AN INFORMATION COMPARISON OF CONVENTIONAL AND ADAPTIVE TESTS IN THE MEASUREMENT OF CLASSROOM ACHIEVEMENT

Isaac I. Bejar
David J. Weiss
Kathleen A. Gialluca

RESEARCH REPORT 77-7
OCTOBER 1977

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

Prepared under contract No. N00014-76-C-0627, NR150-389
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Approved for public release; distribution unlimited.
Reproduction in whole or in part is permitted for
any purpose of the United States Government.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER Research Report 77-7	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

An Information Comparison of Conventional and Adaptive Tests in the Measurement of Classroom Achievement

5. TYPE OF REPORT & PERIOD COVERED

Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Isaac I. Bèjar, David J. Weiss, and Kathleen A. Gialluca

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

October 1977

13. NUMBER OF PAGES

38

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 Defense Advanced Research Projects Agency, Navy Personnel Research and Development Center, Army Research Institute, Air Force Human Resources Laboratory, and 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)

testing
achievement testing
computerized testing
adaptive testing
sequential testing
branched testing
individualized testing
tailored testing
programmed testing
response-contingent testing
automated testing
item characteristic curve theory

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

The information provided by typical and improved conventional classroom achievement tests is compared with the information provided by an adaptive test covering the same subject matter. Both tests were administered to over 700 students in a general biology course. Using the same scoring method, adaptive testing was found to yield substantially more precise estimates of achievement level than the classroom test throughout the entire range of achievement, while at the same time reducing the length of the test. The comparison of the improved conventional test with the stradaptive test indicated also that the scores derived from the adaptive test were more precise even in the range of achievement where the improved test was designed to be optimal. An analysis of the effects of expanding an adaptive test item pool indicates that improved precision of measurement can result from the addition to the pool of only slightly more discriminating items. A comparison of response pattern information values (observed information) with test information values (theoretical information) shows that the observed information consistently underestimates theoretical information, although the pattern of results from the two procedures is quite similar. It is concluded that the adaptive measurement of classroom achievement results in scores which are less likely to be confounded by errors of measurement and, therefore, are more likely to reflect a testee's true level of achievement. In addition, the reduction in number of test items administered by the adaptive measurement of achievement can result in additional time spent in instruction.

CONTENTS

Acknowledgements

Data utilized in this report were obtained from volunteer students in General Biology, Biology 1-011, at the University of Minnesota, during fall quarter 1976 and winter quarter 1977; appreciation is extended to these students for their participation in this research. The cooperation of Kathy Swart and Norman Kerr of the General Biology staff in providing access to the students, as well as their encouragement and important contributions of time and ideas to this research, are also deeply appreciated.

Technical Editor: Barbara Leslie Camm

AN INFORMATION COMPARISON OF CONVENTIONAL AND ADAPTIVE TESTS IN THE MEASUREMENT OF CLASSROOM ACHIEVEMENT

Achievement testing consists of mapping an individual's proficiency level onto an observable indicator of proficiency. This mapping is accomplished by means of a testing procedure. Two of the characteristics defining a testing procedure (Sympson, 1975) are the nature of the items in the test and the way in which the test items are administered. Both of these characteristics are potentially important factors in determining how accurately the observable indicator will reflect the individual's underlying proficiency level.

Given an item type, there are basically two ways of administering a test --individually or in groups. In group testing everyone answers the same set of test items; in individualized or adaptive testing everyone receives a different set of items, and the difficulty of a test is dynamically tailored to the ability level of the testee. The psychometric advantages and disadvantages of these two modes of administration have been the subject of research in recent years (Weiss, 1976; Weiss & Betz, 1973). Results of this research suggest that adaptive testing is superior to group (conventional) testing in terms of precision of measurement (McBride & Weiss, 1976; Vale & Weiss, 1975b; Weiss, 1976), test-taking motivation (Betz & Weiss, 1976), and potential to eliminate bias (Pine & Weiss, 1976).

Virtually all of this research is based on ability measurement rather than achievement measurement. The question which arises, therefore, is whether or not similar benefits would accrue in achievement testing. Since achievement testing can be conceptualized in several ways (Green, 1974), however, a general answer to this question may not be possible. For example, mastery testing (Block, 1971) is an approach to achievement testing which is currently receiving attention from both practitioners and theoreticians. The purpose of mastery testing is to classify individuals into two states: mastery and non-mastery. Because of the instructional philosophy behind mastery testing, there is likely to be a lack of variability in performance at the time of testing on a given instructional unit; and as a result, it becomes profitable to tailor the length of a test rather than its difficulty. Ferguson (1969) has demonstrated the feasibility of implementing such a testing system.

However, when instruction is likely to result in substantial variation with respect to achievement in the population being tested, the procedures for adaptive ability testing become relevant for achievement testing, provided that the same response models which apply in ability testing are also applicable in the measurement of achievement. In a previous report Bejar, Weiss, and Kingsbury (1977) established the plausibility of that assumption in a college instructional setting. The purpose of this study is to investigate in that same setting the performance of an adaptive testing model designed for ability measurement in comparison to classroom examinations covering the same course content.

Comparing testing procedures is difficult (Sympson, 1975) since different procedures usually differ in more than one respect. Comparisons between testing procedures are further complicated by the criteria for evaluation (Weiss & Betz, 1973). Reliability and correlational indices have been used to compare testing procedures in many live data investigations (e.g., Betz & Weiss, 1975; Vale & Weiss, 1975a) and in some simulation investigations (e.g., Jensema, 1976, pp. 82-89). Such comparisons are less than optimal. By summarizing all the data in one single value, important information is likely to be lost (Samejima, 1977).

A more appropriate evaluative criterion for comparing testing procedures is psychometric information. Unlike reliability and correlational indices, information is an index of the precision of measurement at all levels of the trait being measured. Information functions are particularly useful in comparing test models analytically. Bejar (1975) used information functions to compare the dichotomous, graded, and continuous response models; Hambleton and Traub (1971) used them to compare several logistic test models. Because the comparison was among models in these cases, the use of information functions was appropriate.

The comparison of the same model under two modes of administration (conventional and adaptive) is of interest in research on adaptive testing. In this research (e.g., McBride & Weiss, 1976; Vale & Weiss, 1975a) information functions have been computed by monte carlo procedures. The relative efficiency of the two modes of test administration has then been determined by the ratio of the information functions. The results of such comparisons, however, are theoretical predictions which should be verified empirically.

Research comparing conventional and adaptive testing,using information as the evaluation criterion, has been based almost exclusively on monte carlo simulated data. These simulation studies suggest that adaptive testing yields more precise scores than conventional testing; they are not entirely generalizable, however, since they are based on data that fit the model perfectly. There has been only one study based on data from live testees which used information as an evaluative criterion (Brown & Weiss, 1977); however, it was a real-data simulation study (Weiss & Betz, 1973, pp. 11-12) which did not involve the actual adaptive administration of test items to testees.

Purpose

The major aim of the present investigation was to compare an adaptive achievement test to a conventionally-administered classroom test, using information as the evaluative criteria. In contrast to previous investigations, the measure of information used was derived from live test administration of both the adaptive and conventional tests. Because classroom examinations are seldom designed to be psychometrically optimal, the adaptive test was also compared to an improved conventional test which was constructed from the same item pool. In addition, the data provided an opportunity to study the effects of expansion of the adaptive test item pool on its information characteristics.

Method

Data for this study were obtained from students enrolled in a large introductory Biology course at the University of Minnesota (see Bejar et al, 1977). Two midquarter examinations and a final examination are administered in the course. Although each midquarter examination covers several content areas, a single dimension has been shown to account for performance on the examinations (Bejar et al, 1977). In addition to the classroom examinations, volunteers completed two computer-administered adaptive tests which covered the same content as the midquarter examinations. The data analyzed for each student consisted of scores on the two classroom midquarter examinations and the corresponding scores on the first and second midquarter adaptive tests. The results are based on a comparison of the levels of information associated with these scores.

Subjects

Volunteers were recruited during the fall and winter quarters of the 1976-77 academic year. Each quarter an information sheet was distributed to all the students in the class which invited them to participate in the research project. For participating in the first midquarter adaptive test, participants received one point which was to be added to their course grade; and for participating in the second midquarter adaptive test, they received two points. During fall quarter 394 students participated in the first midquarter adaptive testing and 386 participated in the second midquarter adaptive testing; during winter quarter the corresponding numbers were 317 and 349, respectively.

Procedure

For both the first and second midquarter administrations, the volunteer students were given three tests in the following order: 1) an adaptive verbal ability test, 2) the multiple-choice adaptive biology test based on the content covered in the classroom midquarter examinations, and 3) a test consisting of specially designed biology items. In the present report, only the data from the adaptive biology tests were analyzed.

The three tests were administered by means of cathode ray terminals (CRT) connected to a Hewlett-Packard real-time computer system. Instructional screens explaining the operation of the equipment were presented prior to testing (DeWitt & Weiss, 1974). A proctor was present in the testing room at all times to assist students with the equipment. Each test item was presented separately at the rate of 960 characters per second on the CRT screen. Students responded by pressing the key corresponding to the chosen alternative. During the fall quarter administration, feedback was provided after each response, i.e., each student was informed whether or not he/she had answered each test item correctly. During the winter quarter administration, immediate feedback was not provided. There were no time limits imposed on any of the tests. At the completion of testing, students received a printed report which listed questions answered incorrectly and provided the correct answers.

Adaptive Test

Item pools. The development of the item pools used in this study has been described by Bejar et al., 1977. The answer sheets for two midquarter examinations from two previous academic quarters were used as raw data for obtaining the item parameters -- discrimination (a), difficulty (b), and guessing (c) -- of the item characteristic curves for the items. From the fall quarter administration 114 items were available, which covered the contents of the first classroom test; the pool for the second test contained 112 items.[1] From the winter administration 44 items were added to the first test pool, and 49 were added to the second test pool. There was thus a total of 158 items in the first test item pool and 161 in the second test pool.

To construct item pools which could be used for administration of stradaptive tests (Vale & Weiss, 1975a,b; Weiss, 1973), each of the two pools was structured by forming nine strata of increasing difficulty. Mean stratum difficulties were chosen so that there would be approximately the same number of items per stratum. Within each stratum the items were ordered in terms of their discriminations unless it resulted in items covering the same content area appearing consecutively. Appendix Tables A and B show the nine strata into which the first and second test item pools were structured.

Effects of expanding the item pool. In a conventional test the distribution of item parameters will determine the characteristics of scores derived from that test. Similarly, in adaptive testing the characteristics of the items in the item pool should influence the characteristics of the scores. The theoretical research on this question (Jensema, 1976, pp. 82-89), however, suggests that improving the item pool has little effect on precision of measurement.

The question of improving the item pool in adaptive testing was examined by Jensema, using a simulation study with Owen's (1975) Bayesian adaptive strategy. Two kinds of pools were studied: one in which a=1.0 and c=.25 for all items and one in which a=2.0 and c=.20. The distribution of b's within the two pools was the same. Jensema's conclusion, that improving the item pool has no effect on the accuracy of estimating θ, is counter-intuitive. One potential problem with Jensema's study is that the dependent variable used was the correlation of $\hat{\theta}$ and θ, which may not be sufficiently sensitive to detect changes in precision. Furthermore, the composition of the pools used by Jensema were atypical, since all the items were assumed to have the same discrimination. Consequently, his results lack generalizability.

The present data permit a more realistic assessment of the effects of item pool characteristics on the precision of adaptive test scores. Specifically, the response vector information functions computed for both adaptive tests in the winter data were based on an enlarged version of the fall item pool. The items that were added to both pools consisted of those items administered in the fall classroom test for which it was possible to obtain item parameter estimates.

[1] Bejar et al. (1977) reported that the second midquarter item pool contained 123 items; the 112-item pool resulted from the removal of 11 items which were administered in a special format as the third test.

Table 1 shows the mean and standard deviation of item parameter estimates for both fall item pools and the same statistics for the items added to form the winter pool. For the first test the mean of the added items was somewhat lower than the items in the fall pool. For the second test the added items were, on the average, slightly more discriminating. In terms of difficulty, the added items in the first test pool were, on the average, .10 easier. In the second test pool, the added items were only .02 easier. Appendix Tables A and B show that the added items were well distributed across the nine strata of the stradaptive test pools. In addition, within strata, the new items were well distributed in their order of administration. Average stratum discriminations were higher for the improved (winter) pool for only three of the nine strata in the Test 1 pool (Table A) and six of the nine strata in the Test 2 pool (Table B). In no case were the differences in mean discrimination very large.

Table 1
Mean and Standard Deviation of Item Parameter Estimates for Fall Item Pool and for Items Added to Winter Item Pool for Adaptive Tests 1 and 2

		a		*b*		*c*	
Test and Pool	Number	Mean	S.D.	Mean	S.D.	Mean	S.D.
Test 1							
Items in Fall Item Pool	114	1.21	.46	.18	1.22	.25	.09
Items Added for Winter	44	1.15	.37	.08	1.12	.30	.06
Test 2							
Items in Fall Item Pool	112	1.20	.40	.16	1.16	.27	.09
Items Added for Winter	49	1.22	.40	.14	1.23	.29	.07

Implementation. One of the advantages of the stradaptive testing strategy is that prior information can be used to select the stratum from which the first item is administered. In this study the entry point was selected by the student; at the beginning of each stradaptive test students were asked to state their grade-point-average (GPA) by selecting one of nine equally-spaced GPA intervals from 2.00 to 4.00 (DeWitt & Weiss, 1974, p. 49). On the assumption that overall GPA was related to biology achievement levels, students with the highest GPAs began the stradaptive test with an item at the most difficult stratum (Stratum 9), while those with the lowest GPA began with an item at the least difficult stratum (Stratum 1).

A variable criterion was used to terminate testing on the stradaptive test. After a student answered five items in a stratum, if he/she answered 20% or fewer correctly, testing was terminated. If testing was not terminated by this criterion after 50 items had been administered, no further items were administered.

The branching strategy used in the stradaptive test was: 1) if the current item was answered incorrectly or skipped, to administer the next unadministered item from the next easier stratum, or 2) if the current item was answered correctly, to administer the next unadministered item from the next more difficult stratum.

Conventional Tests

Classroom tests. The classroom examinations each quarter included 55 items, which the course staff selected by a combination of pedagogical criteria and procedures from traditional test theory. Their aim in constructing these tests was to produce a "good" test for purposes of course grading. Students were instructed to answer 50 items of their choice. For purposes of this research, however, the tests were shorter than 50 items, since item parameter estimates were not available for some of the items.

The item parameter estimates for the items in Fall Tests 1 and 2 (F1 and F2) are in Appendix Table C; those for Winter Tests 1 and 2 (W1 and W2) are in Appendix Table D.

Improved tests. A major problem in comparing testing procedures is that their inherently dissimilar characteristics frequently make equitable comparisons difficult. The problem can be alleviated by allowing each strategy to function optimally while equating the testing procedures on relevant characteristics. The classroom exams were not expected to be psychometrically optimal; therefore, it was necessary to compare the stradaptive tests with an improved conventional test drawn from the same item pool. The winter item pool contained all the items available; therefore, only the winter data were used in the construction of the improved conventional tests.

The improved conventional test was designed to use the most discriminating items in the item pool in order to measure individual differences in the range of course achievement within which differential grades would be assigned. That is, it was assumed that below a given level of "passing" the course, further differentiations among students were unnecessary; above that level, it was desirable to differentiate as accurately as possible among the students in order to assign differential grades. To permit a psychometrically meaningful comparison with the adaptive test, the improved conventional tests were also designed to be equivalent to the adaptive test in terms of levels of item discrimination and number of items administered.

A comparison of the mean discriminations for the original winter quarter classroom tests with the item pools used for the stradaptive test showed that the mean for the stradaptive pool was a=1.19 for the W1 item pool and a=1.21 for the W2 item pool. Mean discriminations for the winter classroom tests were 1.09 and 1.14, respectively. The comparison between the item discriminations of the two testing strategies is complicated, however, by the way items are selected for administration in the stradaptive test. Since the items in each stratum in the stradaptive pool were ordered by their discriminations and the branching strategy is designed to administer the earlier items in the strata first, the mean discrimination of the stradaptive item pools will be lower than the mean discrimination of items administered in most stradaptive tests.

To provide a fair comparison between the adaptive and the conventional tests, it would be necessary to construct a conventional test "matching" the item discriminations in the adaptive test. This is difficult to implement, however, since the discriminations in each administration of the adaptive test will differ. Instead, the improved conventional tests were designed to provide a comparison which would not favor the adaptive test in terms of mean item discrimination.

The improved conventional tests for each of the two midquarters were constructed by selecting the items which appeared first in the strata of the stradaptive pool; these were generally the most discriminating items in the strata. The number of items selected was based on the overall mean test length for the stradaptive test. The items which were ordered first in the top seven strata of the stradaptive tests were selected to constitute the improved conventional tests. Only seven strata were used rather than nine, so that the improved conventional test would be somewhat peaked. Its precision would thus be concentrated in the range of achievement most relevant for instructional decisions. The improved conventional tests consisted of 24 items each for both the first and second tests administered in the winter quarter; they were based on a stradaptive test with a maximum of 30 items which had a mean test length of approximately 24 items.

The item parameters for the items constituting the two improved conventional tests are shown in Appendix Table E. The first 21 items comprise the first three items in Strata 3 through 9 for both tests. In the improved conventional tests the last four items were the fourth items in Strata 7 through 9. These items had mean discrimination values of 1.73 and 1.76, respectively, for the two midquarter examinations; for the stradaptive pools the mean discriminations were 1.19 and 1.21.

Because of the way the stradaptive item pool is structured and the way stradaptive test items are selected, the mean discrimination of the improved conventional test would be equal to or greater than that of any stradaptive test. The mean discrimination of the two testing procedures would be equal solely for a testee whose stradaptive test response record included only the items in the improved conventional test. For any testee whose responses on the stradaptive test required administration of items farther down the strata than those used by the improved conventional test, the mean discrimination would be lower than that of the conventional test. Since the majority of stradaptive response records utilize items beyond the third item in the strata, the stradaptive tests generally would use items of lower average discrimination than would the improved conventional test.

Scoring

All tests were scored by maximum likelihood estimation, specifying Birnbaum's (1968) three-parameter logistic model as the response model. The item parameter estimates were edited by the scoring program so that the maximum value of the discrimination parameter (*a*) was set to 2.5, the maximum absolute value of the difficulty parameter (*b*) was set to 3.00, and the maximum value of the guessing parameter (*c*) was set to .35. In estimating achievement levels, omitted items were not scored as incorrect; they were merely ignored.

Information

Definitions. Equation 1 gives the *test information function* of a test consisting of *n* items in relation to the logarithm of the likelihood function of response pattern *v* (see Samejima, 1969):

$$I(\theta) = -E\left[\frac{\partial^2 log\ L_v(\theta)}{\partial\theta^2}\right], \qquad [1]$$

where $L_v(\theta)$ is the likelihood function, and

v is the pattern of correct/incorrect responses to a set of test items. That is, information is the (negative) *expected* value of the second derivative of the log likelihood function. "Psychometric information," defined in this way, is identical to Fisher's concept of information (cf. Edwards, 1971).

In this study the comparison between the conventional and adaptive tests was based on *observed* information functions. These were computed from the item responses given by each testee. The observed value of information, as opposed to the expected value, is the value of the second derivative of the log-likelihood function at a testee's estimated value of θ. That is,

$$\hat{I}(\hat{\theta}) = -\left[\frac{\partial^2 log\ L_v(\theta)}{\partial\theta^2}\right]_{\theta=\hat{\theta}} . \qquad [2]$$

Equation 2 defines the response vector counterpart of Samejima's (1969, Ch. 6) *item response information function* which she has called the *response pattern information function* (Samejima, 1973).

For the 3-parameter-logistic model, $\hat{I}(\hat{\theta})$ is given by

$$\hat{I}(\hat{\theta}) = D^2\sum\frac{a_g^2\ e^{x_g}}{[1+e^{x_g}]^2} - D^2\sum\frac{a_g^2 u_g c_g e^{x_g}}{[c_g+e^{x_g}]^2} , \qquad [3]$$

where D = 1.7

a_g = the estimate of the discriminating power of the item

c_g = the estimate of the lower asymptote of the item characteristic curve

$x_g = Da_g(\theta-b_g)$

b_g = the estimate of the difficulty of the item

$u_g = \begin{cases}1 \text{ if item is answered correctly}\\0 \text{ if item is answered incorrectly.}\end{cases}$

It is clear from Equation 3 that for a single item, $\hat{I}(\hat{\theta})$ takes one of two values, namely

$$\hat{I}_u(\hat{\theta}) = \frac{D^2a_g^2e^{x_g}}{\left[1+e^{x_g}\right]^2} - \frac{D^2a_g^2c_g e^{x_g}}{\left[c_g+e^{x_g}\right]^2} , \text{ if } u_g = 1 \qquad [4]$$

$$\hat{I}_u(\hat{\theta}) = \frac{D^2 a_g^2 e^{x_g}}{[1+e^{x_g}]^2} , \qquad \text{if } u_g = 0 \qquad . \qquad [5]$$

Equation 4 occurs with probability $P = P_g(\theta) = c_g + (1-c_g)[1+e^{-x_g}]^{-1}$, while Equation 5 occurs with probability $Q = Q_g(\theta) = 1 - P_g(\theta)$. Thus the expected value of $\hat{I}_{u_g}(\hat{\theta})$ (i.e., $\hat{I}_g(\hat{\theta})$), is

$$\hat{I}_g(\hat{\theta}) = (1-P)\frac{D^2 a_g^2 e^{x_g}}{[1+e^{x_g}]^2} + P\left[\frac{D^2 a_g^2 e^{x_g}}{[1+e^{x_g}]^2} - \frac{D^2 a_g^2 c_g e^{x_g}}{\left[c_g+e^{x_g}\right]^2}\right] \qquad [6]$$

$$= \frac{D^2 a_g^2 e^{x_g}}{[1+e^{x_g}]^2} - \frac{P D^2 a_g^2 c_g e^{x_g}}{[c_g+e^{x_g}]^2} ,$$

which is the usual item information function evaluated at $\hat{\theta}$ (see Birnbaum, 1968, Eq. 20.4.20). The sum of the $I_g(\hat{\theta})$ across all items administered in a test at a given value of $\hat{\theta}$ is $I(\hat{\theta})$, which is the theoretical test information function based on estimated values of θ. Brown and Weiss (1977)used the evaluation and summation of item response information functions by Equation 6 at an estimated value of θ in their live-data simulation study to obtain estimated information curves; their $\hat{\theta}$'s, however, were based on a Bayesian scoring routine.

Both $\hat{I}(\hat{\theta})$ and $I(\hat{\theta})$ depend on the item parameter estimates a, b, and c. However, $I(\hat{\theta})$ is one step further removed from the data, since it does not allow the observed response pattern of correct and incorrect response to dictate its value, whereas $\hat{I}(\hat{\theta})$ does. In theory $\hat{I}(\hat{\theta})$ may be considered an estimate of $I(\hat{\theta})$, which is easily obtained during the estimation of θ by the Newton-Raphson procedure, requiring both the first and second derivative of the log-likelihood function. The value of the second derivative of the log-likelihood function at the last iteration is $\hat{I}(\hat{\theta})$.

Computation. Using the maximum likelihood scores computed for each testee on the conventional and adaptive tests, information was computed for each testee during the scoring process by evaluating the second derivative of the log-likelihood function at the final estimated value of θ, based on test items actually administered. The response vector information curves for a given testing strategy were then obtained by grouping students on their estimated achievement ($\hat{\theta}$) in intervals of .20 from -2.00 to +2.00. The mean response vector information for students within a given interval of $\hat{\theta}$ was assigned to the midpoint of that interval. All information values presented below have been multiplied by 1/2.89.

Comparison. No studies have been reported which utilized the item response pattern information function $[\hat{I}(\hat{\theta})]$ computed from live-testing data; therefore,it was appropriate to compare the results of computing information by this method with the information curves derived from the sum of the item information functions. The computation of test information curves from

Figure 1
Observed and Theoretical Test Information Functions for Test F1

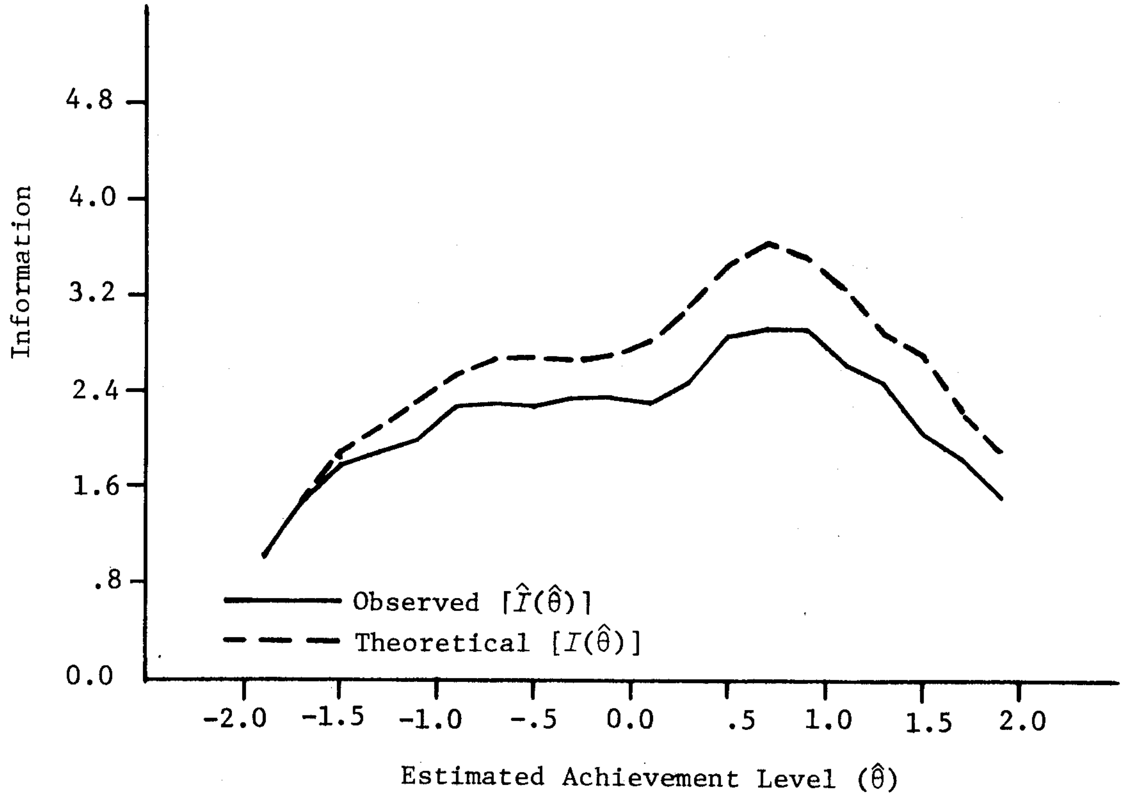
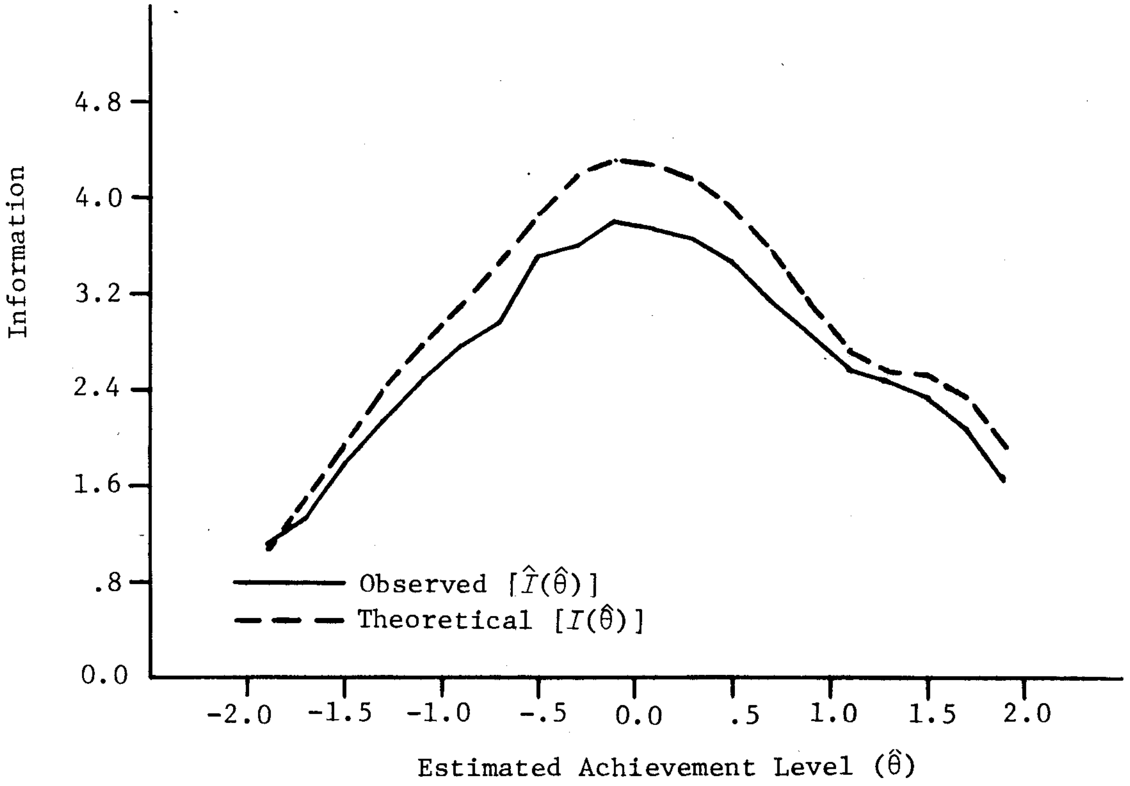


Figure 2
Observed and Theeretical Test Information Functions for Test F2



the sum of item information curves assumes that real testees respond to items in the test in accordance with the item characteristic curve (ICC) model. On the other hand, computing information curves using Equation 3 from the item response pattern of real testees will likely include some error, since all testees do not respond strictly in accordance with the model. A comparison of the two information curves derived from the same set of item responses was, therefore, useful to evaluate the applied usefulness of response pattern information functions, as well as to indicate whether or not the responses of the students to this achievement test were widely discrepant from the ICC model.

Consequently, test information curves were computed from the sum of the item information functions (Equation 6) and from the response pattern information functions (Equation 3), using the responses to the conventional test for each of the four midquarter examinations.

Results

Test Information Versus Response Pattern Information

Figures 1 through 4 show for the four classroom biology examinations the test information curves computed from 1) the sum of the item information, i.e., the theoretical test information function [$I(\hat{\theta})$]; and 2) the response pattern information curves, i.e., the observed test information function [$\hat{I}(\hat{\theta})$]. Data for the test information functions are in Appendix Table F; data for the response pattern information functions are in Tables 2 and 3.

The data for fall quarter (Figures 1 and 2) show that item response pattern information [$\hat{I}(\hat{\theta})$] consistently underestimated the theoretical curve derived by summing the item information functions [$I(\hat{\theta})$]. The difference was fairly consistent throughout the θ range, although for the first test (F1), the discrepancy diminished at the lower end of the θ continuum, where $\theta \leq -1.40$. For both sets of data the largest differences appeared to be at the point of highest information; the magnitude of differences decreased with decreasing information levels.

The winter data (Figures 3 and 4) exhibited the same general pattern of results. It can be seen from Figures 3 and 4 that $\hat{I}(\hat{\theta})$ again underestimated the value of the theoretical test information function. In the first test (W1) there was a marked decrease in the discrepancy between the two curves for those values of θ less than about -1.25; in the second test (W2) data the two curves were closest together at values of θ less than 1.50. The winter data, however, did not fully support the tendency for the two curves to be farthest apart at the point of highest information; this tendency occurred in the W2 data, but not the W1 data.

There were thus three trends common across all four examinations:

1. The observed (response pattern) curve was always an underestimate of the theoretical (test) information curve;
2. The differences between the two information curves tended to diminish, and in some cases disappear, at low levels of θ; and
3. There was a fairly constant difference between the two information curves throughout the range of θ above -1.00.

Figure 3
Observed and Theoretical Test Information Functions for Test W1

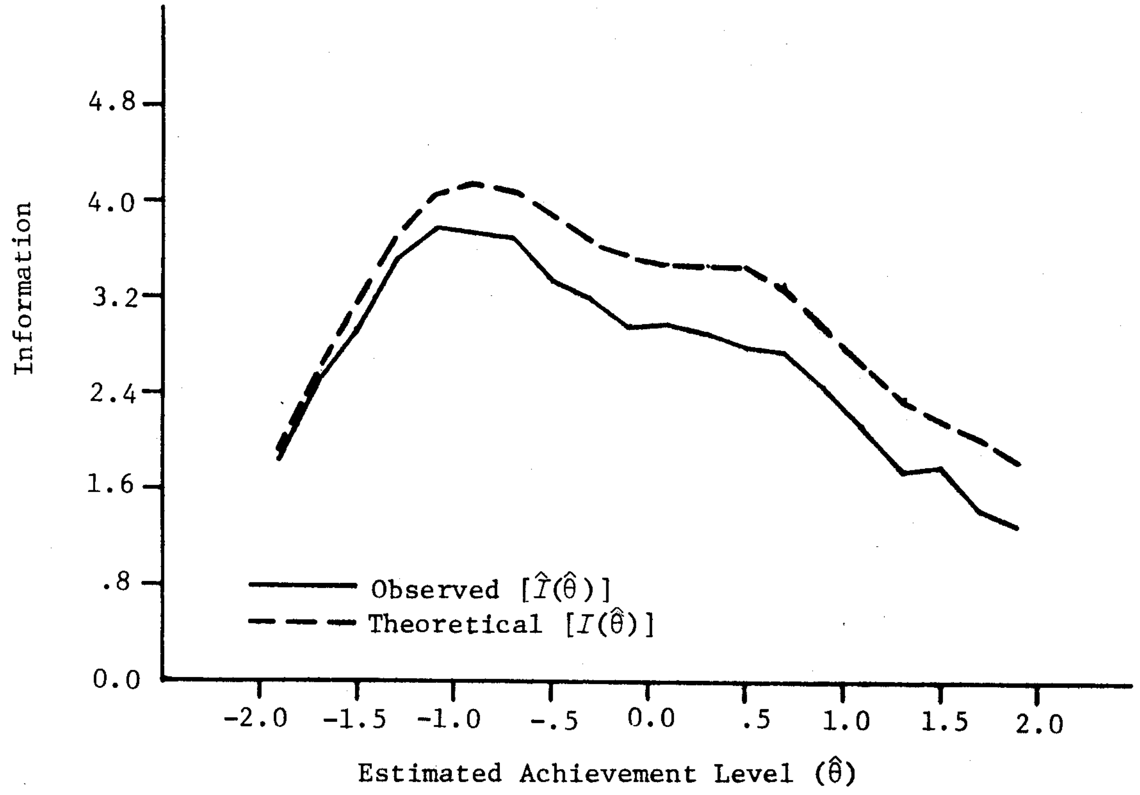
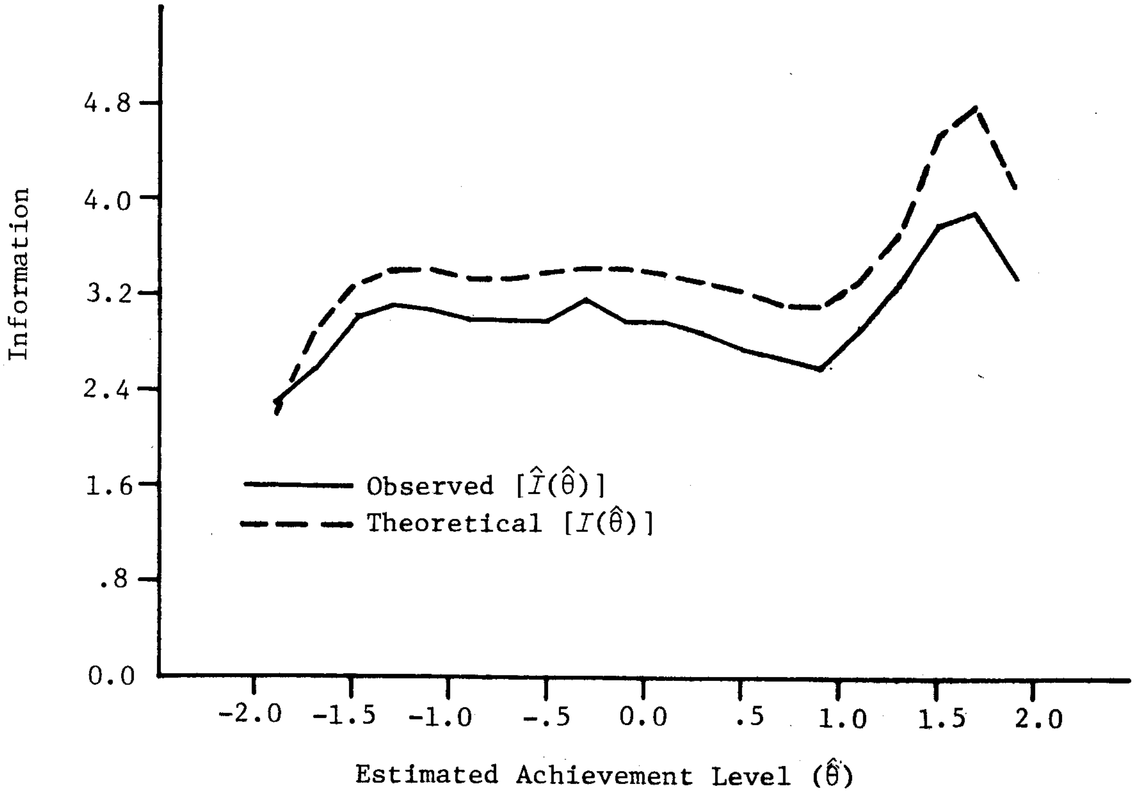


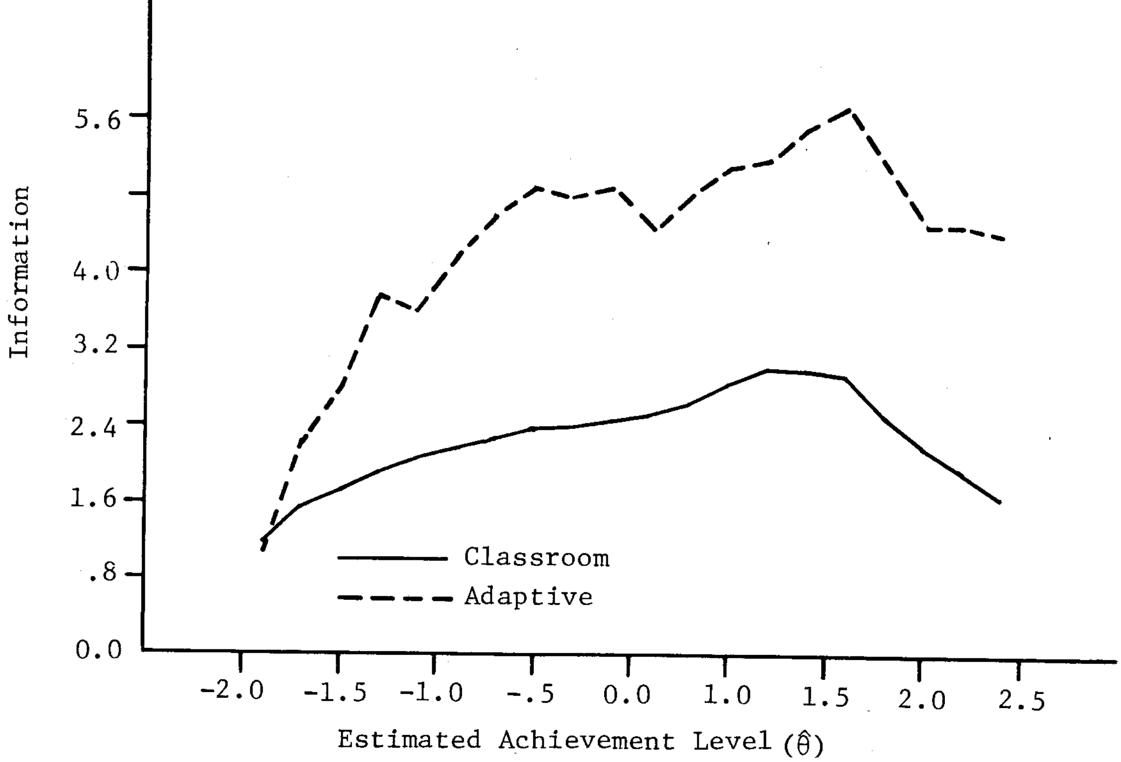
Figure 4
Observed and Theoretical Test Information Functions for Test W2



Adaptive Test Versus Classroom Test

First tests. Table 2 shows the values of observed information (response pattern information) from the first classroom and adaptive tests for fall (F1) and winter (W1). The results are plotted in Figures 5 and 6, which show that for both fall and winter the adaptive test yielded a substantially higher amount of information at all levels of achievement greater than $\hat{\theta}$=-1.5. Because the adaptive test was shorter, on the average, this is particularly significant.

Figure 5
Observed Information Functions for F1 Classroom
and Adaptive Tests



As previously indicated, the number of items was not fixed for the adaptive test. Although the maximum test length for the adaptive test was 50 items, Table 2 shows that on the average students were terminated after 27.2 items in Test F1 and after 31.6 items in Test W1. Excluding students at the extremes of the $\hat{\theta}$ distribution (where the stradaptive test would tend to terminate prematurely because suitable items were not available in the pool), the range of mean number of items to termination on the stradaptive test was 18.9 to 32.5 for Test F1 and 25.9 to 41.1 for Test W1. On the other hand, the mean information values for the classroom test were based on an average of 35 items for Test F1 and 40.5 items for Test W1. (Although the actual classroom test was 50 items long, there were items for which no item parameter estimates were available.)

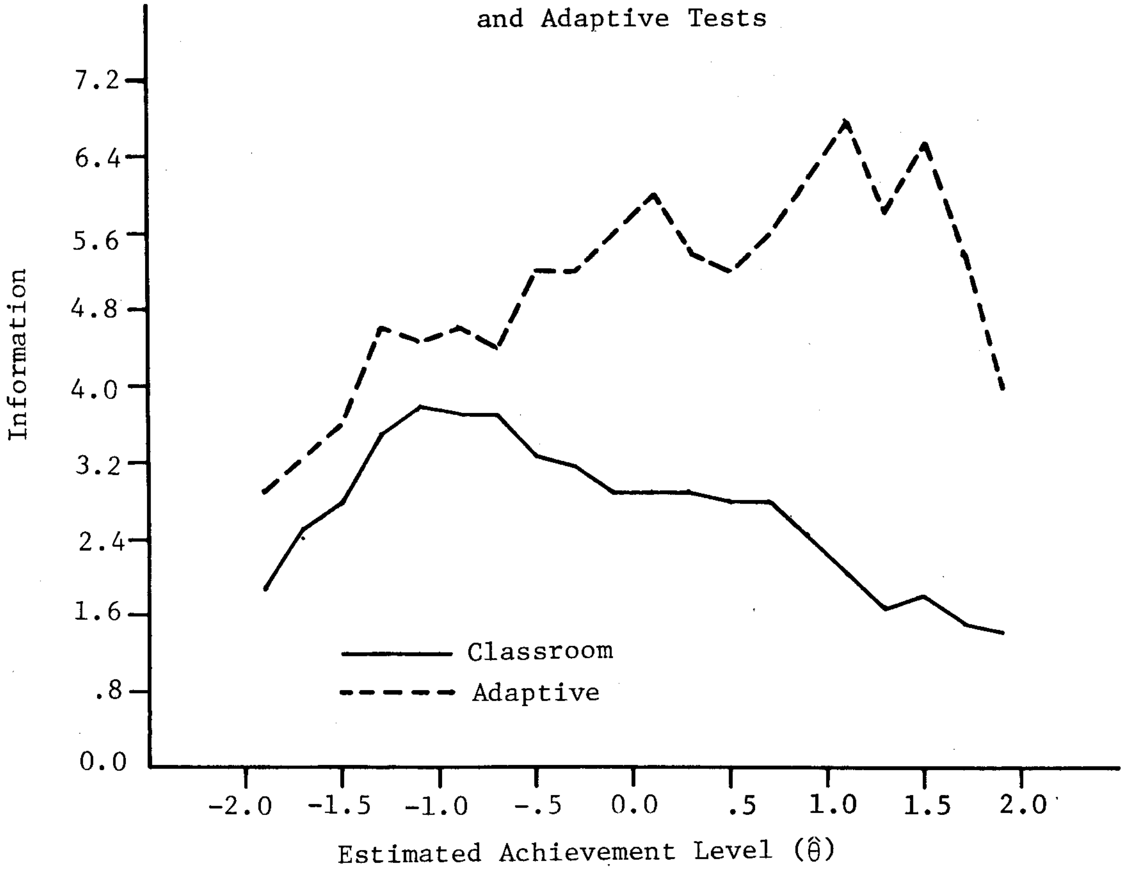
Thus, even though the adaptive test on the average was about eight items shorter, it yielded a much more precise estimate of achievement. For example, at $\hat{\theta}$=.7 (and .9) the F1 classroom test had maximum information of 2.90, whereas at $\hat{\theta}$=.7 the adaptive test had maximum information of 5.07.

Table 2

Number of Testees, Mean Number of Items, and Mean Observed Information $[\hat{I}(\hat{\theta})]$ for Intervals of $\hat{\theta}$ for First Adaptive and Classroom Tests from Fall and Winter Quarters

	Fall Test 1 (F1)						Winter Test 1 (W1)					
$\hat{\theta}$	No. Testees		No. Items		Information		No. Testees		No. Items		Information	
Midpoint	Adap	Class	Adap	Class	Adap	Class	Adap	Class	Adap	Class	Adap	Class
-1.9	5	5	16.8	35.4	1.15	1.03	2	9	27.5	40.4	2.94	1.83
-1.7	8	24	26.6	35.1	2.06	1.48	8	14	32.2	40.5	2.43	2.50
-1.5	13	26	26.3	34.8	2.70	1.77	15	16	28.8	40.6	3.73	2.92
-1.3	15	47	30.8	34.7	3.68	1.88	12	31	34.4	40.8	4.64	3.55
-1.1	14	41	24.5	34.8	3.65	1.99	17	41	34.3	40.7	4.57	3.80
-.9	16	60	31.9	34.9	4.10	2.27	12	64	29.8	40.8	4.63	3.73
-.7	18	53	31.4	34.6	4.50	2.29	20	69	26.1	40.5	4.37	3.68
-.5	19	63	28.6	34.8	4.90	2.26	21	69	33.5	40.6	5.35	3.37
-.3	33	81	22.2	34.8	4.78	2.36	26	72	25.9	40.5	5.36	3.20
-.1	22	64	23.5	35.1	4.74	2.37	24	86	31.2	40.5	5.91	2.94
.1	31	68	25.5	35.0	4.47	2.34	24	83	35.3	40.4	6.13	2.98
.3	32	83	29.0	34.8	4.78	2.48	17	86	30.8	40.3	5.42	2.90
.5	21	68	32.5	35.3	5.05	2.87	16	76	32.1	40.5	5.30	2.78
.7	16	88	31.3	34.9	5.07	2.90	19	68	30.3	40.3	5.54	2.75
.9	17	61	30.0	35.0	5.50	2.90	20	52	36.6	40.6	6.32	2.45
1.1	17	59	29.1	35.0	5.77	2.63	13	54	37.0	40.5	6.74	2.12
1.3	12	49	22.6	35.4	5.26	2.46	10	33	30.8	40.6	5.79	1.78
1.5	10	40	18.9	35.5	4.55	2.05	8	26	41.1	40.5	6.46	1.83
1.7	10	44	29.2	35.3	4.51	1.81	4	27	30.7	40.6	5.52	1.44
1.9	1	20	10.0	35.2	4.43	1.54	4	23	13.2	40.4	4.16	1.30
Total	330	1,044					292	999				
Mean			27.2	35.0	4.53	2.36			31.6	40.5	5.28	2.89

Figure 6
Observed Information Functions for W1 Classroom and Adaptive Tests



The ratio of information values at $\hat{\theta}$=.7 was 1.75. This means that for the conventional test to be equal in precision to the adaptive test at that level of $\hat{\theta}$, it would have to be increased in length by about 75%. This would result in a conventional test of 61 items in order to achieve the same quality of measurement as a stradaptive test with a mean of 31.3 items.

The stradaptive test achieved its highest level of information (5.77) at $\hat{\theta}$=1.1 with an average of 29.1 items; the information provided by the classroom test at $\hat{\theta}$=1.1 was 2.63. The ratio of 2.19 indicates that at this level of $\hat{\theta}$ the classroom test would require 77 items to measure as well as the 29.1-item average adaptive test.

Similar results were observed for the W1 data. At the point where the classroom test was most informative, $\hat{\theta}$=-1.3, the adaptive test was more informative by a factor of 4.64/3.55=1.31 with, on the average, 6.4 fewer items. Thus, at that level of $\hat{\theta}$ the classroom test would require 53 items to measure as precisely as the average 34-item stradaptive test. At the point where the adaptive test was most informative, $\hat{\theta}$=1.1, the improvement factor was 6.74/2.12=3.18. The classroom test,therefore, would require 129 items to measure as precisely as the 37-item adaptive test. Thus, even when comparisons were made at the point of maximum information for the classroom test, the adaptive test was more efficient in terms of information per item. When the comparison was made at the point of maximum information for the adaptive test, the discrepancy in efficiency for the two testing procedures was even greater.

Table 3

Number of Testees, Mean Number of Items,and Mean Observed Information [$\hat{I}(\hat{\theta})$]
for Intervals of $\hat{\theta}$ for Second Adaptive and Classroom Tests from Fall and Winter Quarters

	Fall Test 2 (F2)						Winter Test 2 (W2)					
$\hat{\theta}$	No. Testees		Mean No. Items		Information		No. Testees		Mean No. Items		Information	
Midpoint	Adap	Class	Adap	Class	Adap	Class	Adap	Class	Adap	Class	Adap	Class
-1.9	6	10	26.8	37.0	1.90	1.14	1	8	17.0	40.3	1.99	2.29
-1.7	13	23	24.7	37.1	2.55	1.34	3	10	34.0	39.7	2.02	2.57
-1.5	7	33	34.1	36.8	3.45	1.79	9	21	30.4	40.0	3.41	3.05
-1.3	14	26	34.5	37.5	3.63	2.13	9	29	29.8	39.9	3.88	3.11
-1.1	23	29	33.8	36.8	4.09	2.50	15	38	27.6	40.2	3.91	3.09
-.9	24	45	34.2	37.0	4.23	2.77	13	41	37.7	40.5	4.52	2.98
-.7	17	64	32.7	37.1	4.83	2.97	29	50	33.9	40.2	4.88	2.99
-.5	21	62	24.3	36.8	2.69	3.53	22	68	38.4	40.5	4.91	2.98
-.3	22	84	29.9	37.2	3.26	3.60	20	83	27.2	40.3	4.88	3.17
-.1	35	97	27.6	37.4	3.19	3.82	36	75	25.8	40.0	4.08	2.98
.1	29	94	26.7	37.2	3.11	3.76	34	76	28.2	40.2	4.80	2.98
.3	22	101	32.7	37.2	3.82	3.66	26	63	27.0	40.3	4.41	2.88
.5	21	90	37.6	37.3	4.24	3.49	29	76	33.4	40.6	5.21	2.76
.7	27	70	35.4	37.4	4.44	3.15	15	62	35.8	40.5	5.86	2.67
.9	16	71	28.7	37.7	3.87	2.86	17	56	34.1	40.1	5.08	2.56
1.1	14	57	30.7	37.8	4.13	2.56	16	58	35.6	40.5	4.76	2.96
1.3	24	51	37.8	37.6	4.69	2.45	14	60	39.4	40.4	4.51	3.27
1.5	5	45	34.8	37.6	4.32	2.36	18	59	38.0	40.6	4.74	3.78
1.7	6	27	36.6	37.6	4.75	2.04	6	40	30.1	40.8	4.44	3.85
1.9	3	20	50.0	37.9	6.35	1.66	2	16	50.0	40.6	5.68	3.35
Total	349	1,099					334	989				
Mean			31.7	37.3	3.79	3.06			32.0	40.3	4.64	3.03

Second tests. Table 3 shows the number of testees, the mean number of items, and the mean information as a function of $\hat{\theta}$ for the second classroom and adaptive tests administered during fall (F2) and winter (W2). Estimated information curves are plotted for these tests in Figures 7 and 8. Figure 7 shows that for the F2 data, the adaptive test was generally superior to the classroom test. In the interval from $\hat{\theta}$=-.50 to $\hat{\theta}$=.20, however, the classroom test yielded higher levels of information.

Figure 7

Observed Information Functions for F2 Classroom and Adaptive Tests

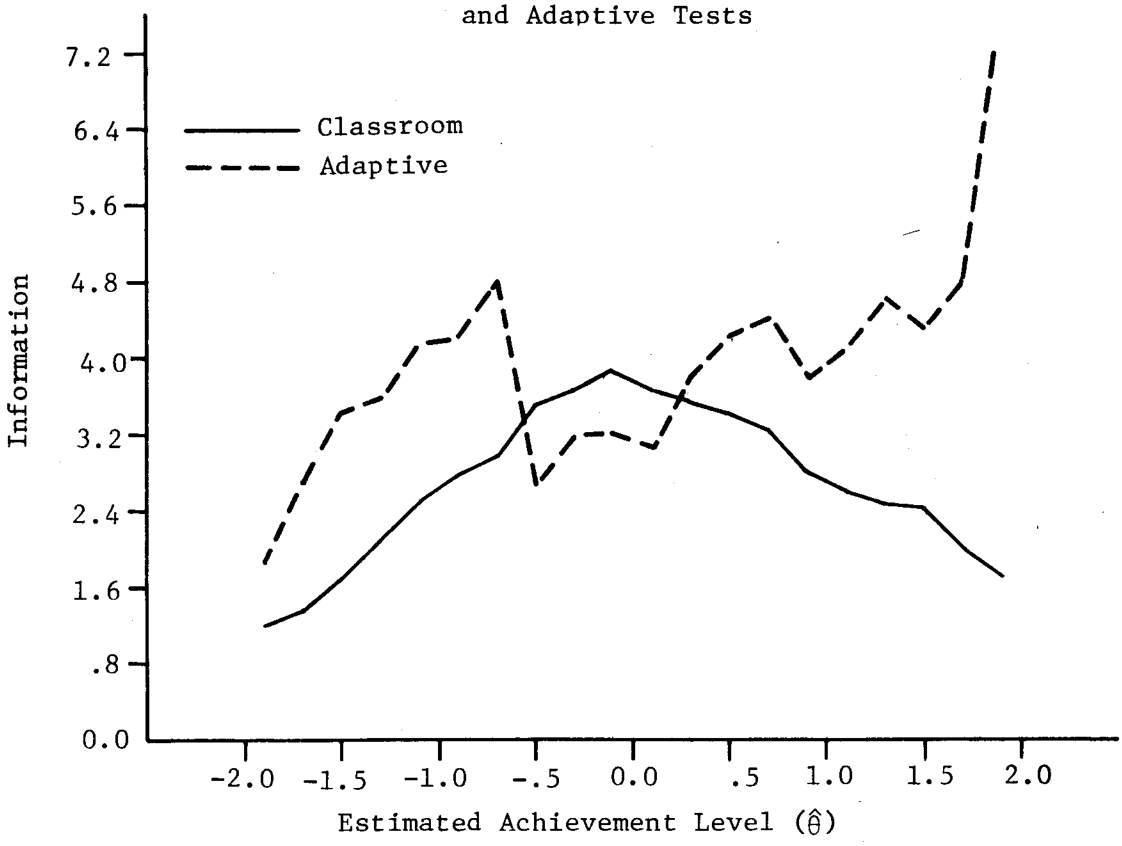
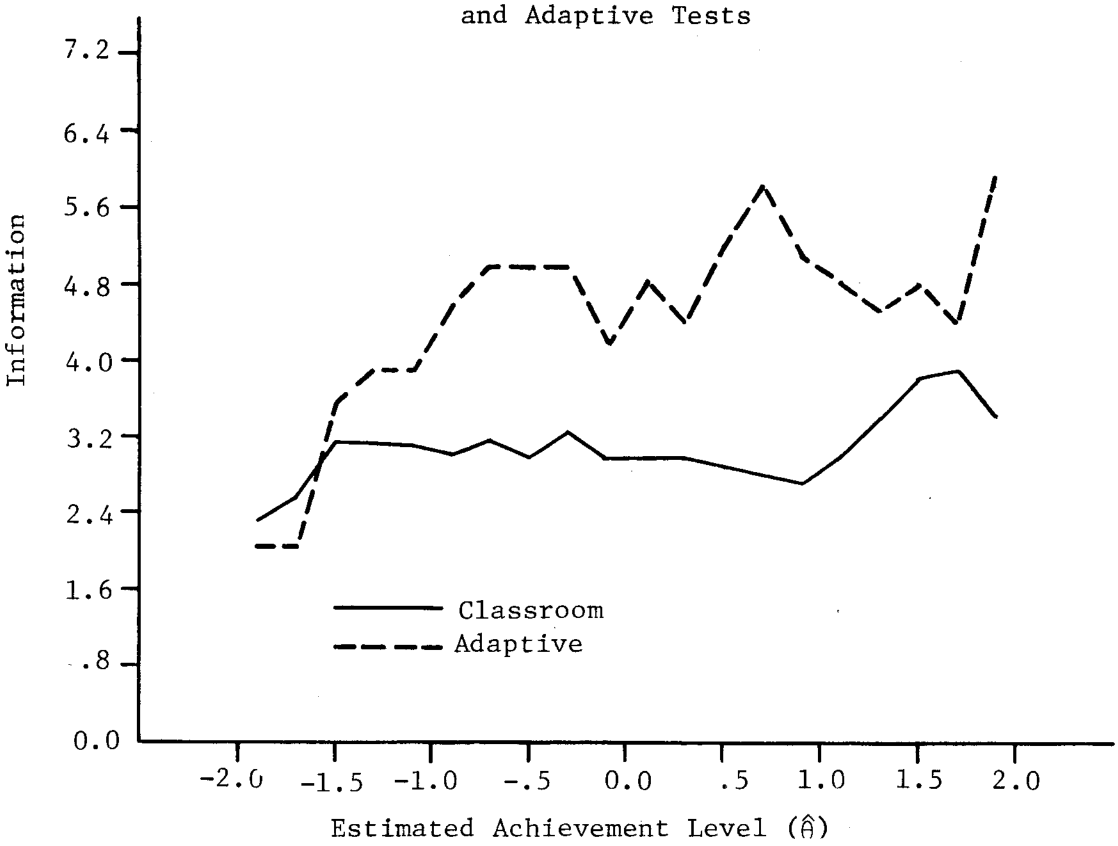


Figure 8 shows the results for W2. For all $\hat{\theta}$ values greater than -1.5,the information provided by the adaptive test was substantially higher than that of the classroom test; this was similar to the findings for F1 and W1. The adaptive test thus provided better measurement throughout the $\hat{\theta}$ range in three of the four tests.

There are two explanations for the adaptive test providing less information than the conventional test for the F2 data in a narrow range around the mean of the $\hat{\theta}$ distribution. First, as Appendix Table C shows, the F2 classroom test was a considerably more peaked test than the F1, W1, and W2 classroom tests. Peaked tests tend to have peaked information functions (Lord, 1970), since they concentrate all their measurement efficiency near one point on the θ continuum.

A more important explanation, however, is seen in Table 3. In the range of $\hat{\theta}$=-.5 to .10, the mean adaptive test length was substantially below the mean classroom test length. Dividing the information at each of these values of $\hat{\theta}$ by their corresponding test length indicates that

Figure 8
Observed Information Functions for W2 Classroom and Adaptive Tests



the mean information per item was higher for the adaptive test than for the classroom test. For example, at $\hat{\theta}$=-.5 the mean information per item was .11 for the adaptive test and .09 for the classroom test. Thus, while observed mean information was lower for the adaptive F2 data, this was merely an artifact and was attributable to the termination rule employed in the test, which resulted in very short tests in the $\hat{\theta}$ range of -.5 to .20.

Summary. The results from both test administrations show that when differences in test length were taken into account, the adaptive tests yielded substantially more precise estimates of achievement than any of the conventional tests at all levels of achievement. The results, summarized in Table 4, were equally favorable to adaptive testing when all $\hat{\theta}$ levels were combined. As shown in Table 4, the information across levels of $\hat{\theta}$ for the F1 data was 4.53 for the adaptive test and 2.36 for the classroom test with test lengths of 27.2 and 35.0 items, respectively. The information ratio of 1.92 in favor of the adaptive test implies that the classroom test would require 67 items in order to measure as precisely as the average 27-item adaptive test. The results for the other three tests (W1, F2, and W2) also showed the overall superiority of the adaptive test while reducing test length. The smallest improvement was for the F2 data; the ratio of mean information for the F2 test was 1.24 in favor of the adaptive test, implying that the conventional test would require 46 items to measure as well as an average 32-item adaptive test. This

Table 4
Mean Information and Mean Test Length for Fall and Winter Adaptive and Classroom Tests

	Mean Information			Mean Test Length		
Test	Adap.	Class.	Ratio[a]	Adap.	Class.	Difference[b]
F1	4.53	2.36	1.92	27.2	35.0	7.8
W1	5.28	2.89	1.83	31.6	40.5	8.9
F2	3.79	3.06	1.24	31.7	37.3	5.6
W2	4.64	3.03	1.53	32.0	40.3	8.3

[a] Adaptive divided by Classroom

[b] Classroom minus Adaptive

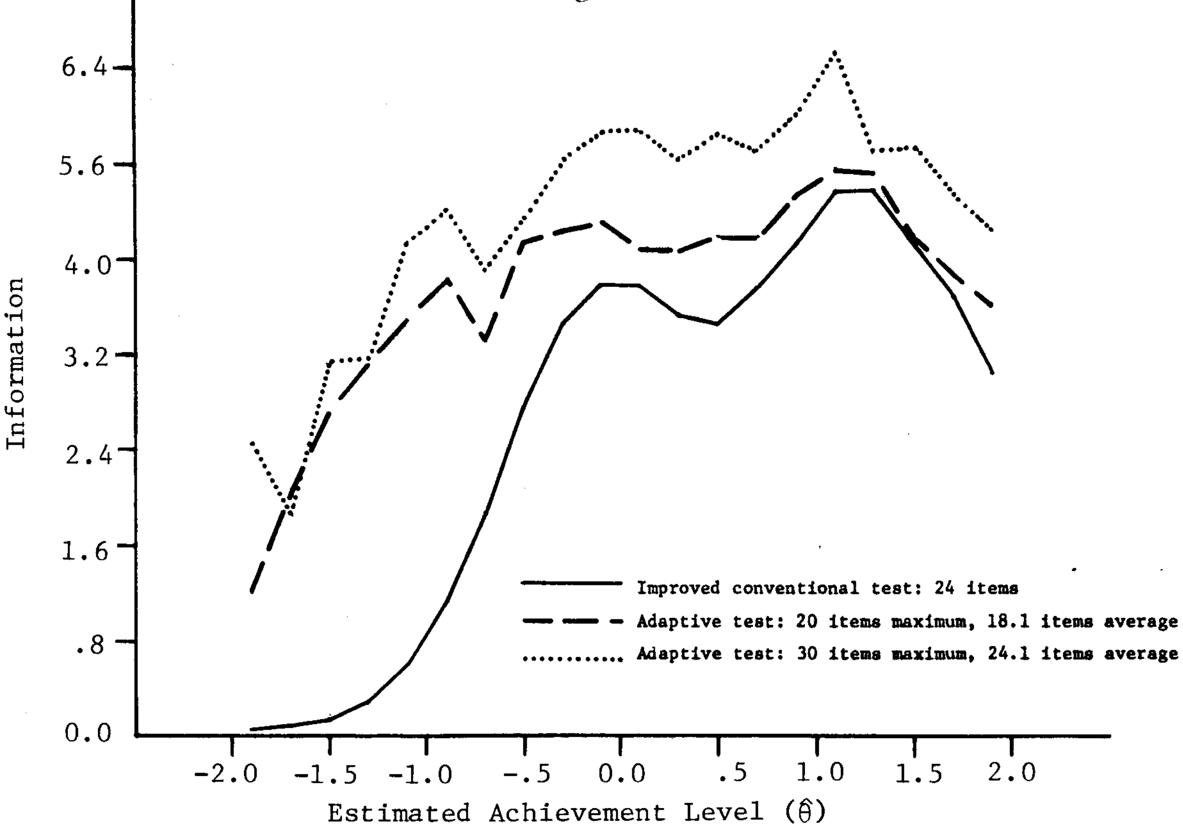
represents a reduction of 30% in classroom test length attributable to adaptive testing,while achieving equivalent average precision with the peaked classroom test.

Adaptive Versus Improved Conventional Test

Test W1. Since the improved conventional test was not actually administered, it was not possible to compute its response vector information function. Instead, mean values of the test (theoretical) information function were computed at 20 levels of $\hat{\theta}$ using Equation 6. The obtained values are in Appendix Table G, which also shows the mean values of response pattern information for the adaptive W1 test, rescored using maximum test lengths of 40, 30, and 20 items. Based on the data in Table G, Figure 9 shows the corresponding response pattern information curves for the stradaptive test at 20- and 30-item maximum lengths and the test information curves for the improved conventional test.

As Figure 9 shows, test information for the improved conventional test was very low for the low levels of achievement. Since there were no items in this test with difficulties less than b=-.65, this was to be expected. The significant comparison between the two testing strategies is for $\hat{\theta}$ values greater than approximately -.40, as indicated by the vertical dashed line in Figure 9. Within this range, both the adaptive tests had maximum information at $\hat{\theta}$=1.10, while the information curve for the improved conventional test was almost at its peak. The mean response vector information for the 20-item maximum length adaptive test at $\hat{\theta}$=1.10 was 4.78; the corresponding value of information for the improved conventional test was 4.52. This represents a 6% increase in information, with an average decrease of 5 items. A more significant comparison can be made with the 30-item maximum adaptive test, since, on the average, it was 24 items long and therefore the same length as the improved conventional test. Throughout the range of $\hat{\theta}$, as well as in the range in which the improved conventional test was designed to function optimally, the value of response vector information for the 30-item maximum adaptive test was substantially higher than that for the improved conventional test (see Figure 9). Specifically, at $\hat{\theta}$=1.3, where the improved conventional test had the highest information, the 30-item maximum adaptive test had at least 7% more information. At that specific value of $\hat{\theta}$, the

Figure 9
Mean Information as a Function of Estimated Achievement Level for Improved Conventional Test and Adaptive Test at Two Test Lengths for W1 Tests

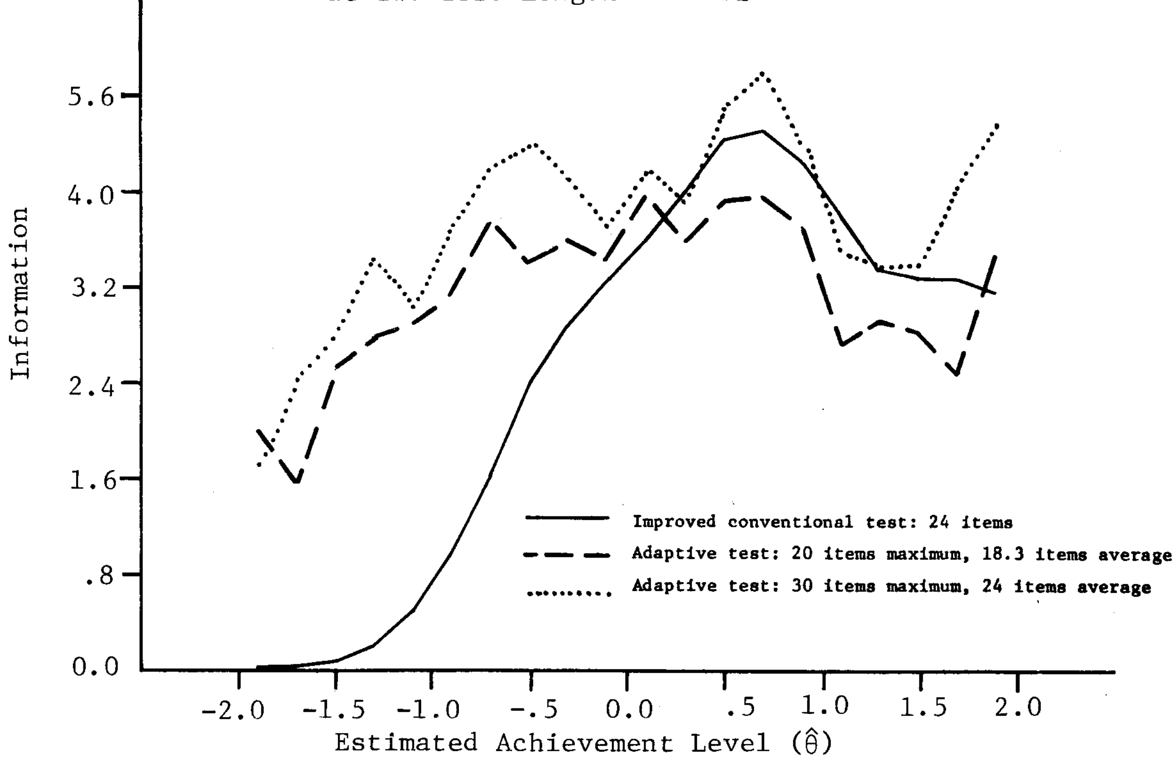


mean test length for the stradaptive test was 22.8 items, or 1.2 items shorter than the improved conventional test. The improved conventional test would then require 25.7 items in order to measure as precisely as the average 22.8-item adaptive test. Thus, with test length and average item discriminations equal, the adaptive process resulted in measurement of higher precision.

Test W2. Appendix Table H shows the values of the theoretical test information functions for the improved conventional test, as well as the mean values of response vector information for the adaptive test rescored with maximum lengths of 40, 30, and 20 items. The information curves based on these data for the conventional test and for the 20- and 30-item adaptive test are plotted in Figure 10.

The information for the improved conventional test was again very low for $\hat{\theta}<-1.00$ (see Figure 10), because of the way in which items were selected; the lowest difficulty level for an item in the conventional test was b=-.61. For $\hat{\theta}$ values in the range providing an equitable comparison of the two testing procedures, the information values for the improved conventional test were higher than those for the adaptive test with a maximum length of 20 items, for $\hat{\theta}>.20$. However, the mean number of items for the adaptive test at these levels of $\hat{\theta}$ was always less than 20, or four to eight items shorter than that of the improved conventional test.

Figure 10
Mean Information as a Function of Estimated Achievement Level for Improved Conventional Test and Adaptive Test at Two Test Lengths for W2 Tests



The comparison of the information curves for the 24-item conventional test with that of the maximum 30-item adaptive test provided a comparison of the two testing procedures which is equated for mean number of items, since under these conditions an average of 24 items was administered in the adaptive test. In the relevant range of $\hat{\theta}$, the adaptive test provided generally higher levels of information, except at $\hat{\theta}$=.3 and $\hat{\theta}$ =1.1, where information provided by the conventional test was slightly higher, and at $\hat{\theta}$=1.3,where both testing procedures provided equal levels of information (see Figure 10). The adaptive test administered two fewer items, on the average, at $\hat{\theta}$=.3 than the improved conventional test; at the other two values of $\hat{\theta}$ the number of items administered was the same.

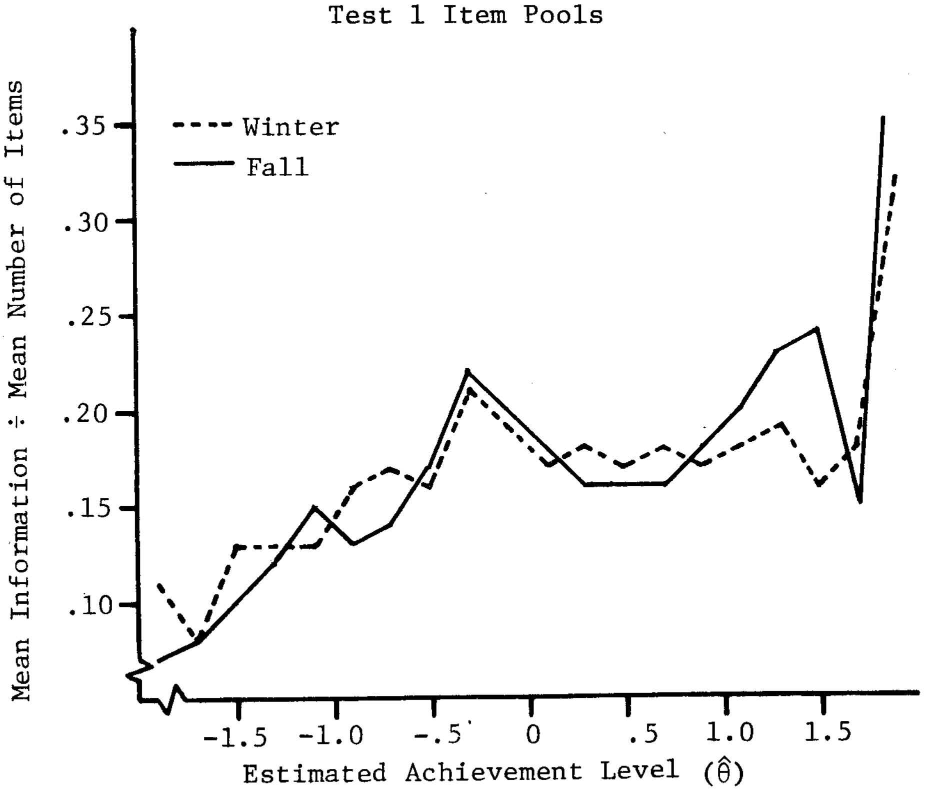
Summary. The comparisons between the improved conventional tests and the adaptive tests showed that 1) improved adaptive tests provided higher levels of information with fewer items than the conventional test and 2) adaptive tests provided generally higher levels of information with approximately the same mean number of items. The comparisons were based on tests with comparable values of item discriminations, although the discriminations in the improved conventional test were generally higher than the mean discriminations of the adaptive tests. One additional factor further influenced these comparisons in favor of the conventional test:

data on which the comparisons were based were the theoretical information functions for the conventional test and the observed (response vector) information functions for the adaptive test. As Figures 1 to 4 show, the theoretical information values consistently over-estimated the observed information values. Thus, the information values for the conventional tests are probably higher than they would be had they been computed from the response vectors of actual testees. As a result, it can be concluded that when adaptive and conventional tests are matched in terms of test length and average item discriminations, the adaptive test results in consistently higher levels of information. The improvement in precision resulting from adaptive testing is a function of the process of selecting test items appropriately matched to the testee's estimated level of achievement.

Effect of Expanding the Item Pool

First tests. The response vector information curves for Tests F1 and W1 are in Figures 1 and 3, respectively; mean information values are in Table 2. As Table 2 shows, however, mean test lengths, as well as mean information for the two tests differed. Both mean test length and mean information were higher for the W1 tests which utilized the enlarged item pool. Consequently, a direct comparison between the two information curves would be confounded by test length.

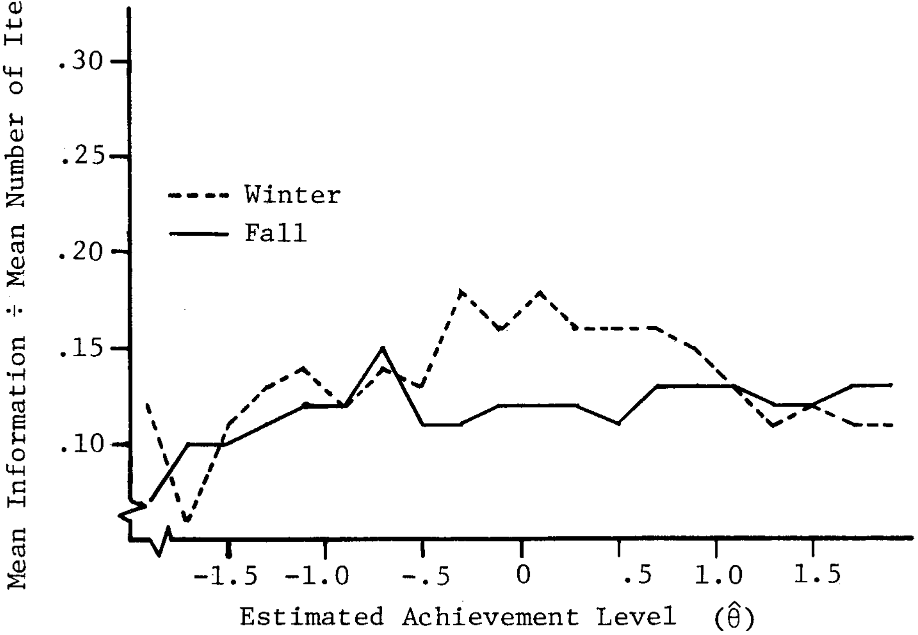
Figure 11
Mean Information Divided by Mean Number of Items
for Fall and Winter Adaptive Tests Using
Test 1 Item Pools



To determine whether or not there had been an improvement in information beyond that attributable to increased test length, the mean response pattern information at each level of $\hat{\theta}$ was divided by the corresponding mean test length. These data are shown in Appendix Table I and the resulting curves are plotted in Figure 11. The two curves differed very little until the point at which $\hat{\theta}$=.10. Thereafter and until the point at which $\hat{\theta}$=.70, the winter data provided slightly more information. After this point the winter pool failed to provide levels of information as high as those of the fall pool. In terms of overall information, however, there was no increase in mean information from fall to winter.

Second tests. Mean values of response vector information for the fall and winter are shown in Table 3, and information curves are plotted in Figures 2 and 4. Figure 12 shows the two information curves equated for mean test length at each interval of $\hat{\theta}$; numerical values are in Appendix Table I.

Figure 12
Mean Information Divided by Mean Number of Items
for Fall and Winter Adaptive Tests Using
Test 2 Item Pools



The winter pool provided higher levels of information than the fall pool at almost levels of $\hat{\theta}$ (see Figure 12). The differences were particularly large in the interval $\hat{\theta}$=-.5 to $\hat{\theta}$=1.10. The mean response vector information values equated for test length across all levels of $\hat{\theta}$ for fall and winter were .12 and .15, respectively; their ratio was 1.25, which represented a 25% increase in information attributable to the expanded item pool with test length held constant.

Summary. These results show the expected outcome. That is, the improvement in precision of measurement as a function of enlarging the item pool depends on the nature of the items added to the pool. For the first tests, the additional items were slightly less discriminating than the items already in the pool; therefore, using the enlarged winter quarter pool did not provide precision of measurement which was appreciably better. For the second tests, however, the items added to form the winter pool were, on the average, slightly more discriminating than the items already in the pool. Scores derived from the enlarged winter item pool were thus more precise than those from fall.

Summary and Conclusions

This report compared the information provided by typical classroom achievement tests and improved conventional tests with levels of information provided by adaptive achievement tests measuring the same course material. The evaluation criterion was response pattern information, a measure of information which can be used with data obtained from live test administration. A comparison of results from the computation of response pattern information with theoretical test information indicated that the response pattern information levels were consistently lower than the test information levels for a given set of items. Presumably, this was because testees were not responding exactly as predicted by the item characteristic curve model. However, the shapes of the information curves provided by the two methods of computing information were very similar. This suggests that response pattern information is useful as a substitute for the theoretical test information function; it is easily computed as part of the maximum likelihood scoring procedure, and it reflects the characteristics of live testing data (a characteristic which is useful in empirical research).

As expected, the adaptive testing of classroom achievement yielded substantially more precise estimates of achievement than the conventional classroom achievement tests. This improvement was evident in several tests; it was reflected globally, as well as at all levels of achievement, when test length was taken into account. However, the results indicated that the degree of improvement of the adaptive test over the conventional classroom test depended upon the psychometric characteristics of the conventional test. For example, the comparison of the F1 classroom test with the F1 adaptive test showed a large advantage in favor of the adaptive test, since the items in the classroom test were well distributed throughout the range of achievement. On the other hand, at some levels of $\hat{\theta}$ the F2 classroom test provided higher levels of information than the stradaptive test. In terms of mean information per item, however, the stradaptive test was still superior to the classroom test.

This finding serves to illustrate the possibility that within a restricted range of θ, a conventional test can provide higher levels of information than an adaptive test unless certain precautions are taken in the administration of the adaptive test. One such precaution would be not to administer too few items. That is, in some circumstances the termination rule used in the stradaptive test should be modified to insure administration of a minimum number of items. A better solution, however, would be to continue testing until a pre-specified level of information is reached for every individual (Samejima, 1977). A positive byproduct of this solution would be to insure a high and horizontal information function for the adaptive test, i.e., equal precision of measurement at all levels of achievement.

On the other hand, these data also reflect the dilemma encountered in the construction of fixed-length conventional tests. Such tests can be peaked so that the item difficulties are concentrated in a given region of θ; the result will be a test providing high levels of information in a restricted range of θ and low levels elsewhere. Alternatively, the fixed number of items can be distributed in difficulty over the range of θ (as in the F1 test used in this study); the result is a horizontal, but low, information function. The test constructor cannot construct a conventional achievement test with an information function which is both high and flat, unless an inordinate number of test items is administered. Adaptive testing, however, provides a ready solution to this problem, which is confronted whenever there is considerable variability among students in degrees of achievement resulting from instruction.

Because the classroom tests had not been constructed to be psychometrically optimal, the information provided by the stradaptive tests was compared to that provided by improved conventional tests which were derived from the stradaptive tests' item pools. The improved conventional tests consisted of items with discriminations at least as high as, and typically higher than, those in the adaptive tests and were the same length as the adaptive tests. No response pattern information function was associated with the improved conventional test, since it had not actually been administered. The test information function associated with the improved conventional test was, therefore, compared to the response pattern information function associated with the stradaptive test, at maximum test lengths of 30 and 20 items. Results indicated that the adaptive test yielded generally higher levels of information than the improved conventional test.

These findings indicate that adaptive testing not only was superior to typical achievement classroom tests, but also was superior to a conventional test which was designed to make best use of the same item pool to measure individual differences in achievement levels within a specified range. The adaptive test both provided scores of higher precision and reduced the number of items administered. The conclusion derived from comparison with the improved conventional test is conservative, since response vector information in the present data consistently underestimated test information. In other words, had the improved conventional test been administered and its response pattern information computed, the adaptive test with a maximum length of 20 items would, in all probability, have been found to be substantially more informative.

Contrary to previous research (Jensema, 1975), it was found that an expanded item pool can improve the precision of measurement of scores derived from it by adaptive testing. Jensema's findings were based on a situation in which the items added to the pool were identical, with respect to all three ICC parameters, to the items already in the pool. The results of the present study indicate that even when the added items were only slightly more discriminating, the addition of new items to the adaptive testing pool had a fairly substantial effect, globally, as well as at most levels of achievement, on the precision of measurement of scores derived from the pool.

This investigation has thus shown adaptive achievement testing to be a feasible approach to the measurement of achievement; compared to conventional tests, adaptive achievement testing yields considerably more precise estimates of achievement, even when conventional tests are designed to take maximum advantage of the items in the pool. In order to exploit the advantages of adaptive achievement testing to its fullest, however, it will be necessary to build a closer psychometric interface between instruction and testing. Reduction in testing time by means of adaptive testing is meaningless if the result is solely early dismissal from examinations. Rather, what is needed is to link adaptive testing with an adaptive instructional context, so that reductions in testing time can be used in increased instructional activity.

Atkinson (1976) has described several examples of adaptive computer-based instruction. These systems are adaptive not only because they sequence instruction differently for each student, but also because they differentially allot instructional time to students in order to maximize specified objectives. Differentially allotting instructional time will, in all probability, preserve individual differences in achievement.

This approach to testing and instruction contrasts with the current emphasis on mastery learning and testing (Block, 1971). Mastery testing, along with related approaches, is based on the conception that if instructional time is long enough, every student will attain the same degree of achievement. Although this may be true in principle, an increasing amount of research suggests that individual differences persevere even when instructional time is allowed to vary (Cronbach & Snow, 1977). The implications for instruction and measurement are obvious: An unequivocally useful system of adaptive instruction and achievement testing must be able to consider individual differences rather than attempt to create student homogeneity. It seems that adaptive testing can meet that challenge.

REFERENCES

Atkinson, R. C. Adaptive instructional systems: Some attempts to optimize the learning process. In D. Klahr (Ed.), Cognition and Instruction. Hillsdale, NJ: Erlbaum, 1976.

Bejar, I. I. An investigation of the dichotomous, graded,and continuous response level latent trait models. Unpublished dissertation, University of Minnesota, 1975.

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, 1977.

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A013185)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. AD A027170)

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.

Block, J. H. (Ed.). Mastery Learning: Theory and Practice. New York, NY: Holt, Rinehart,& Winston, 1971.

Brown, J. B., & 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, 1977.

Cronbach, L. J., & Snow, R. E. Aptitudes and Instructional Methods. New York, NY: Halsted Press, 1977.

DeWitt, L. J., & Weiss, D. J. A computer software system for adaptive ability measurement (Research Report 74-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1974. (NTIS No. AD 773961)

Edwards, A. W. F. Likelihood: An account of the statistical concept of likelihood and its application to scientific inference. London: Cambridge University Press, 1972.

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

Green, D. R. The aptitude-achievement distinction. *Proceedings of the Second CTB/McGraw-Hill Conference on Issues in Educational Measurement.* Monterey, CA: CTB/McGraw-Hill, 1974.

Hambleton, R. M., & Traub, R. E. Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 273-281.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing*. Washington, DC: U.S. Civil Service Commission, 1976.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York, NY: Harper & Row, 1970.

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, 1976. (NTIS No. AD A022964)

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

Pine, S. M., & Weiss, D. J. *Effects of item characteristics on test fairness* (Research Report 76-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NTIS No. A035393)

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

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, 1973, *38*, 221-234.

Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, *1*, 223-247.

Sympson, J. B. Problem: Evaluating the results of computerized adaptive testing. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects* (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (NTIS No. AD A018758)

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, 1975. (a) (NTIS No. AD A018758)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1975. (b) (NTIS No. AD A020961)

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

Weiss, D. J. Computerized ability testing: 1972-1975 (Final Report). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976. (NITS No. AD A024516)

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973. (NTIS No. AD 757788)

APPENDIX

Table A
Item Number, Discrimination (*a*), Difficulty (*b*), and Guessing (*c*)
Parameters for Items in the Midquarter 1 Stradaptive Item Pool

Item	*a*	*b*	*c*
Stratum 9			
(15 items)			
3209	2.50	2.29	.29
3417	2.50	3.00	.35
3033	1.54	2.44	.35
3440*	1.52	2.00	.30
3251	2.50	2.39	.35
3406	1.31	2.48	.35
3045	1.02	2.48	.27
3242	.94	2.40	.35
3407	1.02	2.41	.29
3263*	.99	2.29	.35
3241	.91	2.09	.17
3414	.88	2.29	.32
3402	.83	2.44	.35
3247	.82	2.42	.35
3228	.67	2.49	.31
Mean (F)	1.34	2.43	.32
Mean*(W)	1.33	2.39	.32
Stratum 8			
(20 items)			
3409	2.50	1.28	.00
3234	2.50	1.73	.00
3018	.89	1.25	.35
3204	1.14	1.66	.35
3422	1.47	1.50	.35
3411	1.36	1.23	.35
3250	.91	1.94	.29
3206	.74	1.51	.21
3410	1.30	1.34	.31
3429	1.25	1.24	.28
3419	1.23	1.48	.25
3421	1.17	1.15	.35
3436*	1.12	1.59	.35
3271*	.95	1.32	.30
3061*	.95	1.57	.30
3427	.92	1.51	.26
3449*	.91	1.26	.14
3063*	.91	1.51	.35
3074*	.84	1.79	.35
3420	.68	1.62	.35
Mean (F)	1.29	1.46	.26
Mean*(W)	1.19	1.47	.27
Stratum 7			
(20 items)			
3408	2.50	1.05	.31
3437	1.95	.66	.28
3258	1.24	.81	.35
3432	1.72	.67	.35
3048	1.35	.66	.33
3413	1.40	.76	.35
3448*	1.40	.73	.30
3439*	1.36	.64	.32
3219	1.23	.62	.21
3072*	1.02	.65	.32
3277*	1.00	1.04	.35
3035	.90	.68	.28
3433	1.35	.86	.30
3447*	1.18	.93	.32
3064*	.94	.86	.24
3230	.90	.87	.35
3444*	.88	.78	.35
3012	.75	.80	.35
3260	.71	.84	.28
3056*	.71	.89	.26
Mean (F)	1.28	.78	.31
Mean*(W)	1.22	.79	.31
Stratum 6			
(19 items)			
3047	1.66	.44	.29
3079*	1.61	.27	.35
3213	.93	.52	.35
3041	1.51	.23	.35
3062*	1.47	.43	.30
3405	1.40	.55	.32
3445*	1.19	.44	.34
3218	.82	.58	.12
3019	1.31	.29	.29
3207	.70	.46	.28
3431	.70	.28	.34
3000	1.24	.52	.35
3046	1.18	.24	.22
3042	1.15	.37	.27
3050	1.13	.35	.18
3066	1.05	.53	.31
3034	1.01	.37	.28
3262	.81	.47	.35
3438	.70	.21	.27
Mean (F)	1.13	.40	.28
Mean*(W)	1.14	.40	.29
Stratum 5			
(15 items)			
3282*	2.06	-.02	.35
3220	1.79	-.03	.26
3005	1.43	.11	.35
3425	1.36	.17	.23
3053	1.12	.12	.00
3214	1.12	.03	.23
3412	1.12	.19	.35
3051	1.29	.21	.28
3279*	.99	.01	.28
3403	.99	.18	.19
3069*	.88	-.01	.35
3211	.88	.01	.13
3002	.82	.13	.14
3426	.68	.07	.22
3423	.66	.16	.27
Mean (F)	1.11	.11	.22
Mean*(W)	1.15	.09	.24
Stratum 4			
(13 items)			
3256	2.31	-.33	.26
3430	1.15	-.30	.29
3031	1.47	-.33	.35
3254	3.38	-.17	.22
3237	1.54	-.37	.18
3404	.65	-.29	.35
3244	1.35	-.44	.23
3058*	1.05	-.43	.35
3240	.98	-.28	.15
3268*	.97	-.28	.18
3208	.76	-.16	.12
3006	.77	-.37	.33
3259	.69	-.41	.20
Mean (F)	1.27	-.31	.25
Mean*(W)	1.23	-.32	.25
Stratum 3			
(19 items)			
3021	1.96	-.49	.21
3217	1.06	-.48	.14
3052	1.71	-.93	.00
3055*	1.71	-.65	.24
Stratum 3, con t.			
3215	1.59	-.82	.23
3011	1.32	-.86	.20
3435*	.83	-.61	.35
3216	1.27	-.62	.18
3054*	1.29	-.93	.31
3221	1.25	-.52	.17
3049	1.15	-.71	.18
3255	1.14	-.72	.26
3067*	1.07	-.76	.21
3246	1.10	-.72	.28
3022	1.01	-.48	.30
3272*	1.06	-.81	.35
3017	.99	-.58	.16
3076*	.94	-.73	.21
3224	.80	-.50	.37
Mean (F)	1.26	-.65	.20
Mean*(W)	1.22	-.68	.22
Stratum 2			
(20 items)			
3023	2.40	-1.15	.35
3202	1.81	-.99	.21
3415	.85	-.96	.35
3245	1.34	-.96	.21
3236	1.26	-1.20	.33
3020	1.23	-1.28	.17
3028	1.12	-1.26	.35
3226	1.09	-.98	.20
3210	1.04	-1.22	.35
3239	1.04	-1.13	.21
3013	1.00	-.97	.35
3267*	1.02	-1.22	.23
3257	.98	-1.02	.25
3070*	.95	-1.28	.22
3036	.92	-1.18	.16
3014	.86	-1.24	.14
3060*	.86	-1.31	.29
3274*	.85	-1.05	.26
3238	.82	-1.06	.21
3032	.77	-1.06	.27
Mean (F)	1.16	-1.10	.26
Mean*(W)	1.11	-1.13	.26
Stratum 1			
(17 items)			
3077*	2.50	-1.39	.20
3027	1.67	-1.38	.35
3443*	1.07	-1.64	.35
3249	.91	-1.69	.17
3428	.90	-1.56	.35
3073*	1.43	-1.57	.31
3205	1.25	-1.53	.19
3078*	1.24	-1.65	.35
3057*	1.20	-1.35	.26
3065*	1.17	-1.66	.35
3235	1.15	-1.40	.28
3029	1.13	-1.50	.28
3201	1.07	-1.34	.23
3008	.96	-1.75	.18
3252	.79	-1.77	.35
3003	.96	-1.76	.34
3044	.87	-1.42	.15
Mean (F)	1.06	-1.55	.26
Mean*(W)	1.19	-1.55	.28

Note. Items with asterisks are those which were added to the pool Winter quarter. All other items were in the pool both Fall and Winter quarters.

Table B
Item Number, Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Items in the Midquarter 2 Stradaptive Item Pool

Item	*a*	*b*	*c*
Stratum 9			
(18 items)			
3831	2.50	1.96	.06
3690	2.50	2.36	.24
3833*	2.50	2.85	.35
3904	2.45	1.48	.28
3805	2.50	2.38	.35
3698	2.11	2.82	.35
3901	1.55	2.62	.35
3835*	1.21	2.28	.35
3620	2.04	2.97	.35
3697	1.56	3.00	.35
3810	.92	2.20	.27
3664	1.11	1.60	.35
3625	.98	1.66	.35
3622	.95	2.53	.35
3841*	.87	2.13	.35
3651	.95	2.30	.35
3728*	.91	2.55	.35
3712*	.75	1.64	.30
Mean (F)	1.70	2.31	.31
Mean*(W)	1.58	2.30	.32
Stratum 8			
(18 items)			
3615	1.69	1.17	.29
3916	1.39	1.14	.35
3673	1.51	1.11	.31
3804	.95	1.42	.35
3733*	1.24	1.40	.35
3719*	1.18	1.08	.31
3921*	.91	1.23	.29
3827	.87	1.35	.35
3716*	1.14	1.14	.27
3642	1.11	1.11	.24
3902	.73	1.49	.29
3627	1.03	1.07	.35
3681	1.03	1.54	.35
3676	.89	1.51	.25
3644	.88	1.25	.35
3717*	.83	1.25	.35
3670	.80	1.11	.35
3647	.79	1.14	.35
Mean (F)	1.05	1.26	.32
Mean*(W)	1.05	1.25	.32
Stratum 7			
(15 items)			
3743*	2.14	.68	.32
3661	1.90	.68	.32
3674	1.72	.63	.26
3909	1.34	.77	.35
3662	1.54	.93	.27
3654	1.51	.84	.21
3669	1.45	.70	.32
3623	1.42	.74	.31
3912	.95	.70	.19
3734*	.89	.96	.35
3700	.84	.85	.30
3659	1.37	.67	.29
3635	1.17	.66	.35
3612	1.12	.75	.35
3616	.86	.62	.25
Mean (F)	1.32	.73	.29
Mean*(W)	1.35	.75	.30
Stratum 6			
(20 items)			
3707*	1.75	.55	.31
3746*	1.59	.43	.30
3806	1.57	.48	.35
3925*	1.14	.48	.35
3658	1.24	.32	.35
3905	.98	.35	.20
3738*	1.34	.40	.35
3605	1.22	.57	.34
3815	.95	.58	.35
3611	1.22	.39	.32
3675	1.21	.40	.28
3820	.92	.38	.12
3665	1.19	.54	.22
3709*	1.19	.30	.35
3724*	1.14	.37	.30
3819	.76	.53	.35
3918*	.66	.35	.23
3614	.79	.46	.35
3923*	.63	.38	.31
3626	.65	.52	.25
Mean (F)	1.06	.46	.29
Mean*(W)	1.11	.44	.30
Stratum 5			
(15 items)			
3742*	1.89	.27	.35
3745*	1.58	-.07	.20
3720*	1.45	.26	.29
3607	1.38	.09	.35
3811	1.15	.22	.35
3908	1.15	.07	.31
3649	1.32	.11	.22
3632	1.23	.27	.35
3718*	1.22	.16	.33
3629	1.11	-.03	.35
3732*	.96	-.01	.35
3633	.94	-.08	.35
3609	.78	.18	.35
3730*	.75	.01	.10
3618	.64	-.05	.00
Mean (F)	1.08	.09	.29
Mean*(W)	1.17	.09	.28
Stratum 4			
(19 items)			
3744*	1.94	-.35	.30
3708*	1.62	-.20	.16
3631	1.53	-.18	.35
3814	1.26	-.32	.35
3903	1.21	-.43	.31
3671	1.51	-.14	.26
3701	.82	-.15	.35
3643	1.40	-.50	.25
3914	.98	-.39	.16
3693	1.13	-.24	.24
3725*	1.09	-.52	.24
3710*	1.02	-.33	.30
3653	.83	-.51	.33
3660	.78	-.39	.14
3922*	.64	-.26	.30
3606	.71	-.22	.14
3663	.69	-.17	.33
3696	.68	-.35	.00
3656	.63	-.31	.34
Mean(F)	1.01	-.31	.25
Mean*(W)	1.08	-.31	.26
Stratum 3			
(17 items)			
3634	1.79	-.58	.30
3739*	1.68	-.61	.35
3809	1.27	-.61	.35
3924*	1.13	-.79	.18
3672	1.57	-.80	.15
3737*	1.41	-.66	.34
3915	1.08	-.61	.16
3640	1.43	-.69	.35
3906	.87	-.66	.14
3812	.82	-.63	.13
3682	1.33	-.72	.34
3637	1.29	-.73	.28
3636	1.24	-.63	.27
3641	1.20	-.65	.22
3711*	1.05	-.56	.35
3608	1.04	-.78	.16
3705*	.87	-.58	.14
Mean (F)	1.24	-.67	.24
Mean*(W)	1.25	-.66	.25
Stratum 2			
(20 items)			
3735*	1.63	-.94	.35
3648	1.59	-.96	.33
3807	1.52	-1.10	.17
3907	1.43	-1.08	.35
3704*	1.39	-1.13	.23
3655	1.37	-.90	.35
3813	1.20	-.97	.17
3919*	1.30	-.98	.21
3680	1.33	-1.01	.16
3808	.99	-1.00	.30
3686	1.26	-.88	.29
3721*	1.23	-1.20	.22
3821	.90	-.92	.35
3679	1.21	-.94	.17
3685	1.19	-1.01	.16
3668	.97	-.87	.14
3684	.86	-.85	.14
3703*	.83	-1.16	.21
3617	.79	-1.11	.14
3713*	.75	-1.18	.33
Mean (F)	1.19	-.97	.23
Mean*(W)	1.19	-1.01	.24
Stratum 1			
(19 items)			
3741*	1.63	-1.56	.35
3910	1.58	-1.59	.21
3692	1.53	-1.28	.35
3825	1.09	-1.38	.34
3639	1.47	-1.80	.35
3638	1.35	-1.54	.21
3913	1.31	-1.31	.19
3837*	1.09	-1.59	.25
3715*	1.16	-1.63	.26
3920*	1.12	-1.34	.23
3842*	1.01	-1.55	.35
3695	1.09	-1.73	.22
3731*	1.05	-1.67	.35
3832	.99	-1.74	.32
3838*	.99	-1.68	.35
3613	.86	-1.74	.33
3683	.85	-1.31	.14
3657	.81	-1.74	.35
3610	.80	-1.33	.14
Mean (F)	1.14	-1.54	.26
Mean*(W)	1.15	-1.55	.28

Note. Items with asterisks are those which were added to the pool Winter quarter. All other items were in the pool both Fall and Winter quarters.

Table C
Item Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Classroom Tests F1 and F2

F1				F2			
Item No.	*a*	*b*	*c*	Item No.	*a*	*b*	*c*
3060	.86	-1.31	.29	3922	.64	-.26	.30
3067	1.07	-.76	.21	3904	2.45	1.58	.28
3065	1.17	-1.66	.35	3918	.66	.35	.23
3056	.71	.89	.26	3921	.91	1.23	.29
3063	.91	1.51	.35	3919	1.30	-.98	.21
3073	1.43	-1.57	.31	3920	1.12	-1.34	.23
3058	1.05	-.43	.35	3923	.63	.38	.31
3274	.85	-1.05	.26	3924	1.13	-.79	.18
3271	.95	1.32	.30	3801	.80	-.17	.35
3055	1.71	-.65	.24	3841	.87	2.13	.35
3072	1.02	.65	.32	3838	.99	-1.68	.35
3057	1.20	-1.35	.26	3833	2.50	2.85	.35
3064	.94	.86	.24	3837	1.09	-1.59	.25
3069	.88	-.01	.35	3835	1.21	2.28	.35
3054	1.29	-.93	.31	3641	1.20	-.65	.22
3066	1.05	.53	.31	3708	1.62	-.20	.16
3268	.97	-.28	.18	3718	1.22	.16	.33
3267	1.02	-1.22	.23	3728	.91	2.55	.35
3272	1.06	-.81	.35	3665	1.19	.54	.22
3070	.95	-1.28	.22	3730	.75	.01	.10
3008	.96	-1.75	.18	3719	1.18	1.08	.31
3019	1.31	.29	.29	3705	.87	-.58	.14
3062	1.47	.43	.30	3713	.75	-1.18	.33
3061	.95	1.57	.30	3703	.83	-1.16	.21
3262	.81	.47	.35	3709	1.19	.30	.35
3263	.99	2.29	.35	3707	1.75	.55	.31
3447	1.18	.93	.32	3721	1.23	-1.20	.22
3443	1.07	-1.64	.35	3717	.83	1.25	.35
3438	.70	.21	.27	3715	1.16	-1.63	.26
3448	1.40	.73	.30	3716	1.14	1.14	.27
3435	.83	-.61	.35	3720	1.45	.26	.29
3439	1.36	.64	.32	3744	1.94	-.35	.30
3436	1.12	1.59	.35	3745	1.58	-.07	.20
3449	.91	1.26	.14	3746	1.59	.43	.30
3440	1.52	2.00	.30	3711	1.05	-.56	.35
3437	1.95	.66	.28	3710	1.02	-.33	.30
3427	.92	1.51	.26	3724	1.14	.37	.30
3445	1.19	.44	.34	3725	1.09	-.52	.24
3444	.88	.78	.35	3731	1.05	-1.67	.35
				3712	.75	1.64	.30
				3704	1.39	-1.13	.23
Mean	1.09	.11	.29	Mean	1.17	.07	.28

Table D
Item Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameters for Classroom Tests W1 and W2

W1				W2			
Item No.	*a*	*b*	*c*	Item No.	*a*	*b*	*c*
3287	.85	-1.28	.13	3750	.93	-1.79	.34
3292	.68	1.39	.35	3926	.93	-1.56	.16
3219	1.23	.62	.21	3845	1.71	.26	.29
3290	1.16	-.57	.20	3763	1.23	1.95	.28
3214	1.12	.03	.23	3762	1.97	-1.56	.17
3268	.97	-.28	.18	3772	.74	-.84	.35
3289	1.14	-1.45	.35	3759	.99	-.14	.21
3293	.96	-1.30	.14	3768	1.11	-1.55	.17
3291	.65	.52	.35	3756	1.10	-.21	.28
3249	.91	-1.69	.17	3749	1.05	-1.77	.22
3083	1.05	-.90	.13	3757	1.18	-1.60	.18
3090	1.48	-1.65	.18	3755	1.03	-.12	.16
3054	1.29	-.93	.31	3747	1.11	-1.69	.18
3084	1.22	-1.06	.15	3753	.91	-.55	.17
3092	.98	-.65	.15	3654	1.51	.84	.21
3082	1.05	2.27	.35	3673	1.51	1.11	.31
3011	1.32	-.86	.20	3716	1.14	1.14	.27
3095	.79	-1.20	.12	3700	.84	.85	.30
3085	1.16	-1.81	.35	3773	1.69	1.62	.27
3423	.66	.16	.27	3748	.85	1.31	.35
3453	1.19	.48	.22	3766	1.12	1.41	.35
3456	1.03	2.71	.35	3760	1.28	-1.58	.18
3454	1.10	2.66	.35	3758	.89	-1.45	.15
3460	1.99	1.59	.34	3703	.83	-1.16	.21
3452	.75	1.98	.31	3853	1.05	.12	.17
3406	1.31	2.48	.35	3854	1.03	-.19	.31
3461	.94	1.51	.35	3852	.69	-1.78	.35
3457	.90	1.87	.28	3850	.89	1.83	.35
3459	.84	-.29	.26	3851	.76	.18	.23
3407	1.02	2.41	.29	3752	1.24	-.50	.19
3458	1.46	-1.10	.15	3769	1.15	-.39	.16
3432	1.72	.67	.35	3751	.80	1.91	.35
3455	.96	-.61	.31	3770	2.50	1.73	.00
3420	.68	1.62	.35	3622	.95	2.53	.35
3433	1.35	.86	.30	3761	.84	1.27	.32
3412	1.12	.19	.35	3767	1.02	-.04	.30
3462	1.31	-1.03	.17	3930	1.21	-.44	.35
3285	.79	-.60	.11	3904	2.45	1.58	.28
3294	.76	-.68	.19	3918	.66	.35	.23
3041	1.51	.23	.35	3903	1.21	-.43	.31
3091	1.64	.58	.30	3928	1.00	.65	.35
3089	.92	-.37	.30	3929	.96	-1.76	.22
3093	.75	-.94	.11	3813	1.20	-.97	.17
3096	1.48	-1.48	.16	3927	1.01	-1.34	.16
3086	.74	-.67	.35				
Mean	1.09	.08	.25	Mean	1.14	-.06	.25

Table E

Item Discrimination (*a*), Difficulty (*b*), and Guessing (*c*) Parameter Estimates for Items in the Improved Conventional Tests Derived from the Item Pools for Tests W1 and W2

W1				W2			
Item Number	*a*	*b*	*c*	Item Number	*a*	*b*	*c*
3209	2.50	2.29	.29	3831	2.50	1.96	.06
3417	2.67	3.00	.35	3690	2.50	2.36	.24
3033	1.54	2.44	.35	3833	2.50	2.85	.35
3440	1.52	2.00	.30	3904	2.45	1.58	.28
3409	2.50	1.28	.00	3615	1.69	1.17	.29
3234	2.50	1.73	.00	3916	1.39	1.14	.35
3018	.89	1.25	.35	3673	1.51	1.11	.31
3204	1.14	1.66	.35	3804	.95	1.42	.35
3408	2.50	1.05	.31	3743	2.14	.68	.32
3437	1.95	.66	.28	3661	1.90	.68	.32
3258	1.24	.81	.35	3674	1.72	.63	.26
3432	1.72	.67	.35	3909	1.34	.77	.35
3047	1.66	.44	.29	3707	1.75	.55	.31
3079	1.61	.27	.35	3746	1.59	.43	.30
3213	.93	.52	.35	3806	1.57	.48	.35
3282	2.06	-.02	.35	3742	1.89	.27	.35
3220	1.79	-.03	.26	3745	1.58	-.07	.20
3005	1.43	.11	.35	3720	1.45	.26	.29
3256	2.31	-.33	.26	3744	1.94	-.35	.30
3430	1.15	-.30	.29	3708	1.62	-.20	.16
3031	1.47	-.33	.35	3631	1.53	-.18	.35
3021	1.96	-.49	.21	3634	1.79	-.58	.30
3217	1.06	-.48	.14	3739	1.68	-.61	.35
3055	1.71	-.65	.24	3809	1.27	-.61	.35
Mean	1.74	.73	.28	Mean	1.76	.66	.30

Table F
Theoretical Test Information Values for First and Second Classroom Tests for Fall and Winter Quarters

$\hat{\theta}$	Test 1		Test 2	
Midpoint	Fall	Winter	Fall	Winter
-1.90	1.11	1.89	1.13	2.26
-1.70	1.46	2.55	1.54	2.86
-1.50	1.80	3.18	1.97	3.26
-1.30	2.09	3.70	2.39	3.41
-1.10	2.33	4.04	2.77	3.41
-.90	2.52	4.16	3.11	3.37
-.70	2.64	4.09	3.47	3.38
-.50	2.66	3.89	3.86	3.40
-.30	2.63	3.67	4.18	3.42
-.10	2.66	3.49	4.33	3.42
.10	2.83	3.42	4.31	3.39
.30	3.13	3.41	4.18	3.33
.50	3.46	3.40	3.94	3.23
.70	3.63	3.28	3.57	3.15
.90	3.55	3.00	3.13	3.15
1.10	3.25	2.66	2.74	3.35
1.30	2.89	2.38	2.56	3.87
1.50	2.56	2.20	2.54	4.57
1.70	2.28	2.04	2.37	4.79
1.90	1.93	1.84	1.96	4.13

Table G

Mean Test Length and Mean Value of Information at Intervals of $\hat{\theta}$ for 24-Item Improved Conventional Test and Adaptive Test Rescored at Three Maximum Test Lengths for Test W1

	Improved	Adaptive Test					
	Conventional	40 Items Maximum		30 Items Maximum		20 Items Maximum	
$\hat{\theta}$	Test	Mean	Mean	Mean	Mean	Mean	Mean
Midpoint	Information	No. Items	Information	No. Items	Information	No. Items	Information
-1.9	.03	31.3	2.41	27.6	2.41	19.0	1.20
-1.7	.07	29.6	2.19	26.5	1.88	18.8	2.05
-1.5	.15	26.6	3.51	22.8	3.13	15.0	2.74
-1.3	.31	32.5	3.97	25.9	3.19	18.9	3.10
-1.1	.61	28.5	4.44	22.7	4.14	16.9	3.44
-.9	1.12	28.2	4.55	24.5	4.41	17.7	3.82
-.7	1.89	23.4	4.27	21.8	3.89	17.3	3.31
-.5	2.76	29.8	4.70	23.3	4.35	17.4	4.13
-.3	3.46	24.4	5.18	21.4	4.81	17.4	4.22
-.1	3.77	28.3	5.54	24.8	5.01	18.3	4.28
.1	3.77	28.1	5.12	24.2	5.03	18.6	4.07
.3	3.65	29.9	5.74	24.0	4.85	18.0	4.07
.5	3.61	28.4	5.21	26.0	5.00	19.4	4.17
.7	3.76	30.2	5.37	26.0	4.91	19.4	4.18
.9	4.13	29.5	5.63	25.7	5.20	19.3	4.53
1.1	4.52	30.1	6.13	25.9	5.73	19.0	4.78
1.3	4.55	29.8	5.70	22.8	4.88	18.8	4.74
1.5	4.17	33.6	5.79	25.5	4.90	18.4	4.17
1.7	3.65	28.6	5.04	24.7	4.61	16.6	3.86
1.9	3.07	18.6	4.61	18.8	4.29	16.6	3.61
Mean	2.65	28.4	4.93	24.1	4.49	18.1	3.94

Table H

Mean Test Length and Mean Value of Information at Intervals of $\hat{\theta}$ for 24-Item Improved Conventional Test and Adaptive Test Rescored at Three Maximum Test Lengths for Test W2

$\hat{\theta}$ Midpoint	Improved Conventional Test Information	Adaptive Test					
		40 Items Maximum		30 Items Maximum		20 Items Maximum	
		Mean No. Items	Mean Information	Mean No. Items	Mean Information	Mean No. Items	Mean Information
-1.9	.01	17.0	1.99	23.5	1.70	17.0	1.99
-1.7	.03	34.5	2.68	27.6	2.45	20.0	1.59
-1.5	.10	26.7	2.91	22.2	2.77	17.0	2.50
-1.3	.23	24.8	3.64	23.9	3.45	19.0	2.77
-1.1	.51	24.4	3.48	19.0	2.99	16.4	2.90
-.9	.98	33.7	4.32	26.8	3.73	18.9	3.16
-.7	1.64	32.1	4.45	26.6	4.23	19.2	3.66
-.5	2.34	32.0	4.49	26.3	4.36	19.2	3.39
-.3	2.90	25.9	4.55	22.6	4.12	17.8	3.54
-.1	3.28	22.9	3.85	21.1	3.70	18.3	3.45
.1	3.61	26.1	4.64	23.2	4.16	17.9	3.94
.3	4.03	26.0	4.14	22.0	3.90	18.5	3.59
.5	4.44	29.5	4.94	25.4	4.72	19.0	3.91
.7	4.57	31.7	5.60	27.0	5.01	19.1	3.94
.9	4.27	29.4	4.71	24.2	4.41	18.1	3.72
1.1	3.74	30.6	4.12	24.2	3.50	18.4	2.71
1.3	3.35	30.1	3.85	24.3	3.35	17.6	2.95
1.5	3.26	30.2	4.11	24.0	3.35	17.6	2.80
1.7	3.29	31.4	4.33	22.0	4.02	15.5	2.48
1.9	3.18	40.0	4.10	30.0	4.58	20.0	3.46
Mean	2.49	28.6	4.31	24.0	3.96	18.3	3.41

Table I
Mean Information Divided by Mean Number of Items at Levels of $\hat{\theta}$ for the Adaptive Tests

$\hat{\theta}$	Adaptive Test 1		Adaptive Test 2	
Midpoint	Fall	Winter	Fall	Winter
-1.9	.07	.11	.07	.12
-1.7	.08	.08	.10	.06
-1.5	.10	.13	.10	.11
-1.3	.12	.13	.11	.13
-1.1	.15	.13	.12	.14
-.9	.13	.16	.12	.12
-.7	.14	.17	.15	.14
-.5	.17	.16	.11	.13
-.3	.22	.21	.11	.18
-.1	.20	.19	.12	.16
.1	.18	.17	.12	.17
.3	.16	.18	.12	.16
.5	.16	.17	.11	.16
.7	.16	.18	.13	.16
.9	.18	.17	.13	.15
1.1	.20	.18	.13	.13
1.3	.23	.19	.12	.11
1.5	.24	.16	.12	.12
1.7	.15	.18	.13	.11
1.9	.44	.32	.13	.11
Mean	.17	.17	.12	.15