A VALIDITY COMPARISON OF ADAPTIVE AND CONVENTIONAL STRATEGIES FOR MASTERY TESTING

G. Gage Kingsbury
and
David J. Weiss

RESEARCH REPORT 81-3
SEPTEMBER 1981

COMPUTERIZED ADAPTIVE TESTING LABORATORY
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 Office of Scientific Research, Air Force Human Resources Laboratory, and the Office of Naval Research, and monitored by the Office of Naval Research.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 81-3	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* A Validity Comparison of Adaptive and Conventional Strategies for Mastery Testing		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR*(s)* G. Gage Kingsbury and David J. Weiss		8. CONTRACT OR GRANT NUMBER*(s)* N00014-79-C-0172
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, Minnesota 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.: 6115N Proj.: RR042-04 T.A.: RR042-04-01 W.U.: NR 150-433
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE September, 1981 13. NUMBER OF PAGES 25
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* 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 Office of Scientific Research, Air Force Human Resources Laboratory, 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)*

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

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

Conventional mastery tests designed to make optimal mastery classifications were compared with fixed-length and variable-length adaptive mastery tests in terms of validity of decisions with respect to an external criterion measure. Comparisons between the testing procedures were made across five content areas in an introductory biology course from tests administered to over 400 volunteer students. The criterion measure used was the student's final standing in the course, based on course examinations and laboratory

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

grades. Results indicated that the adaptive test resulted in mastery classifications that were more consistent with final class standing than those obtained from the conventional test. This result was observed within individual content areas and for discriminant analysis classifications made across content areas. This result was also observed for two scoring procedures used with the conventional test (proportion-correct and Bayesian scoring). Results also indicated that there was no decrement in the performance of the adaptive test when a variable termination rule was implemented. This variable termination rule resulted in test lengths which were, on the average, 74% to 88% shorter than the original adaptive tests. Further analyses explicated the manner in which the adaptive tests administered differed from the conventional test for each content area as a function of achievement level. This evidence was used to explain why the adaptive tests resulted in more valid decisions than the conventional procedure, in spite of the fact that the type of conventional test used here was the most informative test concerning the mastery cutoff. It is concluded that variable-length adaptive mastery tests can provide more valid mastery classifications than "optimal" conventional mastery tests while reducing test length an average of 80% from the length of the conventional tests.

Contents

A Validity Comparison of Adaptive and Conventional Strategies for Mastery Testing

The adaptive mastery testing (AMT) procedure developed by Kingsbury and Weiss (1979) is designed to make high-precision classifications concerning students' mastery of specific content areas within a course of instruction. The procedure is also intended to minimize the number of test questions needed to make these classifications in order to increase the amount of class time available for actual instruction. The AMT procedure makes use of item response theory (IRT; Lord, 1980; Lord & Novick, 1968) to adapt the test items administered to suit each student. The AMT procedure was compared in monte carlo simulation (Kingsbury & Weiss, 1980) to a sequential decision procedure developed by Wald (1947) and to a conventional mastery decision procedure. This simulation indicated that the AMT procedure resulted in the most valid mastery classifications of the three methods across most conditions examined.

The present study was designed to further investigate the properties of the AMT procedure and to compare it with a conventional mastery test with optimal information characteristics. This comparison is of interest for practical, as well as theoretical, reasons. If it were found that a conventional test with certain design characteristics could make mastery classifications as well as or better than the AMT procedure, it would probably be more economical to employ the conventional paper-and-pencil testing procedure in most classroom situations (although the rapid proliferation of inexpensive computers is quickly reducing the economic advantage of paper-and-pencil testing).

This study was designed to address three basic questions concerning the performance of these testing procedures within the context of a live-testing situation, using currently available items for which IRT parameter values had previously been estimated. The first question addressed was whether or not the testing procedure chosen made a difference in terms of the set of test items given to the students. Obviously, if the AMT procedure were to select the same items as the conventional test for most of the students, the AMT item selection procedure would be an unnecessary addition to the testing situation in the classroom. To address this question, the overlap in tests generated by the two procedures was examined as a function of achievement level. In addition, the theoretical information available from the questions administered by the two testing strategies was examined as a function of achievement level.

The second question addressed in this study concerned the criterion-related validity of the mastery classifications made by the two testing procedures. To the extent that one testing procedure results in mastery classifications more adequately reflecting some real criterion of performance, that procedure could be designated as a more valid testing paradigm.

The final question concerned the effect of the variable termination criterion for the AMT procedure. This termination criterion is based on the use of Bayesian confidence intervals with certain characteristics and should result in shorter overall test lengths. It is of some practical interest to determine how much test length would be reduced by the use of the AMT termination procedure in

a live-testing situation. In addition, it might be expected that the variable termination criterion would affect the validity of decisions made by the AMT procedure. The strength of this expected effect was also examined.

Method

Subjects

Data were obtained from students enrolled in an introductory biology course at the University of Minnesota during fall quarter 1979. Volunteers were recruited to take experimental computerized tests, covering the same material as would be covered in course examinations, prior to their classroom midquarter and final exams. Administration of the computerized tests began three weeks prior to the actual classroom exams. Students received one point, which was added to their final course grade, for taking one computerized test and an additional two points for participating in both the midquarter and final computerized testing sessions. Students were assigned sequentially to either an adaptive or a conventional testing condition. From the testing session prior to the midquarter, conventional test data were obtained from 237 students and adaptive test data from 237 students. From the testing sessions prior to the final exam, conventional test data were obtained from 226 students and adaptive test data were obtained from 226 students.

In addition to the computerized test data collected from these students, classroom exam and laboratory scores were also available for most of these students. These classroom scores were used in the analysis of the criterion-related validity of the various testing procedures. For this analysis of criterion-related validity, both classroom data and computerized testing data were available for 214 students in the conventional testing condition during the first testing session (prior to the midquarter exam), 213 students in the adaptive testing condition during the first testing session, 209 students in the conventional testing condition during the second testing session (prior to the final exam), and 219 students in the adaptive testing condition during the second testing session.

Test Administration

After assignment to either the conventional or the adaptive testing condition, the student was administered two or three subtests, which were administered by a cathode-ray terminal linked to a minicomputer system. During the first testing session, students took three 20-item subtests designed to evaluate their knowledge of the Chemistry, Cell Structure, and Energy content areas, which were taught in the biology class prior to the midquarter exam. During the second testing session, students were administered two 20-item subtests that were designed to evaluate their knowledge of the Genetics and Reproduction-/Embryology content areas, which were taught in the biology class following the midquarter exam and prior to the final exam.

Each of the questions administered to the students during these experimental testing sessions was in four-alternative multiple-choice format. The pools of test questions had been gathered from the questions that had been used in classroom examinations previously and therefore were representative of the content being taught in the classroom.

Item Pools

The five item pools developed to measure student achievement in the five content areas of interest were composed of examination questions that were administered in the general Biology course during the 1975-1976 and 1976-1977 academic years. The items were parameterized within their respective content areas using the procedure described by Urry (1976). This procedure estimates the discrimination power (a), difficulty (b), and guessing level (c) parameters required for the use of the three-parameter logistic IRT model (Lord, 1980; Lord & Novick, 1968). This calibration procedure is described in detail by Bejar, Weiss, and Kingsbury (1977). The sample sizes used for parameter estimation varied from approximately 800 to 1,200 students. Final item pool sizes ranged from 51 items, for the Reproduction/Embryology content area, to 87 items, for the Energy content area. Item identification numbers and IRT item parameter estimates for the items used in each content area item pool are shown in Appendix Tables A through E.

Classroom Mastery

The validity criterion for evaluation of the mastery classifications made by the two testing strategies was a student's course grade, as determined by the sum of a student's midquarter classroom exam score, final classroom exam score, and a laboratory grade. The maximum score obtainable was 100 points on each for a possible total of 300 points.

For each student the total score was evaluated to determine his/her mastery status on the classroom criterion. A student was declared a master on the classroom mastery criterion if he/she received at least 240 out of the possible 300 points. This criterion corresponds to the 80% cutoff between grades of C and B for classroom performance. By the comparison of the students against this classroom mastery level, an independent evaluation of students' mastery status was obtained that was used to examine the criterion-related validity of each of the experimental testing strategies.

Test Construction

Mastery level. In order to examine how well each testing strategy made mastery classifications, it was necessary to establish a reasonable level of performance that would be comparable to the classroom mastery level. It would then be necessary to construct the various experimental subtests so that they would be maximally efficient for making classifications at the specified mastery level.

For a conventional test using proportion-correct scoring, this 80% correct mastery level (as used in the classroom) would be sufficient for use in making mastery classifications. When an IRT scoring procedure is to be used, the mastery level must be converted from the proportion-correct metric to the latent achievement metric for each content area. Consequently, for each of the five content areas, the 80% criterion was converted to the achievement (θ) metric by use of the test characteristic curve (TCC) for the content area item pools, as described by Kingsbury and Weiss (1979). The θ value on the achievement metric that would most likely correspond to the 80% correct mastery level for each content area is shown in Table 1, along with the subject matter designation of each content area.

Table 1
Subject Matter Included in Each Content Area,
and the Achievement Level Used
as the Mastery Level for Each Content Area

Content Area	Subject Matter	Mastery Level on the θ Metric
1	Chemistry	.27
2	Cell Biology	.23
3	Energy	.79
4	Genetics	.73
5	Reproduction/Embryology	.65

Adaptive tests. The adaptive subtests administered to the students assigned to the adaptive testing condition followed the AMT paradigm described by Kingsbury and Weiss (1979) with one exception. As in the earlier study, a student's achievement level was estimated following his/her response to each test question using Owen's Bayesian scoring algorithm (Owen, 1969). The student's achievement level estimate was then used to select the next item to be administered. Each item remaining in the content area item pool was evaluated in terms of its theoretical information (Birnbaum, 1968), and the item that was capable of providing the most information at the student's current achievement level estimate was chosen to be administered next. In the original AMT paradigm, items were administered to a student until a decision concerning the student's mastery level could be made with a certain degree of confidence, and then the test was terminated. In this study a fixed subtest length of 20 items was used for each content area subtest. Analyses were designed, in part, to test the desirability of the use of the variable-termination rule versus fixed termination in this live-testing application of AMT. This procedure also permitted comparison of adaptive and conventional tests of the same test length.

Each student began each of the content area subtests with a Bayesian prior distribution for his/her achievement level, which had a variance of 1.0 and a mean that was equal to the mastery level for the content area in question. This was equivalent to making the assumption that it was equally probable that a student was a master or a nonmaster.

Conventional tests. For a one-point classification problem like the one involved here, the optimal conventional test is made up of that set of k items that provides the most information in the vicinity of the achievement level chosen as the cutting score θ_m, where item information is defined as in Birnbaum (1968, Equation 20.4.16) and evaluated at $\theta = \theta_m$ (Lord, 1980). To operationalize this design, each item for a particular content area was evaluated in terms of its theoretical information at the mastery level for the content area. The 20 most informative items at the mastery level were chosen to serve as the conventional test questions for that content area. The order of administration of the items to students within each content area was arbitrary, although each student in the conventional testing condition received the questions in the same order. The parameter estimates for the items that made up the conventional tests for each content area are designated in Appendix Tables A through E.

Scoring

Two scores were obtained for each adaptive subtest: the achievement level estimate ($\hat{\theta}$) following administration of the 20th item and the achievement level estimate at the item at which a 95% Bayesian confidence interval surrounding that estimate did not include the mastery cutoff on the achievement continuum. (For a more detailed description, see Kingsbury & Weiss, 1979 pp. 6-8.) For the conventional subtests two scores were computed: the proportion of the subtest items answered correctly and the Bayesian estimate of achievement level obtained using program Lindsco (Bejar & Weiss, 1979) for each subtest.

For both the adaptive and conventional tests, a mastery classification for each student was made for each subtest. If a student's achievement level estimate was greater than or equal to the appropriate mastery level, he/she was declared a master; if the student's achievement level estimate was less than the mastery level, he/she was declared a nonmaster.

Analyses

Comparison of the Tests Given

To determine whether the two testing strategies resulted in the administration of significantly different tests, the percentage of items administered within the 20-item AMT that also appeared in the conventional test was calculated for each person who took the adaptive tests. This was done separately for each of the five content area subtests. The percentage of overlap between the two types of tests was then plotted as a function of the estimated achievement level. These plots were smoothed by dividing the achievement level continuum into 20 approximately equal intervals and by plotting the mean percentage of overlap observed for all individuals whose achievement level estimate fell into each interval.

To determine the effect of the variable termination criterion on the performance of the AMT procedure, frequency distributions were compiled within each content area, showing the number of students for whom the AMT procedure would have reached its termination point as a function of the number of items administered. The percentage of students for whom the AMT procedure reached a confident mastery classification at or before the completion of the 20-item adaptive test within each content area was also determined.

To further compare the tests given by the conventional and adaptive testing strategies, information functions were calculated for each of the testing strategies within each content area. For each of the conventional tests, the function calculated was simply the theoretical test information function (Birnbaum, 1968) within each subtest, which is the sum of the item information functions for the 20-item tests. For the adaptive tests, the information functions were approximated by calculating for each person the sum of the item information functions for the items administered, evaluated at the final achievement level estimate. These information values were then plotted using the smoothing procedure described above. Adaptive test information functions were calculated for the fixed 20-item test length and for the variable-termination condition.

Comparison of Test Validities

As a preliminary test of the validity of each of the four classification procedures--mastery status estimated (1) from the conventional test using the proportion-correct score, (2) from the conventional test using the Bayesian score, (3) from the AMT procedure with the variable-termination criterion, and (4) from the AMT procedure with the fixed, 20-item, test length--Pearson product-moment (phi) correlations were calculated between the mastery status estimated by the classification procedure and the mastery status observed on the classroom performance criterion measure (0 = nonmaster, 1 = master). This was done for each classification procedure, for each content area. In addition, the frequencies of false mastery classifications and false nonmastery classifications were calculated for each classification procedure within each content area.

To further examine the validities of the mastery estimation strategies, discriminant function analysis (Tatsuoka, 1971) was used to combine the separate content area mastery classifications to more accurately predict the global classroom mastery status criterion. First, groups of 100 students were drawn from each testing condition within each testing session. A discriminant function was calculated for each of these development groups.

For the first testing session, a student's mastery status estimates from each of the three content area subtests taken were used as predictors in a discriminant function to estimate the student's classroom mastery status. For the second testing session, the two content area subtest mastery levels were used to estimate the student's overall classroom mastery status. A different prediction equation was developed for each different classification procedure. These functions were then applied to the remainder of the appropriate testing groups in order to cross-validate the discriminant functions. Frequencies and types of classification errors made by the discriminant functions for each of the testing procedures within each testing session were determined for both the development and validation groups.

As a final validity comparison, a discriminant function analysis was conducted on the subgroups of students who took the same type of test (adaptive or conventional) during both testing sessions. This analysis used the mastery classifications made in all five content areas to predict the classroom mastery level. A one-group discriminant analysis was used here because sample sizes were too small to allow for a development group and a cross-validation group. Again, frequencies and types of classification errors made were examined for each testing procedure.

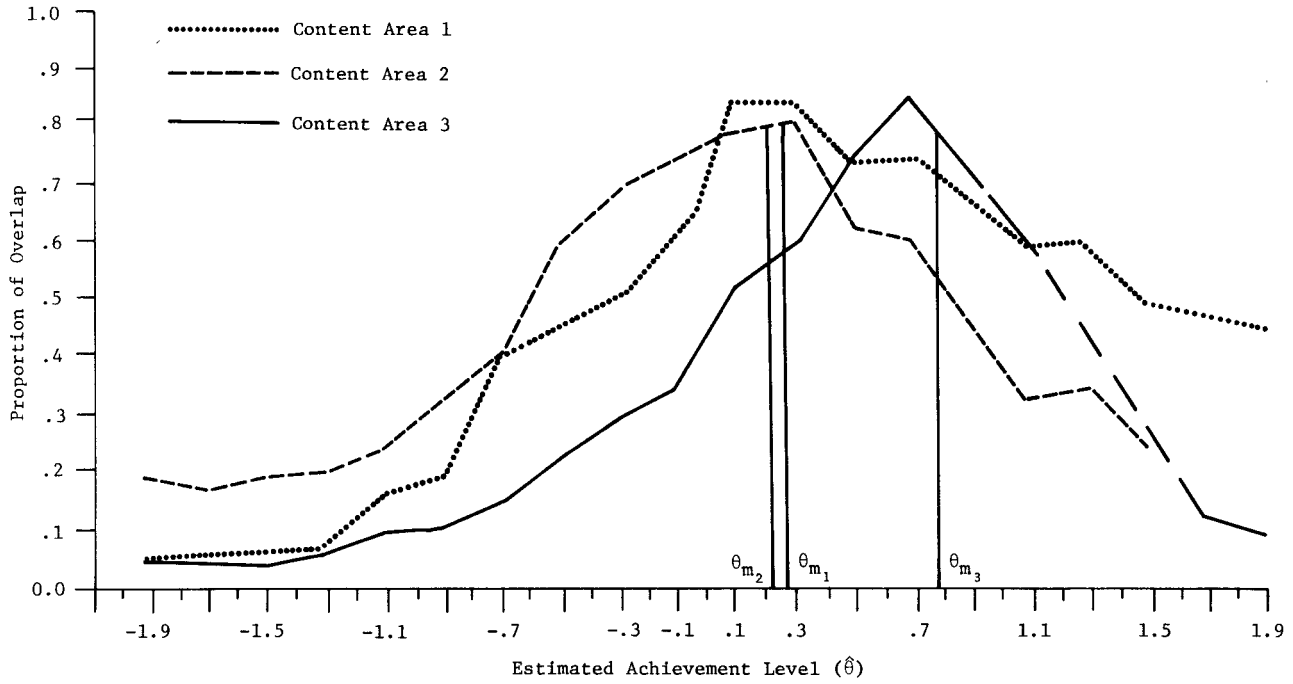
Results

Comparison of the Tests Given

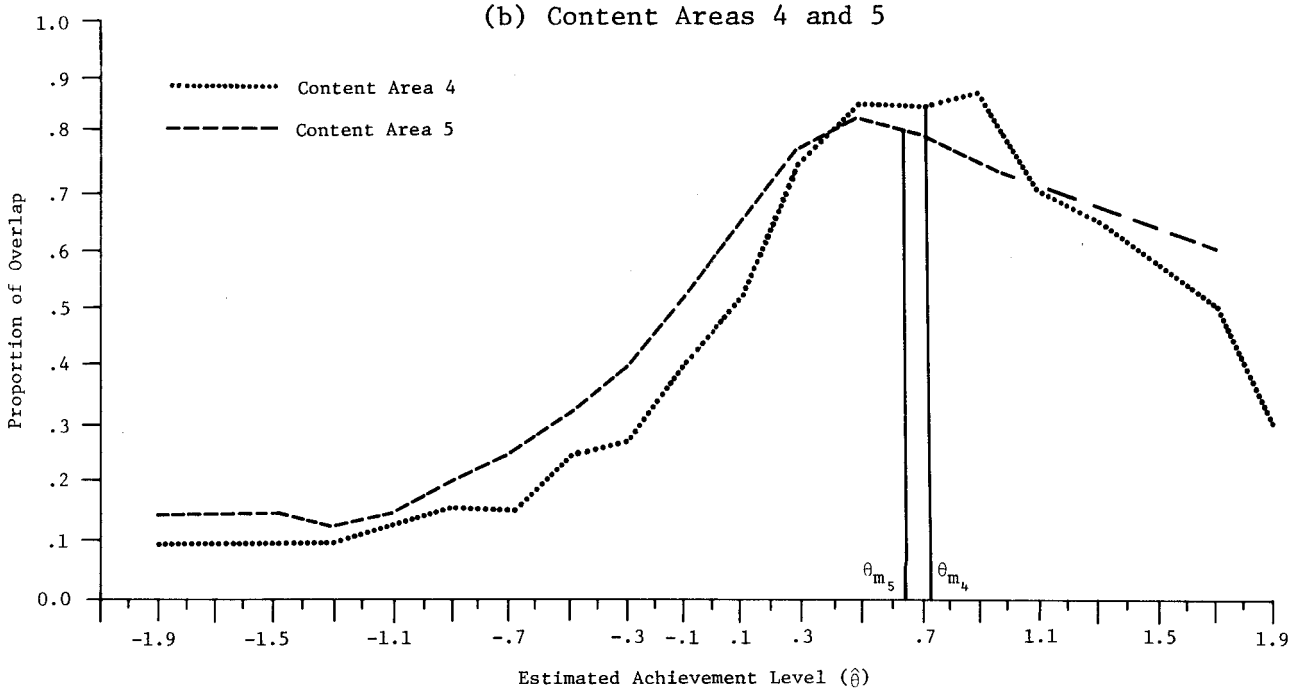
Test overlap. Figure 1 shows the percentage of items administered to students taking adaptive tests that also appeared on the corresponding conventional tests for the first and second testing sessions (Figures 1a and 1b, respectively). The percentage of overlap is shown as a function of achievement level, as estimated by the adaptive testing procedure after 20 items were administered. For each content area subtest the achievement level used as the mastery cutoff level is indicated.

Figure 1
Proportion of Items from the Conventional Test That Were Administered to Students Taking the Adaptive Test as a Function of Achievement Level, for Each Content Area (Mastery Levels are Indicated as θ_{m_1} to θ_{m_5}).

(a) Content Areas 1, 2, and 3



(b) Content Areas 4 and 5



As Figure 1 shows, for each content area the relationship between the percentage of overlap and the achievement continuum is a unimodal function, peaked at moderate achievement levels and much lower at more extreme achievement levels. Across all content areas the highest proportion of overlap was observed for Content Area 4, and was .88 at an achievement level of approximately $\hat{\theta} = .9$ (Figure 1b). The lowest peak overlap observed for any content area was .80, for Content Area 2 at an achievement level of approximately $\hat{\theta} = .3$ (Figure 1a). For these levels of maximum overlap, then, the 20-item adaptive subtests administered an average of 16 to 18 items that appeared on the conventional subtests.

The lowest level of overlap observed was .05, for Content Area 3 at achievement levels of approximately $\hat{\theta} = -1.9$ to -1.5 and for Content Area 1 at an achievement level of approximately $\hat{\theta} = -1.9$. For these very low achievement levels, the average overlap between the 20-item adaptive and conventional subtests was about one item.

Figures 1a and 1b show that the maximum overlap within each content area was observed at an achievement level that was quite close to the mastery level for the content area. In Content Areas 1 through 3, the mastery level was within the range on the achievement continuum that, upon application of the smoothing procedure, was equivalent to the achievement level having the highest level of overlap between the adaptive and conventional subtests. Content Area 4, for which the mastery level and the peak of overlap were observably different, had a mastery level of .73 and an observed overlap peak that occurred at an achievement level of approximately .9. For Content Area 5 the mastery level was .65, whereas the highest observed proportion of overlap occurred at an achievement level of approximately .5. In each of these two content areas, the observed difference between the mastery level and the approximate achievement level at which the highest amount of overlap occurred between the conventional and adaptive tests was less than .2 units on the achievement continuum (about 1/20th of the effective score range for this group of students).

Thus, these data show that for those achievement level estimates in the immediate neighborhood of the mastery level for any particular content area, the adaptive procedure resulted in tests that, on the average, were quite similar to the conventional tests (differing by only a very few items). At the other extreme, for achievement level estimates quite discrepant from the mastery level, the adaptive testing procedure resulted in tests that, on the average, were very different from the conventional tests (having only a very few items in common).

Effect of variable termination. A "high-confidence" classification is made when the Bayes confidence interval around an individual's estimated achievement level fails to include the prespecified mastery cutoff. Table 2 shows the mean test length needed to make a high-confidence classification and the percentage of students for whom a high-confidence classification was made at or before the end of the 20-item adaptive subtest within each content area. It can be seen from these data that the mean number of items required to make a confident classification ranged from 2.30, in Content Area 5, to 5.23 in Content Area 2. These means imply a corresponding reduction in the length of the average test of from 73.8% to 88.5% of the original 20-item test length.

In addition, Table 2 indicates that the percentage of students for whom the

Table 2
Summary Statistics for Number of Items Administered and Percentage of Students for Whom a High-Confidence Classification Was Made by the AMT Procedure Using a Variable Test Length

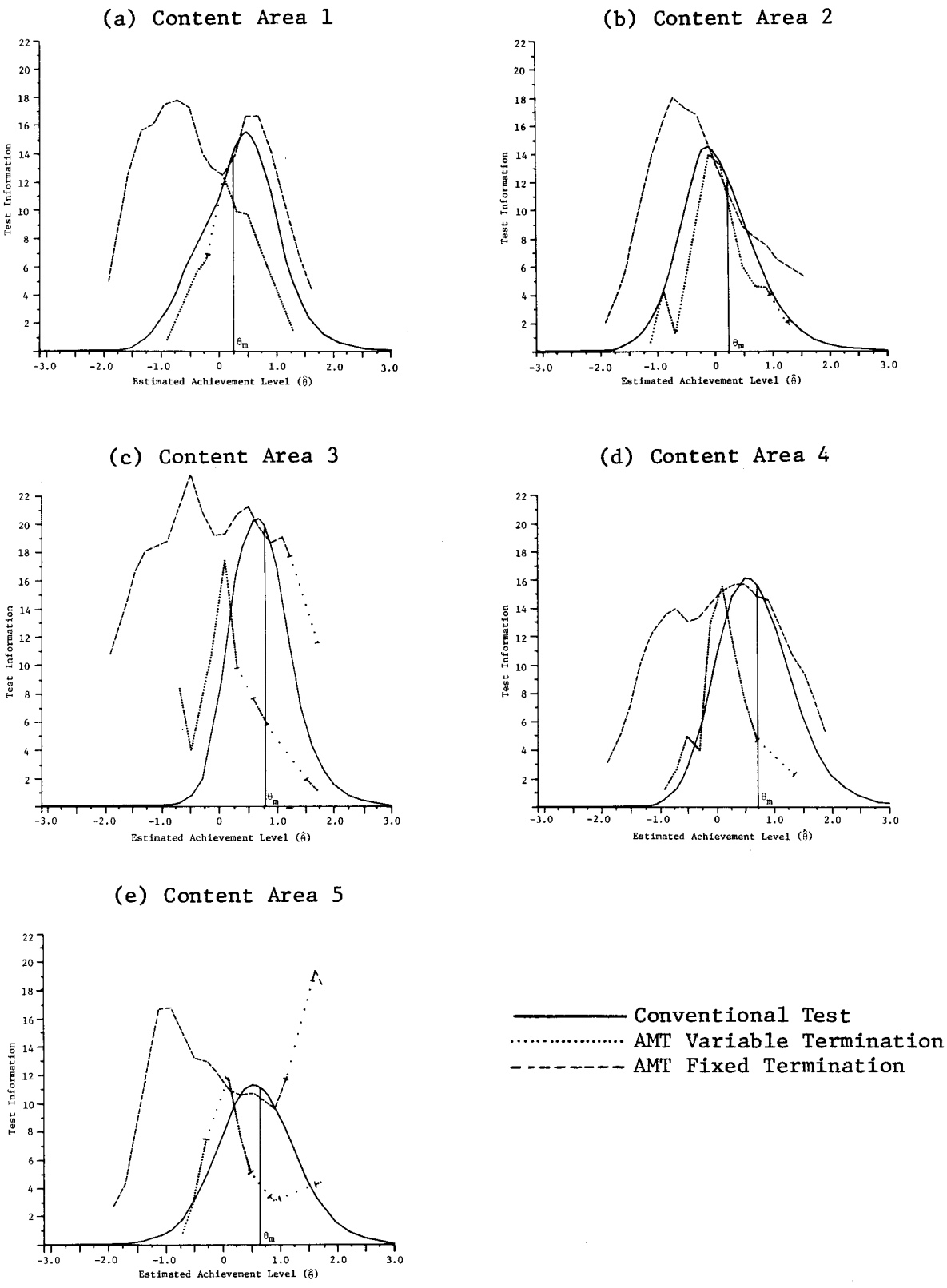
Content Area	Number of Students	Number of Items				Percentage of High-Confidence Classifications
		Mean	Min.	Max.	Standard Deviation	
1	236	5.15	2	20	4.18	98.3
2	236	5.23	1	20	4.47	98.7
3	236	3.57	2	20	1.98	99.6
4	224	3.85	2	20	2.33	99.6
5	224	2.30	1	20	2.54	99.6

AMT procedure was able to make a confident classification in 20 items or less ranged from 98.3% for Content Area 1, to 99.6% for Content Areas 3, 4, and 5. These results indicate that less than 2% of the students needed a test of more than 20 items for the adaptive procedure to make a confident classification in any content area. Appendix Table F shows the percentage of students for whom the AMT procedure reached its termination criterion for each test length within each content area. In each content area the same general pattern of results was observed. The great majority (more than 70%) of the students reached the termination criterion with the administration of 1 to 5 items. The remaining students were fairly evenly divided among the longer test lengths of from 6 to 20 items.

Information. Figure 2 shows, for each of the five content areas, the information functions that were observed for the conventional test, the adaptive test with a fixed length (20 items), and the adaptive test with a variable-length termination condition. Numerical values from which these figures were obtained are shown in Appendix Tables G, H, and I. Mean information for the adaptive tests was plotted as a function of the final achievement level estimate obtained using that strategy. The values on the abscissa represent achievement level estimates grouped in intervals with a range of $\pm.1$ around the plotted achievement level. For the conventional tests, theoretical test information functions are plotted. (Dotted lines in these figures indicate regions of the θ continuum for which no data values were available for that strategy)

In each content area the adaptive test with 20 items resulted in more achievement level estimates with higher levels of information than either of the other two strategies, except near the cutoff level between mastery and nonmastery, at which the conventional test provided slightly more information. For each subtest the conventional test provided maximum information very close to that subtest's mastery cutoff score. This was as expected, since the conventional tests were developed by selecting those 20 items that provide the most information at the mastery cutoff, thereby concentrating the test's efficiency near one point. Except for being slightly less efficient than the conventional test at the mastery cutoff, the adaptive strategy with a 20-item termination provided more precise estimates than the conventional strategy, particularly at the lower end of the achievement continuum.

Figure 2
Test Information for Conventional Test
and Fixed- and Variable-Length Adaptive Mastery Tests
as a Function of Estimated Achievement Level



For Subtests 1 and 2 the conventional test provided higher mean information values than the variable-termination adaptive strategy at all points along the achievement continuum. For Subtests 3, 4, and 5, the conventional test and variable-termination adaptive testing strategy fluctuated as to which provided more information. Generally, the variable-termination adaptive strategy provided more information than the conventional test at the lower portion of the achievement continuum, while the conventional test provided more information at higher achievement levels. It was shown above, though, that the variable-length adaptive testing procedure resulted in tests that were much shorter (2 to 5 items, on the average) than the conventional test (20 items). The higher information levels obtained from the conventional test are, at least partly, a function of the difference in test lengths. The variable-termination adaptive testing strategy provided less information at each achievement level than its 20-item counterpart because it usually consisted of far fewer items.

It should be noted that the information curves for the adaptive subtests were computed by determining the mean information for students whose achievement level estimates fell within certain ranges of the achievement continuum. The conventional test information functions are theoretical and are evaluated at each point along the achievement continuum. Thus, some differences noted between the adaptive tests and the conventional tests may be a function of both the curve-smoothing procedure used with the adaptive tests, and the differences between use of estimated versus "true" achievement levels.

Comparison of Test Validities

Subtest validities. Table 3 shows the phi correlations between each individual's mastery status (master = 1; nonmaster = 0) as estimated from the experimental subtests given in each content area and as observed in classroom performance. These correlations were calculated for each of the four testing strategies. All of the coefficients observed were significantly different from zero ($p < .05$) except for the correlations for the Content Area 5 conventional test scored using the Bayesian scoring system ($p = .066$) and the Content Area 5 variable length adaptive test (also $p = .066$). Coefficients ranged from .102 for the variable length adaptive test for Content Area 5 to .393 for the fixed length adaptive test for Content Area 1, indicating low to moderate validity for each of the content area subtests as predictors of the global classroom mastery criterion. For Content Areas 2 and 4 the AMT procedure with variable termination produced the highest correlations of the four testing methods between estimated and criterion mastery status ($r = .324$ and .391, respectively), although the variable-termination procedure administered only about one-quarter as many items as the other procedures. For Content Areas 1 and 3 the AMT procedure with fixed test length resulted in the highest correlations ($r = .393$ and .388, respectively). For Content Area 5 the conventional test with proportion-correct scoring resulted in the highest correlation ($r = .167$). None of the correlation coefficients within any one content area differed significantly from one another.

Table 4 shows the percentage of total correct and incorrect mastery classifications and the percentage of correct and incorrect mastery and nonmastery classifications made by each testing strategy within each content area. Table 4 shows that for all content areas the vast majority of classifications made by each testing strategy were nonmastery classifications. Performance of the stu-

Table 3
Phi Correlation (r) Between the Criterion Mastery Status and Estimated Mastery Status, and Number of Subjects (N) Within Each Content Area for Each Testing Strategy

Content Area	Testing Strategy and Score											
	Conventional						AMT					
	Proportion Correct			Bayesian			Variable Termination			Fixed Length		
	N	r	p*	N	r	p*	N	r	p*	N	r	p*
1	214	.313	<.001	214	.332	<.001	213	.371	<.001	213	.393	<.001
2	214	.313	<.001	214	.211	.001	213	.324	<.001	213	.281	<.001
3	214	.218	.001	214	.118	.043	213	.208	.001	213	.238	<.001
4	209	.226	<.001	209	.239	<.001	219	.391	<.001	219	.388	<.001
5	209	.167	.008	209	.105	.066	219	.102	.066	219	.156	.010

*Probability of rejecting null hypothesis of zero correlation.

dents on the classroom mastery criterion resulted in 49.0% of the students in the total testing sample attaining mastery status, leaving 51.0% of the students with nonmastery status. For the experimental subtests, though, the percentage of students estimated to have achieved mastery status (averaged across the five content areas, weighted by sample size) was 9.0% for the proportion-correct scoring of the conventional subtests, 6.1% for the Bayesian scoring of the conventional subtests, 11.8% for the adaptive subtests with variable test length, and 9.5% for the adaptive subtests with fixed test length. These low percentages, however, may be due to an artifact of the methodology used for this study. As noted above, students were given their experimental tests in the weeks immediately before their classroom exams. It is quite reasonable to assume that the students had not yet studied for their exam when these tests were given and therefore were functioning at a lower achievement level then they finally demonstrated in their classroom performance. Although this may affect the absolute performance levels of the students, it should have no effect on the relative performance of the testing strategies.

The lowest total error rate (total incorrect classifications) observed in Table 4 is 32.9%, for the AMT procedure with a fixed test length for Content Area 1. The highest total error rate observed was 51.4%, for the conventional procedure with Bayesian scoring for Content Area 3. Across content areas, the conventional test with proportion-correct scoring made 457 incorrect classifications out of 1,060 total classifications (43.1% incorrect classifications). The conventional testing procedure with Bayesian scoring resulted in 482 incorrect classifications out of 1,060 total classifications (45.5% incorrect classifications). The AMT procedure with variable termination resulted in 416 incorrect classifications out of 1,077 total classifications (38.6% incorrect classifications). Finally, the AMT procedure with fixed test length made 421 incorrect classifications out of 1,077 total classifications (39.1% incorrect classifications). Since the five content areas differed in terms of difficulty, content, and contribution to the classroom mastery criterion, no single content area subtest was expected to adequately predict the global classroom performance criterion. Consequently, the error rates observed for the various testing strategies

Table 4
Percentage of Correct and Incorrect Mastery and Nonmastery Classifications Made by Each Testing Strategy Within Each Content Area

Content Area and Classification	Testing Strategy and Score			
	Conventional		AMT	
	Proportion Correct	Bayesian	Variable Termination	Fixed Length
Content Area 1				
Correct Non-Mastery	45.3	45.8	50.7	51.6
Incorrect Non-Mastery	41.1	41.1	30.5	31.0
Correct Mastery	12.6	12.6	16.0	15.5
Incorrect Mastery	.9	.5	2.8	1.9
Total Correct	57.9	58.4	66.7	67.1
Total Incorrect	42.0	41.6	33.3	32.9
Content Area 2				
Correct Non-Mastery	42.1	44.4	52.1	52.6
Incorrect Non-Mastery	34.6	44.4	35.2	38.0
Correct Mastery	19.2	9.3	11.3	8.5
Incorrect Mastery	4.2	1.9	1.4	.9
Total Correct	61.3	53.7	63.4	61.1
Total Incorrect	38.8	46.3	36.6	38.9
Content Area 3				
Correct Non-Mastery	45.8	45.8	53.1	53.5
Incorrect Non-Mastery	47.2	50.9	41.8	41.8
Correct Mastery	6.5	2.8	4.7	4.7
Incorrect Mastery	.5	.5	.5	0
Total Correct	52.3	48.6	57.8	58.2
Total Incorrect	47.7	51.4	42.3	41.8
Content Area 4				
Correct Non-Mastery	53.1	53.1	48.9	50.2
Incorrect Non-Mastery	42.6	42.1	31.5	34.2
Correct Mastery	4.3	4.8	17.4	14.6
Incorrect Mastery	0	0	2.3	.9
Total Correct	57.4	57.9	66.3	64.8
Total Incorrect	42.6	42.1	33.8	35.1
Content Area 5				
Correct Non-Mastery	53.1	53.1	50.2	51.1
Incorrect Non-Mastery	44.5	45.9	46.1	46.6
Correct Mastery	2.4	1.0	2.7	2.3
Incorrect Mastery	0	0	.9	0
Total Correct	55.5	54.1	52.9	53.4
Total Incorrect	44.5	45.9	47.0	46.6

within content areas were rather high, as expected.

Discriminant function analysis within testing sessions. Table 5 shows the percentages of incorrect classifications made using discriminant function analysis within the development samples from each testing session and for each testing strategy. For Testing Session 1, subtest classifications from Content Areas 1, 2, and 3 were used in the discriminant function to predict each student's

Table 5

Percentage of Incorrect Mastery Classifications Made by the Discriminant Function of Content Area Mastery Classifications from Each Testing Strategy During Each Testing Session, for the Development Group (N=100)

Testing Strategy and Score	Testing Session 1			Testing Session 2		
	Incorrect Non-Mastery	Incorrect Mastery	Total Incorrect	Incorrect Non-Mastery	Incorrect Mastery	Total Incorrect
Conventional						
Proportion Correct	28	3	31	40	0	40
Bayesian Score	35	1	36	41	0	41
AMT						
Variable Termination	24	2	26	29	4	33
Fixed Length	27	1	28	30	2	32

Table 6

Percentage of Incorrect Mastery Classifications Made by the Discriminant Function of Content Area Mastery Classifications from Each Testing Strategy, During Each Testing Session, for the Cross-Validation Group, and the Number of Students (N) in Each Group

Testing Strategy and Score	Testing Session 1				Testing Session 2			
	N	Incorrect Non-Mastery	Incorrect Mastery	Total Incorrect	N	Incorrect Non-Mastery	Incorrect Mastery	Total Incorrect
Conventional								
Proportion Correct	114	28.1	7.9	36.0	109	42.2	0.0	42.2
Bayesian Score	114	36.0	4.4	40.4	109	43.1	0.0	43.1
AMT								
Variable Termination	113	29.2	5.3	34.5	119	33.6	.8	34.4
Fixed Length	113	31.0	4.4	35.4	119	36.1	0.0	36.1

classroom mastery status. For Testing Session 2, subtest classifications from Content Areas 4 and 5 were used as predictors. The coefficients used for each discriminant function are shown in Appendix Table J. From Table 5, it may be seen that the total error percentages for Testing Session 1 ranged from 26% for the AMT procedure with variable termination to 36% for the conventional test with Bayesian scoring. For Testing Session 2 the total error percentages ranged from 32% for the AMT procedure with fixed test length, to 41% for the conventional test with Bayesian scoring. For both testing sessions, the two AMT procedures each resulted in lower error rates than either of the conventional test strategies. The two AMT procedures resulted in very similar total error percentages across both testing sessions.

Table 6 shows the error percentages that resulted when the discriminant functions were applied to classify the remainder of the testing sample for each testing strategy, for both testing sessions. The AMT procedure with variable termination resulted in the lowest total error rates (34.5% in Session 1 and 34.4% in Session 2). The AMT procedure with fixed test length resulted in the second lowest total error rates (35.4% in Session 1 and 34.4% in Session 2). The conventional test with proportion-correct scoring gave the third lowest error rates (36.0% in Session 1 and 42.2% in Session 2). The highest total error rates noted for the cross-validation group were observed for the use of the conventional test with Bayesian scoring (40.4% in Session 1 and 43.1% in Session 2). As in the development group, the two AMT procedures differed very little in terms of total error rates for the cross-validation groups. The differences in total error percentages for the two AMT procedures were .9 percentage points for Session 1 and 1.7 percentage points for Session 2.

Discriminant functions across testing sessions. Table 7 shows the percentage of incorrect decisions made by the discriminant functions developed for each testing strategy from the mastery classifications made in each of the five content areas, for individuals who were administered the same type of test during both testing sessions. The discriminant function coefficients used to make these mastery classifications are also shown in Appendix Table J. Table 7 shows that the percentage of incorrect nonmastery classifications (false nonmastery) was much higher than the percentages of incorrect mastery classifications. This trend was earlier observed for the other discriminant function analyses. In

Table 7
Percentage of Incorrect Mastery Classifications Made by the Discriminant Function of Content Area Mastery Classifications from Each Testing Strategy, for Students Who Took the Same Type of Test during Both Testing Sessions (N=89)

Testing Strategy and Score	Incorrect Nonmastery	Incorrect Mastery	Total Incorrect
Conventional			
Proportion Correct	25.8	5.6	31.4
Bayesian Score	37.1	0	37.1
AMT			
Variable Termination	23.6	3.4	27.0
Fixed Length	24.7	2.3	27.0

examining the total percentage of incorrect classifications made by the discriminant functions for each testing strategy, the trends noted in the earlier analyses are seen quite clearly. The lowest total percentage of incorrect classifications observed was 27.0%, for both of the AMT procedures. The conventional test strategy with proportion-correct scoring misclassified 1.16 times as many students as either of the AMT procedures (31.4% of the students), while the conventional test strategy with Bayesian scoring misclassified one-third more students than either AMT procedure (37.1% of the students).

Discussion and Conclusion

Two major conclusions result from this study:

1. In each of the discriminant analyses and in the majority of the individual subtest comparisons, the adaptive testing procedure, with either a fixed or variable test length, resulted in a consistently higher proportion of correct classifications concerning mastery status then did the conventional testing procedure with either scoring strategy when classroom performance was used as a criterion measure.

2. The variable test length condition used with the adaptive testing procedure resulted in test lengths that were, on the average, about 80% shorter then the fixed test length, but no consistent differences in criterion-related validity were found between the adaptive testing procedures that used fixed test length and variable test length.

Although these conclusions appear to contradict previous psychometric theory--that the single most useful type of test to use when making mastery classifications for a group of people is a test that concentrates its measurement precision within the immediate neighborhood of the mastery cutoff level (Birnbaum, 1968, pp. 450)--they really serve as an adjunct to previous findings. Birnbaum's demonstration of the superiority of the peaked test dealt with a single test administered to a group of students. The AMT strategy implemented in this study administers different tests to different individuals within a group of students, depending on the individuals' responses to the test questions. Thus, the use of the AMT strategy allows for an entire class of mastery tests to be used to make mastery classifications. One member of this class is the best peaked test that can be constructed from the item pool for each individual. In fact, in the analysis of test overlap, it was found that for students whose achievement level estimates were quite close to the mastery cutoff, the AMT procedure administered tests that had, on the average, 80% to 90% of the items that appeared on the conventional peaked test. However, when a student's achievement level differed from the mastery cutoff level, the AMT procedure tended to administer tests that had fewer items in common with the best peaked test.

This process of giving tests adapted to different individuals has the effect of increasing the variance of the observed achievement level estimates, thus making differences between students (or between a single student and the mastery criterion) more obvious. In this study, for each of the five content area subtests the adaptive testing procedure (with either the fixed or variable test lengths) resulted in greater score variance than did the conventional testing procedure when Bayesian scoring was used to equate scoring methods. The mean score variance observed for the conventional test with Bayesian scoring

(across content areas) was .237; while for the AMT procedure with variable test length, the mean score variance was .359; and for the AMT procedure with fixed test length, the mean score variance was .506. Thus, the AMT procedure, with or without the variable test length, spread out student achievement level estimates and allowed a more accurate assessment of student mastery status.

This study has thus demonstrated that the AMT procedure resulted in consistently more accurate estimation of students' mastery status within a course of instruction than did the best available conventional test peaked at the mastery level. Further, it was shown that the use of the AMT's variable termination capability did not significantly reduce the validity of the mastery level estimates obtained, while it reduced the mean test length by approximately 80%. It is interesting that these findings were noted even for proportion-correct scoring of the conventional test, which had its scoring method in common with the criterion measure. This common scoring method may explain the observation that the proportion-correct scoring of the conventional subtests resulted in slightly higher percentages of correct mastery classifications (using classroom performance as a criterion) than did Bayesian scoring of the same tests, in most of the subtest comparisons and in all of the discriminant analyses.

The variable termination AMT procedure has been shown here to be an efficient way of reducing test length while producing mastery classifications of comparable or higher quality than those made by conventional mastery tests constructed to maximize accuracy of mastery classifications. Given the proliferation of microcomputers in instructional setting, the AMT procedure should find application in many large-scale instructional settings in which conventional mastery testing is currently being used.

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, February 1979.

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.

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.

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.

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

Lord, F. M. Application of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

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

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

Tatsuoka, M. M. Multivariate analysis: Techniques for educational and psychological research. New York: Wiley, 1971.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (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)

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

Appendix: Supplementary Tables

Table A
Item Numbers and Estimates of Item Discrimination (a), Item Difficulty (b), and Lower Asymptote (c) for Each Item Used in the Adaptive Testing Pool for the Chemistry Content Area, and Items Comprising the Conventional Test

Item Number	a	b	c	Item Number	a	b	c
3000	1.76	.87	.37	3052	.95	.18	.49
3003	1.47	-1.66	.32	3053	1.08	1.32	.49
3005*	1.49	-.26	.26	3054	1.78	-.71	.34
3008	1.36	-1.45	.30	3055	2.36	-.60	.23
3009	2.21	-.82	.16	3056	1.30	1.12	.43
3010	1.05	.44	.51	3057	1.50	-1.10	.28
3011	1.60	-.68	.27	3058	.92	-.93	.30
3012	1.26	.66	.37	3060	1.20	-1.16	.26
3013	1.39	-1.12	.29	3061	.99	1.69	.35
3014	1.35	-.85	.24	3062*	2.22	.49	.35
3016	1.39	-1.41	.49	3064	1.12	.93	.30
3018	.80	1.02	.42	3065	1.66	-1.57	.36
3019*	1.55	.33	.32	3066	1.31	.63	.38
3020	1.61	-1.09	.27	3067*	1.46	-.29	.32
3022	.77	-.66	.15	3069	1.00	-.18	.44
3025	1.03	-1.67	.43	3070	1.07	-1.19	.23
3028	1.72	-1.02	.26	3072*	1.56	.64	.38
3031*	1.54	-.56	.30	3073	1.54	-1.26	.36
3032	1.29	-1.04	.35	3075	.97	-.55	.49
3033	2.38	2.66	.63	3078	1.85	-1.50	.29
3034*	1.34	.40	.38	3082	1.02	1.93	.51
3036*	1.42	-.57	.37	3083*	1.43	-.60	.30
3038	1.98	-.78	.28	3084*	1.73	-.55	.43
3041*	1.94	.36	.42	3085	1.70	-1.70	.42
3042*	1.52	0	.27	3086	1.04	-.46	.44
3044	1.13	-1.19	.23	3087	.93	-1.27	.23
3045	3.00	2.70	.65	3088	1.05	-.21	.48
3046*	1.42	.24	.30	3089*	1.67	.63	.39
3047*	2.11	.39	.31	3090	2.05	-1.56	.32
3048	1.76	.77	.40	3092*	1.50	-.05	.43
3049*	1.30	-.36	.38	3095	1.46	-1.03	.20
3050*	2.14	.64	.40	3096	2.34	-1.35	.27
3051*	2.21	.27	.35	3097*	1.86	.77	.33

*Item was administered on the conventional test.

Table B
Item Numbers and Estimates of Item Discrimination (a), Item Difficulty (b), and Lower Asymptote (c) for Each Item Used in the Adaptive Testing Pool for the Cell Content Area, and Items Comprising the Conventional Test

Item Number	a	b	c	Item Number	a	b	c
3201	1.46	-1.16	.31	3245	2.08	-.87	.34
3202	2.15	-.66	.44	3247	2.33	2.37	.74
3205	1.72	-1.29	.28	3248	.81	-.40	.33
3206	1.34	1.76	.47	3249	1.48	-1.06	.32
3208*	.80	-.19	.21	3250	.99	2.36	.41
3209	2.63	2.93	.72	3252	1.34	-1.55	.47
3210	1.74	-1.31	.28	3259	.95	.28	.42
3211*	1.48	.63	.43	3260	1.24	1.33	.51
3212	1.07	-.85	.45	3261	1.43	.59	.67
3214*	1.86	.24	.37	3262	.98	.55	.52
3216*	1.82	-.29	.32	3263	1.15	2.34	.60
3217*	1.45	-.20	.30	3264	.83	.61	.40
3218	1.29	.76	.36	3265	1.70	-1.38	.68
3219	2.06	.94	.44	3266	1.67	-1.00	.53
3221*	1.99	-.66	.27	3267	1.82	-.70	.40
3222	.99	-.58	.38	3268*	1.80	.38	.48
3223	1.44	-1.54	.59	3270	.94	-.21	.43
3224*	1.14	-.07	.42	3271	1.58	1.84	.57
3226*	1.50	-.22	.37	3272*	2.26	-.48	.54
3227	.97	-1.04	.37	3274*	1.63	-.21	.55
3228	1.27	2.78	.54	3276	1.11	.29	.51
3229	.80	.60	.38	3282*	2.15	-.13	.35
3230	1.47	2.46	.62	3284*	1.17	.12	.51
3232	1.62	.99	.71	3285*	1.37	-.22	.31
3234	2.53	3.01	.59	3286	.91	-1.19	.51
3235	1.97	-1.24	.26	3287	1.63	-.93	.38
3236*	1.35	.03	.46	3289	2.36	-1.21	.66
3237	2.75	-.76	.22	3290*	2.42	-.29	.37
3238*	1.39	-.76	.30	3291	.80	.21	.34
3240*	1.54	.39	.46	3292	2.16	1.60	.69
3241	1.84	1.68	.49	3293	1.58	-1.02	.41
3243	.84	-.56	.40	3294*	1.19	-.69	.30
3244*	1.79	-.28	.32				

*Item was administered on the conventional test.

Table C
Item Numbers and Estimates of Item Discrimination (a), Item Difficulty (b), and Lower Asymptote (c) for Each Item Used in the Adaptive Testing Pool for the Energy Content Area, and Items Comprising the Conventional Test

Item Number	a	b	c	Item Number	a	b	c
3401	.94	1.10	.43	3453*	2.04	.68	.44
3402	1.68	2.17	.55	3454	1.39	2.39	.51
3403*	2.77	.06	.29	3455	1.86	-.73	.37
3404	1.99	-.83	.59	3456	1.30	2.73	.48
3405*	1.30	.52	.35	3457	1.23	1.65	.34
3406	1.42	2.42	.48	3458	2.27	-1.09	.42
3407	1.39	2.22	.49	3459	1.33	-.49	.32
3408*	2.55	.74	.25	3460	2.56	1.39	.34
3409	2.46	2.91	.71	3461	1.18	1.06	.49
3410	2.01	1.41	.43	3462	2.09	-.69	.50
3412	1.46	-.82	.44	3463	2.93	-1.58	.50
3413*	2.22	.60	.52	3464	2.82	-.08	.32
3414	1.66	2.10	.50	3465	1.93	1.18	.62
3415	4.13	-2.27	.12	3466	2.43	-.12	.47
3416	1.49	1.24	.52	3467	1.77	-.44	.48
3418*	1.46	.68	.49	3468	1.43	.96	.58
3419	2.20	1.49	.42	3469	2.38	-.95	.62
3420	1.10	1.62	.47	3470	2.45	-.68	.38
3421	1.22	.07	.32	3471	1.35	-.17	.48
3423*	1.38	.79	.50	3472	2.19	-1.36	.64
3424	1.68	-.19	.59	3473	1.09	.02	.46
3425	4.03	-.49	.00	3474	2.90	2.01	.62
3426	2.28	-.05	.49	3475	1.30	.07	.50
3427	1.36	1.84	.39	3476*	1.96	.12	.49
3428	2.64	-1.44	.64	3477	2.18	-.80	.60
3429*	2.85	.92	.33	3478	1.63	2.01	.63
3431	1.34	.03	.39	3479	1.62	.05	.55
3432	2.36	-.46	.43	3480	1.11	-1.23	.58
3433	1.28	1.28	.37	3482	1.36	.12	.55
3434	.67	.62	.37	3483	3.89	1.14	.43
3435	2.07	-.49	.68	3484*	1.86	.21	.60
3436*	1.74	1.10	.38	3485	3.26	-1.19	.26
3437*	2.73	.45	.22	3486	1.79	1.42	.55
3438*	1.34	.16	.36	3487	2.29	1.64	.63
3439*	2.34	.28	.34	3488	1.93	-.08	.54
3440	1.78	1.93	.38	3489	2.65	-.99	.40
3441	1.31	.19	.61	3490	1.40	-1.39	.55
3443	2.89	-1.33	.78	3491	.93	.24	.51
3444*	1.47	.60	.40	3492*	2.62	.28	.50
3445*	2.04	.36	.40	3493	3.34	-1.85	.22
3447*	1.26	.97	.38	3494	3.27	-1.57	.19
3448*	1.69	.64	.32	3495*	2.07	.81	.53
3449	2.73	2.29	.48	3496	2.54	-1.26	.34
3452	.86	2.24	.39				

*Item was administered on the conventional test.

Table D
Item Numbers and Estimates of Item Discrimination (a), Item Difficulty (b), and Lower Asymptote (c) for Each Item Used in the Adaptive Testing Pool for the Genetics Content Area, and Items Comprising the Conventional Test

Item Number	a	b	c	Item Number	a	b	c
3601	1.08	1.30	.41	3666	.70	1.21	.27
3602	1.15	-1.29	.54	3668	1.16	-.74	.17
3603*	1.29	.41	.28	3669*	1.89	.22	.18
3606	.77	-.27	.13	3671	1.49	-.23	.22
3609	.89	.18	.43	3673	1.44	1.36	.33
3610	.98	-1.10	.17	3674*	1.66	.66	.28
3611*	1.34	.26	.29	3675*	1.30	.48	.33
3614	.66	.36	.35	3679	1.42	-.89	.27
3615*	1.74	1.12	.30	3680	1.59	-.85	.21
3616	.99	1.06	.41	3683	.94	-1.22	.18
3617	.99	-.99	.23	3684	.90	-.69	.21
3618	.98	.13	.41	3685	1.25	-.98	.18
3620	1.92	2.83	.66	3692	1.38	-.98	.33
3621	.98	-.66	.16	3693	1.46	-.18	.33
3622	1.14	2.60	.51	3695	1.23	-1.31	.32
3623*	1.54	.74	.32	3696	.83	-.51	.14
3625	1.15	2.11	.50	3698	2.27	2.45	.60
3627*	1.21	.32	.37	3699	.65	.52	.36
3628*	1.17	.46	.27	3700	1.10	1.03	.35
3630	.68	-.52	.38	3701	.95	-.74	.27
3631	1.73	-.86	.28	3703	1.08	-.70	.27
3632*	1.39	.16	.36	3704	1.59	-1.06	.30
3633	.99	-.98	.30	3707*	1.89	.48	.29
3635	.66	.72	.38	3708	1.57	-.20	.16
3636	1.17	-.49	.17	3709	1.29	.25	.36
3637	1.22	-.62	.18	3710	1.16	-.63	.20
3638	1.70	-1.42	.34	3711	1.31	-.82	.30
3640	1.42	-.67	.40	3712	.84	1.89	.37
3641	1.21	-.61	.23	3713	.74	-.91	.42
3642	1.06	1.17	.26	3715	1.37	-1.50	.34
3646	1.28	.89	.37	3716	1.29	1.27	.35
3648	1.89	-1.08	.32	3717	.90	1.25	.41
3649	1.14	-.03	.21	3718	1.03	.12	.31
3651	1.14	2.18	.53	3719*	1.10	.49	.24
3654*	1.83	.94	.26	3720*	1.48	.18	.26
3656	.67	-.40	.32	3721	1.53	-1.05	.29
3657	.87	-1.67	.38	3728	1.09	2.87	.52
3658*	1.31	.36	.40	3733	1.37	1.26	.39
3661*	1.68	.29	.25	3735	1.42	-1.03	.22
3662*	1.10	.64	.17	3745*	2.01	-.10	.17
3663	.72	-.10	.36	3746*	1.88	.32	.25
3665*	1.43	.87	.33	3751	.85	2.02	.41

*Item was administered on the conventional test.

Table E
Item Numbers and Estimates of Item Discrimination (a), Item Difficulty (b), and Lower Asymptote (c) for Each Item Used in the Adaptive Testing Pool for the Reproduction/Embryology Content Area, and Items Comprising the Conventional Tests

Item Number	a	b	c	Item Number	a	b	c
3804	1.90	1.71	.50	3902	.92	1.74	.40
3806*	2.28	.30	.34	3903	1.30	-.76	.30
3807	3.01	-1.04	.18	3904	2.64	2.68	.54
3812*	1.18	-.05	.36	3905*	2.07	.69	.43
3813	1.69	-.76	.40	3906*	1.08	-.53	.21
3814	1.64	-.47	.44	3907	2.40	-1.06	.68
3815*	1.47	.56	.44	3908*	1.69	.14	.39
3817*	1.12	-.07	.47	3909*	1.58	1.04	.48
3819*	1.47	.54	.49	3910	2.47	-1.47	.43
3820*	1.30	.52	.26	3912*	1.41	1.02	.41
3825	1.98	-1.17	.36	3913	2.41	-1.05	.25
3830	4.13	1.52	.11	3914*	1.79	-.07	.30
3832	1.75	-1.51	.38	3915	2.53	-.33	.24
3833	3.10	2.29	.40	3918*	1.41	.63	.44
3834	1.74	-1.28	.77	3919	2.41	-.49	.49
3835	1.40	2.03	.57	3920	2.05	-1.01	.53
3837	1.60	-.79	.59	3921	1.85	1.52	.53
3838	2.28	-1.36	.61	3922*	1.52	.38	.53
3841	1.20	2.23	.50	3923*	1.41	.61	.52
3847	1.36	-.27	.55	3924*	1.88	-.18	.54
3850	1.79	1.41	.58	3925*	1.68	.74	.46
3851*	1.02	.19	.33	3926	1.67	-1.08	.36
3852	.99	-1.59	.49	3927	1.71	-1.51	.40
3853*	1.30	.34	.37	3928	1.45	-.96	.34
3854*	1.36	-.47	.32	3929	3.43	1.36	.10
3901	2.34	2.59	.52				

*Item was administered on the conventional test.

Table F
Percentage of Students for Whom the Adaptive Testing Procedure Terminated at Each Test Length Within Each Content Area

Number of Items Administered	Content Area				
	1	2	3	4	5
1	0.0	24.2	0.0	0.0	62.5
2	32.6	0.0	31.4	37.5	0.0
3	19.9	23.7	25.0	12.1	21.4
4	13.6	14.0	26.3	23.7	5.4
5	.8	7.6	9.7	12.9	4.5
6	6.8	.8	3.0	4.0	1.8
7	6.4	5.5	.8	3.1	2.2
8	2.5	3.4	.4	2.7	.4
9	1.3	3.4	1.7	1.3	0.0
10	3.0	3.8	.4	.9	0.0
11	4.2	1.3	.4	.4	0.0
12	1.3	4.7	0.0	0.0	.4
13	.8	2.1	.4	.4	0.0
14	2.1	0.0	0.0	.4	0.0
15	1.3	1.3	0.0	0.0	0.0
16	.4	.4	0.0	0.0	.4
17	.4	1.3	0.0	0.0	0.0
18	.8	.8	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	.4
20	1.7	1.7	.4	.4	.4

Table G
Mean and Standard Deviation (SD) of Information Obtained From Variable-Length Adaptive Tests in Each Content Area as a Function of Estimated Achievement Level

Estimated Achievement Level	Content Area 1			2			3			4			5		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
-1.9	0			0			0			0			0		
-1.7	0			0			0			0			0		
-1.5	0			0			0			0			0		
-1.3	0			0			0			0			0		
-1.1	0			42	.79	.00	0			0			0		
-.9	77	.84	.00	8	4.25	.00	0			52	1.30	.00	0		
-.7	20	2.68	1.43	61	1.20	1.45	31	8.28	.00	23	2.61	.00	114	.86	.36
-.5	50	4.30	1.65	52	4.55	3.15	105	4.02	5.48	3	4.90	.78	57	3.32	1.34
-.3	32	6.07	1.53	23	9.19	4.75	19	7.49	4.99	74	4.04	2.67	19	7.27	1.58
-.1	0			6	14.03	.90	9	11.69	1.05	7	12.82	1.70	0		
.1	2	12.46	.40	3	13.28	1.13	3	17.64	3.01	3	15.45	.35	2	11.86	.43
.3	10	9.86	1.12	5	9.25	1.99	11	10.66	2.36	7	11.22	2.36	3	7.76	1.30
.5	6	9.76	.15	10	5.87	.60	0			2	7.54	.00	1	5.09	.00
.7	23	7.54	.41	7	4.56	.14	40	6.72	.00	10	4.71	.15	0		
.9	0			4	4.51	.00	0			0			6	3.20	.00
1.1	0			0			0			0			0		
1.3	17	1.65	.00	16	2.04	.00	0			0			0		
1.5	0			0			19	1.28	.00	45	1.35	.00	0		
1.7	0			0			0			0			24	4.38	.00
1.9	0			0			0			0			0		

Table H
Mean and Standard Deviation (SD) of Information Obtained From 20-Item Adaptive Tests in Each Content Area as a Function of Estimated Achievement Level

Estimated Achievement Level	Content Area 1			2			3			4			5		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
-1.9	2	5.11	.70	2	2.13	1.21	3	10.82	.81	1	3.14	.00	2	2.71	.05
-1.7	14	9.32	1.06	8	4.14	.60	9	13.41	.90	4	4.85	.70	11	4.34	.77
-1.5	31	13.13	.83	12	7.00	4.02	12	16.31	.67	11	7.19	.93	22	8.33	1.18
-1.3	21	15.79	.70	18	10.31	.94	26	18.00	.49	16	10.30	.82	45	12.64	1.35
-1.1	17	16.03	1.86	29	13.74	1.09	29	18.52	.77	22	12.28	.78	47	16.66	1.02
-.9	28	17.54	1.60	23	16.23	1.05	29	18.69	.94	22	13.35	1.08	23	16.72	1.28
-.7	28	17.85	1.54	35	17.91	1.20	36	21.47	1.26	25	13.95	1.09	26	14.96	.79
-.5	22	17.23	1.42	36	17.30	1.39	26	23.57	1.21	26	13.12	1.02	9	13.27	.97
-.3	11	14.12	1.53	22	16.83	1.05	18	20.91	1.85	27	13.36	.76	16	13.05	.87
-.1	14	12.93	1.30	19	14.75	1.77	4	19.09	.44	18	14.29	.58	6	12.17	.88
.1	6	12.11	.95	12	12.74	1.26	10	19.27	1.52	13	15.15	.87	5	10.98	1.43
.3	8	13.89	.86	9	10.88	.96	13	20.65	.52	15	15.67	.66	2	10.62	.44
.5	4	16.64	.42	3	8.80	1.05	8	21.09	.37	7	15.68	.60	5	10.68	.77
.7	8	16.62	.37	3	8.14	.44	4	19.73	.99	8	14.93	.61	2	10.16	.87
.9	7	14.49	.96	1	7.50	.00	3	18.60	1.15	4	14.65	.52	2	9.61	1.16
1.1	7	11.58	.53	3	6.47	.14	3	19.01	.42	1	12.95	.00	2	11.74	.84
1.3	3	8.66	1.36	1	5.97	.00	0			2	10.91	.56	0		
1.5	4	5.78	.41	1	5.39	.00	0			2	9.65	.51	0		
1.7	0			0			2	11.55	.99	1	7.57	.00	1	19.93	.00
1.9	2	3.30	.22	0			2	7.96	3.58	1	5.35	.00	0		

Table I
Theoretical Test Information for Conventional Tests in Each Content Area as a Function of Achievement Level

Achievement Level	Content Area 1	2	3	4	5
-1.9	.07	.15	.00	.01	.08
-1.7	.17	.30	.00	.02	.14
-1.5	.39	.60	.01	.04	.26
-1.3	.85	1.20	.02	.10	.47
-1.1	1.71	2.32	.04	.24	.80
-.9	3.04	4.24	.11	.59	1.33
-.7	4.75	7.12	.29	1.37	2.14
-.5	6.50	10.63	.82	2.92	3.34
-.3	8.10	13.55	2.39	5.47	4.97
-.1	9.79	14.46	5.95	8.80	6.91
.1	12.00	13.39	11.28	12.24	8.91
.3	14.40	11.40	16.47	14.89	10.54
.5	15.54	9.19	19.65	16.06	11.28
.7	14.38	7.02	20.32	15.66	10.99
.9	11.42	5.07	18.64	14.08	9.79
1.1	8.00	3.49	14.50	11.74	7.99
1.3	5.13	2.32	9.63	9.06	6.04
1.5	4.01	1.50	5.89	6.49	4.30
1.7	1.82	.96	3.51	4.38	2.92
1.9	1.06	.62	2.07	2.84	1.92

Table J
Development Group Discriminant Function Weights and Constants Used to Estimate Classroom Mastery Status from Mastery Status Estimated from Each Content Area Test during Each Testing Session and Across Testing Sessions for Each Testing Procedure

Testing Session and Procedure	Content Area 1	2	3	4	5	Constant
Testing Session 1 (N=100)						
Conventional						
Proportion Correct	1.86	1.75	1.45			-.54
Bayesian	2.80	1.48	1.30			-.40
AMT						
Variable Termination	1.89	1.88	.21			-.61
Fixed Length	2.26	1.46	.26			-.55
Testing Session 2 (N=100)						
Conventional						
Proportion Correct				5.92	5.92	-.18
Bayesian				7.17	.00	-.14
AMT						
Variable Termination				2.50	.40	-.60
Fixed Length				2.55	1.56	-.53
Both Sessions (N=89)						
Conventional						
Proportion Correct	1.95	1.41	1.49	-.23	-.08	-.69
Bayesian	2.68	-.06	-1.96	2.22	-1.96	-.49
Adaptive						
Variable Length	1.20	.70	-.87	2.05	-.83	.76
Fixed Length	1.63	.75	-1.41	2.11	.30	-.72