Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population

John T. Martin
James R. McBride
and
David J. Weiss

Research Report 83-1
January 1983

Computerized Adaptive Testing Laboratory
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

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 83-1
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population

5. TYPE OF REPORT & PERIOD COVERED: Technical Report

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): John T. Martin, James R. McBride 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: January 1983

13. NUMBER OF PAGES: 43

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)

Adaptive Testing	Latent Trait Theory
Computerized Testing	Item Response Theory
Tailored Testing	Response Latency
Individualized Testing	Bayesian Testing
Response-Contingent Testing	Owen's Bayesian Procedure

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

A conventional verbal ability test and a Bayesian adaptive verbal ability test were compared using a variety of psychometric criteria. Tests were administered to 550 Marine recruits, half of whom received two 30-item alternate forms of a conventional test and half of whom received two 30-item alternate forms of a Bayesian adaptive test. Both types of tests were computer administered and were followed by a 50-item conventional verbal ability criterion test. The alternate forms of the adaptive test resulted in scores that were

much more similar in means and variances than were the conventional tests for which most means and variances for various test lengths were significantly different. Adaptive testing resulted in significantly higher alternate forms reliability correlations for all test lengths through 19 items; reliability of a 9-item adaptive test was equal to that of a 17-item conventional test. Validity correlations were higher for the adaptive procedure for all test lengths. Validity of an 11-item adaptive test was equal to that of a 27-item conventional test, in spite of lower discriminating items being used, on the average, by the adaptive tests in comparison to the conventional test. Very few of the recruits had difficulty in responding to the computer-administered instructions on use of the testing terminals. Analysis showed some differences in test duration between the two testing strategies; where they occurred, they were explained by the ability level of the examinees, i.e., higher ability examinees who were administered adaptive tests received more difficult items and therefore had significantly longer testing times. Combined with reduced test length for the adaptive test to obtain similar reliabilities and validities to the conventional test, however, the slight increases observed in adaptive testing time were negligible. The data support the feasibility of adaptive testing with military recruit populations and support theoretical predictions of the psychometric superiority of adaptive tests in comparison with number-correct scored conventional tests.

Contents

Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population

Testing theorists have proposed a number of adaptive testing strategies over the last two decades (see Weiss, 1974). Although mechanical selection strategies were dominant at the beginning of the 1970s, they have now been largely replaced by item selection strategies based on item response theory (IRT). In mechanical item selection strategies, items are selected sequentially on the basis of their position in a structured item pool. Hence, at any point in the test, only certain items are available for selection and presentation. IRT-based item selection strategies select items which minimize or maximize some mathematical quantity. Thus, any item in the pool is potentially available for selection. The dominant mathematical item selection strategies are maximum information and Owen's Bayesian procedure.

Maximum information item selection involves selecting at each stage of an adaptive test the test item that has the highest level of psychometric information at the examinee's current ability estimate. This testing strategy has been used in a number of studies (Bejar & Weiss, 1978; Bejar, Weiss, & Gialluca, 1977; Prestwood & Weiss, 1978). It is preferred by some adaptive testing researchers (e.g., Lord, 1976) because it does not make prior judgments as to the distribution of ability in the population. However, others (e.g., Samejima, 1969; Urry, 1977) have claimed that maximum likelihood scoring procedures, which are usually utilized in conjunction with maximum information item selection, implicitly specify a flat prior distribution, and a flat prior distribution of ability would seldom correspond to the actual distribution of ability in the population. Additionally, maximum likelihood estimates for an individual's ability level do not explicity exist when that individual answers all items correctly or all items incorrectly; and, occasionally, maximum likelihood scoring can result in indeterminant ability estimates for an individual on short tests.

For these reasons some adaptive testing researchers combine maximum information item selection with Bayesian scoring procedures. The Bayesian modal procedure (Samejima, 1969) scores response patterns by using the mode of the posterior ability distribution as the estimate of ability, where the initial prior distribution is usually specified as having a mean of 0 and a standard deviation of 1. Owen's (1969, 1975) Bayesian scoring method, which can be combined with maximum information item selection (e.g., Brown & Weiss, 1977; Kingsbury & Weiss, 1979) is similar to Bayesian modal procedures except that ability is estimated by using the mean of the posterior ability distribution. Both Bayesian scoring methods, however, require the assumption of a normal distribution of ability. Owen's Bayesian scoring method, when combined with a Bayesian item selection procedure, provides a fully Bayesian strategy for adaptive test administration (Owen, 1969, 1975) in which items are selected at each stage of the test to minimize the Bayesian posterior variance of the ability estimate.

Research on Owen's Bayesian Adaptive Testing Strategy

Simulation studies. Many simulation studies have shown that Owen's Bayesian adaptive testing strategy results in stable, reliable, and valid scores even

for very short tests (Jensema, 1974, 1976; McBride, 1977; McBride & Weiss, 1976; Urry, 1974). For example, Urry (1974) found that Owen's Bayesian strategy achieved the reliability of a 60-item conventional test in from 10 to 15 items. Urry (1977) found that the validity of scores from Owen's Bayesian procedure for a sample of 57 live examinees was higher than that predicted by theory and by simulation results. However, Urry did not employ any other testing strategies that could be used for comparison with the Bayesian strategy, and his sample was sufficiently small so that the unexpectedly high validities may well have been a sampling artifact.

Gorman (1980) compared three types of conventional tests (strongly peaked, somewhat peaked, and rectangular) to adaptive tests using maximum information item selection and Bayesian modal scoring and to Owen's Bayesian adaptive testing strategy. Using both known and estimated item parameters, he found both Bayesian procedures superior to any conventional procedure on all evaluation criteria, which were (1) the fidelity coefficient (correlation of true and estimated ability scores), (2) conditional bias (mean directional error of ability estimates), (3) conditional accuracy (root mean square error of ability estimates), and (4) conditional precision (derived from the test score information function). He found that Owen's Bayesian procedure provided less bias using estimated item parameters than did the Bayesian modal adaptive or Bayesian-modally-scored conventional strategies. Altogether the Owen procedure provided somewhat better psychometric properties than the Bayes modal procedure. Gorman also found that for all of the adaptive tests evaluated, their superiority over conventional tests increased as a function of item discriminations.

Thus, these simulation studies have shown that Owen's Bayesian adaptive procedure achieves specified levels of measurement precision using far fewer items than conventional testing procedures and results in scores with substantially higher reliability and validity than those from conventional tests of the same length.

Live-testing studies. One of the first reported live-testing studies of Owen's Bayesian adaptive testing strategy (Thompson & Weiss, 1980) was based on a group of about 100 college undergraduates. The study compared criterion-related validity of the adaptive testing strategy with conventional tests administered to another group of students. Correlations of ability estimates with grade-point averages (GPA) were higher for the Bayesian test than for the conventional test. Scores on the Bayesian test correlated significantly higher with high school GPA (r = .51) than did the number-correct score on the conventional test (r = .40), even though the median number of items in the Bayesian test was 12.5% fewer than were administered in the conventional test.

Kingsbury and Weiss (1980) reported the first large-scale investigation of the performance of Owen's Bayesian strategy in live testing. They examined both alternate forms reliability and concurrent validity of Owen's Bayesian strategy in comparison with a conventional ability test. They administered to 472 college students a 120-item conventional criterion test scored by Bayesian methods, two 30-item conventional tests, and two 30-item adaptively administered Bayesian tests. The results were not completely in accord with theoretical expectations. For tests of one and two items in length, the conventional strategy was superior in parallel forms reliability; the adaptive tests achieved higher reliabilities

for test lengths of four to 30 items. However, the conventional strategy achieved consistently higher validities than the Bayesian adaptive strategy.

In a third live-testing study, also using large groups of college students, Johnson and Weiss (1980) compared 30-item conventional, 30-item Bayesian adaptive, and 30-item maximum information tests. They concluded that the alternate forms of the conventional test were more nearly parallel than the alternate forms of either adaptive strategy. Parallel forms reliabilities were similar for the conventional strategy and the two adaptive strategies for tests up to about 10 items in length. After that point, conventional test reliabilities were higher than those of the adaptive strategies.

Three factors may have contributed to these unexpected results: (1) the item pool had fewer items at the extremes of the ability distribution than near the center of the ability distribution, and the items at the extreme were of lower discrimination; (2) error in item parameter estimates may have been of sufficient magnitude to degrade the effectiveness of the adaptive testing strategies; (3) the range of the ability distribution in the college student sample was small. Data presented by Johnson and Weiss (1980) suggest that inadequacies in the item pool might have accounted for the failure of the adaptive tests to perform in accordance with expectations. Their data on conditional errors of measurement show that the standard error of measurement (SEM), which was always lower for the adaptive tests, increased for the maximum information strategy, especially at the lower end of the ability distribution, in contrast to simulation studies which show essentially flat information (and, therefore, SEM) functions. Since the SEM in an adaptive measurement is a joint function of the discrimination of the items and the number of items near the current estimated ability level, the combination of insufficient numbers of items with relatively low levels of item discrimination toward the lower extreme of the ability distribution might have resulted in the poorer performance of the adaptive tests in comparison to the conventional test.

All three of these live-testing studies of the Bayesian adaptive strategy were confounded by the small numbers of examinees on which the item parameters were obtained and by nonoptimal item pools for the adaptive strategy. In addition, because all studies were based on data from college students, restrictions in the range of abilities in the population undoubtedly affected the correlational results. Finally, in the Kingsbury and Weiss (1980) study, method variance might have been partially responsible for the higher correlations of the conventional experimental tests with the conventional criterion tests.

McBride (1980), in a live-testing pilot study on which the present study is based, found that Owen's Bayesian procedure produced verbal ability scores that were more reliable and valid at all test lengths than a conventional ability test. Since he tested Marine recruits, restriction of the ability range should have been less severe than in the case of the college population. He concluded from his data that a fixed-length adaptive test was as reliable as a variable-length adaptive test, and that adaptive tests of about 10 items were sufficiently reliable for military personnel testing purposes. This was the first comparative live-data study that fulfilled theoretical expectations.

Research on Other Aspects of Adaptive Testing

An important aspect of computerized testing is how testing strategy is related to the time it takes examinees to complete a test. It might be expected that for items that are in the middle range of difficulty for an individual examinee, response latencies (and, therefore, total testing time) would be greater than for items that are much too easy or much too difficult for that examinee. Since adaptive testing procedures select items for administration that are near the ability level of the examinee, whereas the conventional strategy does not, there may be differences in response latencies (or total testing time) due to the testing strategy. Using ANOVA, Betz and Weiss (1976) compared mean item latencies employing knowledge of results (KR), test type, and ability level as the independent variables. Although latencies for the stradaptive tests were slightly longer, differences were not statistically significant for test type but were statistically significant for ability level. Waters (1977) found that examinees responding to items in a stradaptive test required about 11% longer ($p \leq .05$) to respond to each item than did examinees who took a conventional test.

Johnson, Weiss, and Prestwood (1981) also found that items on stradaptive tests took examinees an average of 4% longer for fixed-length tests and an average of 11% longer for variable-length tests in comparison with conventionally administered items. They also noted that examinees taking the conventional tests more frequently reported that the items were too easy or too difficult for them, in comparison with those taking stradaptive tests.

Previously published research on computer-administered testing has not addressed the important practical question of whether novices have problems in learning to use the equipment. Such information is particularly important, along with information on the length of time it takes examinees to learn to use the equipment, in evaluations of the feasibilty of adaptive testing in large unselected populations.

Purpose

The present study was undertaken primarily to further study the reliability and validity of Bayesian adaptive tests, in comparison with conventional tests, in a military recruit population. Also of interest were a comparison of the amount of time required for administration of the adaptive and conventional tests and an evaluation of the effectiveness of the instructional sequence for this population.

METHOD

Subjects

Subjects were 553 male Marine recruits from the Marine Corps Recruiting Department (MCRD) in San Diego, California. In contrast to the design of the Kingsbury and Weiss (1980) alternate forms study, in the present study an independent groups design was used in which recruits were sequentially assigned to an adaptive or a conventional testing group. There were 263 recruits in the adaptive test group and 267 in the conventional test group.

Procedures

Testing equipment. Testing was controlled by a Hewlett-Packard real-time minicomputer system located at the University of Minnesota in Minneapolis. A multiplexed leased telephone circuit was connected to four cathode ray terminals (CRTs) operating at 120 characters per second at MCRD. The testing room was continually monitored by a test proctor, who helped the recruits become familiar with the equipment, answered "proctor calls" generated by the testing system, and insured that the equipment was operating satisfactorily.

Instructional sequence. Since the Marine recruit examinees were not expected to be familiar with the operation of a CRT, a sequence of instructional screens was presented to each examinee before beginning test administration. The 15 primary instructional screens, based on those originally described by DeWitt and Weiss (1974) and used for several thousand test administrations since, are shown in Appendix Table A. The instructional screen sequence assisted the recruits in learning to communicate with the computer by requesting that they (1) type a number and press the return key, (2) type "GO" and press the return key, (3) use the shift key, and (4) demonstrate their ability to change a response that was already typed. Appropriate error sequences were provided (see Appendix Tables A and B) to give examinees additional help when needed. Repeated errors resulted in an audible proctor call; when this occurred, the proctor intervened directly to assist the examinee in learning use of the CRT terminal.

After the examinee had demonstrated his understanding of the mechanics of CRT operation, five sample verbal ability items were presented to familiarize him with the item types and formats he would encounter in the experimental and criterion tests. Item types consisted of Sentence Completion, Synonyms, Analogies, and Opposites. The sample items (see Appendix Table A) were chosen to be very easy items that would be likely to be answered correctly by all examinees. If an incorrect answer was given, the examinee was given a second opportunity to answer the question; an incorrect answer the second time the screen was presented led to a proctor call.

Item pool. The items consisted of the same 150 five-alternative multiple-choice verbal ability items used by McBride (1980) in the pilot study. IRT parameters for the items were estimated using Urry's (1976; Gugel, Schmidt & Urry, 1976) OGIVIA program, based on samples of 980 to 2,200 Marine recruits. All item response function (IRF) discrimination parameters were greater than $\underline{a}$ = .80, difficulties were approximately rectangularly distributed between $\underline{b}$ = +2 and -2, and "guessing" parameters were less than $\underline{c}$ = .30. As Appendix Table C shows, the mean discrimination parameter for the pool was a relatively high $\underline{a}$ = 1.24, the mean difficulty was $\underline{b}$ = -.09, while the mean guessing parameter was $\underline{c}$ = .12. The classical item parameters for these 150 items were a mean biserial correlation of .76 and a mean difficulty of $\underline{p}$ = .57.

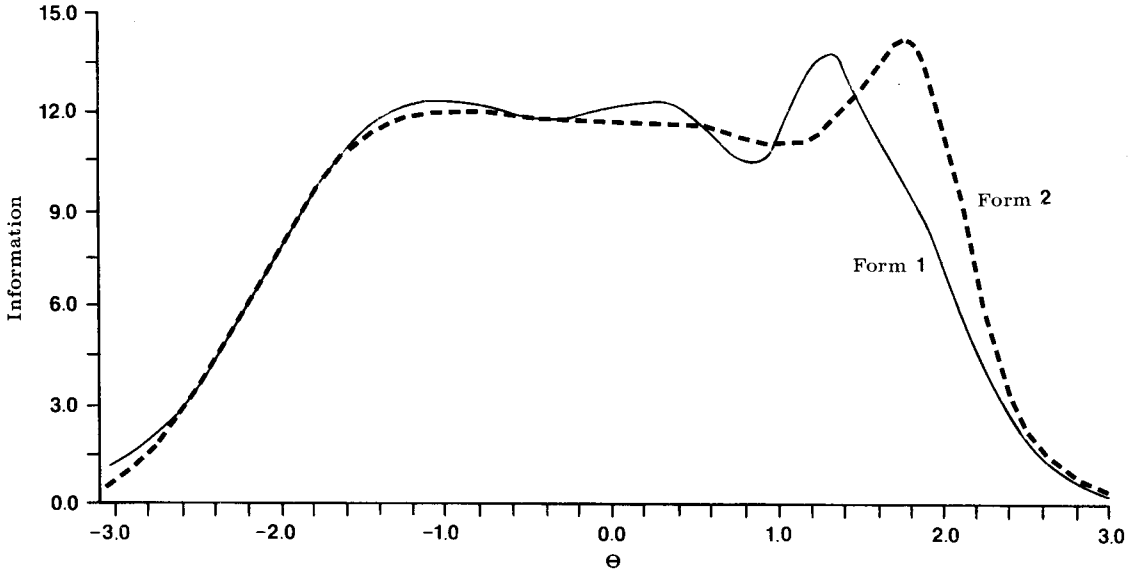
Tests

Experimental tests. The conventional test consisted of two alternate forms, each 30 items in length. Both conventional forms were administered on the CRT at the same time. Items were presented from each form (Forms 1 and 2)

in the repeating order 12212112. The conventional tests were constructed to have a rectangular distribution of item difficulties spanning the difficulty range of the item pool (IRT parameters and classical item parameters for each conventional test item are shown in Appendix Table D). Rectangular conventional tests were employed to equalize measurement precision across ability levels and to be similar to the verbal tests used in the Armed Services Vocational Aptitude Battery. The two forms were constructed to be "weakly parallel" (Samejima, 1977), i.e., to have test information functions that were approximately equal.

To select the items for the conventional tests, the 150 items in the item pool were sorted into five difficulty levels. Six items were selected in a balanced way from each difficulty level for each form of the conventional test, starting with the most discriminating items at each level. This design was used so that more discriminating items would appear earlier in the test than less discriminating items, thus allowing a more meaningful comparison with the adaptive tests, which were expected to select the most discriminating items toward the beginning of the test. This procedure resulted in mean discriminations of $\underline{a}$ = 1.42 for Form 1 and $\underline{a}$ = 1.46 for Form 2, mean difficulties of $\underline{b}$ = -.50 and $\underline{b}$ = -.32 for the two forms, respectively, and mean "guessing" parameters of $\underline{c}$ = .11 for both forms (see Appendix Table D). Figure 1 shows the test information curves for Forms 1 and 2 of the conventional tests. As can be seen, the test construction procedures resulted in very similar information functions for the two forms, thus fulfilling Samejima's (1977) weakly parallel criterion. The conventional tests were scored by number correct at each test length from 1 to 30 items.

Figure 1
Test Information Functions for Forms 1 and 2
of the 30-Item Conventional Tests



Administration of alternate forms of the tests to the adaptive test group was similar to the procedure used with the conventional test group, with the exception that items were selected by means of Owen's (1969, 1975) Bayesian sequential adaptive testing procedure. For each of the two adaptive forms (Form 1

and Form 2) items were independently selected from the item pool in the repeating order 12212112. To operationalize this procedure, as was done by Kingsbury and Weiss (1980) and Johnson and Weiss (1980), one item was selected from the pool as needed and assigned to Form 1 or Form 2 of the adaptive test according to the 12212112 rotational scheme. This procedure was repeated after each item was answered. As with the conventional tests, adaptive tests were 30 items in length, and no item was common to both forms for an individual examinee. The adaptive tests were scored at test lengths from 1 to 30 items by means of Bayesian ability estimates as an integral part of the test administration procedure.

Criterion test. The same 50-item multiple-choice conventional test was used as the criterion test for both the adaptive and conventional test groups. The criterion test was formed by selecting items measuring word knowledge from obsolete forms of the ASVAB. This test contained four-alternative multiple-choice items and was administered on the CRT immediately following administration of the two 30-item experimental tests. The criterion test was scored by number correct.

Data Analysis

Reliability and Validity

Reliability. Following the analysis of Kingsbury and Weiss (1980), Johnson and Weiss (1980), and McBride (1980), reliability was indexed by the correlations between the scores on the alternate forms for tests of each length (1 through 30 items). Because independent groups were used in the present study, observed differences in reliability correlations between the two testing strategies could be tested for statistical significance. After using Fisher's z transformation on the correlations, t tests were computed for differences between the reliabilities of the adaptive test forms and reliabilities of the conventional test forms.

One question of interest in the interpretation of these reliability correlations is the degree to which the alternate forms of the two testing strategies were truly parallel, since in the study by Kingsbury and Weiss (1980) apparent differences were observed in the degree of parallelism for the adaptive and conventional tests. To answer this question, correlated means t tests were computed to examine differences in means and standard deviations of scores of each test length for both testing strategies.

Validity. Scores from forms of every length were correlated with total number-correct scores from the 50-item criterion test, separately for the adaptive and conventional tests, and for Forms 1 and 2 of each test. All possible pairwise comparisons between the adaptive and conventional tests of correlations with the criterion test, for forms of the same length, were tested for differences using t tests. Since the criterion test was the same for both groups, and other sources of variation were controlled, any differences in validities between the testing strategies were due to the testing strategies or to sampling error in the sampling of examinees or abilities.

As is well known, validity is reduced by the unreliability of the measures

employed. The correction for attenuation results in a validity coefficient with the effects of reliability removed. Consequently, attenuation corrected validity coefficients were computed for tests of lengths 5, 10, 15, 20, 25, and 30 items. Reliability was assessed for the criterion test by means of coefficient alpha and parallel forms reliability was used for the experimental tests.

Comparisons of item characteristics. Previous research comparing adaptive and conventional tests (e.g., Kingsbury & Weiss, 1980; Thompson & Weiss, 1980) has frequently used independent item pools for each testing strategy, thus rendering comparisons of the results difficult since observed differences in reliabilities and/or validities may be due to differing item discriminations used for the different testing strategies. Even when the same item pool has been used in independent groups (e.g., Johnson & Weiss, 1980; McBride, 1980), the higher reliabilities and/or validities for the adaptive test may be a result of their selection of the most discriminating items in the pool, resulting in scores based on more discriminating items than for the conventional tests.

To determine whether this occurred with the present data, means and standard deviations of the item parameter estimates for the conventional tests were compared with those for the adaptive tests based on items actually administered by the adaptive procedure. Thus, item parameter descriptive statistics were computed prior to testing for the conventional test forms, but were computed after the data were collected for the adaptive forms.

Testing Time

To compare the amounts of testing time required by conventional and adaptive tests, cumulative item response latencies in seconds (i.e., total testing time excluding instructions) were analyzed using two-way analysis of variance with four levels of ability and the two testing strategies as the independent variables. Ability levels were arbitrarily defined such that Level 1 included examinees of estimated ability below $\hat{\theta} = -1.0$, Level 2 between $\hat{\theta} = -1.0$ and 0, Level 3 between $\hat{\theta} = 0$ and 1.0, and Level 4 above $\hat{\theta} = 1.0$. Ability levels in the conventional test group were defined so as to make the distribution of examinees in the four levels as similar to that of the adaptive test group as possible. Separate analyses were performed for test lengths of 5, 10, 15, 20, 25, and 30 items.

RESULTS

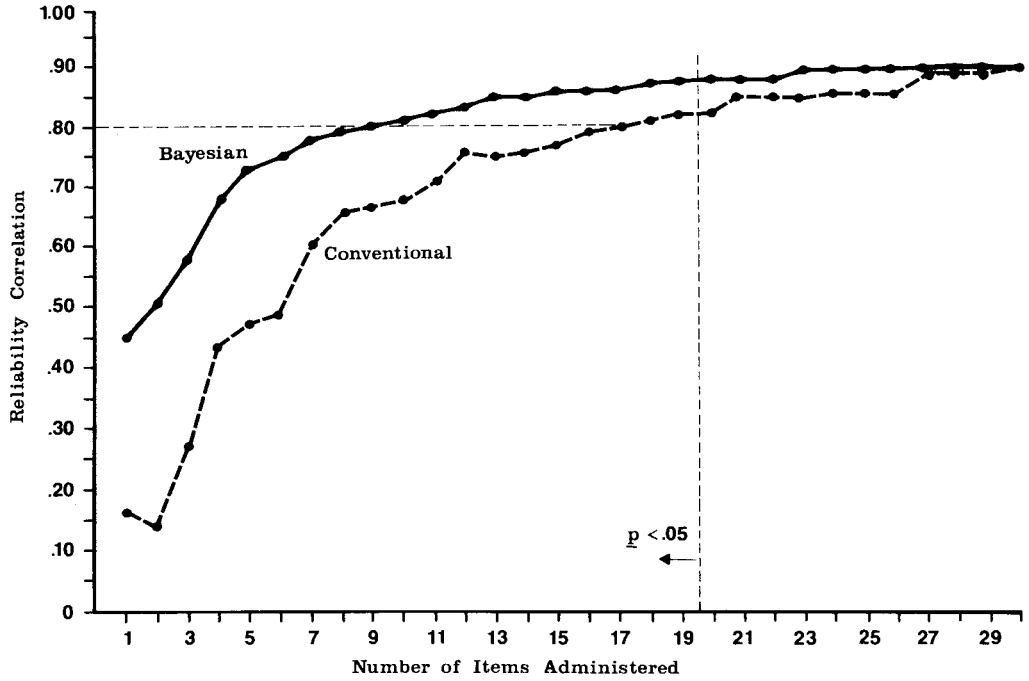
Reliability

Alternate Forms Correlations

Alternate forms correlations were computed using scores on the two forms of the Bayesian adaptive tests and on the two forms of the conventional tests, as a function of test length; these data are plotted in Figure 2 (numerical values are shown in Appendix Tables E and F). As Figure 2 shows, the Bayesian scores for the two adaptive tests correlated .45 after one item, increased rapidly to .78 after 7 items, then increased more slowly to .90 after all 30 items were administered. The scores on the two forms of the conventional test correlated .16 after one item, dropped to .13 after the second item, increased to .76 after

12 items and then more slowly to .89 after all 30 items were administered. After using Fisher's z transformation, t tests for differences between the reliability correlations were computed. These t tests show that for each test length up to 19 items (i.e., values to the left of the vertical dashed line in Figure 2) the adaptive forms correlated significantly higher ($p \leq .05$) with each other than did the conventional forms. Also, for all test lengths, the alternate forms reliabilities of the adaptive tests were higher than the reliabilities of the conventional tests. The horizontal dashed line in Figure 2 also shows that the adaptive test required only 9 items to achieve the same alternate forms reliability (.80) as a 17-item conventional test.

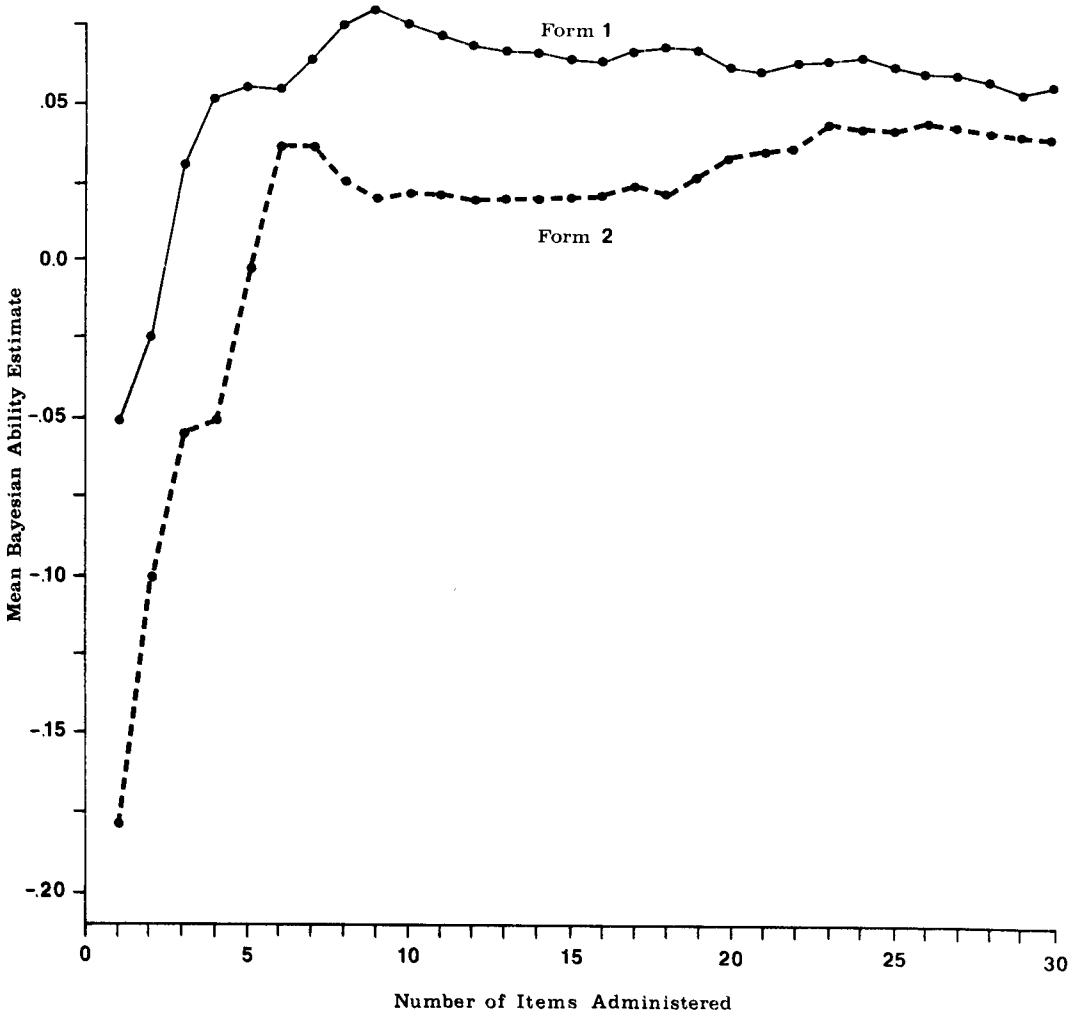
Figure 2
Alternate Forms Reliability Correlations for the Adaptive (N=263) and Conventional (N=267) Tests, as a Function of the Number of Items Administered



Parallelism of the Alternate Forms

Adaptive tests. Means, variances, skewness, and kurtosis statistics for the scores on the two forms of the Bayesian adaptive test are listed in Appendix Table E. Figure 3 shows the mean scores for the two forms of the adaptive test; after the first item the mean Bayesian score for Form 1 was -.05, and for Form 2, it was -.18. Mean scores for both forms rose until the 5th item for Form 1 and the 8th for Form 2, after which they were fairly stable. After the 18th item there is a pronounced trend for the scores from the two forms to further converge. After all 30 items were administered, the mean score on Form 1 was .06, whereas the mean score on Form 2 was .04. Scores were statistically significantly ($p \leq .05$) different, using correlated means t tests only for tests of one item and four items in length. At all other lengths, the adaptive forms showed no significant ($p \leq .05$) differences in Bayesian ability estimates between the two test forms.

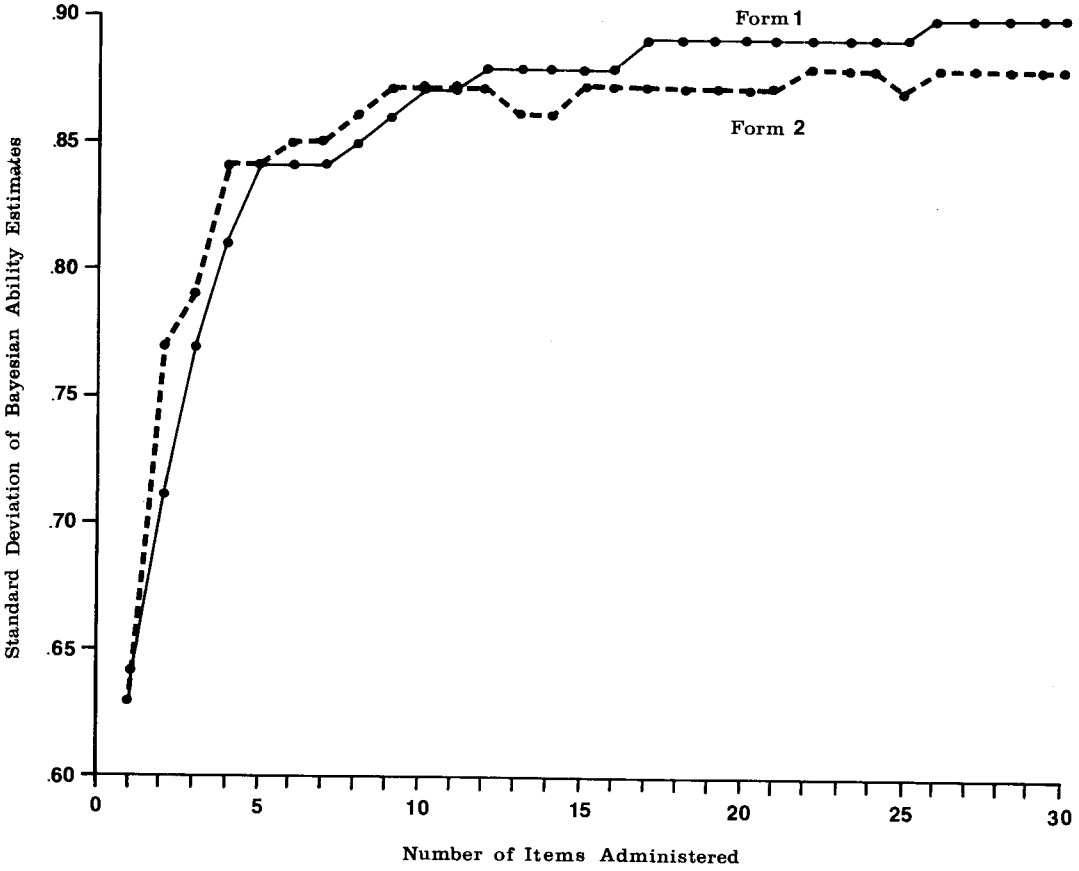
Figure 3
Mean Bayesian Ability Estimates for Forms 1 and 2 of the Adaptive Test, as a Function of Number of Items Administered



A somewhat similar pattern is seen in the standard deviations of the Bayesian adaptive test scores for the two forms (Figure 4; numerical data are in Appendix Table E). SDs after one item were .63 and .64, respectively, rising quickly to .84 after five items. Tests of lengths from 6 to 30 items had score SDs slowly increasing to .90 and .88 for Form 1 and Form 2, respectively. Unlike the means, which tended to converge, the SDs showed a slight divergence with increasing test length. However, using a correlated variances t test (McNemar, 1969, p. 282), none of the differences in variances between the alternate forms were statistically significant at any of the test lengths.

Mean Bayesian posterior variances were highly similar for the two forms for all test lengths, as shown in Table 1. Mean posterior variances after the first item were .59 for Form 1 and .60 for Form 2 and proceeded smoothly to .05 for both forms after 25 items.

Figure 4
Standard Deviations of Bayesian Ability Estimates for Forms 1 and 2 of the Adaptive Test, as a Function of Number of Items Administered



Conventional test. Figures 5 and 6 (and Appendix Table F) show data pertaining to the parallelism of the conventional test. Figure 5 shows that the mean proportion-correct score on Form 1 of the conventional test after 30 items was .65 for Form 1, and .64 for Form 2. Correlated means t tests for score differences between mean number-correct scores on the two alternate forms of the conventional test (see Appendix Table F) showed that the means of the conventional forms were significantly different for 29 of the 30 t tests at a significance level of $p \leq .05$; of these 29, 27 were significantly different at $p \leq .001$. There was no significant difference in mean number-correct scores only at a test length of 14 items. Thus, although the two forms of the conventional test were designed to be weakly parallel (see Figure 1), their mean scores did not meet the classical definition of parallel tests. Unlike the results for the Bayesian adaptive forms, there was little tendency toward score convergence for the two conventional forms with increasing test length, as mean absolute t values remained high through a test length of 29 items.

Figure 6 plots the number-correct standard deviations for the two conventional forms with increasing numbers of items. Form 2 showed somewhat greater standard deviations at almost all test lengths; however, after 28 items the standard deviations converged to .17. In contrast to the adaptive tests, sig-

Table 1
Means and Standard Deviations of Bayesian Posterior Variances for the Two Forms of the Adaptive Tests

Test Length	Form 1 Mean	Form 1 SD	Form 2 Mean	Form 2 SD
1	.59	.045	.60	.052
2	.40	.055	.40	.052
3	.30	.046	.30	.041
4	.24	.037	.23	.024
5	.20	.032	.20	.021
6	.17	.025	.17	.019
7	.15	.022	.15	.017
8	.13	.019	.13	.015
9	.12	.017	.12	.013
10	.11	.014	.11	.011
11	.10	.013	.10	.010
12	.09	.012	.09	.010
13	.09	.011	.09	.009
14	.08	.010	.08	.009
15	.08	.010	.08	.008
16	.07	.009	.07	.008
17	.07	.009	.07	.007
18	.07	.008	.07	.007
19	.06	.008	.06	.007
20	.06	.008	.06	.007
21	.06	.007	.06	.006
22	.06	.007	.06	.006
23	.06	.007	.06	.006
24	.06	.007	.06	.006
25	.05	.007	.05	.006
26	.05	.006	.05	.006
27	.05	.006	.05	.006
28	.05	.006	.05	.006
29	.05	.006	.05	.006
30	.05	.006	.05	.006

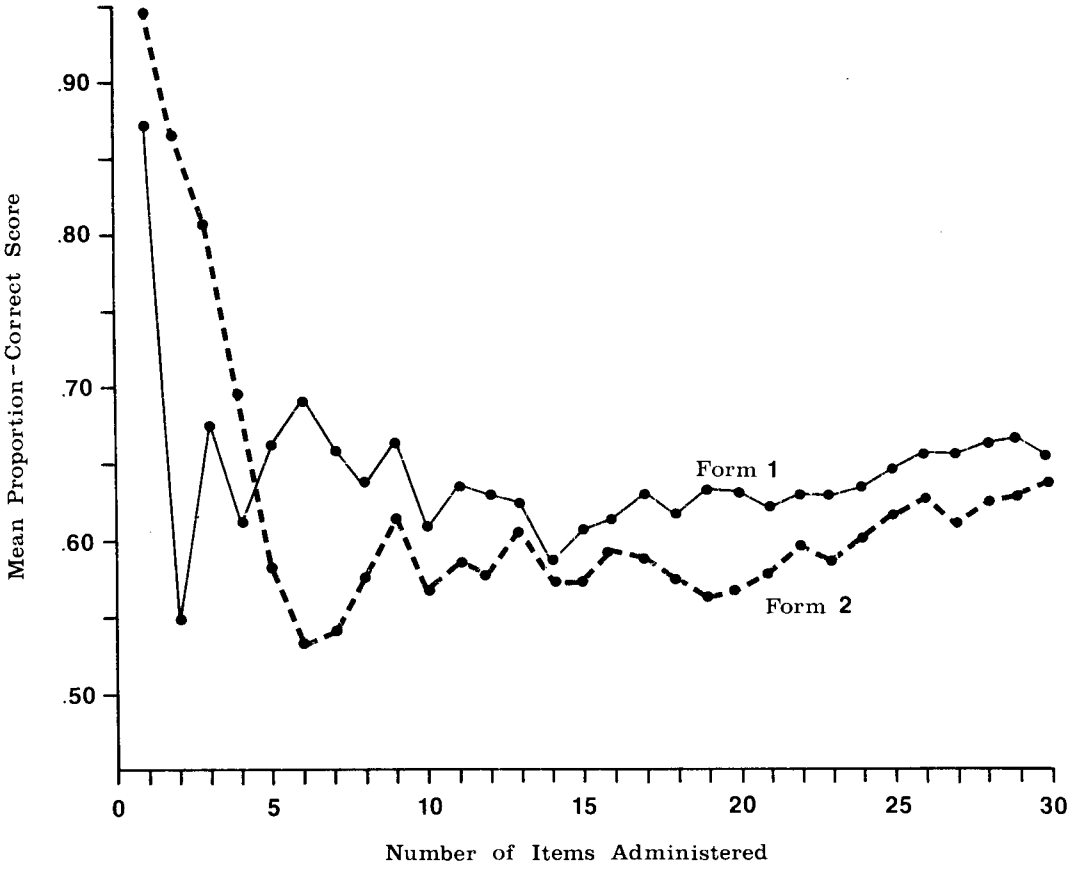
nificant differences in variances of the alternate forms were observed at 22 of the test lengths examined. With the exception of two-item tests, the conventional alternate forms had statistically significant differences in variances at test lengths through 23 items.

Validity

Correlation with Criterion Test Scores

Scores from each form of both the adaptive and conventional tests at lengths from one to 30 items were correlated with number-correct scores on the 50-item criterion test. These correlations are plotted in Figure 7; numerical

Figure 5
Mean Proportion-Correct Scores for Forms 1 and 2 of the Conventional Test, as a Function of Number of Items Administered



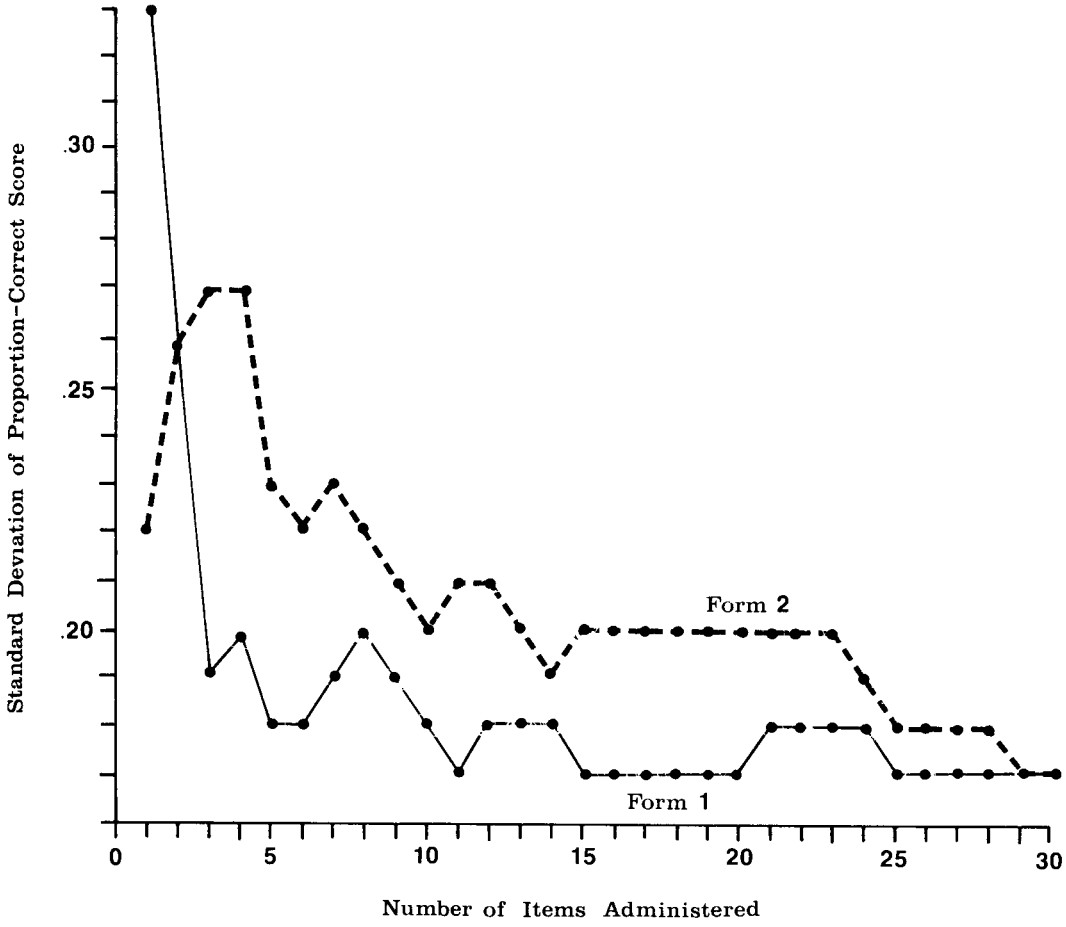
data are in Appendix Table G. Both adaptive forms correlated .39 with criterion test scores after one item was administered, rising to .84 after all 30 items were administered. Scores on the two forms of the conventional tests correlated .28 and .31, respectively, with criterion test scores after one item and .80 and .81, respectively, after 30 items were administered. As shown by the dashed horizontal line in Figure 7, scores on Forms 1 and 2 of the adaptive tests correlated .80 with scores on the criterion test after 10 items and 11 items, respectively, whereas scores on the two forms of the conventional test required 30 items and 28 items, respectively, to achieve the same level of validity.

Appendix Table G also shows results of the pairwise comparisons (between forms of the adaptive and conventional tests) of the correlations with criterion test scores. In all 120 comparisons, scores on the adaptive tests correlated more highly with scores on the criterion test than did scores on the conventional test. Although some of the differences were slight, 43 of them were sufficiently large to be statistically significant at the .05 level. Most of the significant differences occurred at test lengths of 18 items or less.

Attenuation-Corrected Correlations

Table 2 shows validity correlations from Appendix Table G for tests of

Figure 6
Standard Deviations of Proportion-Correct Scores for Forms 1 and 2 of the Conventional Test, as a Function of Number of Items Administered



length 5, 10, 15, 20, 25, and 30 items that have been corrected for attenuation caused by imperfect reliability in both the experimental tests and the criterion test. Alpha reliability for the 50-item criterion test was .85 in both the adaptive and conventional test groups; for these computations for experimental tests of a given length, alternate forms reliabilities were used (Appendix Tables E and F). Overall, scores on the Bayesian adaptive tests showed higher attenuation-corrected validity correlations than did scores on the conventional tests. (The corrected correlation of 1.07 for Form 2 of the conventional test at five items was a result of sampling artifacts). For example, at the 15-item test length, average corrected validities for the adaptive tests were .97; those for the conventional tests averaged .92; at 25 items, average validities were .97 for the adaptive tests and .915 for the conventional tests. The implication of these corrected correlation coefficients seems to be that the ability dimension that was measured by the criterion test was more nearly identical to that measured by the Bayesian adaptive tests than by number-correct scores on the conventional test, i.e., Bayesian adaptive scores contained less error and specific variance than did number-correct scores on conventional tests of the same length.

Figure 7
Validity Correlations with the Criterion Test for Two Forms of the Adaptive Test and Two Forms of the Conventional Test, as a Function of Number of Items Administered

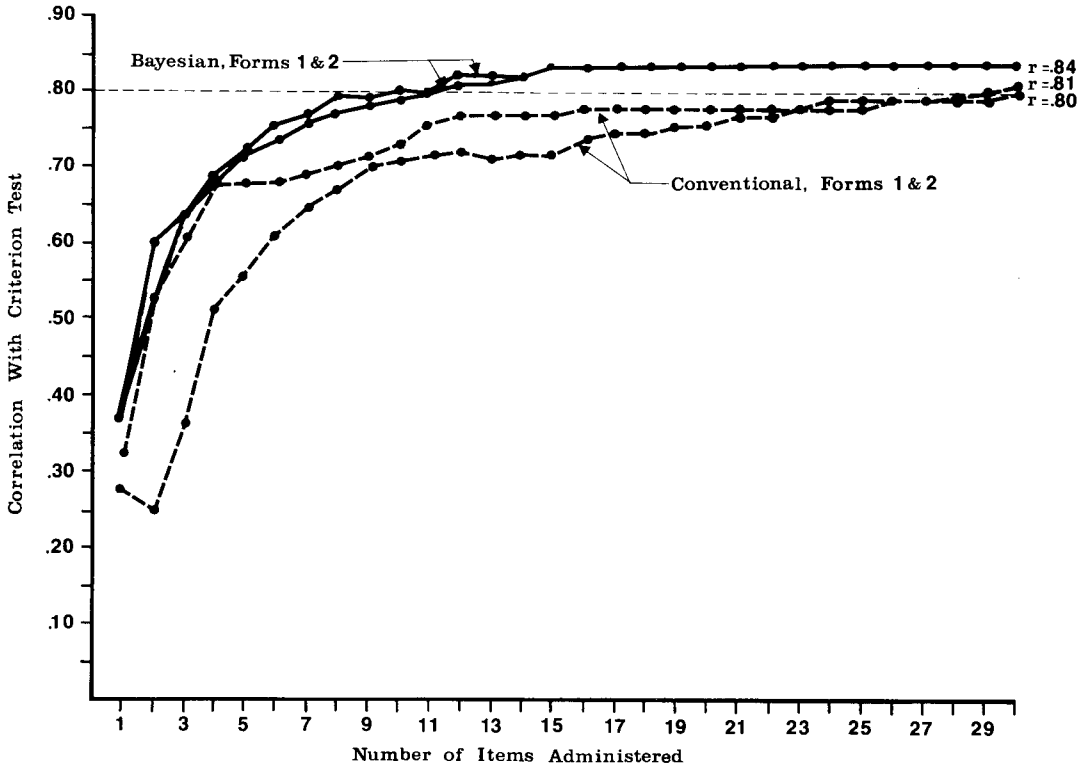


Table 2
Validity Correlations Corrected for Attenuation for Forms 1 and 2 of the Adaptive and Conventional Tests, as a Function of Test Length

Test Length	Adaptive		Conventional	
	Form 1	Form 2	Form 1	Form 2
5	.93	.91	.90	1.07
10	.96	.95	.93	.95
15	.97	.97	.89	.95
20	.97	.97	.92	.93
25	.97	.97	.91	.92
30	.96	.96	.92	.93

Characteristics of Items Administered

Item parameter means for each form of the adaptive and conventional tests are given in Table 3. The mean discrimination (a) parameter for items actually administered averaged across examinees for the two adaptive test forms were $\underline{a}$ = 1.33 and 1.32, respectively; for the conventional forms the mean $\underline{a}$ was 1.42 and 1.46, respectively. Thus, on the average, the conventional tests administered more discriminating items than did the adaptive tests. Table 3 also shows that the mean difficulties of the items administered in the adaptive tests were $\underline{b}$ = .06 for Form 1 and $\underline{b}$ = -.15 for Form 2, whereas those of the conventional test were $\underline{b}$ = -.50 and -.32. Thus, the adaptive tests were, on the average, more difficult than the conventional tests, but their difficulty was closer to the mean for the population on which the items were calibrated. All four tests administered items with mean $\underline{c}$ = .11.

Table 3
Means and Standard Deviations of the Item Parameters for the Adaptive and Conventional Forms

	Adaptive				Conventional			
	Form 1		Form 2		Form 1		Form 2	
Parameter	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$\underline{a}$	1.33	.34	1.32	.35	1.42	.49	1.46	.40
$\underline{b}$	.06	.90	-.15	.93	-.50	1.16	-.32	1.22
$\underline{c}$	.11	.05	.11	.05	.11	.06	.11	.07

Table 4 contains means and standard deviations for the discrimination ($\underline{a}$) parameter for each sequential position of the adaptive and conventional test. Mean $\underline{a}$ values were high in the early part of the adaptive test but decreased steadily with increasing test length. The highest mean $\underline{a}$ (1.984) occurred in the third sequential position, while the lowest (1.077) occurred in the 30th and last sequential position. Thus, the adaptive test used the "best" items in the pool early in the test and, as test length increased, used items of lower discrimination. The pattern was similar but not as smooth for the conventional test, where more highly discriminating items tended to occur earlier in the test. For 20 of the 30 test lengths, the mean item discrimination for the conventional test was higher than those of the adaptive test.

Additional Results

Testing Time

Table 5 presents cumulative item response latencies (i.e., net testing time) in minutes for each form of the adaptive and conventional tests, for each of four ability levels. The adaptive tests consistently resulted in higher mean net testing times than did the conventional tests for the highest ability level group (Level 4). Examinees in the lower half of the ability distribution showed no differences in mean net testing times between adaptive and conventional tests for tests of any length. As tests became longer, differences in mean net testing time increased substantially for examinees of high ability and increased

Table 4
Means and Standard Deviations of the Discrimination (a) Parameter for Each Sequential Position of the Adaptive and Conventional Tests for Both Forms Combined

Test Length	Adaptive Mean	Adaptive SD	Conventional Mean
1	1.470	.050	1.41
2	1.715	.071	2.52
3	1.984	.788	1.58
4	1.660	.398	1.80
5	1.579	.315	2.05
6	1.535	.262	1.42
7	1.534	.369	1.60
8	1.458	.270	1.59
9	1.427	.253	2.29
10	1.400	.277	2.29
11	1.354	.242	1.40
12	1.336	.230	1.45
13	1.330	.237	1.29
14	1.289	.184	1.79
15	1.273	.199	1.28
16	1.244	.212	1.26
17	1.221	.175	1.18
18	1.201	.171	1.46
19	1.204	.177	1.40
20	1.194	.217	1.40
21	1.179	.225	1.37
22	1.169	.220	1.16
23	1.158	.231	1.38
24	1.150	.220	1.20
25	1.132	.205	1.21
26	1.123	.202	.84
27	1.089	.175	1.29
28	1.095	.192	1.14
29	1.088	.187	1.04
30	1.077	.169	1.17

Note. Standard deviations are not presented for the conventional group, since they are based on only two values, one from each form.

somewhat less for examinees of moderately high ability, in favor of the conventional test condition; at the 30-item length, examinees on the adaptive test at the highest level of ability required about 75% more time to respond to the items, on the average, than did examinees on the conventional test. For the combined ability groups at the 30-item length the adaptive test group required 17% longer to respond to the items. Net testing time differences were more pro-

Table 5
Means and Standard Deviations of Total Response Latencies in Minutes (Net Testing Time) for Tests of Lengths 5, 10, 15, 20, 25, and 30 Items for Two Forms of the Adaptive and Conventional Tests at Four Levels of Ability and for Combined Ability Groups

Test Length and Ability Level*	Adaptive						Conventional					
	Form 1			Form 2			Form 1			Form 2		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
5 Items												
1	30	2.80	2.27	35	2.44	1.01	47	2.53	.83	35	2.66	1.04
2	81	1.95	.94	85	1.84	1.17	101	1.96	1.02	55	2.24	.89
3	110	1.78	.76	111	1.90	.87	86	1.70	.69	153	1.93	.94
4	30	2.41	1.10	20	2.03	1.03	17	1.23	.80	8	1.45	.65
Combined	251	2.03	1.18	251	1.97	1.03	251	1.93	.93	251	2.08	.97
10 Items												
1	29	4.72	2.81	32	4.53	2.60	21	4.71	1.28	37	4.48	1.44
2	89	3.77	1.54	86	3.74	1.28	74	4.18	1.91	77	4.48	2.04
3	97	3.89	1.62	103	3.71	1.42	110	3.20	1.08	97	3.53	1.22
4	36	4.06	1.90	30	4.09	1.46	46	2.49	1.12	40	2.53	.92
Combined	251	3.97	1.82	251	3.87	1.60	251	3.48	1.56	251	3.80	1.66
15 Items												
1	32	6.77	3.70	36	6.18	3.19	23	6.57	1.93	43	6.30	1.90
2	85	5.48	1.91	83	5.80	1.96	84	5.68	2.23	69	5.96	2.73
3	102	5.66	1.93	103	5.37	1.80	105	4.52	1.43	101	5.15	1.83
4	32	6.42	2.76	29	6.17	2.13	39	3.38	1.37	38	3.70	1.22
Combined	251	5.84	2.36	251	5.72	2.15	251	4.92	2.00	251	5.35	2.21
20 Items												
1	34	8.56	4.73	34	8.02	4.03	27	8.65	2.84	36	7.88	2.69
2	87	7.32	2.44	80	7.50	2.57	99	7.24	2.57	91	7.89	3.16
3	98	7.76	2.58	108	7.25	2.29	92	5.75	1.83	100	6.57	2.19
4	32	7.74	3.17	29	8.14	3.12	33	3.97	1.10	24	4.37	1.11
Combined	251	7.71	3.00	251	7.54	2.77	251	6.41	2.56	251	7.02	2.78
25 Items												
1	36	10.17	5.33	35	9.74	4.66	32	10.30	3.48	32	9.33	3.29
2	85	9.11	2.82	77	9.06	3.09	81	9.46	3.31	73	9.36	3.68
3	99	9.43	3.10	107	9.05	2.77	110	7.39	2.28	111	7.78	2.65
4	31	9.71	3.58	32	9.86	3.71	28	5.00	1.36	35	5.93	1.94
Combined	251	9.46	3.47	251	9.25	3.30	251	8.16	3.16	251	8.18	3.19
30 Items												
1	34	12.21	6.32	35	11.32	5.25	36	12.41	4.72	30	12.48	5.10
2	88	10.57	3.27	82	10.63	3.50	82	10.39	3.17	77	10.15	3.37
3	96	11.10	3.53	101	10.63	3.19	100	8.38	2.58	109	8.84	2.94
4	33	11.48	3.89	33	11.74	4.09	33	6.29	1.86	35	6.82	2.03
Combined	251	11.11	3.99	251	10.87	3.76	251	9.34	3.57	251	9.40	3.62

*For the adaptive test, Level 1 = $\hat{\theta} \leq -2.0$; Level 2 = $-2.0 \leq \hat{\theta} < 0.0$; Level 3 = $\leq 0.0\ \hat{\theta} < 1.0$; and Level 4 = $\hat{\theta} \geq 1.0$. For the conventional tests, the score distributions were approximately matched to those of the adaptive tests.

nounced with increasing test length, since for the combined group at the 5-item length there were essentially no differences, and at the 10-item length adaptive tests took only 8% longer. For the conventional test group, net testing time was strongly related to ability; as ability level increased, net testing time decreased. This was not the case in the adaptive test group where mean net testing times tended to be greater for the highest and lowest ability levels and somewhat less for middle ability levels.

Table 6
Two-way Analysis of Variance of Net Testing Time
by Ability Level and Testing Strategy (for Data in Table 5)

Test Length and Effect	Form 1				Form 2			
	DF	MS	F	p	DF	MS	F	p
5 Items								
Ability (A)	3	15.2	15.0	.001	3	7.9	8.3	.001
Strategy (S)	1	3.6	3.5	.061	1	2.0	2.1	.145
A × S	3	4.4	4.4	.005	3	2.1	2.2	.088
Residual	494	1.0			494	1.0		
Total	501	1.1			501	1.0		
10 Items								
Ability (A)	3	27.2	10.4	.001	3	26.1	10.9	.001
Strategy (S)	1	21.0	8.0	.005	1	.4	.2	.692
A × S	3	20.0	7.6	.001	3	21.8	9.1	.001
Residual	494	2.6			494	2.4		
Total	501	2.9			501	2.7		
15 Items								
Ability (A)	3	44.1	10.3	.001	3	36.5	8.3	.001
Strategy (S)	1	91.1	21.2	.001	1	15.5	3.5	.062
A × S	3	47.1	11.0	.001	3	29.5	6.7	.001
Residual	494	4.3			494	4.4		
Total	501	5.0			501	4.8		
20 Items								
Ability (A)	3	86.0	12.5	.001	3	44.4	6.2	.001
Strategy (S)	1	203.0	29.4	.001	1	40.3	5.6	.018
A × S	3	73.4	10.6	.001	3	59.2	8.3	.001
Residual	494	6.9			494	7.2		
Total	501	8.2			501	7.8		
25 Items								
Ability (A)	3	103.8	10.6	.001	3	50.0	5.0	.002
Strategy (S)	1	200.9	20.5	.001	1	136.4	13.7	.001
A × S	3	116.0	11.8	.001	3	71.9	7.2	.001
Residual	494	9.8			494	9.9		
Total	501	11.4			501	10.8		
30 Items								
Ability (A)	3	162.5	12.8	.001	3	93.2	7.5	.001
Strategy (S)	1	396.0	31.1	.001	1	251.9	20.2	.001
A × S	3	137.1	10.8	.001	3	119.1	9.5	.001
Residual	494	12.7			494	12.5		
Total	501	15.1			501	14.1		

Table 7

Errors Reported during Introduction to CRT Usage as a Function of the Instructional Screens Which Preceded Them for Adaptive Test (N=263) and Conventional Test (N=267) Groups

Error Screen No.	Test	Instructional Screen Number															
		9981	9101	9102	9103	9985	9105	9987	9211	9212	9215	9217	9219	9221	9222	9224	Total
9001	Adap								5	2		3			1		11
	Con								1			2					3
9035	Adap					2		1									3
	Con																0
9060	Adap					1											1
	Con								1	1	1						3
9061	Adap									1		1					2
	Con								1				1				2
9213	Adap								6	4	1			1			12
	Con								10	7	4			3			24
9214	Adap									21							21
	Con									23							23
9216	Adap										7						7
	Con										11						11
9218	Adap											13					13
	Con											18					18
9220	Adap												6				6
	Con												3				3
9223	Adap														21		21
	Con														35		35
9900	Adap	8															8
	Con	10															10
9901	Adap		12								1						13
	Con		7														7
9902	Adap			8	5		7		3					13		3	39
	Con			9	6		7		10					6		2	40
9904	Adap					106											106
	Con					108											108
9906	Adap							57									57
	Con							64									64
Total	Adap	8	12	8	5	109	7	58	14	28	9	17	6	14	22	3	320
	Con	10	7	9	6	108	7	64	23	31	16	20	4	9	35	2	351
Total Errors		18	19	17	11	217	14	122	37	59	25	37	10	23	57	5	671

Table 6 presents two-way Anova results for the data in Table 5. At the 5-item length the only main effect that was significant for both forms was ability level; at the 10- and 15-item length, Form 1 additionally showed a significant main effect for testing strategy, but this was not a significant main effect for Form 2. For tests longer than 15 items both ability level and testing strategy were significant ($p \leq .06$). For all test lengths (except Form 2 at 5 items) the ability level by testing strategy interaction was significant.

Effectiveness of the Instructions

Table 7 shows for each instructional screen the number of times each error screen was presented. (Instructional screens are in Appendix Table A; error screens are in Appendix Table B.) As can be seen in Table 7, examinees had the greatest difficulty when they were required to use the "SHIFT" key. Instructional Screen 9985 required examinees to change a typed "5" to a "4," which required the use of both the "SHIFT" key and the "RUB(out)" key; this screen resulted in 217 errors. Instructional Screen 9987, which required typing a question mark (again requiring the "SHIFT" key) resulted in 122 errors. Otherwise, there were only scattered errors, mostly in response to the five sample verbal test items (Screens 9212, 9215, 9217, 9219, 9222). The five sample items resulted in 188 errors altogether. An unknown hardware or software problem caused Error Screen 9213 to be presented 16 times in response to Instructional Screen 9211, for which the proper Error Screen was 9902.

Error screens could also occur in response to other error screens. Appendix Table H gives a similar breakdown of these errors. Altogether there were 161 such errors, with 76 of them resulting from Screen 9904 (second attempt to change a response). Since errors resulting in Screens 9060 and 9061 were proctor errors, only 131 of these errors were made by the recruits.

Table 8
Number of Error Screens Encountered
by Examinees During the Instructional
Screen Sequence for Total Group
(N = 531)

	Number of Errors								
Testees	0	1	2	3	4	5	6	7	≥ 8
Number	175	150	45	48	51	22	14	18	8
Percent	.33	.28	.08	.09	.10	.04	.03	.03	.02

Table 8 shows the distribution of the number of errors committed during the instructions. One-third of the examinees had no errors during the instructional sequence, while 28% had only one error. Only 2% of the examinees made eight or more errors while progressing through the instructional sequence. Mean number of errors of any kind per examinee was 1.56.

Means and standard deviations for the time it took the examinees to complete the instructional sequence are given in Table 9. The adaptive test group required 10.60 minutes to complete the sequence, while the conventional test

Table 9
Means and Standard Deviations in Minutes for Time Required to Complete Instructional Sequence

Group	Mean	SD
Adaptive	10.60	3.96
Conventional	10.64	4.27
Total	10.62	4.12

group required 10.64 minutes. The mean time to complete the instructional sequence for the total group was 10.62 minutes.

DISCUSSION AND CONCLUSIONS

Reliability and Validity

The adaptive tests were substantially more parallel than the conventional tests, which may have affected the alternate forms reliability correlations for the conventional tests. For almost all test lengths, score means on the conventional tests were significantly different from each other, whereas significant differences in score means were generally not observed for the adaptive tests. The adaptive tests achieved an alternate forms reliability correlation of .80 after only 9 items; the conventional tests required 17 items to achieve the same reliability. Also, for all test lengths up to 19 items the adaptive tests had significantly higher reliabilities than the conventional tests. Thus, except for the lack of parallelism in the conventional tests, the results of this study support theoretical predictions that fewer items are required to achieve a given level of measurement precision using adaptive, as opposed to conventional, tests.

The reliability of the Bayesian test scores at 30 items was only .04 higher than that at 15 items for these tests, but was .12 higher for the conventional tests. One reason why the reliabilities of the scores from the Bayesian tests did not continue to increase as test length increased may have been the declining discriminations of the items available in the item pool with increasing test length. By contrast, there was greater similarity in the conventional test discriminations throughout.

Correlations with the criterion test were consistently higher for the adaptive tests than for the conventional tests. To achieve a validity correlation of .80 required an average of 10.5 items for the adaptive test scores; however, to achieve the same correlation, an average of 29 conventionally administered items was required. Adaptive test score validities increased rapidly for scores based on tests of up to 15 items in length but showed little improvement after that. Again, this may have been attributable to few items with high discriminations being available for selection in the last half of the adaptive test.

The lower reliabilities of the conventional tests may be one explanation for the validity differences between testing strategies. However, when the validity correlations were corrected for attenuation, validity differences still favored the adaptive strategy. While the reliability differences between the two testing strategies might have been clouded by the less parallel nature of the conventional tests, the validity results were not dependent upon parallelism, since validity correlations were computed separately for each form of each test. Although differences in item discriminations might have caused validity differences, the conventional test item discriminations were generally higher than those of the adaptive test; observed validity differences were in the opposite direction.

The results of this study are contrary to those of Johnson and Weiss (1980) and Kingsbury and Weiss (1980) using somewhat similar research designs. Kingsbury and Weiss (1980) did not employ an independent groups design, and their examinees received 249 items, which may have introduced fatigue effects. Johnson and Weiss (1980) did employ an independent groups design, which should have eliminated any fatigue effects. Both of these studies used college student volunteers; this may have restricted the range of ability and thus affected the resulting correlations. The present study, however, investigated testing strategy effects on Marine recruits, who represent a wider distribution of ability than college students. Also, the items used in this study were parameterized on samples of 980 to 2,200 recruits, which is much larger than were used in the other two studies. Since the size of the parameterization sample is strongly related to the accuracy of the resultant item parameters (Schmidt & Urry, 1976), and since it should be expected that IRT-based item selection and person scoring strategies would be sensitive to the quality of the item parameter estimates, it is likely that differences in the quality of the item parameters led to the different results of these studies. In addition, the experimental conventional tests used by both Kingsbury and Weiss (1980) and Johnson and Weiss (1980) were peaked, as opposed to rectangular, tests, which might also have affected the results in an unknown way. The rectangular tests used in the present study better reflect the types of ability tests currently in use in military testing environments.

Testing Time

Although for some of the shorter test lengths testing times were shorter for the adaptive than for the conventional tests, in the majority of comparisons the adaptive tests required more testing time than the conventional tests. These data support those of Waters (1977) and Johnson, Weiss, and Prestwood (1981), indicating that it takes an individual slightly longer, on the average, to respond to items on adaptive tests than to those on a conventional test. Since items on an adaptive test are selected according to difficulty to be near each person's ability level, the slight increase in testing time must be judged from within the total context of the testing procedure. As was seen, the longer testing times for the adaptive procedure resulted from individuals of high ability receiving items of appropriate difficulty for them; however, in the conventional test, high ability individuals received items that were much too easy for them, as reflected in the very short response latencies.

While items that are far removed in difficulty from an individual's ability

level may require less time for a response, such items offer relatively little that is informative of that individual's status on the trait of interest. Thus, testing time is important only in relation to the psychometric properties of the testing outcome. If it takes 10% longer per item for examinees to respond to items selected by a given testing strategy, but that strategy requires only half as many items to achieve a given level of reliability or validity, then the increased efficiency of that procedure mitigates against the importance of the differences in testing time. For example, after an average of 3.54 minutes of testing, the adaptive group had responded to 9 items and the alternate forms reliability was .800; yet, after the conventional group had taken 10 items in an equivalent amount of time (3.64 minutes), this resulted in an alternate forms reliability correlation of only .675. Similarly, after the same 9 items the adaptive tests had an average validity of .785, but after 10 items the conventional tests had an average validity of only .72. Thus, while the adaptive tests required somewhat more time, on the average, to administer, they obtained given levels of reliability and validity in less time than did the conventional tests.

Effectiveness of the Instructions

Analysis of errors made during administration of the initial instructions indicated that examinees adjusted quite readily to CRT-presented testing. Using the "RUB(out)" key to change a response and using the "SHIFT" key were the only CRT operations that generated many errors. However, even for these operations, after the first error there were relatively few repeated errors. These results demonstrate that previous familiarity with CRT operation is not necessary for military recruits before undertaking a program of computer-administered adaptive testing. The sample items were answered without difficulty by almost all of the recruits. The majority of the instructional screens and the sample items thus appeared to function adequately in preparing the majority of the examinees for the tests.

Conclusions

The results of this study supported the feasibility and psychometric superiority of computer-administered adaptive tests as replacements for paper-and-pencil administered conventional tests in a military testing environment. On an item-for-item basis, the adaptive tests took slightly longer than the conventional tests; but with testing time held constant, the adaptive tests obtained substantially higher levels of both reliability and validity than did the conventional tests. The data showed that to obtain equal reliabilities, adaptive tests could administer 50% fewer items than the conventional tests; adaptive tests could also achieve the same level of validity as the conventional tests using only one-third the number of items, supporting earlier validity data reported by Thompson and Weiss (1980) on college students. The data also showed that using a realistic item pool with good distributions of item parameters, the adaptive tests reached their maximum levels of validity after 15 items had been administered, although reliabilities increased slowly beyond that length; this supports the use of short adaptive tests in practical applications.

REFERENCES

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., & 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.

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976.

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.

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.

Gorman, S. A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy. Unpublished doctoral dissertation, The Catholic University of America, Washington DC, 1980.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary estimation procedure. 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)

Jensema, C. J. The validity of Bayesian tailored testing. Educational and Psychological Measurement, 1974, 34, 757-766.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (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).

Johnson, M. F., & Weiss, D. J. Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Johnson, M. F., Weiss, D. J., & Prestwood, J. S. Effects of immediate feedback and pacing of item presentation on ability test performance and psychological reactions to testing (Research Report 81-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, February 1981.

Kingsbury, G. G., & Weiss, D. J. An adaptive testing strategy for mastery decisions (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1979.

Kingsbury, G. G., & Weiss, D. J. An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests (Research Report 80-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, December 1980.

Lord, F. M. 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).

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J. R. Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

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)

McNemar, Q. Psychological statistics. New York: John Wiley & Sons, 1969.

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

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. The effects of knowledge of results and test difficulty on ability test performance and psychological reactions to testing (Research Report 78-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1978.

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

Schmidt, F. L., & Urry, V. W. Item parameterization procedures for the future. 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).

Thompson, J. G., & Weiss, D. J. Criterion-related validity of adaptive testing strategies. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, June 1980.

Urry, V. W. Computer-assisted testing: The calibration and evaluation of the verbal ability bank (Technical Study 74-3). Washington DC: U. S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, September 1976. (NTIS No. PB-261 694)

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

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.

Waters, B. Discussion. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974.

Table A
Normal Sequence of Instructional Screens,
and Resultant Error Screens

Screen Number and Contents	Error Screens
Screen 9981 The tests you are going to take are being given to you by a computer. The instructions for the tests will appear on this screen. You will be asked some questions at the end of each part of the instructions to be sure that you understand how to answer the test questions. Type your answer on the typewriter keyboard. You must remember two things in order to talk to the computer: 1. Do not type anything until a question mark (?) appears on the screen. 2. Once you have typed an answer, the computer does not receive it until you press the "RETURN" key. Now, the first thing you must do is find the "RETURN" key. This key is the large key near the right-hand end of the second row of keys. Now press the "SPACE BAR" once, followed by the "RETURN" key, to continue the instructions.	9900
Screen 9101 Sometimes the computer will ask you to type a number to give your answer to a question. You will find the number keys on the top row of the keyboard. Just for practice, type the number "3". Be sure to press the "RETURN" key afterward.	9901
Screen 9102 That's good. Sometimes the computer will ask you to type a word rather than a number. For practice, type the word "GO" and press the "RETURN" key to continue the instructions.	9902
Screen 9103 You're doing fine so far. You know how to type words and numbers, and you know that you must press the "RETURN" key to send your answer to the computer. Suppose you make a mistake typing in your answer to a question. You can correct it at any time before you press the "RETURN" key. Type "GO" and press the "RETURN" key to find out how to correct an error.	9902

--continued on next page--

Table A, continued
Normal Sequence of Instructional Screens,
and Resultant Error Screens

Screen Number and Contents	Error Screens
Screen 9985 To correct an answer, hold down the "SHIFT" key while you press the "RUB" (stands for rub-out) key. The "SHIFT" key is the long gray key at either end of the bottom row of keys. The "RUB" key is the second key from the right-hand end of the third row of keys. The computer will respond with a "\" and the blinking light will move down one row. You may then retype your answer. Suppose you typed a 5 where you meant 4 As long as you have not pressed the "RETURN" key, you can correct the error by following the above instructions. To show that you understand how to change answers, change the following "5" to a "4".	9904
Screen 9105 Now you know what to do in case you make a mistake. Sometimes the computer makes a mistake, too (although it hates to admit it), and you can't read the question on the screen. If this happens you can repeat the question by pressing the "SPACE BAR" and then the "RETURN" key. Type "GO" and press "RETURN" to continue.	9902
Screen 9987 Sometimes you may not know the answer to a question and want to skip it. To do this, hold down the "SHIFT" key and type a question mark (?). Since the question mark is the same key as the slash (/), you must hold down the "SHIFT" key while you press the "?" (The "SHIFT" key is the long key at the left-hand end of the bottom row of keys, or third from the right-hand end of the bottom row.) Now go ahead and type a question mark. Don't forget to press the "RETURN" key.	9906

--continued on next page--

Table A, continued
Normal Sequence of Instructional Screens,
and Resultant Error Screens

Screen Number and Contents	Error Screens
Screen 9211 The test you are about to take is a test of your ability with words. Each test question will appear on this screen, followed by four or five possible answers. There is only one correct answer to each question. You must choose the correct answer to each question, and type its number on the keyboard. You may type a "?" if you do not know the answer and do not want to guess. Now type "GO", then press the "RETURN" key to continue.	9902
Screen 9212 There are five different kinds of questions in this test. One kind of question you will be answering in this test is called "opposites". Some examples of opposites are: GOOD is the opposite of BAD. TRUTHFULNESS is the opposite of LYING. Try this example: The opposite of NEAR is: 1. Happy 2. Close 3. Listen 4. Portentous 5. Far Type a number from "1" to "5" and press "RETURN".	9950 9213 9216

--continued on next page--

Table A, continued
Normal Sequence of Instructional Screens,
and Resultant Error Screens

Screen Number and Contents	Error Screens
Screen 9215	9950
That's right.	9213
Now let's try a different type of test question.	9216
In this type of question you must choose the word or phrase which completes a sentence so that it makes sense.	
For example	
I thought he was asleep because his eyes were _____.	
1. dark	
2. shut	
3. dull	
4. gray	
5. heavy	
Type a number from "1" to "5" and press "RETURN".	
Screen 9217	9950
That's right.	9218
In another kind of test question you should choose the word which means the same as the word in CAPITAL letters.	
Let's try an example.	
The word which means the same as PAINFUL is:	
1. Cup	
2. Playful	
3. Sore	
4. Amputated	
5. Smoke	
Type a number from "1" to "5" and press "RETURN".	
Screen 9219	9950
That's right.	9920
Sometimes this same kind of question looks like this:	
We CAPTURED the enemy agent.	
1. Caught	
2. Tried	
3. Scalded	
4. Helped	
Here, again you should type the number of the word that means the same as the word in CAPITAL letters.	
Type a number from "1" to "4" and press "RETURN".	

--continued on next page--

Table A, continued
Normal Sequence of Instructional Screens,
and Resultant Error Screens

Screen Number and Contents	Error Screens
Screen 9221 That's right. The last type of question is based on pairs of words that are related in some way. Your task is to decide how the first word in CAPITAL letters is related to the second word in CAPITAL letters. Next look at the third word in CAPITAL letters, then select an answer that has the same relationship to the third word as the first two words have to each other. For example: OATS is to HORSE as GAS is to CAR. Now type "GO" and press the "RETURN" key to continue.	9902
Screen 9222 Here is a practice question for you to answer: SAILOR is to NAVY as SOLDIER is to 1. Battle 2. Fort 3. Army 4. Regiment 5. War Type a number from "1" to "5" and press "RETURN".	9950 9223
Screen 9224 That's right. You have now completed the sample questions. To start the test, type "GO" and press "RETURN".	9002

Table B
Error Screens Required in the Instructional Sequence,
and their Frequency of Use

Screen Number and Contents	Number of Times Used
Screen 9001 You seem to be having trouble with the instructions or the equipment. Please call the test proctor.	140
Screen 9035 You have reached the time limit of 5 minutes for this screen. Please call the proctor for assistance.	2
Screen 9060 Incorrect input. TRY AGAIN.	22
Screen 9061 Input is still incorrect. Check your instruction manual and try again.	18
Screen 9213 You didn't type a number from "1" to "5". Because this is a very easy sample question, a "?" is not allowed (although you can answer with a "?" on the actual test questions). Please retype your answer following the instructions above.	47
Screen 9214 That's not right. Let's try that question again. The opposite of NEAR is: 1. Happy 2. Close 3. Listen 4. Portentous 5. Far Type a number from "1" to "5" and press "RETURN".	44
Screen 9216 That's not right. Let's try that question again. You should choose the answer that completes the sentence so that it makes sense. I thought he was asleep because his eyes were _____. 1. dark 2. shut 3. dull 4. gray 5. heavy Type a number from "1" to "5" and press "RETURN".	18

--continued on next page--

Table B, continued
Error Screens Required in the Instructional Sequence,
and their Frequency of Use

Screen Number and Contents	Number of Times Used
Screen 9218 That's not right. Let's try that question again. The word that means the same as PAINFUL is: 1. Cup 2. Playful 3. Sore 4. Amputated 5. Smoke Type a number from "1" to "5" and press "RETURN".	31
Screen 9220 That's not right. Let's try that question again. You are to choose that answer which is most similar in meaning to the CAPITALIZED word. We CAPTURED the enemy agent. 1. Caught 2. Tried 3. Scalded 4. Helped Type a number from "1" to "4" and press "RETURN".	9
Screen 9223 That's not right. Let's try that question again. You want to figure out how the first pair of words is related. Then choose a word for the second pair of words so that the second pair is related in the same way as the first pair. SAILOR is to NAVY as SOLDIER is to 1. Battle 2. Fort 3. Army 4. Regiment 5. War Type a number from "1" to "5" and press "RETURN".	49

--continued on next page--

Table B, continued
Error Screens Required in the Instructional Sequence,
and their Frequency of Use

Screen Number and Contents	Number of Times Used
Screen 9900 You found the "RETURN" key, but you typed something other than a space before you pressed it. In order to do well on these tests, it is important that you follow instructions carefully. Now press the "SPACE BAR" once and then the "RETURN" key, to continue.	18
Screen 9901 You didn't type the number "3". Have another practice try. Type the number "1" this time, then press the "RETURN" key.	20
Screen 9902 You didn't type the word "GO". Please try again. Type the word "GO" without any other letters or spaces, and press the "RETURN" key.	79
Screen 9904 You apparently were not successful in correcting the error. Here is another chance to practice. Change the following "7" to a "6". ? 7_	214
Screen 9906 You didn't type a question mark. Remember, you must: 1. Hold down the "SHIFT" key and 2. Press the "?" key. If you don't hold down the "SHIFT" key while you type the question mark, the computer reads a slash (/) and will tell you to try the same question again. Now, once again, type a question mark.	121

Table C

Item Statistics for the 150 Items in the Item Pool, Including Biserial Correlation (Rbis), Point-Biserial Correlation (Rptbis), Proportion Correct (Diff), and IRT Parameters (a, b, c)

Item Number	Rbis	Rptbis	Diff	Item Parameter Estimates a	b	c	Item Number	Rbis	Rptbis	Diff	Item Parameter Estimates a	b	c
1701	.66	.38	.85	.88	-1.51	.07	1770	.80	.51	.75	1.34	- .44	.23
1702	.64	.30	.91	.82	-1.91	.24	1771	.65	.45	.72	.86	- .76	.05
1704	.67	.37	.86	.89	-1.63	.08	1772	.83	.53	.82	1.46	-1.00	.05
1705	.72	.34	.92	1.03	-1.91	.13	1773	.77	.52	.76	1.19	- .80	.05
1706	.69	.32	.93	.96	-1.93	.23	1774	.77	.42	.31	1.20	1.23	.15
1710	.72	.48	.52	1.04	.36	.16	1775	.72	.45	.82	1.05	-1.15	.06
1711	.78	.56	.71	1.26	- .56	.06	1776	.68	.48	.62	.92	- .22	.09
1713	.74	.46	.83	1.09	-1.18	.06	1777	.74	.52	.65	1.11	- .26	.11
1714	.73	.51	.71	1.08	- .56	.08	1778	.77	.53	.73	1.22	- .60	.10
1715	.68	.44	.78	.92	-1.03	.06	1779	.72	.51	.67	1.03	- .40	.08
1716	.79	.57	.64	1.31	- .25	.08	1780	.73	.53	.50	1.06	.24	.08
1719	.82	.61	.61	1.42	- .18	.05	1781	.75	.43	.38	1.15	.99	.18
1720	.76	.54	.67	1.17	- .41	.07	1782	.81	.29	.17	1.38	1.94	.13
1721	.72	.47	.53	1.04	.34	.17	1783	.82	.58	.63	1.41	- .14	.12
1722	.67	.45	.48	.91	.45	.13	1784	.67	.40	.38	.90	1.00	.15
1723	.79	.56	.70	1.28	- .50	.06	1785	.81	.55	.38	1.39	.69	.10
1724	.69	.39	.86	.94	-1.54	.07	1786	.84	.59	.52	1.53	.22	.11
1725	.67	.32	.91	.90	-1.92	.19	1787	.79	.45	.24	1.28	1.34	.10
1726	.78	.37	.93	1.26	-1.84	.09	1790	.95	.22	.11	3.06	1.88	0
1727	.68	.35	.89	.93	-1.86	.09	1791	.83	.57	.42	1.48	.56	.11
1730	.75	.53	.70	1.13	- .52	.07	1792	.91	.37	.14	2.17	1.67	.06
1731	.84	.63	.57	1.52	- .06	.05	1793	.92	.34	.11	2.41	1.77	.04
1732	.78	.55	.44	1.26	.49	.09	1794	.80	.27	.17	1.31	2.00	.14
1733	.79	.57	.51	1.31	.23	.09	1796	.65	.31	.91	.86	-1.91	.23
1735	.73	.52	.49	1.08	.32	.10	1797	.63	.34	.87	.82	-1.79	.08
1743	.76	.45	.28	1.18	1.22	.11	1798	.75	.52	.73	1.14	- .65	.06
1744	.75	.42	.20	1.14	1.56	.08	1799	.71	.41	.85	1.00	-1.43	.07
1746	.71	.37	.26	1.01	1.52	.13	1800	.75	.40	.89	1.14	-1.59	.08
1747	.74	.35	.19	1.09	1.79	.11	1802	.75	.49	.46	1.14	.55	.15
1748	.89	.38	.08	1.92	1.91	.04	1804	.65	.37	.52	.86	.66	.25
1752	.62	.42	.38	.80	.84	.08	1805	.90	.22	.18	2.12	1.94	.14
1754	.72	.36	.15	1.03	1.94	.08	1806	.81	.52	.82	1.41	-1.20	.06
1759	.65	.37	.38	.85	1.14	.17	1807	.85	.57	.77	1.65	- .87	.09
1762	.69	.36	.89	.96	-1.81	.09	1808	.84	.58	.67	1.54	- .39	.15
1763	.66	.46	.70	.88	- .65	.05	1809	.87	.55	.64	1.71	- .13	.23
1765	.67	.39	.85	.91	-1.49	.07	1810	.79	.54	.63	1.30	- .52	.12
1768	.69	.36	.89	.94	-1.75	.08	1811	.83	.50	.53	1.46	.32	.24
1769	.62	.42	.74	.80	- .92	.05	1812	.81	.44	.36	1.39	.94	.18

-continued on next page-

Table C, continued

Item Statistics for the 150 Items in the Item Pool, Including Biserial Correlation (Rbis), Point-Biserial Correlation (Rptbis), Proportion Correct (Diff), and IRT Parameters (a, b, c)

Item Number	Rbis	Rptbis	Diff	Item Parameter Estimates a	b	c	Item Number	Rbis	Rptbis	Diff	Item Parameter Estimates a	b	c
1813	.81	.50	.45	1.40	.54	.18	1882	.77	.43	.89	1.22	-1.88	.07
1814	.86	.40	.29	1.69	1.19	.18	1885	.81	.48	.87	1.36	-1.66	.06
1815	.80	.27	.33	1.32	1.54	.26	1886	.72	.44	.84	1.03	-1.67	.06
1816	.87	.50	.42	1.71	.64	.19	1888	.82	.52	.84	1.45	-1.45	.05
1817	.84	.21	.22	1.61	2.00	.20	1889	.75	.48	.77	1.03	-1.01	.18
1828	.75	.47	.83	1.15	-1.52	.06	1890	.75	.48	.82	1.14	-1.49	.06
1830	.70	.46	.77	.98	-1.31	.05	1891	.70	.49	.70	.98	- .93	.07
1833	.73	.50	.72	1.08	- .89	.11	1892	.80	.53	.80	1.35	-1.28	.05
1834	.68	.48	.62	.92	- .50	.10	1893	.82	.53	.77	1.45	- .95	.17
1837	.75	.51	.75	1.13	-1.13	.05	1894	.77	.42	.89	1.20	-1.94	.07
1838	.64	.37	.85	.82	-2.00	.06	1895	.80	.56	.60	1.33	- .32	.13
1839	.79	.54	.63	1.27	- .41	.15	1896	.69	.47	.69	.95	- .82	.11
1840	.79	.51	.62	1.28	- .26	.21	1897	.70	.50	.47	.99	.10	.09
1841	.75	.50	.67	1.15	- .53	.18	1898	.73	.54	.58	1.07	- .44	.04
1842	.76	.53	.50	1.15	.05	.12	1899	.81	.58	.49	1.39	.03	.10
1843	.77	.53	.72	1.22	- .90	.09	1900	.74	.49	.57	1.11	- .11	.17
1844	.74	.54	.58	1.11	- .37	.08	1901	.81	.55	.40	1.37	.37	.10
1845	.73	.42	.37	1.07	.78	.16	1902	.88	.52	.37	1.82	.58	.15
1846	.73	.26	.32	1.06	1.58	.24	1903	.79	.54	.45	1.30	.23	.12
1847	.67	.35	.43	.89	.81	.22	1904	.87	.58	.40	1.77	.36	.11
1848	.77	.48	.36	1.19	.65	.13	1905	.82	.46	.31	1.46	.88	.14
1849	.76	.48	.44	1.19	.40	.16	1906	.74	.44	.25	1.10	1.10	.09
1850	.86	.38	.19	1.66	1.40	.11	1908	.91	.24	.19	2.24	1.63	.14
1851	.67	.47	.52	.90	- .07	.11	1910	.72	.49	.69	1.05	- .70	.15
1852	.82	.39	.30	1.45	1.09	.18	1911	.76	.45	.86	1.17	-1.73	.06
1853	.80	.57	.38	1.34	.39	.08	1912	.85	.60	.72	1.65	- .85	.07
1854	.86	.49	.51	1.67	.28	.25	1913	.77	.44	.88	1.22	-1.85	.07
1855	.69	.46	.44	.95	.34	.12	1914	.79	.49	.84	1.29	-1.53	.06
1856	.73	.46	.47	1.08	.34	.18	1915	.79	.47	.87	1.30	-1.69	.06
1857	.74	.33	.25	1.11	1.46	.16	1916	.67	.44	.66	.91	- .52	.18
1858	.63	.29	.40	.80	1.20	.24	1917	.75	.47	.82	1.14	-1.51	.06
1859	.74	.31	.24	1.11	1.58	.16	1918	.64	.44	.60	.84	- .42	.12
1861	.66	.29	.22	.88	1.81	.14	1919	.83	.52	.53	1.50	.12	.20
1862	.71	.37	.35	1.00	1.00	.19	1920	.67	.42	.61	.91	- .15	.24
1866	.67	.29	.33	.90	1.41	.21	1923	.96	.23	.30	3.39	1.26	.13
1867	.71	.48	.47	1.02	.22	.13	1924	.91	.12	.31	2.13	1.94	.29
1870	.80	.46	.26	1.34	1.03	.11	Mean	.76	.45	.57	1.24	-.09	.12
1878	.75	.24	.23	1.14	1.88	.19	S.D.	.07	.10	.23	.38	1.17	.06

Table D
Items in the Two Forms of the Conventional Test in Order of Presentation By Form, Biserial Correlations (Rbis), Point-Biserial Correlations (Rptbis), Proportion Correct (Diff), and IRT Parameters (a, b, c)

Form and Item	Rbis	RptBis	Diff	IRT Parameter Estimates a	b	c
Form 1						
1893	.82	.53	.77	1.45	-.95	.17
1923	.96	.23	.30	3.39	1.26	.13
1888	.82	.52	.84	1.45	-1.45	.05
1904	.87	.58	.40	1.77	.36	.11
1705	.72	.34	.92	1.03	-1.91	.13
1828	.75	.47	.83	1.15	-1.52	.06
1731	.84	.63	.57	1.52	-.06	.05
1816	.87	.50	.42	1.71	.64	.19
1890	.75	.48	.82	1.14	-1.49	.06
1793	.92	.34	.11	2.41	1.77	.04
1726	.78	.37	.93	1.26	-1.84	.09
1919	.83	.52	.53	1.50	.12	.20
1723	.79	.56	.70	1.28	-.50	.06
1748	.89	.38	.08	1.92	1.91	.04
1886	.72	.44	.84	1.03	-1.67	.06
1843	.77	.53	.72	1.22	-.90	.09
1768	.69	.36	.89	.94	-1.75	.08
1905	.82	.46	.31	1.46	.88	.14
1806	.81	.52	.82	1.41	-1.20	.06
1811	.83	.50	.53	1.46	.32	.24
1791	.83	.57	.42	1.48	.56	.11
1713	.74	.46	.83	1.09	-1.18	.06
1783	.82	.58	.63	1.41	-.14	.12
1778	.77	.53	.73	1.22	-.60	.10
1894	.77	.42	.89	1.20	-1.94	.07
1796	.65	.31	.91	.86	-1.91	.23
1716	.79	.57	.64	1.31	-.25	.08
1917	.75	.47	.82	1.14	-1.51	.06
1889	.75	.48	.77	1.03	-1.01	.18
1870	.80	.46	.26	1.34	1.03	.11
Mean	.80	.47	.64	1.42	-.50	.11
SD	.07	.09	.25	.49	1.17	.06

-continued on the next page-

Table D, continued
Items in the Two Forms of the Conventional Test in Order of Presentation by Form, Biserial Correlations (Rbis), Point-Biserial Correlations (Rptbis), Proportion Correct (Diff), and IRT Parameters (a, b, c)

Form and Item	Rbis	RptBis	Diff	IRT Parameter Estimates		
				a	b	c
Form 2						
1885	.81	.48	.87	1.36	-1.66	.06
1912	.85	.60	.72	1.65	-.85	.07
1809	.87	.55	.64	1.71	-.13	.23
1902	.88	.52	.37	1.82	.58	.15
1790	.95	.22	.11	3.06	1.88	.00
1814	.86	.40	.29	1.69	1.19	.18
1854	.86	.49	.51	1.67	.28	.25
1772	.83	.53	.82	1.46	-1.00	.05
1892	.80	.53	.80	1.35	-1.28	.05
1792	.91	.37	.14	2.17	1.67	.06
1808	.84	.58	.67	1.54	-.39	.15
1813	.81	.50	.45	1.40	.54	.18
1915	.79	.47	.87	1.30	-1.69	.06
1850	.86	.38	.19	1.66	1.40	.11
1786	.84	.59	.52	1.53	.22	.11
1914	.79	.49	.84	1.29	-1.53	.06
1719	.82	.61	.61	1.42	-.18	.05
1852	.82	.39	.30	1.45	1.09	.18
1812	.81	.38	.36	1.39	.94	.18
1770	.80	.51	.75	1.34	-.44	.23
1711	.78	.56	.71	1.26	-.56	.06
1882	.77	.43	.89	1.22	-1.88	.07
1853	.80	.57	.38	1.34	.39	.08
1911	.76	.45	.86	1.17	-1.73	.06
1913	.77	.44	.88	1.22	-1.85	.07
1702	.64	.30	.91	.82	-1.91	.24
1787	.79	.45	.24	1.28	1.34	.10
1800	.75	.40	.89	1.14	-1.59	.08
1775	.72	.45	.82	1.05	-1.15	.06
1799	.71	.41	.85	1.00	-1.43	.07
Mean	.81	.47	.61	1.46	-.32	.11
SD	.06	.09	.26	.40	1.22	.07

Table E
Descriptive Statistics for the Scores on the Bayesian Adaptive Tests, *t* Values for Differences in Means and Variances Between the Forms, and Correlations Between Scores on the Two Forms, for Test Lengths of 1 to 30 Items

Test Length	Form 1 Mean	SD	Skewness	Kurtosis	Form 2 Mean	SD	Skewness	Kurtosis	*t* Values Mean	Variances	*r*
1	-.05	.64	-.02	-2.02	-.18	.63	.25	-1.96	3.10**	.34	.451
2	-.02	.71	.05	-.85	-.10	.77	.07	-1.13	1.72	-1.37	.505
3	.03	.77	.13	.13	-.06	.79	.08	-.41	1.92	-.58	.571
4	.05	.81	-.02	-.12	-.05	.84	.04	-.30	2.41	-.98	.672
5	.06	.84	-.02	-.04	-.00	.84	.04	-.21	1.50	.00	.733
6	.05	.84	-.11	-.28	.04	.85	-.03	-.30	.46	-.26	.751
7	.06	.84	-.12	-.27	.04	.85	-.11	-.37	.76	-.10	.777
8	.07	.85	-.05	-.26	.03	.86	-.15	-.34	1.40	-.10	.787
9	.08	.86	-.07	-.36	.02	.87	-.07	-.30	1.77	-.23	.800
10	.08	.87	-.05	-.31	.02	.87	-.04	-.20	1.52	-.06	.808
11	.07	.87	-.05	-.26	.02	.87	.00	-.18	1.55	-.07	.822
12	.07	.88	-.08	-.27	.02	.87	-.04	-.12	1.56	.27	.833
13	.07	.88	-.08	-.28	.02	.86	-.05	-.10	1.56	.52	.845
14	.07	.88	-.09	-.22	.02	.86	-.06	-.15	1.66	.64	.853
15	.06	.88	-.10	-.26	.02	.87	-.06	-.19	1.57	.50	.858
16	.06	.88	-.10	-.25	.02	.87	-.01	-.21	1.51	.40	.860
17	.07	.89	-.10	-.23	.02	.87	.01	-.22	1.49	.77	.864
18	.07	.89	-.08	-.25	.02	.87	-.00	-.21	1.68	.67	.869
19	.07	.89	-.09	-.23	.03	.87	-.01	-.17	1.48	.60	.871
20	.06	.89	-.07	-.23	.03	.87	-.02	-.15	1.05	.57	.875
21	.06	.89	-.08	-.24	.03	.87	-.03	-.12	.99	.69	.877
22	.06	.89	-.06	-.25	.04	.88	-.03	-.14	1.08	.67	.884
23	.06	.89	-.07	-.27	.04	.88	-.04	-.15	.80	.47	.887
24	.06	.89	-.06	-.26	.04	.88	-.06	-.13	.84	.44	.887
25	.06	.89	-.09	-.25	.04	.87	-.05	-.14	.78	.72	.888
26	.06	.90	-.08	-.26	.04	.88	-.05	-.15	.62	.80	.889
27	.06	.90	-.08	-.24	.04	.88	-.05	-.15	.66	.85	.891
28	.06	.90	-.09	-.23	.04	.88	-.05	-.16	.65	.65	.893
29	.05	.90	-.07	-.24	.04	.88	-.05	-.17	.54	.65	.894
30	.06	.90	-.08	-.24	.04	.88	-.04	-.18	.59	.78	.897

*Differences statistically significant at $p \leq .05$.
**Differences statistically significant at $p \leq .01$.

Table F

Descriptive Statistics for Number-Correct and Proportion-Correct Scores for Forms 1 and 2 of the Conventional Test, t Values for Differences in Means and Variances Between the Forms, and Correlations Between Scores on the Two Forms, for Test Lengths of 1 to 30 Items

Test Length	Form 1 Number Correct Mean	SD	Skew-ness	Kur-tosis	Form 1 Proportion Correct Mean	SD	Form 2 Number Correct Mean	SD	Skew ness	Kur-tosis	Form 2 Proportion Correct Mean	SD	t Values Means	Variances	r
1	.87	.33	-2.25	3.08	.87	.33	.95	.22	-4.04	14.42	.95	.22	-3.30***	6.87***	.162
2	1.10	.52	.13	.59	.55	.26	1.72	.52	-1.68	1.97	.86	.26	-14.83***	-.06	.130
3	2.04	.57	-.48	1.89	.68	.19	2.40	.80	-1.17	.58	.80	.27	-7.10***	-5.62***	.271
4	2.45	.81	-.14	.10	.61	.20	2.77	1.06	-.57	-.39	.69	.27	-5.18***	-4.92***	.427
5	3.31	.92	-.36	.33	.66	.18	3.88	1.13	-.47	-.31	.58	.23	6.60***	-3.88***	.463
6	4.16	1.08	-.68	.75	.69	.18	4.19	1.34	-.22	-.50	.53	.22	12.71***	-4.04***	.487
7	4.68	1.35	-.51	.18	.67	.19	4.78	1.57	-.27	-.49	.54	.23	11.19***	-3.14**	.607
8	5.09	1.57	-.29	.03	.64	.20	5.63	1.75	-.50	-.15	.58	.22	5.50***	-2.31*	.659
9	5.99	1.67	-.37	.16	.67	.19	6.53	1.89	-.68	.25	.61	.21	5.11***	-2.80**	.675
10	6.08	1.77	-.12	.28	.61	.18	6.67	2.01	-.40	.06	.57	.20	4.42***	-2.99**	.694
11	7.02	1.83	-.21	.30	.64	.17	7.44	2.28	-.54	.07	.59	.21	5.89***	-5.14***	.709
12	7.55	2.10	-.10	-.12	.63	.18	7.93	2.52	-.40	-.19	.58	.21	6.18***	-4.63***	.756
13	8.13	2.36	-.07	-.35	.63	.18	8.85	2.59	-.45	-.02	.60	.20	2.53*	-2.26*	.748
14	8.20	2.45	.10	.23	.59	.18	9.00	2.71	-.27	-.12	.57	.19	1.83	-2.53*	.760
15	9.08	2.52	.06	-.16	.61	.17	9.57	3.01	-.27	-.29	.57	.20	4.32***	-4.55***	.771
16	9.81	2.74	-.04	-.26	.61	.17	10.48	3.11	-.34	-.14	.59	.20	2.79**	-3.41***	.789
17	10.72	2.82	-.10	-.21	.63	.17	10.98	3.39	-.28	-.29	.59	.20	5.98***	-5.01***	.798
18	11.09	3.08	-.03	-.30	.62	.17	11.30	3.60	-.17	-.39	.57	.20	6.16***	-4.41***	.816
19	12.00	3.21	-.15	-.21	.63	.17	11.67	3.83	-.09	-.40	.56	.20	10.02***	-5.11***	.824
20	12.60	3.45	-.11	-.40	.63	.17	12.35	4.06	-.11	-.46	.57	.20	9.06***	-4.81***	.831
21	13.03	3.75	-.04	-.54	.62	.18	13.12	4.23	-.14	-.49	.58	.20	6.57***	-3.71***	.846
22	13.84	3.86	-.08	-.55	.63	.18	14.04	4.29	-.16	-.45	.59	.20	5.80***	-3.23**	.850
23	14.45	4.11	-.12	-.60	.63	.18	14.43	4.52	-.11	-.46	.58	.20	7.00***	-3.05**	.853
24	15.18	4.31	-.15	-.56	.63	.18	15.39	4.58	-.15	-.38	.60	.19	5.40***	-1.94	.857
25	16.14	4.33	-.16	-.52	.65	.17	16.38	4.60	-.17	-.37	.62	.18	5.17***	-1.88	.857
26	17.07	4.42	-.20	-.48	.66	.17	17.27	4.69	-.19	-.38	.63	.18	5.48***	-1.91	.862
27	17.66	4.67	-.21	-.49	.65	.17	17.45	4.83	-.12	-.39	.61	.18	8.12***	-1.09	.869
28	18.52	4.81	-.26	-.44	.66	.17	18.39	4.91	-.17	-.31	.62	.18	7.58***	-.72	.875
29	19.35	4.96	-.30	-.40	.67	.17	19.22	5.04	-.22	-.21	.63	.17	7.57***	-.53	.881
30	19.57	5.16	-.24	-.44	.65	.17	20.11	5.14	-.26	-.17	.64	.17	3.06**	.15	.886

*Differences statistically significant at $p \leq .05$.
**Differences statistically significant at $p \leq .01$.
***Differences statistically significant at $p \leq .001$.

Table G
Pearson Product-Moment Correlations of Scores on Forms 1 and 2 of the Adaptive Test (A1, A2) and the Conventional Test (C1, C2) with Number-Correct Scores on the 50-Item Criterion Test, and Results of Tests of the Significance of Differences in Pairs of Correlations

Test Length	Adaptive		Conventional		Significant Differences			
					A1 vs.		A2 vs.	
	Form 1	Form 2	Form 1	Form 2	C1	C2	C1	C2
1	.39	.39	.28	.31				
2	.60	.53	.25	.52	*		*	
3	.64	.63	.37	.61	*		*	
4	.69	.68	.51	.67	*		*	
5	.73	.72	.56	.67	*		*	
6	.76	.74	.61	.67	*	*	*	
7	.77	.76	.65	.68	*	*	*	
8	.79	.77	.67	.70	*	*	*	
9	.79	.78	.70	.71	*	*	*	
10	.80	.79	.71	.73	*		*	
11	.80	.80	.71	.76	*		*	
12	.82	.81	.72	.77	*		*	
13	.82	.81	.71	.77	*		*	
14	.82	.82	.72	.77	*		*	
15	.83	.83	.72	.77	*		*	
16	.83	.82	.74	.78	*		*	
17	.83	.82	.75	.78	*		*	
18	.83	.82	.75	.78	*		*	
19	.83	.83	.76	.78	*		*	
20	.83	.83	.77	.78				
21	.83	.83	.77	.78				
22	.84	.83	.77	.78	*			
23	.84	.83	.78	.78	*			
24	.84	.83	.78	.79				
25	.84	.84	.78	.79	*			
26	.84	.84	.79	.79				
27	.84	.84	.79	.79				
28	.84	.84	.79	.80				
29	.84	.84	.79	.80				
30	.84	.84	.80	.81				

*$p \leq .05$.

Appendix Table H
Error Screens Reported During Introduction to CRT Usage
as a Function of the Error Screens Preceding Them for
the Adaptive (N=263) and Conventional (N=267) Test Groups

Resulting Error Screen	Group	Original Error Screen 9900	9901	9902	9904	9906	9213	9214	9216	9218	9220	9223	Total
9001	Adap	1	7	5	33	9	2	1	1	1	1	3	64
	Con			4	34	7	1		3	2		11	62
9060	Adap		1		3			1			1	1	7
	Con				3	2		1				4	10
9061	Adap				1			2				1	4
	Con			2	2	1		2	1			2	10
9213	Adap			2									2
	Con			1				1					2
Total	Adap	1	8	7	37	9	2	4	1	1	2	5	77
	Con			7	39	10	1	4	4	2		17	84
Total Errors		1	8	14	76	19	3	8	5	3	2	22	161