DOCUMENT RESUME

ED 191 882 TM 800 506

AUTHOR Thompson, Janet G.; Weiss, David J.
TITLE Criterion-Related Validity of Adaptive Testing Strategies. Research Report 80-3.
INSTITUTION Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Jun 80
CONTRACT N00014-76-C-0243
NOTE 40p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; College Entrance Examinations; *Computer Assisted Testing; *Correlation; Grade Point Average; Higher Education; Item Banks; Latent Trait Theory; *Multiple Choice Tests; *Predictor Variables; Scoring Formulas; *Test Validity; Vocabulary Skills
IDENTIFIERS *Adaptive Testing; American College Testing Program; Bayesian Adaptive Ability Testing; Peaked Ability Tests; Stradaptive Tests

ABSTRACT

The relative validity of adaptive and conventional testing strategies using non-test variables as one set of external criteria was investigated. A total of 101 college students completed both a variable length stradaptive test and peaked conventional test; a second group of 131 college students completed a variable length Bayesian adaptive test and the same peaked conventional test. All tests were computer-administered and consisted of five-alternative multiple-choice vocabulary items. Test scores were correlated with high school and college grade point averages and American College Testing Service Program subtest scores. The data showed generally higher criterion-related validities for the adaptive tests as compared to the conventional tests. In comparing the two adaptive testing procedures, the data suggested that the stradaptive test scored by mean difficulty methods resulted in more valid ability estimates than the Bayesian adaptive test. Conclusions must be considered tentative until supported by additional research. (RL)

CRITERION-RELATED VALIDITY OF ADAPTIVE TESTING STRATEGIES

Janet G. Thompson
and
David J. Weiss

RESEARCH REPORT 80-3
JUNE 1980

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

This research was supported by funds from the Navy Personnel Research and Development Center, 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 80-3	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Criterion-Related Validity of Adaptive Testing Strategies		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Janet G. Thompson and David J. Weiss		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0243
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.: NR150-382
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE June 1980 13. NUMBER OF PAGES 31
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 Navy Personnel Research and Development Center, and the Office of Naval Research, and monitored by the Office of Naval Research.

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

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

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

Criterion-related validity of two adaptive tests was compared with a conventional test in two groups of college students. Students in Group 1 (N=101) were administered a stradaptive test and a peaked conventional test; students in Group 2 (N=131) were administered a Bayesian adaptive test and the same peaked conventional test. All tests were computer-administered multiple-choice vocabulary tests; items were selected from the same pool, but there was no overlap of items between the adaptive and conventional tests within each group. The stradaptive test item responses were scored using four different

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)

methods (two mean difficulty scores, a Bayesian score, and maximum likelihood) with two different sets of item parameter estimates, to study the effects on criterion-related validity of scoring methods and/or item parameter estimates. Criterion variables were high school and college grade-point averages (GPA), and scores on the American College Testing Program (ACT) achievement tests.

Results indicated generally higher validities for the adaptive tests; at least one method of scoring the stradaptive tests resulted in higher correlations than the conventional test with seven of the eight criterion variables (and equal correlations for the eighth), even though the stradaptive test administered over 25% fewer items, on the average, than did the conventional test. The stradaptive test obtained a significantly higher correlation with overall college GPA (r=.27) than did the conventional test; when math GPA was partialled from overall GPA, the maximum correlation for the stradaptive test with an average length of 29.2 items was r=.51, while the 40-item conventional test correlated only .36. The data showed generally higher criterion-related validities for the mean difficulty scores on the stradaptive test in comparison to the Bayesian and maximum likelihood scores; the different item parameter estimates had no effect on validity, resulting in scores that correlated .98 with each other.

Although the mean length of the Bayesian adaptive test was 48.7 items, the median number of items (35) was less than that of the 40-item conventional test. Ability estimates from this adaptive test also correlated higher with seven of the eight criterion variables than did scores on the conventional tests, although none of the differences were statistically significant.

These data indicate that adaptive tests can achieve criterion-related validities equal to, and in some cases significantly greater than, those obtained by conventional tests while administering up to 27% fewer items, on the average. The data also suggest that latent-trait-based scoring of stradaptive tests may not be optimal with respect to criterion-related validity. Limitations of the study are discussed and suggestions are made for additional research.

CONTENTS

CRITERION-RELATED VALIDITY OF ADAPTIVE TESTING STRATEGIES

Adaptive administration of ability and achievement tests promises considerable improvement in the measurement of individual differences. Some of these advantages were demonstrated in a series of theoretical studies by Lord (e.g., Lord, 1969, 1971a, 1971b) illustrating the potential of adaptive tests for measurement with more equal precision throughout the range of measured ability than was possible with conventional tests of comparable length. Later simulation studies (e.g., Betz & Weiss, 1974, 1975; McBride & Weiss, 1976; Vale & Weiss, 1975b) that further varied the characteristics of item pools used for adaptive tests and conventional comparison tests supported these theoretical results, demonstrating that in comparison to conventional tests, adaptive tests can measure with greater precision for a fixed number of items or with equal precision but using considerably fewer items. This finding has been observed in the measurement of both ability and achievement (e.g., Bejar & Weiss, 1978; Bejar, Weiss, & Gialluca, 1977; Brown & Weiss, 1977; Gialluca & Weiss, 1979).

Early live-testing studies comparing adaptive and conventional tests sought evidence for increased precision of measurement in higher levels of reliability. Because of problems in computing indices of internal consistency for adaptive tests, these studies used test-retest reliability over relatively short time intervals to demonstrate higher levels of precision for adaptive tests. Data supporting this hypothesis were obtained in a number of studies on the measurement of ability (Betz & Weiss, 1973, 1975; Larkin & Weiss, 1974, 1975; Vale & Weiss, 1975) and achievement (e.g., Koch & Reckase, 1979).

Although considerable research has thus been concerned with investigating the increased precision of adaptive versus conventional tests, the validity of adaptive testing procedures has also been of concern. The majority of validation evidence has derived from computer simulation studies. In these studies, true ability (or achievement) level is known, and an item characteristic curve (ICC) model in conjunction with a set of ICC item parameters, a testing strategy, and a scoring method is used to generate an estimated ability level. The estimated ability level can then be correlated with the true, or generated, ability level to yield an index of the validity or fidelity (Green, 1976) of measurement. This correlation indicates how well the true ability level can be recaptured by the combination of item pool, testing strategy, and scoring method. Data from a number of such simulation studies (e.g., Betz & Weiss, 1974, 1975; Urry, 1970; Vale & Weiss, 1975) indicate higher levels of validity for adaptive tests in comparison with conventional tests.

The validity of adaptive tests has also been investigated in terms of correlations of adaptive test scores with scores on conventional tests. Early studies of this type were real-data simulation studies in which the administration of an adaptive test was simulated using a set of item responses obtained from the prior administration of a conventional test; items from the conventional test were "re-administered" to the same testee in an adaptive sequence, and the validity of the procedure was determined by correlation of the score on the adaptive test with the score on the parent conventional test (e.g., Cleary, Linn, & Rock, 1968a, 1968b; Krathwohl & Huyser, 1956). This procedure is not really a demonstration of val-

idity, however, since the obtained correlation is merely a part-whole correlation that will reach a value of 1.0 when the adaptive test administered includes all items in the conventional test.

In other validity studies (e.g., Bayroff & Seeley, 1967; Hansen, 1969) two independent tests measuring the same ability--one adaptive and one conventional--were administered to the same group of testees. The validity of the adaptive test was then evaluated by the correlation of scores on the two tests. Although this approach implements currently accepted definitions of concurrent validity, it is insufficient evidence for the validity of the adaptive procedure. The problem with this method lies in evaluating the appropriate degree of correlation to be expected between the two measurements (Weiss & Betz, 1973). A very high correlation between the two test scores would indicate that the two tests were measuring equivalently; yet a demonstration of equivalent measurement is not a demonstration of the improvement of adaptive testing over conventional testing. If the correlation between scores on the two tests is not very high, however, the question of which procedure is measuring better can be raised. Thus, this approach to studying validity results in an unresolvable dilemma.

As a partial resolution, the relative construct validity of adaptive versus conventional testing strategies has been studied (Bejar & Weiss, 1978). Although this approach is useful, it requires the precise specification of a nomological net for its implementation and may not always result in clearly interpretable results because of the measurement properties of other variables in that net.

For practical applications of adaptive testing, criterion-related validity evidence will be most appropriate. However, the literature to date includes very few criterion-related validity studies. Angoff and Huddleston (1958), using real-data simulations, were the first to study the criterion-related validity of an adaptive test. They examined the correlations with grade-point averages of several two-stage tests in comparison to several conventional tests using items administered to about 6,000 students from the College Entrance Examination Board's Scholastic Aptitude Test. Their results indicated that the narrow-range (peaked) second-stage tests of their simulated two-stage tests had slightly higher validities than did the wide-range (rectangular) conventional tests constructed from the same item pool.

Linn, Rock, and Cleary (1969) also studied the criterion-related validity of adaptive and conventional tests. Their study used scores on the College Board Achievement Tests in American History and English Composition, with the verbal-/mathematics tests of the Preliminary Schclastic Aptitude Test as external criteria. The verbal portion of the School and College Aptitude Tests and the Sequential Tests of Educational Progress were administered to 4,885 testees and then, using real-data simulation techniques, were rescored for approximately two-thirds of the group for whom criterion information was available, using five different adaptive testing procedures. The conventional comparison test was created from the same 190-item pool.

Linn et al. (1969) found that the adaptive tests had higher correlations with the criterion tests than did the conventional tests shortened to the length of the adaptive tests. This study had the limitation of using a simulated adaptive testing administration mode rather than live adaptive administration. This makes it difficult to generalize the results to testees actually taking adaptive tests where interaction effects may exist between testee response, item selection, and item

order. Also, this study was confounded by item overlap between the conventional and adaptive tests.

Waters (1974, 1976), in his adaptive test validation study, also correlated scores on adaptive and conventional tests with another test, which served as an external criterion. His criterion was the Florida 12th Grade Verbal Test scores. Waters divided his testee population into six groups: One group of 55 testees was administered a stradaptive test (Weiss, 1973), and five smaller groups (N = 8, 7, 9, 13, and 10) were each given a different conventional test. One-fifth of the items on the stradaptive test were the same as those on the conventional tests. Although the scores for the five conventional subtests were different, they were normalized and pooled for comparison with stradaptive results.

Waters found restriction in the range of ability level for his sample: Most testees tended to be at the high end of the continuum. His results indicated that none of the stradaptive validity coefficients were significantly different from the conventional test validities; the results did show, however, that the shorter stradaptive test proved more reliable than the longer conventional test. Thus, with fewer items administered, the stradaptive test produced validity coefficients comparable to conventional test validity coefficients.

The Angoff and Huddleston (1958), Linn et al. (1969), and Waters (1974, 1976) studies were all criterion-related validity studies. The Angoff and Huddleston (1958) and the Linn et al. (1969) studies were limited by the tests being scored as if they were administered adaptively, introducing limitations created by the simulation approach, and by some of the same items being used in both tests. Waters' (1974, 1976) study eliminated one of these problems: He used live adaptive testing and did not give the same subjects both the adaptive test and the conventional test, even though one-fifth of the items were common between the two tests. However, since his study was an independent groups design in which the adaptive and conventional tests were administered to different groups of testees, he may have introduced sample-specific error into his research design, particularly because of the relatively small sample sizes used. An additional problem in Waters' study results from the pooling of data from the five conventional subtests given to five different groups of testees and the comparison of the pooled score distributions with the adaptive test score distribution.

A problem characteristic of both the Linn et al. (1969) and the Waters (1974, 1976) studies was the use of scores on a conventional test as an external criterion. Since one of the predictors was also a conventional test, this could have introduced method variance in the correlation of the conventional predictor test scores with the conventional criterion test scores, thus conceivably inflating these validity coefficients. If such method variance was present, it would not have similarly inflated the validity coefficients for the adaptive tests, possibly masking gains in relative validity due to adaptive testing. The Angoff and Huddleston (1958) study, however, used grade-point average as the criterion but did not use actual adaptive test administration.

Purpose

The present study was designed to investigate the relative validity of adaptive and conventional testing strategies using non-test variables as one set of external criteria. The study was similar to Waters' (1974, 1976) study in that the adaptive tests were computer-administered; it was similar to the Linn et al. (1969)

study in that each group of testees took both an adaptive and a conventional test, but there was no overlap in the item pools used for the two testing strategies.

METHOD

Two adaptive testing strategies were compared to a conventional ability test in terms of criterion-related validity for two separate groups of students. In one group students completed both a variable length stradaptive test and a peaked conventional test; in the second group students completed a variable length Bayesian adaptive test (Owen, 1975) and the same peaked conventional test. All tests were computer-administered and consisted of five-alternative multiple-choice vocabulary items. Test scores from each of the tests were correlated with high school grade-point average, University of Minnesota grade-point average, and scores on the American College Testing Program subtests.

Subjects and Data Collection

Group 1 testees were administered the stradaptive test and the conventional test. Volunteer testees were college students attending classes at the University of Minnesota. Most were juniors, seniors, or graduate students enrolled in psychology courses at the time of testing. A total of 101 students had usable data for this study. Data were collected during the winter (51.5%) and spring (48.5%) quarters of 1973. All students were given the conventional test followed by the stradaptive test or vice versa. The order in which the tests were given was alternated to control for sequence effects. Both tests were given in a single administration.

Students in Group 2 were administered the Bayesian adaptive test and the conventional test. Forty-three percent of the students in this group were given the tests during spring quarter of 1973; the other 57% were administered the test during winter quarter of 1974. As in Group 1 all testees were college student volunteers attending classes at the University of Minnesota; most were juniors, seniors, or graduate students enrolled in psychology courses at the time of testing. A total of 131 subjects had usable data. Testees were alternately given the conventional test followed by the Bayesian adaptive test or vice versa.

All items given were multiple-choice vocabulary items selected from the same item pool (McBride & Weiss, 1974). Item pools for the stradaptive and Bayesian tests utilized a subpool that excluded the 40 items in the conventional test. All tests were presented using cathode-ray-terminals (CRTs) acoustically coupled to a time-shared computer. Items were presented with a number representing the correct alternative; testees answered by typing the number of their choice. If testees did not know the answer and did not wish to guess, they were instructed to respond with a question mark. Items answered with a question mark were scored as incorrect. Tests were preceded by instructions on how to use the CRT; basic biographical data were also collected on the CRT prior to test administration (see DeWitt & Weiss, 1974).

Testing Strategies

Stradaptive Test

Item branching. The stradaptive test item pool consisted of 141 items strati-

fied into 9 strata, or peaked item pools, each varying in level of difficulty. Stratum 9 contained items of the highest difficulty level, and Stratum 1 included items of the lowest difficulty level. Entry points for selection of the first item to be administered to a testee were based on the student's reported grade-point average (GPA), as shown in Figure 1. Following entry into the stradaptive structure, an up-one, down-one branching rule was used. That is, a testee was administered the next unadministered item from the next lower stratum, or difficulty level, following an incorrect answer or the next unadministered item from the next higher stratum, or difficulty level, following a correct answer. Question mark responses, which were treated as incorrect responses, caused the testee to be branched to the next easier stratum.

Figure 1
Stradaptive Test Entry Point Question

IN WHICH CATEGORY IS YOUR CUMULATIVE GPA TO DATE?	Entry Stratum (Not Seen by Testee)
1. 3.76 to 4.00	9
2. 3.51 to 3.75	8
3. 3.26 to 3.50	7
4. 3.01 to 3.25	6
5. 2.76 to 3.00	5
6. 2.51 to 2.75	4
7. 2.26 to 2.50	3
8. 2.01 to 2.25	2
9. 2.00 or less	1

ENTER THE CATEGORY (1 THROUGH 9) AND PRESS THE "RETURN" KEY.

The stradaptive test was variable length. Testing was terminated when a ceiling stratum was identified for a testee (Weiss, 1973). The ceiling stratum was identified as the stratum in which the proportion of correct responses made by the testee was .20 or less, following the administration of five items in that stratum. This is the proportion of correct answers expected as a result of random guessing on five-alternative multiple-choice items. If a ceiling stratum was not identified after 75 items had been administered, testing was terminated.

Item pool. Appendix Table A shows the item pool used for the stradaptive test. Strata included from a minimum of 10 items in Stratum 9, the most difficult stratum, to a maximum of 36 items in Stratum 2. The item pool was structured and item selection implemented using a set of item characteristic curve (ICC) item parameters available at the time that tests were administered; these are referred to in Table A as original parameters. As described by Prestwood and Weiss (1977), these parameters were later recalculated for scoring purposes. All ICC item parameter estimates were based on conversions of the classical difficulty and discrimination parameters to the ICC metric, as described by McBride and Weiss (1974) and Prestwood and Weiss (1977). ICC lower asymptote (c, or guessing) parameters were set at .20 for all items.

Scoring. The stradaptive test was scored by a number of different scoring methods in order to compare the relative validity of different ways of scoring the

same pattern of item responses. Scoring methods that used the ICC item parameters were applied using both the original and revised item parameters to determine the effects of the item parameter revision on score validity.

Stradaptive test responses were scored for ability level with two scoring methods that used only some of the information in the ICC item parameters. The mean difficulty of all items administered (Mean Difficulty Administered) score was expected to provide more stable ability estimates because it used difficulty information from all the items administered to a testee. A potential deficiency of this score is that it is affected by inappropriate entry points. For example, if a testee begins the test with items from a stratum of much higher difficulty level than his/her ability, he or she will have taken more unnecessarily difficult items than if the test had been begun with items of appropriate difficulty. Thus, the Mean Difficulty Administered score would be higher than warranted for the testee. To eliminate this problem, the mean difficulty of items answered correctly (Mean Difficulty Correct) score was also computed. This score does not take into account spuriously administered items of high difficulty unless they are answered correctly. One potential disadvantage, however, is that it ignores information from items not answered correctly.

ICC-based scoring methods (Bejar & Weiss, 1979), which utilize not only the testee's entire response pattern but also the difficulties, discriminations, and guessing parameters of all the items administered to a testee, should provide optimal scoring of any response pattern. To compare the relative validity of these scoring methods, both Maximum Likelihood and Owen's (1975) Bayesian scoring methods were used to score the stradaptive test item responses. Bejar and Weiss (1979) have provided descriptions and computer programs for these scoring methods.

A problem characteristic of Maximum Likelihood scoring is that a score cannot be determined for testees who answer every item correctly, who answer every item incorrectly, or who have very unusual response patterns (e.g., answering many difficult items correctly and many easy items incorrectly). In these cases the estimation procedure fails to converge, i.e., it converges on plus or minus infinity (Kingsbury & Weiss, 1979). In the stradaptive data, two testees had item response patterns that failed to converge using the Maximum Likelihood scoring procedure. Their test scores, derived from this procedure, were deleted from the data analyses.

The preceding four scores are all "point estimates" of ability level (Weiss, 1973). However, as Trabin and Weiss (1979) have shown, there is additional information in test item response patterns beyond these point estimates. An individual whose response pattern fluctuates between several strata is a more inconsistent responder than one who is administered items from only a few strata adjacent to one another. Consistency among scores indicates either the stability of a testee's ability estimate (Weiss, 1973, p. 26) or the testee's fit to the ICC model. In this study the standard deviation of item difficulties of all items administered (SD Administered) was used as one consistency score. This score was chosen from among the available types of consistency scores to reflect the dispersion of the difficulties of all items administered, not just those items that were answered correctly, in order to make more complete use of the item response patterns available. In addition, the standard error of Owen's Bayesian score (SE Owen's Bayesian) was used as a second consistency score.

Bayesian Adaptive Test

A variable length adaptive test based on Owen's (1975, McBride & Weiss, 197 Bayesian adaptive testing strategy was administered to all testees in Group 2. T item pool for this test consisted of 200 items selected from a larger pool (McBri & Weiss, 1974) after the conventional test items were excluded. Items in the poo ranged in difficulty from $\underline{b}$=-3.19 to $\underline{b}$=2.95; all items had $\underline{a}$ values of .40 or greater (see Appendix Table B). Items were selected and scored using only the original item parameters.

The Bayesian adaptive test was begun with differential prior ability estimat ($\hat{\theta}$), as shown in Figure 2. The prior $\hat{\theta}$s shown in Figure 2 for each of the levels of student-reported grade-point average (GPA) were chosen to reflect a positive level of correlation between GPA and vocabulary ability as measured by the adapti test; the relatively lower $\hat{\theta}$ values for higher GPAs were designed to take into account chance successes resulting from guessing. The relatively large variances of the prior $\hat{\theta}$ values were chosen to reflect a high degree of uncertainty about t prior ability estimates, so as not to assume a very high positive correlation be- tween GPA and vocabulary ability. Testing was terminated either when the varianc of the posterior ability estimate was .09 or less, reflecting a standard error of of .03 or less, or when a maximum of 135 items had been administered.

Figure 2
Bayesian Test Entry Point Question

IN WHICH CATEGORY IS YOUR CUMULATIVE GPA TO DATE?	Initial Values Set for Bayesian Ability Estimate ($\hat{\theta}$) and Variance of $\hat{\theta}$ (Not Seen By Testee)	
	$\hat{\theta}$	Variance of $\hat{\theta}$
1. 3.76 to 4.00	1.23	3.5
2. 3.51 to 3.75	.77	3.0
3. 3.26 to 3.50	.50	2.5
4. 3.01 to 3.25	.18	2.0
5. 2.76 to 3.00	.09	2.0
6. 2.51 to 2.75	-.31	2.5
7. 2.26 to 2.50	-.56	3.0
8. 2.01 to 2.25	-.85	3.5
9. 2.00 or less	-1.41	4.0

ENTER THE CATEGORY (1 THROUGH 9) AND PRESS THE "RETURN" KEY.

Conventional Test

The same 40-item peaked conventional test was administered to the groups of students who took the stradaptive and Bayesian tests. Items were selected based o a proportion correct of about .60, in order to adjust the average difficulty of th items for guessing and high biserial correlations with total score.

Appendix Table C shows the ICC item discrimination and difficulty parameter estimates for items in the conventional test. The standard deviation of the item difficulties for this test was .11, which was considerably lower than those of

either the stradaptive or Bayesian test item pools. The average item discrimination of the stradaptive pool ($\underline{a}$=.745 for the original parameters) was slightly higher than that of the conventional test ($\underline{a}$=.543), as was the average discrimination of the Bayesian pool ($\underline{a}$=.796). The conventional test was scored by counting the number of correct answers (Number Correct score); omitted answers were scored as incorrect.

Criterion Variables

Because the tests being investigated were verbal ability tests, the criterion variables were chosen to reflect this ability. Four different variables were obtained from student records, but not all variables could be obtained for every student in the two groups:

1. High school GPA (HS-GPA);
2. University of Minnesota overall GPA (UM-OGPA);
3. University of Minnesota math GPA (UM-MGPA), which was used to partial out the effects of numerical ability resulting in a partial GPA (UM-PGPA); and
4. American College Testing Program (ACT) test scores.

All GPAs were calculated by assigning the following numerical values to letter grades: A=4, B=3, C=2, D=1. HS-GPA was calculated as the overall GPA of the students when they were sophomores through seniors in high school; UM-OGPA was computed as the overall college GPA of the students through the spring of 1976; and UM-MGPA was derived from the GPA of all math classes taken by the students at the University of Minnesota.

The ACT battery was administered to the students in either their junior or senior years of high school. The test is designed to measure a student's ability to perform "typical intellectual tasks asked of college students." The ACT resulted in five scores: English, mathematics, social science, natural science, and a composite score.

Data for two of the criterion variables were available prior to test administration (HS-GPA and ACT scores). Data for the other two criteria were gathered after the students had taken the conventional and adaptive tests.

Data Analysis

Comparison of the Adaptive and Conventional Tests

The adaptive and conventional tests were designed to compare the respective criterion-related validities of the testing strategies against the four external criteria. Comparative validity assessments were of specific interest. Predictor variables used were the ability estimates from both adaptive tests and the conventional test. Consequently, Pearson product-moment correlations were calculated between ability estimates derived from the adaptive tests and the four external criteria and between the conventional test and these four measures.

In addition, the mean, median, standard deviation, skewness, and kurtosis were calculated for all predictor variables and the criterion variables. Although ability estimates derived from the different test administration strategies and scoring methods could not be evaluated on how closely they reflected the true underlying ability distribution because this distribution was not known for the testees, these

data provided a relative comparison of how the different testing strategies and scoring methods described the individual differences among the students tested.

Correlations between Stradaptive and Conventional Test Scores

To determine whether the adaptive and conventional tests were measuring the same ability, ability estimates from the adaptive tests were correlated with scores from the 40-item conventional test for all examinees who completed both tests. Correlations were calculated using both original and revised item parameters for all stradaptive scoring methods.

These data also provided intercorrelations among scores on the stradaptive test for both the original and revised item parameters. This comparison provided information on the effects of using the original item parameters. Correlations of these scores with the criterion variables also permitted evaluation of the effect of the different item parameter estimates on criterion-related validity.

Test Length versus Ability

Ability estimates from both the Bayesian and stradaptive tests were correlated with test length. For the stradaptive test this analysis was performed to determine if the scoring method interacted with item pool characteristics, resulting in different correlations for the various scores and test lengths. These correlations were also computed for scores derived from the two different sets of item parameters.

RESULTS

Characteristics of Score Distributions

Table 1 shows descriptive statistics for scores for all tests administered in both groups.

Conventional test. The 40-item conventional test performed almost identically in both groups; there was no significant difference in the mean test scores for the two groups. The average number-correct scores (Number Correct) were 22.60 and 22.82, with standard deviations of 8.33 and 9.01 in Group 1 and Group 2, respectively. These mean scores were very close to the predicted means for the group on which the test was constructed. Neither score distribution was significantly skewed, although both distributions were significantly platykurtic, indicating a flatness in the scores in comparison to a normal distribution.

Stradaptive test. The stradaptive test administered an average of 29.29 items, with a median of 21. The distribution of number of items administered (Number Administered) was significantly positively skewed and leptokurtic, indicating a distribution that was more peaked than a normal distribution, with a few very long test lengths. The distribution of number-correct scores (Number Correct) for the stradaptive test was skewed similarly to that of Number Administered but with a mean of 14.90 and a median of 11.20. Both the means and medians indicate that, on the average, the stradaptive test functioned almost optimally, administering to the average student items that were answered correctly about 50% of the time. The average Number Administered in the stradaptive test was 25% lower than the 40-item length of the conventional test.

Table 1
Descriptive Statistics for Scores from Conventional, Stradaptive, and Bayesian Adaptive Tests

Test and Score	N	Mean	Median	SD	Skew	Kurtosis
Conventional Test						
Number Correct						
Group 1	100	22.60	21.50	8.33	.13	-1.08*
Group 2	131	22.82	22.60	9.01	.04	-1.09*
Stradaptive Test (Group 1)						
Number Administered	101	29.29	21.00	24.03	2.50**	7.08**
Number Correct	101	14.90	11.20	12.04	2.31**	6.58**
Original Item Parameters						
Mean Difficulty Administered	101	.26	.17	1.00	.15	-.71
Mean Difficulty Correct	101	-.10	-.18	1.04	.28	-.62
Owen's Bayesian	101	-.18	-.30	.94	.31	.17
Maximum Likelihood	100	-.05	-.30	1.14	.81**	.78
SD Administered	101	.73	.72	1.19	.68**	1.31
SE Owen's Bayesian	101	.41	.39	.15	1.28**	2.52**
Revised Item Parameters						
Mean Difficulty Administered	101	.68	.57	1.10	.16	-.75
Mean Difficulty Correct	101	.26	.17	1.12	.31	-.58
Owen's Bayesian	101	.23	.12	1.08	.41*	.05
Maximum Likelihood	99	.30	.20	1.11	.49*	.08
SD Administered	101	.84	.80	.23	.47*	.28
SE Owen's Bayesian	101	.32	.29	.21	4.47**	23.86**
Bayesian Adaptive Test (Group 2)						
Number Administered	131	48.75	35.00	29.71	.90**	-.04
Number Correct	131	25.56	16.42	19.36	1.83**	4.03**
Bayesian Ability Estimate	131	.36	.06	1.17	.34	-.62
Variance of Ability Estimate	131	.08	.08	.02	6.78**	48.04**

*Statistically different from zero at $p < .05$.
**Statistically different from zero at $p < .01$.

Mean ability scores using the original item parameters were similar for Mean Difficulty Correct (-.10), Maximum Likelihood (-.05), and Owen's Bayesian (-.18) scoring methods; as expected, the average Mean Difficulty Administered scores were different from the other scores, due to some inappropriately high entry point estimates. Owen's Bayesian score resulted in the lowest mean ability estimate (-.18); median ability estimates for Owen's Bayesian and Maximum Likelihood scores were identical (-.30). All ability estimate distributions were positively skewed, although only the Maximum Likelihood score was significantly skewed. The distributions of the two latent-trait-based scores were leptokurtic, whereas the mean difficulty scores were platykurtic; however, none of these kurtosis values were significantly different from a normal distribution. In contrast to Number Correct from the conventional test, three of the four stradaptive ability scores using the original item parameters better approximated a normal distribution. Both the SD Administered and SE Owen's Bayesian scores resulted in positively skewed and peaked distributions.

Using the revised item parameters, the four stradaptive ability scores showed nearly equal standard deviations and positive skew. Owen's Bayesian score and the

Maximum Likelihood score had significant positive skew ($p < .05$). The mean difficulty scores were platykurtic, but not significantly so, whereas the Bayesian and Maximum Likelihood estimates did not deviate from normal kurtosis. All medians of the ability estimates were smaller than their corresponding means. Again, the Mean Difficulty Administered score had a higher mean (and median) than did the other three ability scores. The SD Administered score and the SE Owen's Bayesian score had similar distributions with the revised parameters as they did with the original item parameters. Both means and medians of all scores computed using the revised item parameters were consistently higher than they were using the original item parameters.

Bayesian adaptive test. Mean test length for the Bayesian adaptive test was 48.75 items, an increase of 8.75 items (22%) over the length of the 40-item conventional test. The median test length for this test, however, was 35 items, a 12.5% reduction from the conventional test length. Thus, some of the Bayesian adaptive tests were quite long, resulting in a positively skewed distribution of Number Administered (50 students answered more than 50 items, and 19 students answered more than 80 items). These long test lengths were probably due to the large prior variances used in selecting the first item for the Bayesian test in conjunction with the small posterior variance used to terminate the test. Both the mean and median of the Number Correct in the Bayesian test (25.56 and 16.42, respectively) show that the Bayesian test operated properly in administering items at a difficulty level so that about 50% of the items administered were answered correctly.

The Bayesian ability estimates were distributed normally with slight, but non-significant, platykurtosis. The variance of the ability estimates had a very peaked distribution, with a significant positive skew.

Criterion Variable Distributions

Table 2 presents descriptive statistics for the criterion variables for both groups. The means for HS-GPA in both groups were higher than means of either UM-OGPA or UM-MGPA, which were nearly equal both within groups and between groups. The distributions of HS-GPA and UM-OGPA had significant negative skew in Group 1, but skew was not significant in Group 2. None of the GPA distributions differed significantly from normality in terms of kurtosis in either group, although there was a slight tendency toward platykurtosis. The standard deviations for all GPAs were very similar.

ACT mean scores ranged from 22.00 to 26.61 and were essentially equivalent for the two groups. Standard deviations varied from 3.52 to 6.47 and were also comparable for the two groups. All ACT scores were negatively skewed, with several significantly so. There was a general tendency for ACT scores to be leptokurtically distributed, although most did not differ significantly from normal in terms of kurtosis. None of the differences in mean scores between the two groups on any of the criterion variables were statistically significant ($p < .05$).

Test Score Correlations

Stradaptive and conventional tests. Product-moment intercorrelations among the four stradaptive ability estimates and the corresponding consistency scores are shown in Table 3. Intercorrelations are shown between scores derived from the original item parameters and the revised item parameters of the stradaptive test, and with Number Correct on the conventional test. Also included are the students'

Table 2
Descriptive Statistics for Criterion Variables

Group and Criterion	N	Mean	Median	SD	Skew	Kurtosis
Group 1						
HS-GPA	56	3.12	3.15	.68	-.72*	-.21
UM-MGPA	77	2.81	3.00	.83	-.41	-.62
UM-OGPA	101	2.80	2.90	.73	-.76**	.37
ACT Score						
English	55	22.00	21.95	3.52	-.33	.26
Mathematics	55	25.98	27.25	6.47	-.91**	.35
Social Science	55	24.93	25.42	4.50	-.82**	1.43*
Natural Science	55	25.42	25.57	5.76	-.51	-.83
Composite	55	24.76	25.00	4.30	-.46	-.50
Group 2						
HS-GPA	71	3.17	3.14	.55	-.49	.01
UM-MGPA	104	2.71	2.67	.76	-.08	-.39
UM-OGPA	131	2.81	2.83	.60	-.22	-.47
ACT Score						
English	72	22.08	22.30	4.23	-.21	1.76
Mathematics	71	26.10	26.89	5.41	-.78	.47
Social Science	71	24.79	26.00	5.04	-1.11**	.72
Natural Science	71	26.61	27.91	5.00	-1.55**	3.18**
Composite	71	24.99	25.44	3.93	-.77	.25

*Statistically different from zero at $p < .05$.
**Statistically different from zero at $p < .01$.

reported GPAs used as an entry point to the stradaptive test, and Number Administered and Number Correct in the stradaptive test.

Although there were nonsignificant correlations between the entry point and Number Administered and Number Correct, the latter two variables correlated .97. This high correlation resulted from the lack of very difficult items in the stradaptive test (e.g., Stratum 9, the most difficult stratum had only 10 items), which resulted in the inability of the test to locate a ceiling stratum for students with very high ability. Thus, for these students, the test would continue administering items that were answered correctly.

Using both the original and revised item parameters, the entry point variable (reported GPA) had moderate and significant correlations with all ability scores; the lowest were $r = .31$ and .26 with Owen's Bayesian score for the original and revised parameters, respectively. Entry point data correlated highest ($r = .45$ and .46) with the Mean Difficulty Administered score. Although the entry point data correlated nonsignificantly with the SD Administered consistency score, the SE Owen's Bayesian consistency score correlated significantly ($r = .33$ and .44) with entry point data. This latter result, however, is likely a result of the same factors that resulted in the correlation of .97 between Number Correct and Number Administered. Stradaptive entry point data also correlated $r = .34$ with Number Correct on the conventional test, whereas neither Number Administered nor Number Correct in the stradaptive test correlated significantly with Number Correct on the conventional test.

Table 3

Intercorrelations of Scores from Stradaptive and Conventional Tests (N=101)

Test and Score	Score 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Stradaptive Test															
1. Entry Point (Reported GPA)															
2. Number Administered	-.11														
3. Number Correct	-.18	.97													
Original Item Parameters															
4. Mean Difficulty Administered	.46	-.07	.04												
5. Mean Difficulty Correct	.43	-.06	.06	1.00											
6. Owen's Bayesian	.31	-.09	.04	.96	.97										
7. Maximum Likelihood	.34	-.05	.04	.96	.96	1.00									
8. SD Administered	.01	.47	.50	.11	.09	.06	.01								
9. SE Owen's Bayesian	.33	-.34	-.30	.75	.75	.73	.78	-.41							
Revised Item Parameters															
10. Mean Difficulty Administered	.45	-.07	.04	1.00	.99	.95	.95	.11	.75						
11. Mean Difficulty Correct	.42	-.06	.05	1.00	1.00	.96	.96	.09	.76	.99					
12. Owen's Bayesian	.26	-.11	.04	.96	.97	.98	.97	.07	.74	.96	.97				
13. Maximum Likelihood	.33	-.06	.03	.96	.96	.98	.98	.03	.78	.96	.96	.97			
14. SD Administered	-.15	.58	.61	.33	.31	.26	.21	.94	-.19	.33	.30	.27	.23		
15. SE Owen's Bayesian	.44	-.52	-.45	.38	.39	.36	.44	.40	.72	.38	.40	.38	.38	-.32	
Conventional Test															
16. Number Correct	.34	-.07	.03	.85	.85	.82	.80	.16	.61	.84	.85	.82	.79	.35	.31

Correlations > ±.30 are significant at $p < .001$; > ±.23 are significant at $p < .01$; > ±.16 are significant at $p < .05$.

For both the original and revised item parameters, all stradaptive ability estimates correlated .96 or higher. Mean Difficulty Correct correlated .97 with Owen's Bayesian score in both cases and .96 with the Maximum Likelihood score; Mean Difficulty Administered correlated .96 with these two scores in both cases, and Owen's Bayesian and Maximum Likelihood scores correlated 1.00 and .97. These results show that the simple average difficulty scores ordered students almost identically with the more complex latent-trait based scores.

The only obvious effect of revising the item parameter estimates was on the correlations of the consistency scores with the ability scores. Using the original item parameters, the SD Administered score correlated nonsignificantly with all ability scores, and the SE Owen's Bayesian score correlated from .73 to .78 with ability scores. For these same variables, using the revised item parameters, both the SD Administered and SE Owen's Bayesian scores correlated significantly with the ability scores, but correlations ranged only from .23 to .40. The effect of the revised parameter estimates on these two consistency scores is seen in the correlation of .94 between original and revised parameter estimates for the SD Administered score, whereas the relevant correlation for the SE Owen's Bayesian score was only .72.

Revision of the item parameter estimates had no important effect on the ability scores. Intercorrelations of ability estimates using the two sets of item parameter estimates ranged from .95 to 1.00; correlations computed between the same ability score using the two sets of item parameter estimates were .98 or 1.00. These correlations were as high as the intercorrelations of different types of ability estimates using a common set of item parameters.

Convergent validity of the stradaptive ability scores is indicated by their relatively high correlations with the conventional test. These correlations, which were not affected by use of the different item parameter estimates, ranged from .79 to .85, with a tendency for the non-latent-trait-based scores to correlate higher with conventional test scores than did the scores using latent trait scoring methods. Correlations of the consistency scores with conventional test scores differed for the two kinds of item parameter estimates.

Bayesian and conventional tests. Product-moment correlations of scores from the Bayesian adaptive test and the conventional test are shown in Table 4. Number Administered in the Bayesian test correlated highest (r=.90) with Number Correct in that test. This resulted from a lack of highly discriminating items of high difficulty in the Bayesian item pool, similar to the correlation of the same variables in the stradaptive test. Therefore, more items of low discrimination were necessary to reach the fixed posterior variance termination criterion for high ability students than for low ability students, for whom more highly discriminating items were available. This is further supported by the correlation between Number Correct and the Bayesian ability estimate (r=.89) and between the Bayesian ability estimate and Number Administered (r=.84). A high and significant correlation (r=.85) was observed between the Bayesian ability estimate and the conventional test Number Correct score, indicating that they were both measuring the same trait. Bayesian test length (which, because of its high correlation with the Bayesian ability estimate, essentially measured ability level) correlated moderately (r=.59) with Number Correct on the conventional test, whereas Number Correct on the two tests correlated .72. The variance of the Bayesian ability estimate, which was essentially fixed for all but the very high ability testees (for whom there were not suffi-

ciently discriminating items available), correlated essentially zero with all variables.

Table 4
Intercorrelations of Bayesian Adaptive and Conventional Test Scores (N=131)

	Score			
Test and Score	1	2	3	4
Bayesian Adaptive Test				
1. Number Administered				
2. Number Correct	.90			
3. Ability Estimate	.84	.89		
4. Variance of Ability Estimate	-.07	-.08	.18	
Conventional Test				
Number Correct	.59	.72	.85	.10

Note. Correlations >.28 significant at $p < .001$; >.17 significant at $p < .05$.

Intercorrelations of Criterion Variables

Table 5 shows the intercorrelations of the three GPA variables and the five ACT scores for the two groups. As expected, the highest intercorrelations with each group were between the four subscores of the ACT and the ACT composite. HS-GPA was most highly correlated with the ACT math score in Group 1 and with UM-OGPA in Group 2, and UM-OGPA was most highly correlated with the ACT composite score in both groups. UM-MGPA correlated highest with the ACT social science score, and UM-OGPA correlated highest with the ACT composite score (r=.43 and .51) for Group 1. Both UM-MGPA and UM-OGPA correlated highest among the ACT scores with the ACT composite (r=.33 and .53) for Group 2. The three GPA measures appear to have provided different criterion information than the ACT scores, whereas the ACT composite score provided much of the same information as the four ACT subscores from which it was derived (r=.78 to .89 in Group 1 and .74 to .86 in Group 2).

Correlations of Test Scores and Criterion Variables

Stradaptive versus conventional. Table 6 shows the validity correlations for the stradaptive and conventional testing strategies. For all three GPA variables, the best predictor was reported college GPA, the stradaptive entry point information. In the prediction of HS-GPA, the conventional test Number Correct score correlated .40 and the stradaptive ability scores correlated from .41 to .45, with essentially no difference between scores derived from the two sets of item parameter estimates. Using both sets of item parameter estimates, Mean Difficulty Administered achieved the highest validity. In predicting UM-MGPA, Number Correct on the conventional test correlated .31, and the best of the adaptive scores (Mean Difficulty Administered, using the revised parameters) correlated .32. Again, the Mean Difficulty Administered score obtained the highest correlation among the stradaptive scoring methods, closely followed by Mean Difficulty Correct; the two latent-trait-based scoring methods—Bayesian and Maximum Likelihood—resulted in lower validities.

Table 5
Intercorrelations of Criterion Variables For Both Groups

Group and Criterion Variable	N	Criterion Variable: GPA HS	GPA UM-M	GPA UM-O	ACT Score: English	Math	Social Science	Natural Science
Group 1								
HS-GPA	56							
UM-MGPA	77	.46						
UM-OGPA	101	.63	.67					
ACT Score								
English	55	.57	.39	.44				
Math	55	.71	.31	.49	.58			
Social Science	55	.49	.43	.43	.63	.61		
Natural Science	55	.43	.22*	.37	.61	.71	.64	
Composite	55	.66	.40	.51	.78	.88	.83	.89
Group 2								
HS-GPA	71							
UM-MGPA	104	.46						
UM-OGPA	131	.61	.78					
ACT Score								
English	72	.40	.19	.41				
Math	71	.55	.27	.37	.40			
Social Science	71	.46	.20*	.47	.60	.40		
Natural Science	71	.46	.31	.41	.47	.61	.63	
Composite	71	.58	.33	.53	.74	.77	.82	.86

*All correlations are statistically different from 0.0 ($p < .05$) except those with an *.

Table 6

Correlations of Criterion Variables with Scores from Stradaptive and Conventional Tests

Test and Score	Criterion Variable: GPA HS (N=56)	GPA UM-M (N=77)	GPA UM-O (N=101)	ACT Score English (N=55)	ACT Score Math (N=55)	ACT Score Social Science (N=55)	ACT Score Natural Science (N=55)	ACT Score Composite (N=55)
Stradaptive Test								
Entry Point (Reported GPA)	.49**	.57**	.59**	.53**	.38**	.44**	.28**	.46**
Number Administered	-.17	.02	-.09	-.31**	-.38**	-.14	-.26*	-.33**
Number Correct	-.16	-.03	-.13	-.28**	-.37**	-.12	-.23*	-.30**
Original Item Parameters								
Mean Difficulty Administered	.45**	.30**	.27**	.61**	.41**	.58**	.53**	.59**
Mean Difficulty Correct	.44**	.29**	.25**	.60**	.39**	.57**	.51**	.58**
Owen's Bayesian	.43**	.18	.14	.60**	.39**	.55**	.54**	.58**
Maximum Likelihood	.41**	.24*	.17*	.57**	.37**	.54**	.52**	.56**
SD Administered	.03	.10	-.05	-.19	-.07	-.06	-.14	-.13
SE Owen's Bayesian	.36**	.28**	.21*	.53**	.39**	.52**	.47**	.54**
Revised Item Parameters								
Mean Difficulty Administered	.44**	.32**	.27**	.51**	.40**	.58**	.52**	.59**
Mean Difficulty Correct	.43**	.30**	.25**	.60**	.38**	.57**	.51**	.58**
Owen's Bayesian	.43**	.24*	.19*	.62**	.38**	.57**	.53**	.58**
Maximum Likelihood	.41**	.25*	.18*	.58**	.36**	.52**	.54**	.56**
SD Administered	.12	.20*	.04	-.01	.06	.10	-.01	.04
SE Owen's Bayesian	.24*	.29**	.18*	.30**	.23*	.35**	.27*	.35**
Conventional Test								
Number Correct	.40**	.31**	.14	.62**	.40**	.54**	.52**	.58**

*Statistically different from zero at $p \leq .05$.

**Statistically different from zero at $p \leq .01$.

The most striking differences in validity between the adaptive and conventional tests were obtained on the UM-OGPA criterion (for which the largest sample size was available). Number Correct on the conventional test correlated .14 with UM-OGPA, which was not significantly different from zero. By contrast, using the revised item parameters, the correlations of all stradaptive scores were significantly different from zero, ranging from $r=.18$ to .27. Using the original parameters, three of the four stradaptive score correlations were significantly different from zero, the exception being the Bayesian score. Thus, the best stradaptive scoring method (Mean Difficulty Administered) accounted for 3.7 times the amount of criterion variance than did the conventional test Number Correct score; the second best stradaptive scoring method (Mean Difficulty Correct) accounted for 3.2% more common variance. It should also be recalled that the stradaptive test administered 25% fewer items, on the average, than did the conventional test. Thus, the higher validities were obtained despite shorter test lengths.

Correlations of stradaptive and conventional test scores with ACT scores were similar to the correlations of the stradaptive and conventional scores with HS-GPA and UM-MGPA. For all but ACT English, one or more of the stradaptive test scores correlated higher than did the conventional test score: For ACT English, Number Correct on the conventional test correlated .62, as did Owen's Bayesian score on the stradaptive test with revised parameter estimates. The largest difference in correlations between the conventional test and the stradaptive test was with ACT social science; the conventional test Number Correct score correlation of .54 was exceeded by all but Maximum Likelihood scoring of the stradaptive test, with correlations ranging from .55 to .58. In almost every case where stradaptive score validities exceeded those of the conventional test, highest correlations were obtained with the Mean Difficulty Administered score. Lowest correlations between stradaptive scores and ACT scores were generally obtained with the Maximum Likelihood scoring method.

Results of significance tests on the differences in the validity correlations shown in Table 6 indicated the following statistically significant differences:

1. Mean Difficulty Administered, using both original and revised item parameters correlated significantly higher ($p < .05$) with UM-MGPA than did either Owen's Bayesian score or the Maximum Likelihood score. Number Correct on the conventional test correlated significantly higher ($p < .05$) with this criterion variable than did the Bayesian score using the original item parameters.
2. Mean Difficulty Correct, also using both sets of item parameters, correlated significantly higher ($p < .05$) with UM-MGPA than did the Bayesian score; but it was not significantly higher than the Maximum Likelihood score. Using the original item parameters, the Maximum Likelihood score correlated higher with UM-MGPA than did the Bayesian score.
3. Mean Difficulty Administered and Mean Difficulty Correct correlated higher ($p < .01$) with UM-OGPA than did the Bayesian score, the Maximum Likelihood score, or the Number Correct score on the conventional test, for both the original and revised parameters.
4. Mean Difficulty Administered correlated significantly ($p < .05$) higher with ACT social science than did the Maximum Likelihood score using the revised item parameters.

The data in Table 6 show that none of the ability test scores correlated highly with UM-OGPA; the highest correlation was $r=.27$. Since UM-OGPA was an aver-

age across a wide variety of classes, frequently including substantial nonverbal material, high correlations with the vocabulary tests would not be expected. To determine whether the vocabulary tests correlated in the typically observed range with a relevant GPA variable, the effect of the mathematics grade on UM-OGPA was eliminated by computing the partial correlations of test scores with UM-OGPA, thus partialling out the effects of UM-MGPA. These results are shown in Table 7.

Table 7
Intercorrelations of UM-OGPA and UM-PGPA with Scores from Stradaptive and Conventional Tests, Partialling Out UM-MGPA

	Criterion Variable	
Test and Score	UM-OGPA (N=101)	UM-PGPA (N=71)
Stradaptive Test		
Original Item Parameters		
Mean Difficulty Administered	.27**	.51**
Mean Difficulty Correct	.25	.49**
Owen's Bayesian	.14	.43**
Maximum Likelihood	.17	.43**
SD Administered	-.05	.10
SE Owen's Bayesian	-.21	.36**
Revised Item Parameters		
Mean Difficulty Administered	.27**	.50**
Mean Difficulty Correct	.25**	.49**
Owen's Bayesian	.19*	.44**
Maximum Likelihood	.18*	.45**
SD Administered	.04	.19
SE Owen's Bayesian	.18*	.18
Conventional Test		
Number Correct	.14	.36**

*Statistically different from zero at $p < .05$.
**Statistically different from zero at $p < .01$.

As Table 7 shows, the partial correlations of all scores with GPA were higher than were the original correlations. All ability estimate scores were significantly correlated with UM-PGPA, using both original and revised item parameters for the stradaptive test. In addition, the correlation of Number Correct on the conventional test with UM-PGPA was also statistically different from zero. Correlations of the stradaptive scores with UM-PGPA were still substantially higher than Number Correct on the conventional test; the best stradaptive score (Mean Difficulty Correct with original item parameters) accounted for 26% of criterion variance, whereas Number Correct on the conventional test accounted for only 13% of criterion variance.

Bayesian versus conventional. Table 8 presents validity correlations for the Bayesian adaptive and conventional tests obtained from Group 2. On the average, the Bayesian ability estimate correlated more highly with the external criteria than did Number Correct on the conventional test. The Bayesian score correlated significantly higher (at $p < .05$) with HS-GPA than did the conventional test score

Table 8
Correlations of Criterion Variables with Scores from Bayesian Adaptive Test and Conventional Test

Criterion Variable	N	Bayesian Test				Conventional Test
		Number Administered	Number Correct	Ability Estimate	Variance of Ability Estimate	Number Correct
GPA						
HS	71	.44**	.46**	.51**	.09	.40**
UM-M	104	.23**	.20**	.22**	-.10	.16
UM-O	131	.12	.08	.16*	.13	.13
ACT Score						
English	72	.42**	.41**	.48**	.12	.50**
Math	71	.28**	.32**	.34**	.10	.33**
Social Science	71	.43**	.48**	.62**	.17	.59**
Natural Science	71	.40**	.40**	.50**	.15	.41**
Composite	71	.49**	.51**	.62**	.16	.57**

*Statistically different from zero at $p < .05$.
**Statistically different from zero at $p < .01$.

(r=.51 versus r=.40). UM-MGPA was also more accurately predicted by the Bayesian score (r=.22) than by the conventional test score (r=.16), but the difference was not statistically significant. No significant differences (at $p < .05$) were found between the validity coefficients for the Bayesian ability estimate and the conventional test Number Correct score in predicting UM-OGPA and the five ACT scores. However, with the exception of ACT English, the ability scores from the Bayesian adaptive test correlated higher with the criterion variables than did the score on the conventional test.

Table 9
Correlations of UM-OGPA and UM-PGPA with Scores from the Bayesian Adaptive and Conventional Tests, Partialling Out UM-MGPA

Test and Score	Criterion Variable	
	UM-OGPA (N=131)	UM-PGPA (N=100)
Bayesian Test		
Ability Estimate	.16*	.47*
Conventional Test		
Number Correct	.13	.44**

*Statistically different from zero at $p < .05$.
**Statistically different from zero at $p < .01$.

Correlations of the Bayesian ability estimate and Number Correct score on the conventional test with UM-PGPA are shown in Table 9. As was found in the Group 1

data, partialling out the effects of UM-MGPA resulted in higher correlations of both test scores with the GPA variable. Correlations for both test scores increased .31, and both partial correlations were significantly different from zero. However, there still were no significant differences between the validity correlations for the two tests.

DISCUSSION AND CONCLUSIONS

Testing Strategies

The major finding of this research was that the stradaptive and Bayesian adaptive testing strategies could predict to external criterion measures as accurately, and in some cases more accurately, as could the conventional test. In achieving these equal or higher levels of validity, the stradaptive test used approximately 25% fewer items, on the average, than did the conventional test. The Bayesian adaptive test used 20% more items, on the average, than the conventional test to achieve the same validity, although the median number of items administered in the Bayesian test was 12.5% fewer than in the conventional test. There were no significant differences between the stradaptive and Bayesian tests in terms of their correlations with the external criterion variables. The stradaptive test, using the Mean Difficulty Administered and Mean Difficulty Correct scores, predicted to overall college GPA at a significantly higher level than did the conventional test.

It may be argued that the differences in observed validities between the adaptive and conventional tests are a function of the higher item discriminations of items administered in the adaptive test and, consequently, that a comparison between the two testing strategies that does not equate for discriminations is unfair to the conventional test. What this criticism ignores, however, is that selecting items of high discriminations from a large pool is one of the important advantages of adaptive testing and can not be denied to the procedure.

A conventional test constructed to have discriminations equal to those items selected by the adaptive test would have at a specific point on the ability scale (1) good fidelity and poor bandwidth if it were a peaked test or (2) good bandwidth and poor fidelity if it had a rectangular distribution of item difficulties (McBride, 1976). Either test would correlate poorly with a criterion variable if there were any range of individual differences in the group being measured. Thus, the adaptive test is designed to resolve this bandwidth-fidelity dilemma by administering to each individual a test of high fidelity (high item discriminations) at or near the individual's estimated ability level (i.e., in a narrow bandwidth) with the location of the high fidelity measurement adapted to each testee.

This argument regarding higher levels of validity for adaptive tests attributable to higher item discriminations also does not take into account the somewhat different findings obtained with the overall college GPA variable between the stradaptive and Bayesian adaptive tests. Both adaptive tests tend to select the most discriminating items in the pool that are closest to the individual's ability level. Given that the average discriminations for the two adaptive procedures were similar, the significant differences between them in predicting overall college GPA in relation to the conventional test must have been due to their item selection procedures, their scoring methods, or the interaction of these two test characteristics.

Scoring Methods

The data in Table 6 suggest that the differences in the validities of the adaptive test relative to overall college GPA might have been due to scoring methods. On the average, the two mean difficulty scores used on the stradaptive test data had the highest correlations with all criterion variables. These two scores, in comparison to the Bayesian and Maximum Likelihood scores, are relatively simple scores that do not use complex latent-trait-based calculations. The simple average difficulty scores also do not utilize in their calculation the differing discriminations of items administered. The effect may be a score that is less sample-specific in that it is not optimized using explicit weights for both difficulty and discrimination. Similar to multiple-regression-weighted-composites, such optimally weighted scores may be sample-specific (in this case, highly dependent on the particular pattern of item responses and the specific values of the item parameter estimates), resulting in lower correlations with complex external criterion variables such as GPA. Another explanation may be that the latent-trait item discrimination parameter is related to the first principal component of an item set; and its use in scoring may result in a "factor pure" score that would correlate lower with an external criterion (which, like GPA, is likely not to be factorially pure) than would a score that is factorially somewhat more complex.

It may also be argued that the higher validities obtained for the adaptive test using the overall college GPA criterion was partially the result of the use of estimated GPA to begin testing in the stradaptive test. This argument does not take into account, however, the fact that the entry point information is not explicitly incorporated into the stradaptive test mean difficulty scores; it serves only as a means of selecting the first item to be administered. After that item, all subsequent item selection is based on the pattern of responses given by the individual. Entry point information in the stradaptive test might have a minor effect on the Mean Difficulty Administered score to the extent that the entry point is an accurate estimate of the ability being measured (Table 3 shows that it correlated .34 with conventional test scores and from .26 to .46 with adaptive test scores); but it would have no direct effect on Mean Difficulty Correct scores, since they are solely a function of ability level. In addition, this argument would not explain the lower validity correlations for the Bayesian test as compared to the stradaptive test, since the entry point (reported GPA) was explicitly included in scoring the Bayesian test as a consequence of its use as a differential prior ability estimate.

Data in Table 3 show that the simpler mean difficulty scores, however, conveyed almost the same information as the more complex latent trait scores; mean difficulty scores correlated .96 to .97 with Bayesian and Maximum Likelihood scores. The higher validities for the mean difficulty scores for most criteria, in conjunction with these high correlations, suggest that the mean difficulty scores from the stradaptive test may be as good for practical purposes as more complex scoring methods. These results support those of Vale and Weiss (1975a, 1975b) who, using other criteria and comparisons, concluded that Mean Difficulty Correct was a very useful scoring method for stradaptive tests. Further research would be desirable to determine if these simpler scoring methods might be useful in other adaptive tests.

The data in Table 3 also show correlations of .97 and 1.00 between Bayesian and Maximum Likelihood ability estimates. These correlations, based on response records averaging about 30 items, are slightly higher than the correlation of .95

obtained by Kingsbury and Weiss (1979) in their comparison of Bayesian and Maximum Likelihood logistic scoring of achievement test data using the three-parameter model.

Item Parameter Estimates

The data comparing the two sets of item parameter estimates used to score the stradaptive test by Bayesian and Maximum Likelihood methods were motivated by a desire to examine the generality of the finding by Prestwood and Weiss (1977) that the parameter estimation procedure suggested by Urry (1976), which corrected the biserial correlations for guessing, produced scores that were essentially linear transformations of the scores obtained by using parameter estimates that did not. The data presented in Table 3 support the earlier conclusion. Correlations between ability estimates based on the two sets of item parameter values were .98 for the two latent-trait scoring methods. The validity data (Table 6) also show no general differences in correlations of Bayesian and Maximum Likelihood scores with the criterion variables when the scores were obtained from the original and revised item parameter estimates; there were, however, slightly higher correlations with the two college GPA variables when the new parameters were used, with the differences tending to be larger for the Bayesian score. None of the differences between validity correlations based on the two sets of item parameter estimates were, however, statistically significant. The data, therefore, support the conclusion that the two sets of item parameter estimates are essentially linear transformations of each other, since they performed essentially equivalently in this study and correlated highly in both the present study and the Prestwood and Weiss (1977) study.

Reported GPA

A minor finding from this study indicates that self-reports of college GPA have a degree of validity. Data in Table 6 show that GPA reported in the intervals shown in Figure 1 correlated .59 with overall college GPA as obtained from university records. These data suggest that, even when obtained under volunteer research conditions, some confidence can be had in student-reported GPAs. The data also show significant correlations of reported college GPA with ACT scores--correlations which in some cases were not substantially different from those obtained from the verbal ability tests administered.

Conclusions

The data show generally higher, and in some cases significantly higher, criterion-related validities for the adaptive tests as compared to the conventional tests. There is some suggestion in the data that scoring of the ability test item responses by the Bayesian and Maximum Likelihood latent-trait scoring methods may have reduced the validities of the adaptive test. In comparing the two adaptive testing procedures, the data suggest that the stradaptive test scored by mean difficulty methods results in more valid ability estimates than the Bayesian adaptive test.

This study has been one of the first evaluations of the criterion-related validity of adaptive testing strategies. Thus, these conclusions must be considered tentative until supported by additional research. Characteristics of the item pools, decisions made in implementation of the adaptive strategies, design of the conventional test, and characteristics of the sample may all have affected the results. Yet the obtained findings are consistent with a wide range of related re-

search using different samples, tests, and procedures, which shows important gains in measurement precision and accuracy realized by the use of adaptive, as opposed to conventional, testing strategies.

REFERENCES

Angoff, W. H., & Huddleston, E. M. The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test (Statistical Report SR-58-21). Princeton, NJ: Educational Testing Service, June 1958.

Bayroff, A. G., & Seeley, L. C. An exploratory study of branching tests (Technical Research Note 188). Washington, DC: U.S. Army Behavioral Science Research Laboratory, Military Selection Research Division, June 1967. (NTIS No. AD 655263)

Bejar, I. I., & Weiss D. J. A construct validation of adaptive achievement testing (Research Report 78-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1978.

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. (NTIS No. AD A067752)

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A047495)

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1973. (NTIS No. AD 768993)

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1974. (NTIS No. AD A001230)

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, July 1975. (NTIS No. AD A013185)

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A046062)

Cleary, T. A., Linn, R. L., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187. (a)

Cleary, T. A., Linn, R. L., & Rock, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (b)

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, January 1974. (NTIS No. AD 773961)

Gialluca, F. A., & Weiss, D. J. Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1979. (NTIS No. AD A080956)

Green, B. F., Jr. Discussion. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Hansen, D. N. An investigation of computer-based science testing. In R. C. Atkinson & H. A. Wilson (Eds.), Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.

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

Koch, W. R., & Reckase, M. D. Problems in application of latent trait models to tailored testing (Research Report 79-1). Columbia: University of Missouri, Department of Educational Psychology, 1979.

Krathwohl, D. R., & Huyser, R. J. The sequential item test (SIT). American Psychologist, 1956, 2, 419.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1974. (NTIS No. AD 783553)

Larkin, K. C., & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing (Research Report 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1975. (NTIS No. AD A006733)

Linn, R. L., Rock, D. A., & Cleary, T. A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Lord, F. M. Robbins-Monro procedures for tailored testing (ETS RB-69-18). Princeton, NJ: Educational Testing Service, March 1969.

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813. (a)

Lord, F. M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242. (b)

McBride, J. R. Bandwidth, fidelity, and adaptive tests. In T. J. McConnell, Jr. (Ed.), CATC 2/1975. Atlanta, GA: The Atlanta Public Schools, 1976.

McBride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement (Research Report 74-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1974. (NTIS No. AD 78189)

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

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

Prestwood, J. S., & Weiss, D. J. Accuracy of perceived test-item difficulties (Research Report 77-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, May 1977. (NTIS No. AD A041084)

Trabin, T. E., & Weiss, D. J. The person response curve: Fit of individuals to item characteristic curve models (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1979.

Urry, V. W. A monte carlo investigation of logistic mental test models (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts International, 1971, 31, 6319B. (University Microfilms No. 71-9475)

Urry, V. W. Ancillary estimates 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)

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

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, December 1975. (NTIS No. AD A020961) (b)

Waters, B. K. An empirical investigation of the stradaptive testing model for the measurement of human ability. Unpublished doctoral dissertation, Florida State University, 1974.

Waters, B. K. An empirical investigation of Weiss' stradaptive testing model. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

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

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, February 1973. (NTIS No. AD 757788)

Table A

Original and Revised Discrimination (a) and Difficulty (b) Parameter Estimates for Items in the Stradaptive Test Item Pool by Stratum, In Order of Administration

Item Number	Original a	Original b	Revised a	Revised b	Item Number	Original a	Original b	Revised a	Revised b	Item Number	Original a	Original b	Revised a	Revised b	Item Number	Original a	Original b	Revised a	Revised b	Item Number	Original a	Original b	Revised a	Revised b
Stratum 9 (10 Items)					400	.34	1.61	.93	1.68	289	.35	-.05	.48	.69	85	.76	-1.07	.93	-.67	76	.56	-2.19	.62	-1.75
664	3.00	3.26	.84	2.95	541	.32	1.63	3.00	1.16	593	.32	-.13	.56	.55	173	.76	-1.43	.88	-1.06	641	.52	-1.89	.58	-1.40
309	3.00	3.26	.48	2.47	581	.31	1.36	1.26	1.21	Mean	.69	-.01	1.42	.47	637	.75	-1.41	.88	-1.02	95	.51	-2.20	.56	-1.71
298	3.00	3.26	.43	2.62	Mean	.46	1.33	2.03	1.50	SD	.27	.19	.81	.20	204	.73	-1.15	.88	-.74	108	.47	-1.71	.54	-1.16
388	3.00	3.26	.43	2.86	SD	.11	.18	1.05	.25	Stratum 4 (30 Items)					285	.71	-1.42	.84	-1.02	649	.44	-2.21	.49	-1.65
385	3.00	3.26	.42	2.35	Stratum 6 (19 Items)					501	1.82	-.70	1.20	-.55	227	.71	-1.63	.81	-1.25	214	.42	-2.08	.48	-1.49
627	3.00	3.26	.42	2.67	337	.98	.23	3.00	1.18	522	.92	-.73	1.06	-.39	640	.67	-1.47	.78	-1.06	642	.42	-1.80	.48	-1.19
245	3.00	2.63	.38	2.32	296	.91	.34	3.00	.67	270	.86	-.52	1.22	-.14	189	.66	-1.60	.76	-1.19	141	.42	-1.83	.48	-1.21
193	3.00	2.07	.35	2.37	114	.77	.65	3.00	.96	535	.86	-.68	.77	-.37	515	.62	-1.33	1.08	-.71	276	.41	-2.12	.45	-1.53
364	3.00	3.26	.33	3.11	294	.70	.79	3.00	1.07	91	.83	-.59	1.13	-.20	110	.58	-1.04	.70	-.54	546	.32	-1.92	.56	-.80
533	.63	2.15	.21	2.50	321	.63	.79	3.00	1.00	128	.82	-.75	1.07	-.36	185	.57	-1.17	.68	-.68	26	.31	-1.84	.36	-1.02
Mean	2.76	2.97	.43	2.62	651	.56	.49	1.09	.89	143	.77	-.57	1.04	-.15	235	.56	-1.27	.66	-.78	Mean	.84	-1.93	.97	-1.54
SD	.75	.49	.16	.27	666	.55	.42	1.00	.85	130	.75	-.85	.95	-.44	293	.56	-1.07	.67	-.57	SD	.38	.22	.46	.30
Stratum 8 (15 Items)					299	.52	.98	1.77	1.16	188	.71	-.47	.97	-.04	222	.54	-1.02	.65	-.50	Stratum 1 (35 Items)				
672	.85	1.89	3.00	3.26	302	.51	.37	.85	.85	104	.68	-.40	.84	.05	53	.52	-1.01	.64	-.48	71	3.00	-2.42	3.00	-2.32
166	.64	2.03	3.00	3.26	375	.49	.46	.83	.93	149	.67	-.91	.83	-.46	671	.52	-1.31	.62	-.79	7	3.00	-2.42	3.00	-2.32
662	.57	1.93	3.00	3.26	111	.48	.46	.82	.94	123	.67	-1.00	.82	-.56	112	.52	-1.30	.61	-.78	64	3.00	-2.45	3.00	-2.36
328	.54	2.31	3.00	3.26	215	.48	.65	.91	1.07	37	.67	-.69	.86	-.24	117	.52	-1.19	.62	-.66	102	3.00	-2.45	3.00	-2.36
381	.51	1.79	3.00	2.36	231	.45	.78	.87	1.19	46	.67	-.81	.84	-.36	94	.49	-1.57	.56	-1.02	28	3.00	-2.72	3.00	-2.63
336	.49	2.05	3.00	3.26	656	.44	.71	.82	1.15	365	.66	-.56	.88	-.11	287	.44	-1.27	.52	-.65	25	3.00	-2.72	3.00	-2.63
273	.49	1.79	3.00	2.14	238	.43	.65	.76	1.13	154	.66	-.58	.87	-.12	371	.38	-1.59	.44	-.92	42	3.00	-2.72	3.00	-2.63
115	.45	1.88	3.00	2.02	164	.41	.62	.69	1.14	203	.65	-.84	.82	-.38	155	.34	-1.35	.40	-.57	14	1.79	-2.67	2.21	-2.46
180	.43	2.07	3.00	2.63	397	.37	.83	.65	1.34	33	.64	-.85	.80	-.39	157	.32	-1.08	.38	-.25	24	1.59	-2.54	1.75	-2.37
274	.42	2.13	3.00	2.72	341	.37	.75	.63	1.28	372	.27	-.41	.35	.56	Mean	.73	-1.28	.90	-.84	11	1.48	-2.81	1.75	-2.58
297	.40	2.31	3.00	3.26	216	.37	.92	.67	1.40	183	.60	-.94	.73	-.45	SD	.27	.18	.35	.24	9	1.29	-2.46	1.45	-2.24
253	.39	1.65	2.32	1.44	Mean	.55	.65	1.44	1.06	145	.59	-.41	.79	.09	Stratum 2 (36 Items)					99	1.26	-2.78	1.24	-2.67
383	.36	1.82	2.11	1.52	SD	.18	.18	.99	.18	588	.53	-.89	.47	-.46	196	1.76	-1.99	2.13	-1.79	70	1.16	-2.47	1.29	-2.24
659	.35	2.27	3.00	1.95	Stratum 5 (25 Items)					391	.48	-.53	.62	.06	13	1.54	-1.78	1.89	-1.55	206	1.01	-2.43	1.11	-2.19
360	.34	2.18	3.00	1.71	630	1.31	-.05	3.00	.28	292	.48	-.58	.61	.01	138	1.52	-2.22	1.73	-2.02	124	1.01	-2.87	1.09	-2.64
Mean	.48	2.01	2.89	2.54	347	1.07	.14	3.00	.49	502	.41	-.40	.73	.22	190	1.46	-1.68	1.82	-1.44	65	.96	-2.94	1.02	-2.71
SD	.13	.21	.27	.70	272	.98	-.13	1.96	.22	318	.40	-.36	.53	.31	84	1.43	-1.87	1.70	-1.64	181	.94	-2.83	1.02	-2.58
Stratum 7 (23 Items)					283	.97	.15	3.00	.49	355	.40	-.58	.51	.10	27	1.23	-1.92	1.43	-1.68	68	.93	-2.74	1.01	-2.48
120	.72	1.07	3.00	1.46	568	.91	-.08	1.63	.29	116	.38	-.38	.49	.33	96	1.14	-1.88	1.13	-1.72	105	.91	-2.88	.98	-2.63
319	.62	1.49	3.00	2.14	266	.87	.16	2.12	.51	252	.32	-.34	.42	.47	125	1.10	-2.13	1.24	-1.88	126	.88	-2.54	.96	-2.27
652	.60	1.33	3.00	1.66	329	.87	-.21	1.42	.18	54	.30	-.67	.38	.20	129	1.08	-1.64	1.27	-1.35	80	.79	-2.55	.86	-2.25
359	.58	1.54	3.00	2.07	161	.86	-.25	1.38	.13	Mean	.65	-.66	.79	-.17	22	1.07	-2.23	1.20	-1.97	198	.74	-2.81	.80	-2.50
288	.56	1.11	3.00	1.26	254	.86	.21	2.28	.55	SD	.28	.19	.24	.39	101	1.02	-1.67	1.17	-1.40	5	.69	-2.50	.75	-2.16
152	.55	1.40	3.00	1.61	315	.83	.17	1.85	.52	Stratum 3 (36 Items)					44	.99	-1.71	1.15	-1.41	89	.67	-2.82	.72	-2.49
162	.52	1.17	3.00	1.22	599	.81	-.23	1.63	.16	191	1.40	-1.51	1.75	-1.26	158	.98	-2.26	1.08	-2.00	184	.67	-2.54	.73	-2.19
140	.52	1.30	3.00	1.38	340	.78	.30	1.92	.65	194	1.35	-1.23	1.79	.96	134	.96	-2.21	1.07	-1.94	31	.66	-2.50	.72	-2.14
263	.51	1.38	3.00	1.47	301	.76	.08	1.38	.47	36	1.23	-1.08	1.64	.79	127	.93	-1.66	1.08	-1.35	63	.64	-2.51	.69	-2.14
378	.49	1.44	3.00	1.48	56	.75	-.29	1.11	.14	51	1.16	-1.33	1.43	-1.04	186	.92	-1.65	1.07	-1.34	106	.62	-2.39	.67	-2.01
291	.44	1.31	1.64	1.35	60	.66	.24	1.23	.64	40	1.02	-1.34	1.24	-1.03	90	.82	-1.65	.94	-1.31	202	.57	-2.58	.62	-2.17
217	.43	1.25	1.25	1.38	271	.53	.33	.89	.80	87	.99	-1.10	1.24	.76	66	.80	-2.32	.87	-2.02	131	.36	-3.80	.60	-2.58
304	.42	1.00	.89	1.34	377	.43	-.23	.59	.39	199	.92	-1.42	1.09	-1.09	83	.77	-1.80	.88	-1.45	628	.52	-2.73	.57	-2.29
660	.41	1.01	.83	1.37	506	.43	-.09	.81	.58	43	.91	-1.21	1.11	-.86	559	.62	-1.80	.62	-1.68	82	.50	-2.77	.54	-2.31
668	.39	1.26	.93	1.49	538	.42	.15	1.18	.52	109	.89	-1.06	1.11	-.70	34	.74	-1.93	.83	-1.58	93	.48	-2.68	.52	-2.18
168	.37	1.36	.91	1.55	133	.41	-.09	.57	.56	103	.89	-1.34	1.06	-1.00	262	.70	-2.29	.77	-1.93	643	.44	-2.56	.49	-2.03
159	.36	1.24	.77	1.56	629	.40	-.26	.53	.42	47	.87	-1.31	1.04	-.96	311	.66	-1.83	.75	-1.43	81	.41	-2.95	.44	-2.39
562	.35	1.60	3.00	1.22	655	.39	.08	.55	.74	239	.77	-1.10	.94	-.71	88	.63	-1.75	.71	-1.33	Mean	1.28	-2.67	1.36	-2.38
107	.35	1.21	.69	1.59	324	.37	.09	.52	.77	86	.77	-1.55	.89	-1.19	232	.59	-1.70	.67	-1.25	SD	.93	.26	.92	.20
260	.34	1.47	.71	1.82																				

Table B
Item Discrimination (a) and
Difficulty (b) Parameter Estimates for
the Bayesian Adaptive Test Item Pool.

Item No.	a	b	Item No.	a	b	Item No.	a	b	Item No.	a	b
100	.56	-3.55	95	.51	-2.20	87	.99	-1.10	302	.51	.37
187	.45	-3.53	76	.56	-2.19	36	1.23	-1.08	666	.55	.42
8	.93	-3.42	125	1.10	-2.13	293	.56	-1.07	111	.48	.46
135	.40	-3.34	276	.41	-2.12	85	.76	-1.07	375	.49	.46
16	.70	-3.26	214	.42	-2.08	109	.89	-1.06	651	.56	.49
151	.41	-3.19	196	1.76	-1.99	110	.58	-1.04	164	.41	.62
17	.68	-3.19	34	.74	-1.93	222	.54	-1.02	215	.48	.65
121	.70	-3.11	27	1.23	-1.92	53	.52	-1.01	114	.77	.65
131	.56	-2.98	641	.52	-1.89	123	.67	-1.00	238	.43	.65
81	.41	-2.95	96	1.14	-1.88	183	.60	-.94	656	.44	.71
65	.96	-2.94	84	1.43	-1.87	149	.67	-.91	337	.98	.73
105	.91	-2.88	311	.66	-1.83	130	.75	-.85	341	.37	.75
124	1.01	-2.87	141	.42	-1.83	33	.64	-.85	231	.45	.78
181	.94	-2.83	642	.42	-1.80	203	.65	-.84	294	.70	.79
89	.67	-2.82	83	.77	-1.80	46	.67	-.81	321	.63	.79
198	.74	-2.81	13	1.54	-1.78	128	.82	-.75	397	.37	.83
11	1.48	-2.81	88	.63	-1.75	37	.67	-.69	216	.37	.92
99	1.26	-2.78	108	.47	-1.71	91	.83	-.59	299	.52	.98
82	.50	-2.77	44	.99	-1.71	154	.66	-.58	304	.42	1.00
68	.93	-2.74	232	.59	-1.70	292	.48	-.58	660	.41	1.01
628	.52	-2.73	190	1.46	-1.68	143	.77	-.57	120	.72	1.07
42	3.00	-2.72	101	1.02	-1.67	365	.66	-.56	288	.56	1.11
28	3.00	-2.72	127	.93	-1.66	391	.48	-.53	162	.52	1.17
25	3.00	-2.72	90	.82	-1.65	270	.86	-.52	217	.43	1.25
93	.48	-2.68	186	.92	-1.65	188	.71	-.47	140	.52	1.30
14	1.79	-2.67	129	1.08	-1.64	145	.59	-.41	291	.44	1.31
202	.57	-2.58	227	.71	-1.63	209	.64	-.40	652	.60	1.33
643	.44	-2.56	189	.66	-1.60	104	.68	-.40	263	.51	1.38
80	.79	-2.55	94	.49	-1.57	116	.38	-.38	152	.55	1.40
184	.67	-2.54	86	.77	-1.55	318	.40	-.36	378	.49	1.44
126	.88	-2.54	191	1.40	-1.51	56	.75	-.29	319	.62	1.49
24	1.59	-2.54	640	.67	-1.47	629	.40	-.26	359	.58	1.54
63	.64	-2.51	173	.76	-1.43	161	.86	-.25	381	.51	1.79
5	.69	-2.50	199	.92	-1.42	377	.43	-.23	273	.49	1.79
31	.66	-2.50	285	.71	-1.42	329	.87	-.21	115	.45	1.88
70	1.16	-2.47	637	.75	-1.41	272	.98	-.13	672	.85	1.89
9	1.29	-2.46	40	1.02	-1.34	133	.41	-.09	662	.57	1.93
102	3.00	-2.45	103	.89	-1.34	630	1.31	-.05	166	.64	2.03
64	3.00	-2.45	51	1.16	-1.33	301	.76	.08	336	.49	2.05
206	1.01	-2.43	47	.87	-1.31	655	.39	.08	180	.43	2.07
71	3.00	-2.42	671	.52	-1.31	324	.37	.09	274	.42	2.13
7	3.00	-2.42	112	.52	-1.30	347	1.07	.14	297	.40	2.31
106	.62	-2.39	235	.56	-1.27	283	.97	.15	328	.54	2.31
66	.80	-2.32	287	.44	-1.27	266	.87	.16	385	.42	2.35
262	.70	-2.29	194	1.35	-1.23	315	.83	.17	309	.48	2.47
158	.98	-2.26	43	.91	-1.21	264	.86	.21	298	.43	2.62
22	1.07	-2.23	117	.52	-1.19	60	.66	.24	627	.42	2.67
138	1.52	-2.22	185	.57	-1.17	340	.78	.30	388	.43	2.86
649	.44	-2.21	204	.73	-1.15	271	.53	.33	664	.84	2.95
134	.96	-2.21	239	.77	-1.10	296	.91	.34	290	.42	3.38

Table C
Item Discrimination (a) and Difficulty (b) Parameters for the Items in the Conventional Test, in Order of Administration

Item Reference No.	a	b
58	.482	-.957
221	.647	-.740
307	.562	-.836
386	.697	.136
211	.609	-.720
224	.543	-.785
390	.627	-.731
667	.568	-.726
156	.647	-.631
208	.582	-.681
234	.512	-.687
52	.606	-.282
137	.400	-.739
176	.338	-.897
207	.602	-.526
218	.332	-.928
205	.472	-.618
382	.638	-.481
342	.774	.172
265	.772	.173
645	.501	-.320
661	.579	-.296
670	.620	-.282
327	.571	-.248
50	.505	-.234
144	.627	-.184
369	.562	-.215
233	.468	-.172
139	.417	.189
633	.501	-.078
146	.607	.000
295	.474	-.035
113	.609	.247
267	.436	.188
59	.637	.173
147	.383	1.152
174	.638	1.156
242	.310	.979
306	.490	.969
367	.377	.978
Mean	.543	-.188
SD	.112	.593

Previous Publications

*Proceedings of the 1977 Computerized Adaptive Testing Conference. July 1978.

Research Reports

80-2. Interactive Computer Administration of a Spatial Reasoning Test. April 1980.
Final Report: Computerized Adaptive Performance Evaluation. February 1980.
80-1. Effects of Immediate Knowledge of Results on Achievement Test Performance and Test Dimensionality. January 1980.
79-7. The Person Response Curve: Fit of Individuals to Item Characteristic Curve Models. December 1979.
79-6. Efficiency of an Adaptive Inter-Subtest Branching Strategy in the Measurement of Classroom Achievement. November 1979.
79-5. An Adaptive Testing Strategy for Mastery Decisions. September 1979.
79-4. Effect of Point-in-Time in Instruction on the Measurement of Achievement. August 1979.
79-3. Relationships among Achievement Level Estimates from Three Item Characteristic Curve Scoring Methods. April 1979.
Final Report: Bias-Free Computerized Testing. March 1979.
79-2. Effects of Computerized Adaptive Testing on Black and White Students. March 1979.
79-1. Computer Programs for Scoring Test Data with Item Characteristic Curve Models. February 1979.
78-5. An Item Bias Investigation of a Standardized Aptitude Test. December 1978.
78-4. A Construct Validation of Adaptive Achievement Testing. November 1978.
78-3. A Comparison of Levels and Dimensions of Performance in Black and White Groups on Tests of Vocabulary, Mathematics, and Spatial Ability. October 1978.
78-2. The Effects of Knowledge of Results and Test Difficulty on Ability Test Performance and Psychological Reactions to Testing. September 1978.
78-1. A Comparison of the Fairness of Adaptive and Conventional Testing Strategies. August 1978.
77-7. An Information Comparison of Conventional and Adaptive Tests in the Measurement of Classroom Achievement. October 1977.
77-6. An Adaptive Testing Strategy for Achievement Test Batteries. October 1977.
77-5. Calibration of an Item Pool for the Adaptive Measurement of Achievement. September 1977.
77-4. A Rapid Item-Search Procedure for Bayesian Adaptive Testing. May 1977.
77-3. Accuracy of Perceived Test-Item Difficulties. May 1977.
77-2. A Comparison of Information Functions of Multiple-Choice and Free-Response Vocabulary Items. April 1977.
77-1. Applications of Computerized Adaptive Testing. March 1977.
Final Report: Computerized Ability Testing, 1972-1975. April 1976.
76-5. Effects of Item Characteristics on Test Fairness. December 1976.
76-4. Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. June 1976.
76-3. Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. June 1976.
76-2. Effects of Time Limits on Test-Taking Behavior. April 1976.
76-1. Some Properties of a Bayesian Adaptive Ability Testing Strategy. March 1976.
75-6. A Simulation Study of Stradaptive Ability Testing. December 1975.
75-5. Computerized Adaptive Trait Measurement: Problems and Prospects. November 1975.
75-4. A Study of Computer-Administered Stradaptive Ability Testing. October 1975.
75-3. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975.
75-2. TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations. March 1975.
75-1. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. February 1975.
74-5. Strategies of Adaptive Ability Measurement. December 1974.
74-4. Simulation Studies of Two-Stage Ability Testing. October 1974.
74-3. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974.
74-2. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974.
74-1. A Computer Software System for Adaptive Ability Measurement. January 1974.
73-3. The Stratified Adaptive Computerized Ability Test. September 1973.
73-2. Comparison of Four Empirical Item Scoring Procedures. August 1973.
73-1. Ability Measurement: Conventional or Adaptive? February 1973.

Copies of these reports are available, while supplies last, from:
Computerized Adaptive Testing Laboratory
Psychometric Methods Program, Department of Psychology
N660 Elliott Hall, University of Minnesota
75 East River Road, Minneapolis, Minnesota 55455