1.82	1.76
6	1.30	1.39	1.68	1.72
7	1.24	1.33	1.58	1.62
8	1.16	1.28	1.49	1.58
9	1.15	1.25	1.38	1.56
10	.97	1.20	1.31	1.48

Table B-2
Frequency Distributions of Number of Items Administered by AMT Procedure from Test 11 by Mastery Subgroup for Each Mastery Level (*P*=.7, .8, and .9)

	Group											
Number of Items Administered	Total			Mastery			Nonmastery			High Confidence		
	P=.7	*P*=.8	*P*=.9	*P*=.7	*P*=.8	*P*=.9	*P*=.7	*P*=.8	*P*=.9	*P*=.7	*P*=.8	*P*=.9
3	43	54	45	39	39		4	15	45	43	54	45
4	1	1	24				1	1	24	1	1	24
5	36	1	10	36				1	10	36	1	10
6	1		3				1		3	1		3
7	10	3	17	10		8		3	9	10	3	17
8	13	2	2	13	1			1	2	13	2	2
9	3		2				3		2	3		2
10			1						1			1
11	7		6	7					6	7		6
12	1		3				1		3	1		3
13	3	2		3				2		3	2	
14	1	2	1	1					1	1	2	1
15		3	2		2			1	2		3	2
16	1	1	1	1				1	1	1	1	1
17	1	7	4		6		1	1	4	1	7	4
18	7		2	7					2	7		2
19	4	6	1	1			3	6	1	4	6	1
20	4			4						4		
21		1	3					1	3		1	3
22		2	2		2				2		2	2
23	1		3	1					3	1		3
24	7	9		7	8			1		7	9	
25	55	105	67	44	68	26	11	37	41	10	6	

Table B-3
Frequency Distributions of Number of Items Administered by AMT Procedure
From Test 31 by Mastery Subgroup for Each Mastery Level (*P*=.7, .8, and .9)

	Group											
Number of Items Administered	Total			Mastery			Nonmastery			High Confidence		
	P=.7	*P*=.8	*P*=.9	*P*=.7	*P*=.8	*P*=.9	*P*=.7	*P*=.8	*P*=.9	*P*=.7	*P*=.8	*P*=.9
3	7	7	86				7	7	86	7	7	86
4	27	1	11	27				1	11	27	1	11
5	4	15	7				4	15	7	4	15	7
6	8	6	6	7			1	6	6	8	6	6
7	10	10	4	10	7			3	4	10	10	4
8	6	4	2	5			1	4	2	6	4	2
9	6	3	1	5			1	3	1	6	3	1
10	7	6	7	5	4		2	2	7	7	6	7
11	1	10	4		3		1	7	4	1	10	4
12	1	4	3	1	1			3	3	1	4	3
13	2	8	2	2				8	2	2	8	2
14	5	1	2	4			1	1	2	5	1	2
15	3	2		3				2		3	2	
16	5	2	2	5				2	2	5	2	2
17	2	6		1	5		1	1		2	6	
18	3	6		1	5		2	1		3	6	
19	4	4	1	1	3		3	1	1	4	4	1
20	2	1		1			1	1		2	1	
21	3	4		2			1	4		3	4	
22			2						2			2
23	5	2		4			1	2		5	2	
24	1	5	1	1	2			3	1	1	5	1
25	1	1		1				1		1	1	
26												
27	1	1		1	1					1	1	
28		1	1		1				1		1	1
29	3	1			1		3			3	1	
30	1		2				1		2	1		2
31	2	1		1			1	1		2	1	
32		1						1			1	
33		1	1		1				1		1	1
34			2						2			2
35	1			1						1		
36	1	2	2				1	2	2	1	2	2
37		1	2					1	2		1	2
38	78	83	49	43	40	28	35	43	21			

Table B-4

Mean Information ($\bar{I}$) Obtained by AMT and Conventional Testing Procedures for Tests 11 and 31 At Three Mastery Levels (P=.7, .8, and .9) for Trainees with Various Achievement Level Estimates ($\hat{\theta}$), and Number of Trainees (N) at Each Achievement Level

		Test 11								Test 31							
				AMT								AMT					
$\hat{\theta}$ Range		Conventional		(P=.7)		(P=.8)		(P=.9)		Conventional		(P=.7)		(P=.8)		(P=.9)	
Lo	Hi	$\bar{I}$	N	$\bar{I}$	N	$\bar{I}$	N	$\bar{I}$	N	$\bar{I}$	N	$\bar{I}$	N	$\bar{I}$	N	$\bar{I}$	N
-2.000	-1.800	12.15	1														
-1.799	-1.600	11.88	5	7.42	2					23.44	2	19.58	1				
-1.599	-1.400	11.59	3	10.27	7	6.20	7	4.47	4	18.08	1	10.10	8	4.82	4		
-1.399	-1.200	10.58	3	10.46	4	4.46	12	3.19	7	15.53	10	9.36	12	7.23	10	8.00	6
-1.199	-1.000	8.87	10	8.93	6	6.56	4	1.60	4	12.31	9	9.19	11	6.57	8		
-.999	-.800	7.47	4	7.68	6	7.17	11			10.11	18	10.07	7	4.71	22	10.36	4
-.799	-.600	6.61	12	6.63	10	6.69	10	2.34	47	9.30	21	9.35	21	5.80	28	2.17	25
-.599	-.400	5.41	10	5.51	9	5.56	10	2.53	25	8.79	15	8.78	15	7.61	12	1.24	24
-.399	-.200	4.65	14	4.73	9	4.68	18	3.05	20	8.46	19	8.43	17	8.40	17	1.70	58
-.199	.000	4.04	21	3.79	16	4.03	18	3.68	10	8.06	18	8.06	15	8.04	17	3.96	22
.001	.200	3.73	15	2.69	24	3.72	20	3.64	15	7.58	12	7.57	4	7.51	12	6.10	11
.201	.400	3.72	19	2.21	47	3.71	9	3.73	13	6.96	15	5.83	15	6.95	12	6.89	4
.401	.600	4.91	19	4.69	4	4.60	9	4.56	12	6.48	16	3.88	23	6.50	14	6.51	8
.601	.800	8.86	22	8.80	11	8.37	16	9.41	9	6.05	7	2.60	18	6.00	6	6.00	6
.801	1.000	17.01	16	14.40	1	15.81	9	16.09	9	5.78	6	2.93	5	5.61	9	5.74	6
1.001	1.200	16.45	9			13.94	3	17.33	7	5.24	11	1.97	27	4.67	13	5.18	10
1.201	1.400	6.35	3					6.10	2	4.89	6			3.51	8	4.86	6
1.401	1.600	3.70	3	1.29	39	1.29	39	3.16	3	4.05	4			2.20	7	3.84	3
1.601	1.800	1.63	5					1.41	8	2.94	5					3.41	1
1.801	2.000															2.13	5