# An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests

G. Gage Kingsbury
and
David J. Weiss

RESEARCH REPORT 80-5
DECEMBER 1980

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

This research was supported by funds from the Air Force Office of Scientific Research, the Army Research Institute, the Air Force Human Resources Laboratory, and the Office of Naval Research, and monitored by the Office of Naval Research.

Unclassified

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

# REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Research Report 80-5

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*
An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR*(s)*
G. Gage Kingsbury and David J. Weiss

8. CONTRACT OR GRANT NUMBER*(s)*
N00014-76-C-0243
N00014-79-C-0172

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
PE: 6115N Proj: RR042-04
TA: RR042-04-01
WU: NR150-382, 150-433

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
December 1980

13. NUMBER OF PAGES
20

14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*

15. SECURITY CLASS. *(of this report)*
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

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

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

This research was supported by funds from the Army Research Institute, the Air Force Human Resources Laboratory, the Air Force Office of Scientific Research, 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)*

Computerized Testing
Adaptive Testing
Tailored Testing
Sequential Testing
Individualized Testing
Response-Contingent Testing
Latent Trait Test Theory
Item Characteristic Curve Theory
Item Response Theory
Bayesian Testing
Ability Testing

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

Two 30-item alternate forms of a conventional test and a Bayesian adaptive test were administered by computer to 472 undergraduate psychology students. In addition, each student completed a 120-item paper-and-pencil test, which served as a concurrent validity criterion test, and a series of very easy questions designed to detect students who were not answering conscientiously. All test items were five-alternative multiple-choice vocabulary items.

Unclassified

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

Reliability and concurrent validity of the two testing strategies were evaluated after the administration of each item for each of the tests, so that trends indicating differences in the testing strategies as a function of test length could be detected. For each test, additional analyses were conducted to determine whether the two forms of the test were operationally alternate forms.

Results of the analysis of alternate-forms correspondence indicated that for all test lengths greater than 10 items, each of the alternate forms for the two test types resulted in fairly constant mean ability level estimates. When the scoring procedure was equated, the mean ability levels estimated from the two forms of the conventional test differed to a greater extent than those estimated from the two forms of the Bayesian adaptive test.

The alternate-forms reliability analysis indicated that the two forms of the Bayesian test resulted in more reliable scores than the two forms of the conventional test for all test lengths greater than two items. This result was observed when the conventional test was scored either by the Bayesian or proportion-correct method.

The concurrent validity analysis showed that the conventional test produced ability level estimates that correlated more highly with the criterion test scores than did the Bayesian test for all lengths greater than four items. This result was observed for both scoring procedures used with the conventional test.

Limitations of the study, and the conclusions that may be drawn from it, are discussed. These limitations, which may have affected the results of this study, included possible differences in the alternate forms used within the two testing strategies, the relatively small calibration samples used to estimate the ICC parameters for the items used in the study, and method variance in the conventional tests.

# CONTENTS

# An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests

The potential advantages of the use of computerized adaptive testing to more effectively assess individuals' ability levels have been pointed out by a number of researchers (e.g., Lord, 1977a; Urry, 1977; Weiss, 1974; Weiss & Betz, 1973). The most widely used approaches to adaptive testing use item characteristic curve or item response theory (IRT; Lord & Novick, 1968) to adapt a test given to an individual to his or her trait level by administering items with characteristics that allow very efficient measurement. A number of research studies have been concerned with how well the potential advantages of adaptive testing are borne out in live studies (e.g., Bejar & Weiss, 1978; Betz & Weiss, 1975; Larkin & Weiss, 1974; Thompson & Weiss, 1980).

One type of procedure used for adapting test characteristics is a Bayesian algorithm for adaptive testing developed by Owen (1969, 1975). This Bayesian procedure has been studied in both monte carlo simulation and live studies (e.g., Jensema, 1974; McBride & Weiss, 1976, 1977; Urry, 1971), which have attempted to explicate the properties of the testing strategy. In many of these studies, however, testing strategies were evaluated using criteria derived from IRT rather than using classical reliability and validity concepts. The present study investigated how well this Bayesian adaptive testing procedure performed relative to a conventional testing strategy, using two classical psychometric indices of test performance.

## Method

Owen's Bayesian adaptive testing strategy was compared in two ways to a conventional testing strategy. After each item was administered, the two testing strategies were compared in terms of (1) their alternate-forms reliability and (2) their concurrent validity.

### Subjects

The subjects taking part in this study were 472 undergraduate students at the University of Minnesota. These students volunteered to take part in the study as partial fulfillment of the requirements of the general psychology course in which they were enrolled. Subjects were recruited and tested during the winter and spring academic quarters of 1976.

### Test Administration

Each volunteer took each of the following vocabulary ability tests during the testing session:

1. A 120-item conventional test administered in paper and pencil format.
2. Two 30-item conventional tests administered by computer and designed to be parallel tests.

3. Two 30-item Bayesian adaptive tests administered by computer and designed to be parallel tests.
4. Three 3-item "catch trials" administered by computer and consisting of extremely easy questions.

All of the 249 items administered during the testing session were five-alternative multiple-choice items.

Each student began the testing session by taking the 120-item conventional test. Scores on this test served as the criterion against which the relative validities of the two types of computer-administered tests were judged.

Following the criterion test, the order of administration of the two types of computer-administered tests was counterbalanced. Half of the students were given the two parallel forms of the Bayesian adaptive test, followed by the two forms of the conventional test; the other half received the conventional tests first, followed by the Bayesian tests.

For both the conventional and Bayesian tests, the two parallel forms were administered as close to simultaneously as possible. To operationalize this, an ABBA rotation was used; that is, one item was administered from Form A to begin the test, followed by two items from Form B, followed by two items from Form A. For each individual the prior distribution specified at the beginning of each of the Bayesian test forms had a mean of 0.0 and a standard deviation of 1.0.

Three catch trials consisting of three very easy items each were included during the computer-administered testing period. These catch trials were designed to identify students who were exceptionally careless, who deliberately responded incorrectly, or who did not understand the instructions. Once these individuals were identified, they would be marked as having inappropriate response patterns.

The catch trial items were not separated in any way from the actual tests. The first catch trial consisted of the first three items administered by the computer. The second catch trial occurred at the middle of the computerized test session (i.e., between the two different types of computer-administered tests). The third catch trial consisted of the last three items administered by the computer.

## Test Design and Scoring

Criterion test. The criterion test administered to the students consisted of 120 vocabulary questions taken from Part III of Forms 2A, 2B, 3A, and 3B of the Cooperative School and College Ability Tests (SCAT I). This test was a portion of the item pool described by Lord (1977b) as a broad-range item pool for the measurement of verbal ability. The items were five-alternative multiple-choice questions which had been extensively normed and for which item parameter estimates from the three-parameter normal ogive IRT model were available. The parameter estimates for the items making up the criterion test are shown in Appendix Table A. The criterion test was scored using Owen's IRT-based Bayesian scoring method.

Conventional tests. The two conventional test forms were designed to be parallel tests, peaked at an average ability level. An item pool, which contained 577 five-alternative multiple-choice vocabulary items (McBride & Weiss, 1974), was available for use. For each of these items, estimates of the a (item discrimination) and b (item difficulty) parameters, which had earlier been calculated using Jensema's (1976) approximation procedure, were available. Since each of the items had five choices, the estimate of c (the lower asymptote parameter) had been set at .20 for each item. The method used to calculate item parameter estimates, corrected for guessing, is described by Prestwood and Weiss (1977).

From this large item pool, 120 items were selected that had the highest available information at the ability level ($\theta$) of 0.0 with difficulty estimates between -1.0 and +1.0. These 120 items were further subdivided into two 60-item pools equated for available information at $\theta$=0.0. One of these 60-item pools was used as a portion of the Bayesian testing pool (described below), and the other was used to construct the two alternate forms of the conventional test.

The two 30-item forms of the conventional test were constructed from the 60-item pool in order to equate as closely as possible the amount of information available at $\theta$=0.0 in each form after each item was administered. Thus, the first item chosen for Form A was the most informative item at $\theta$=0.0, the next two most informative items at $\theta$=0.0 were chosen to serve as the first two items of Form B, then the next two most informative items were chosen as the next two items for Form A, and so on until the last item in the 60-item pool was chosen to serve as the last item of Form A. The parameter estimates for the items making up each of the conventional test forms are shown in Appendix Table B in the order of their administration.

Conventional tests were scored by proportion correct at each test length from 1 to 30 items. In addition, to maximize comparability with the IRT-scored Bayesian adaptive test, the conventional tests were also scored by Owen's Bayesian scoring method, and scores were recorded at all test lengths.

Bayesian adaptive tests. The two Bayesian adaptive test forms both drew items from a single 180-item pool in the ABBA fashion described above. For any one individual, a given item appeared only on one form (if at all); but across individuals, a single item might have appeared on Form A for one person, Form B for another person, and neither form for a third person.

Sixty of the items in the 180-item Bayesian item pool came from the 60-item pool developed as described above. The additional 120 items were selected from the remainder of the original 577-item pool. The items that were chosen were 6 groups of 20 items each that provided the most information at 6 ability levels ($\theta$=-2.0, -1.5, -1.0, 1.0, 1.5, 2.0). The parameter estimates for the 180 items in the final Bayesian testing pool are shown in Appendix Table C.

The Bayesian adaptive test ability estimates were recorded for each of the two dynamically administered parallel forms at each test length from 1 to 30 items.

## Analyses

Catch trial analysis. Prior to all other analyses, those subjects who failed to correctly answer at least seven of the nine items administered during the catch trials were removed from further analyses. This was intended to identify those subjects who incorrectly answered these extremely easy items, thus indicating that they either misunderstood the instructions, were deliberately answering incorrectly, or were careless. Once these individuals were identified, a more detailed analysis of their response patterns was planned to determine whether the catch trials had performed successfully (i.e., had detected individuals with very inconsistent response patterns).

Correspondence of test forms. The two forms of the conventional test were designed to measure vocabulary ability in the same manner with approximately the same precision, especially for individuals with average ability levels ($\theta$=0.0). To determine whether the design had been satisfactorily achieved, three criteria were used. First, the theoretical test information functions (Birnbaum, 1968) for the two forms were calculated and inspected for differences in their general shape and in the amount of information available at $\theta$=0.0. (The theoretical information function serves as an upper bound to the amount of information which may be recovered from the items. The actual information recovered is a function of the scoring procedure employed.)

The second criterion was the mean Bayesian ability estimate computed within the testing sample after each item was administered within each test form. This was a reasonable criterion because at every test length the test forms were designed to measure the same ability with an equal degree of precision. To the extent that the two forms did not produce the same mean ability estimate for the same group of people, it could be concluded that the two test forms were not measuring in the same manner.

The third criterion used to evaluate the equivalence of the two conventional test forms was the mean proportion of items answered correctly within the testing sample after each item was administered within each test form. The rationale behind this criterion was the same as that used for the second criterion, except that the more widely used proportion-correct scoring system was used here in place of the Bayesian ability estimation procedure.

For the Bayesian test forms, the item selection procedure used in this study was designed to result in two test forms that measured the same ability with approximately the same precision after each item was administered by the two forms. To determine the effectiveness of this design in terms of equalizing the two Bayesian test forms, the first and third criteria used for the analysis of the conventional test forms were inappropriate. The first criterion was inappropriate since the theoretical test information functions for the two forms would be different for each person taking the adaptive tests; and the third was inappropriate because the observed proportion correct is not an estimate of an individual's true ability level within the context of an adaptive test. Consequently, for the Bayesian test forms the equivalence of the two forms was examined by observing the differences in the mean Bayesian ability estimate obtained from the two test forms, following the administration of each item to the students.

Alternate forms reliability. The two testing strategies were compared in terms of the alternate forms reliability of the ability level estimates obtained for individuals from the two alternate test forms. Specifically, Pearson product-moment correlations were calculated between the ability level estimates obtained from the alternate test forms at all test lengths from 1 to 30 items. For the conventional test, two different ability level estimates were available--proportion correct and Bayesian. Therefore, two different alternate-forms reliability coefficients were computed at each conventional test length.

Concurrent validity. Bayesian ability level estimates were obtained for each subject based on their responses to the 120-item paper-and-pencil criterion test. Correlations between the ability level estimates obtained from the various computer-administered tests and the criterion test ability estimates were calculated at each possible test length, for each computer-administered test form.

For the Bayesian test, 30 validity coefficients were calculated for each of the two test forms. Similarly, for the conventional test, 30 validity coefficients were calculated for each of the four combinations of a scoring strategy and a test form. To facilitate the comparison of the two testing strategies and to attain more stable estimates of validity, validity coefficients that resulted from the alternate forms of the same test type using the same scoring strategy were averaged across test forms at each test length.

## Results

### Catch Trial Analysis

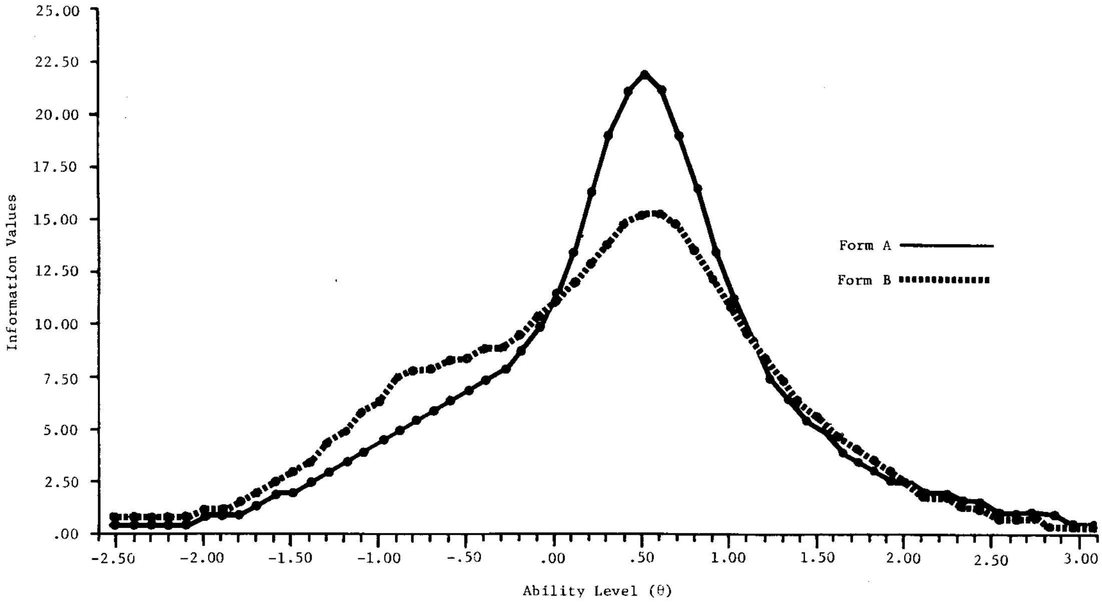
Of the 472 students in the testing sample, none failed to correctly answer at least seven of the catch trial items. Thus, none of the students' response patterns were removed from the data set used in the analyses reported below. In the entire testing sample, 95% of the students answered all nine of the catch trial items correctly. The other 5% of the sample correctly answered eight of the nine catch trial questions. No individual answered less than eight of the questions correctly.

### Correspondence of Test Forms

The theoretical test information functions (i.e., the sums of the item information functions) for Forms A and B of the conventional test are shown in Figure 1. It can be seen from this figure that each of the test forms was fairly sharply peaked. For both forms the information peak was reached between $\theta$=.5 and .6. The information peak calculated for Form A, 21.90 information units (IU), was higher than that for Form B, 15.66 IU. At the ability level at which the two test forms were designed to provide the same amount of information, $\theta$=0.0, Form A had a potential of 11.580 IU, and Form B had a potential of 11.055 IU. In terms of their information potential, the two conventional test forms conformed to their design specifications fairly well and should have resulted in approximately equally precise ability estimates for ability levels near $\theta$=0.0.

Figure 2 shows the mean proportion of correct answers observed within the

Figure 1
Theoretical Information Available from Forms A and B
of the Conventional Test, as a Function of Ability Level



testing sample after each item was administered within the conventional test, for both Forms A and B. It can be seen from this figure that the mean observed proportion correct for each of the test forms varied somewhat for test lengths up to about 10 items. For Form A the highest mean proportion correct (.55) was observed following the administration of the third item, and the lowest mean proportion correct (.32) was observed following the administration of the first item. For Form B the highest and lowest mean proportion-correct values (.57 and .41) were observed following the first and third items, respectively. Following this initial fluctuation, each test form resulted in quite consistent observations of the mean proportion-correct values at all longer test lengths. Following the first 10 items, the highest mean proportion correct observed for Form A was .50 following Item 12, and the lowest was .47 following Item 21. For Form B, after the first 10 items, the highest and lowest mean proportion-correct values were .55 and .52, following Item 22 and Item 17, respectively. Form A resulted in a mean proportion-correct value of .48 after all 30 items were administered, whereas Form B resulted in a value of .53.

Figure 3 shows the mean Bayesian ability level estimate observed across the testing sample within each of the conventional test forms, following the administration of each item. The pattern of Bayesian ability level estimates shown in Figure 3 is very similar to that of the pattern of mean proportion-correct values in Figure 2. As in the proportion-correct analysis, the mean Bayesian ability level estimates for each form were most variable in the first third of the test, becoming much less variable as the test proceeded. For Form A the highest mean Bayesian ability level estimate that was observed was -.13, following the third item, whereas the lowest mean estimate was -.44, following the 18th item. For Form B, the highest mean estimate was .02, after the first item, and the lowest estimate was -.31, following the 15th item. After 30 items were administered for each of the conventional test forms, the mean ability estimate observed was -.40 for Form A and -.28 for Form B.

Figure 2
Mean Proportion of Items Answered Correctly for Two Conventional Test Forms, as a Function of Number of Items Administered

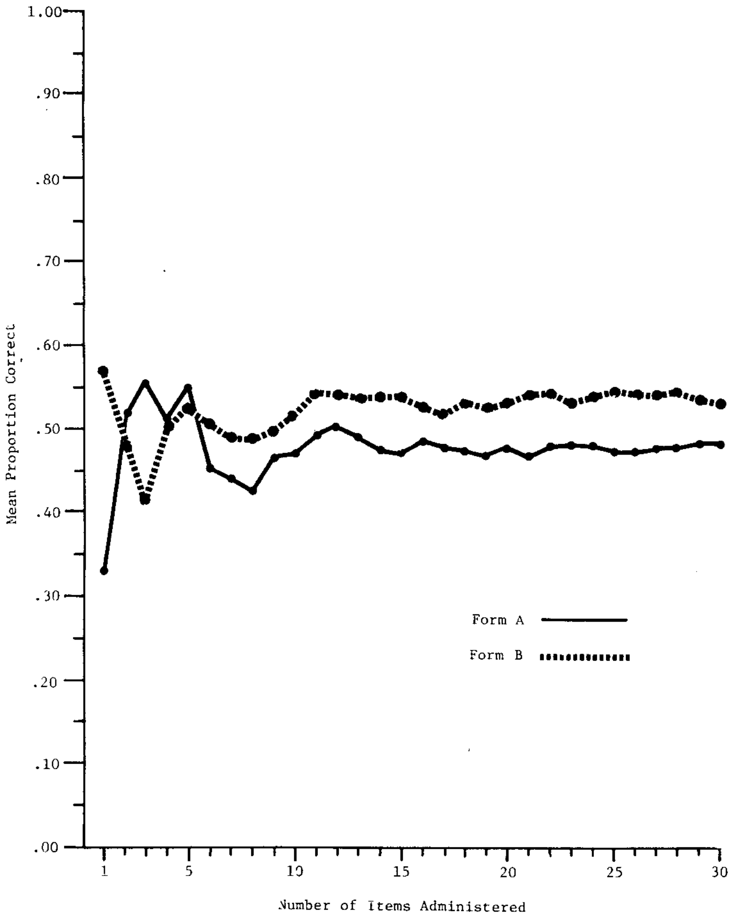
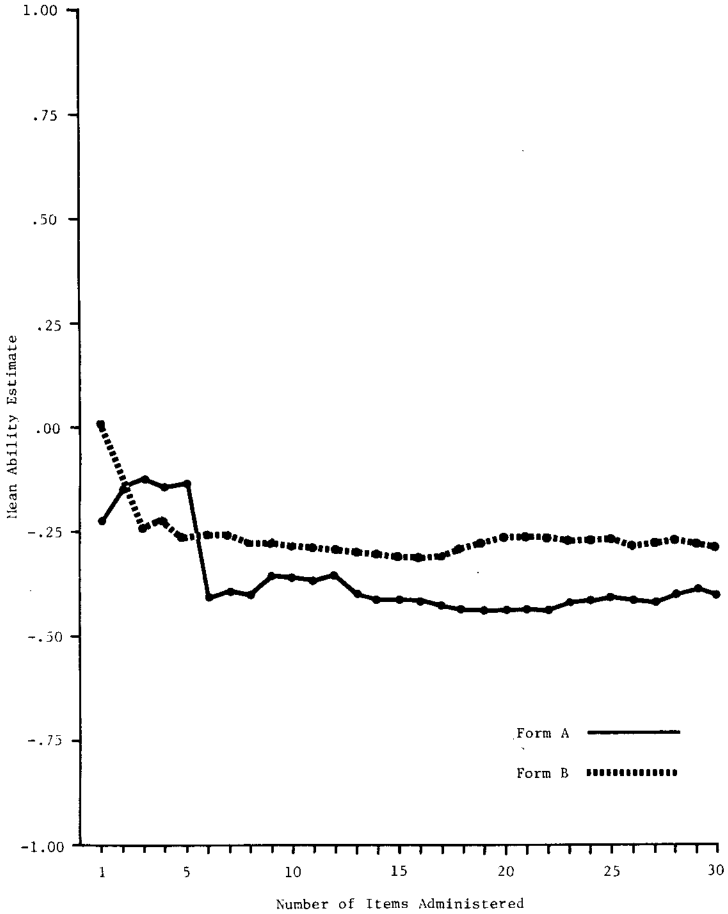


Figure 4 shows the mean Bayesian ability level estimate observed within the testing sample following the administration of each item on each of the forms of the Bayesian adaptive test. For Form A the highest mean ability estimate observed was -.27, following the 13th item. The lowest mean ability estimate for Form A was -.36, after the second item was administered. For Form B the range of the mean ability estimates was from -.03 to -.29. These estimates were observed following the first and last items, respectively. Following the administration of the final item from each of the Bayesian test forms, the mean ability level estimate observed was -.32 for Form A and -.29 for Form B.

## Alternate Forms Reliability

Figure 5 shows the Pearson product-moment correlations between the ability level estimates obtained from the two forms of the conventional test using the Bayesian scoring strategy and proportion-correct scoring strategy and from the two forms of the Bayesian test using the Bayesian scoring strategy (the numerical values are shown in Appendix Table D). These correlations serve as estimates of the alternate-forms reliabilities of the different test types. The most obvious result reflected in this figure is that except for the first two items administered, the Bayesian adaptive test resulted in higher alternate

Figure 3
Mean Bayesian Ability Level Estimates for Two Conventional Test Forms, as a Function of Number of Items Administered



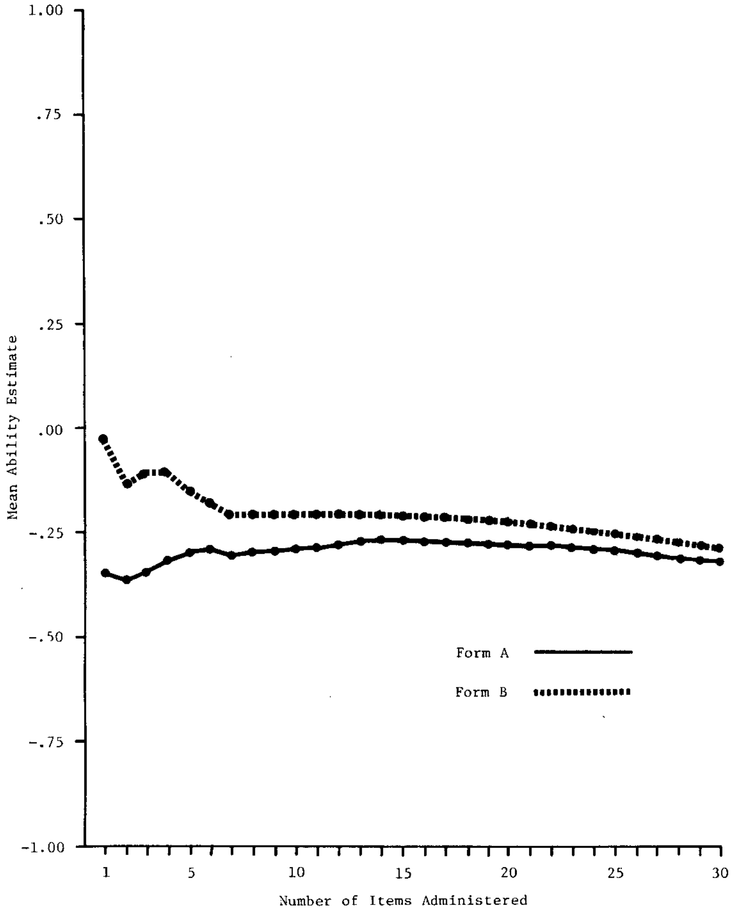
forms reliability than the conventional test at all test lengths, regardless of the scoring method used for the conventional test. Further, the difference in reliability between the two testing strategies increased as the length of the tests increased from 10 to 30 items. Following the administration of the final item, the reliability of the Bayesian test was .920, whereas for the conventional test the reliabilities observed were .879 and .868, respectively, for the proportion-correct and Bayesian scoring strategies.

Another result shown in Figure 5 is that both the Bayesian and proportion-correct scoring strategies resulted in very similar reliabilities for the conventional test. This finding is counter to expectation, since a scoring strategy that uses information concerning differences among the items when scoring should result in more reliable ability level estimates than a scoring system that treats all of the items as if they were the same.

## Concurrent Validity

Figure 6 shows the mean Pearson product-moment correlations between the Bayesian ability level estimates derived for the testing sample from the

Figure 4
Mean Bayesian Ability Level Estimate for Two Bayesian Test Forms, as a Function of Number of Items Administered

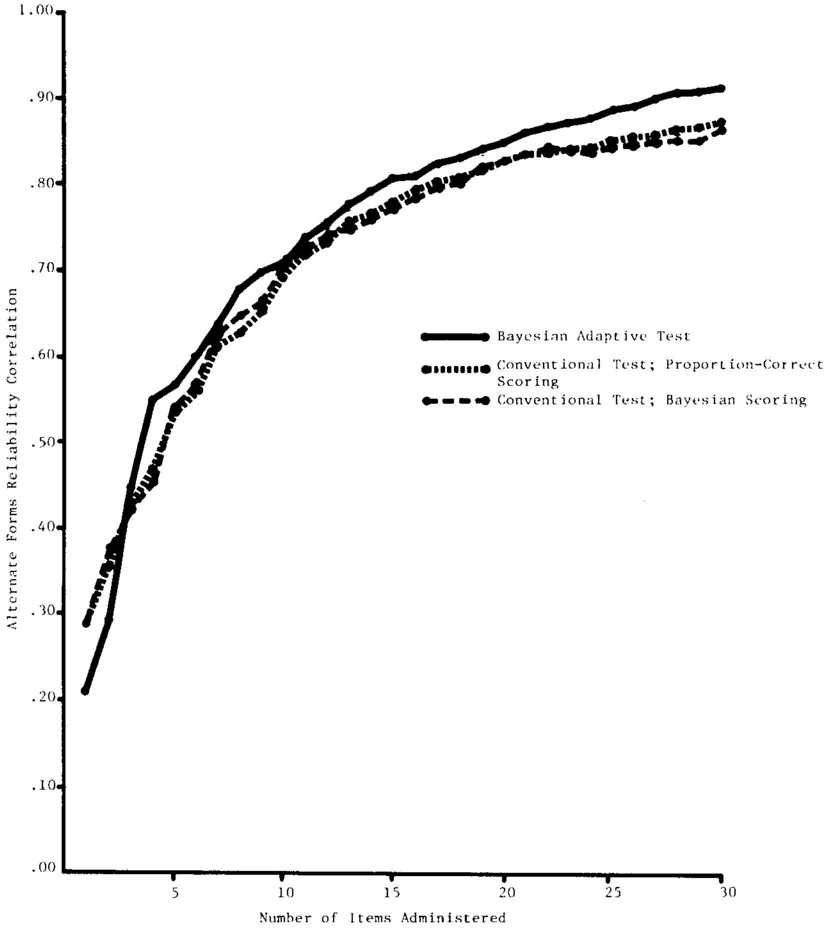


120-item paper-and-pencil criterion test and the ability estimates derived from the Bayesian and conventional tests, across all test lengths (numerical values are shown in Appendix Table E). The conventional test forms were again scored using both the proportion-correct scoring system and the Bayesian scoring system. As was indicated above, the values shown in Figure 6 are mean correlations, averaged across the two forms of the test involved.

From Figure 6, the first trend observed is that for all test lengths greater than four items, the conventional test scores were more highly correlated with the criterion scores than were the scores derived from the Bayesian adaptive test forms. Following the final item, the Bayesian adaptive test scores resulted in a criterion test correlation of .797, the conventional test Bayesian scores had a criterion correlation of .834, and the conventional test proportion-correct scores had a criterion correlation of .841.

A second trend seen is that for the conventional test, the proportion-correct scoring method resulted in scores that had a slightly higher criterion correlation than Bayesian scoring at all test lengths greater than three items. Across all test lengths, the average difference in the criterion correlation was .008, a small but consistent difference.

Figure 5
Alternate Forms Reliability of Ability Level Estimates for the Bayesian Adaptive Test and for the Conventional Test Scored by Proportion-Correct and Bayesian Scoring, as a Function of the Number of Items Administered

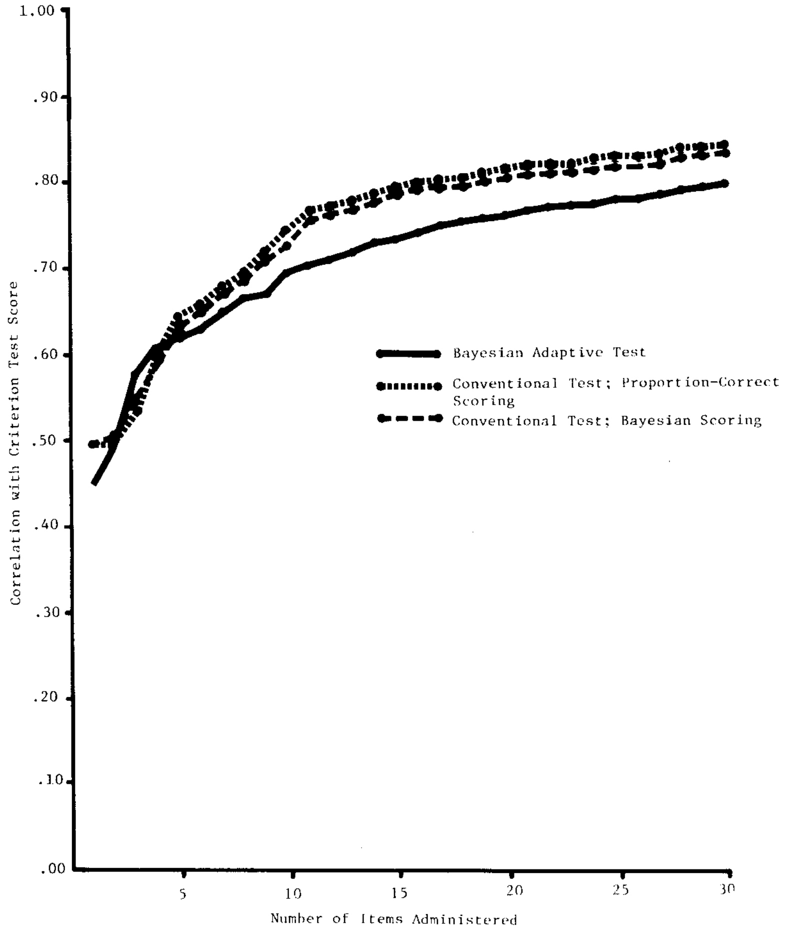


A final trend, which is seen in Figure 6, is that the largest criterion correlation difference between the Bayesian test and the conventional test (using either scoring system) occurred following the administration of the 11th item (.056 with Bayesian scores and .065 with the proportion-correct scores). For longer test lengths the two testing strategies resulted in increasingly similar criterion correlations until, after the last item was administered, the differences in the criterion correlations derived from the Bayesian testing strategy and the conventional testing strategy were .037 (scoring the conventional test by the Bayesian scoring method) and .044 (scoring the conventional test by the proportion-correct scoring method).

## Discussion and Conclusions

The results of this study imply that with the subjects and item pools used the Bayesian adaptive testing strategy results in test scores that are more reliable and less valid than the scores derived from a conventional testing strategy for test lengths greater than about 10 items.

Figure 6
Correlations of Criterion Test Scores with Ability Level Estimates from the Bayesian Adaptive Test and the Conventional Test Scored by Proportion-Correct and Bayesian Scoring, as a Function of the Number of Items Administered (Averaged Across two Test Forms)



To more accurately reflect what has been done in this study, it is important to more closely examine two factors:

1. The correspondence of the alternate forms used for the analysis of alternate-forms reliability with the two testing strategies, and
2. The relative performance of the two scoring methods within the two forms of the conventional test.

## Correspondence of Alternate Forms

Examination of the mean Bayesian ability level estimates obtained from Forms A and B for the two testing strategies (Figures 3 and 4) provides important information. The mean ability level estimates produced by the Bayesian test forms were less disparate than the Bayesian estimates produced by the conventional test forms at almost all test lengths. If perfectly parallel test forms were used, mean ability estimates would differ from one form to the other only by measurement error. With a suitably large testing sample, the mean ability estimates should converge to a common value. To the extent that two forms of a test result in different mean ability level estimates, (1) the two test forms have observable measurement error or (2) the two test forms were not perfectly parallel. Thus, the observation that the forms of the conventional test resulted in mean ability level estimates that were more disparate than those produced by the two forms of the Bayesian test can be attributed to either (1) the conventional test resulting in more measurement error than the Bayesian adaptive test or (2) the Bayesian test forms being closer to parallel than the conventional test forms. Either explanation is feasible, and the available data permit no method for gaining support for one explanation or the other.

It is possible, then, that as with the disparate mean ability estimates, the differential reliability of the scores derived from the two testing strategies can be attributed to either a true difference in the reliabilities of the scores derived from the two testing strategies or to differences in the approximation of the test forms to perfect parallelism. This possibility may limit the confidence that can be placed in the conclusion that the Bayesian testing strategy resulted in more reliable scores than the conventional testing strategy.

## Scoring Methods

The second factor to be taken into account in qualifying the conclusions is the relative performance of the two scoring strategies applied to the two conventional test forms. It has been noted above that the Bayesian and proportion-correct scoring methods resulted in very similar alternate-forms reliability coefficients for the conventional test (as shown in Figure 5).

The Bayesian scoring algorithm uses the item parameter estimates along with the observed pattern of item responses to determine the ability level estimate for each individual. This procedure gives differential weightings to each of the individual's responses, depending on the parameter estimates for the items. To the extent that the items differ from one another in terms of their difficulties, and particularly in terms of their discriminations, these differential item response weightings should reduce the amount of measurement error expected in the individual's ability level estimate. This trend should result in higher alternate-forms reliability for a test when it is scored using the Bayesian procedure than when it is scored using the proportion of correct answers.

This result was not seen in this study, and the reason may be that the parameter estimates used contained too much error to allow the Bayesian scoring procedure to perform at a level of efficiency high enough to result in higher reliabilities than the proportion-correct procedure. This line of argument has been presented by Lord (1979) in a paper that limited itself to the one- and two-parameter logistic models and a maximum likelihood trait level estimator, but the argument is clearly generalizable. If the parameters of a model are estimated using a small group of individuals, the resulting parameter estimates might be sufficiently poor to obviate the gain in precision of measurement (and, hence, reliability) that should be observed with the use of a more sensitive scoring procedure (such as the Bayesian procedure).

For the present study, the mean calibration sample size used for determining the item a and b parameter estimates for the items used in the conventional and Bayesian tests was less than 200, ranging from 61 to 328 subjects. It is not clear whether the calibration sample sizes used were sufficient to adequately estimate the parameters of the response model used for the purposes of this study.

If the subject sample used to calibrate the items in this study was too small to allow calibration that was accurate enough to result in increased reliability with the conventional test, however, these inaccurate parameter estimates would also have affected the performance of the Bayesian testing strategy. If there were inaccuracy in the item parameters, the effect on the Bayesian test would be twofold, decreasing the efficiency of both the item selection procedure and the scoring system. This factor could have caused this study to underestimate the reliability and validity that could be obtained with the Bayesian testing procedure with more accurate item parameter estimates, resulting in greater differences in reliabilities and unknown differences in validities for the two testing strategies.

## Method Variance

There is one additional explanation for the findings of this study, which assumes the accuracy of both the reliability and validity findings observed. This explanation assumes that the validity differential in favor of the conventional test is due to method variance, since both the experimental conventional test and the criterion test were conventional (i.e., nonadaptive tests). If conventional test scores tended to correlate higher with each other than with adaptive test scores due solely to characteristics of the conventional tests, the results of this study would be in accord with such a hypothesis. Both adaptive test theory and prior data suggest that adaptive tests have higher reliabilities than do conventional tests, and the data from this study support this contention. Similarly, a previous study (Thompson & Weiss, 1980), in which conventional tests were not used as a validity criterion, showed higher validities for adaptive tests than for conventional tests. Thus, the lower validities observed in this study for the adaptive tests could have resulted from method variance in the conventional test correlations. Such method variance may be due to the distributional characteristics of the conventional tests, to correlated errors, or to other aspects of the tests constructed and administered by the conventional strategy.

Thus, future research comparing the relative reliabilities and validities of conventional and adaptive testing strategies should carefully balance the correspondence between the alternate forms of the tests and should use large samples of subjects for the calibration of the items used as well as a research design and validity criterion that would minimize the potential effects of method variance on the results.

## References

Betz, N. E., & Weiss, D. J. *Empirical and simulation studies of flexilevel ability testing* (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1975.

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.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

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

Jensema, C. J. A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, 1976, *36*, 705-715.

Larkin, K. C., & Weiss, D. J. *An empirical investigation of computer-administered pyramidal ability testing* (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1974.

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138.

Lord, F. M. A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1977, *1*, 95-100.

Lord, F. M. Small N justifies the Rasch Model. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980.

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

McBride, J. R., & Weiss, D. J. *A word knowledge item pool for adaptive ability measurement* (Research Report 74-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1974.

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.

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. Accuracy of perceived test-item difficulties (Research Report 77-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, May 1977.

Thompson, J. G., & Weiss, D. J. Criterion-related validity of adaptive testing strategies (Research Report 80-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1980.

Urry, V. W. Individualized testing by Bayesian estimation (Research Bulletin 0171-177). Seattle: University of Washington, Bureau of Testing, April 1971.

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

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

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973.

## APPENDIX: SUPPLEMENTARY TABLES

Table A
Item Parameter Estimates for Items on the Criterion Test

| Item | a | b | c | Item | a | b | c | Item | a | b | c |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .830 | -2.852 | .130 | 41 | 1.132 | .182 | .191 | 81 | .821 | -.158 | .139 |
| 2 | .421 | -.820 | .130 | 42 | 1.935 | .540 | .253 | 82 | 1.193 | .638 | .139 |
| 3 | 1.039 | -1.009 | .199 | 43 | 1.151 | -.318 | .150 | 83 | 2.101 | -.069 | .194 |
| 4 | .900 | -1.804 | .130 | 44 | .135 | 4.558 | .150 | 84 | .957 | -.038 | .139 |
| 5 | .621 | -.945 | .130 | 45 | 1.183 | -.026 | .100 | 85 | 1.061 | .406 | .150 |
| 6 | 1.173 | -1.014 | .130 | 46 | .505 | .601 | .150 | 86 | 1.869 | -.038 | .181 |
| 7 | .394 | -.033 | .130 | 47 | .908 | .001 | .150 | 87 | 2.104 | .831 | .054 |
| 8 | 1.381 | .860 | .305 | 48 | 1.623 | .430 | .183 | 88 | 1.157 | .526 | .111 |
| 9 | 1.373 | -.332 | .199 | 49 | 1.726 | .516 | .197 | 89 | 1.882 | 1.471 | .197 |
| 10 | 1.862 | .147 | .084 | 50 | .516 | -.691 | .150 | 90 | 2.104 | 1.271 | .170 |
| 11 | .740 | -1.717 | .130 | 51 | 1.128 | .643 | .203 | 91 | .593 | -2.830 | .150 |
| 12 | 1.472 | .816 | .167 | 52 | 1.611 | .291 | .150 | 92 | 1.289 | -1.394 | .150 |
| 13 | .899 | -.020 | .198 | 53 | .814 | .176 | .150 | 93 | .742 | -.918 | .150 |
| 14 | 1.862 | .992 | .316 | 54 | .605 | .822 | .150 | 94 | .765 | -2.267 | .150 |
| 15 | .544 | -.437 | .130 | 55 | 1.699 | 1.048 | .184 | 95 | 1.047 | -1.009 | .150 |
| 16 | 1.862 | .383 | .197 | 56 | 1.935 | .856 | .100 | 96 | 1.588 | -.416 | .196 |
| 17 | 1.611 | .799 | .130 | 57 | .555 | .674 | .150 | 97 | 1.302 | -.913 | .150 |
| 18 | 1.378 | .352 | .130 | 58 | .747 | .085 | .115 | 98 | 1.347 | -1.569 | .150 |
| 19 | 1.282 | .692 | .179 | 59 | 1.935 | 1.888 | .122 | 99 | .605 | -2.075 | .150 |
| 20 | 1.862 | .522 | .061 | 60 | 1.935 | 1.255 | .110 | 100 | 1.034 | -.266 | .150 |
| 21 | .892 | .376 | .191 | 61 | .908 | -2.746 | .139 | 101 | .884 | -1.016 | .150 |
| 22 | 1.862 | 1.906 | .147 | 62 | .737 | -2.463 | .139 | 102 | 1.068 | .535 | .195 |
| 23 | 1.339 | .225 | .174 | 63 | .516 | -3.818 | .139 | 103 | 1.285 | -.003 | .150 |
| 24 | 1.259 | 1.147 | .199 | 64 | 1.114 | -.952 | .139 | 104 | 1.281 | -1.168 | .150 |
| 25 | 1.523 | .898 | .089 | 65 | .718 | -1.288 | .139 | 105 | 1.083 | -.062 | .150 |
| 26 | 1.862 | .983 | .130 | 66 | .732 | -.817 | .150 | 106 | .501 | -.872 | .150 |
| 27 | .574 | 1.119 | .150 | 67 | 1.604 | -.983 | .139 | 107 | 1.123 | -.250 | .150 |
| 28 | 1.758 | 1.375 | .187 | 68 | 1.498 | -.888 | .139 | 108 | 1.679 | -.279 | .195 |
| 29 | 1.045 | 1.662 | .092 | 69 | 1.005 | -1.084 | .139 | 109 | .713 | -.883 | .150 |
| 30 | 1.862 | 2.620 | .169 | 70 | 1.226 | -.250 | .179 | 110 | 1.557 | -.299 | .150 |
| 31 | .675 | -2.523 | .150 | 71 | .993 | -.991 | .139 | 111 | 1.217 | .724 | .204 |
| 32 | .882 | -2.584 | .150 | 72 | 1.074 | -.697 | .139 | 112 | .877 | -.387 | .100 |
| 33 | .564 | -1.805 | .150 | 73 | 1.914 | -.355 | .254 | 113 | 1.355 | .697 | .210 |
| 34 | .745 | -1.721 | .150 | 74 | 1.513 | -.429 | .169 | 114 | 1.088 | -.027 | .150 |
| 35 | 1.076 | -.285 | .150 | 75 | .697 | -1.095 | .139 | 115 | 1.595 | .177 | .115 |
| 36 | 1.776 | .589 | .150 | 76 | .991 | -.618 | .150 | 116 | 1.782 | -.397 | .195 |
| 37 | .757 | -.070 | .150 | 77 | 2.104 | .054 | .210 | 117 | 1.312 | .243 | .182 |
| 38 | .950 | -.098 | .150 | 78 | 1.931 | .047 | .139 | 118 | .925 | -.413 | .100 |
| 39 | 1.908 | .182 | .210 | 79 | 2.104 | .433 | .248 | 119 | 1.745 | 1.330 | .171 |
| 40 | 1.935 | -.271 | .150 | 80 | 1.105 | -.545 | .081 | 120 | 2.161 | 1.430 | .071 |

Table B
Item Parameter Estimates for Items from
Alternate Forms A and B of the Conventional Test
in Order of Administration (*c*=.20 for All Items)

| Item | Form | *a* | *b* | Item | Form | *a* | *b* |
|---|---|---|---|---|---|---|---|
| 1 | A | 3.000 | .276 | 31 | B | 1.093 | .601 |
| 2 | B | 1.634 | .158 | 32 | A | 1.043 | -.962 |
| 3 | B | 1.627 | .289 | 33 | A | .831 | .171 |
| 4 | A | 1.223 | -.138 | 34 | B | .933 | .467 |
| 5 | A | 1.131 | -.197 | 35 | B | .823 | -.559 |
| 6 | B | 2.120 | .509 | 36 | A | .793 | -.034 |
| 7 | B | 1.644 | -.789 | 37 | A | .887 | .401 |
| 8 | A | 1.854 | .523 | 38 | B | 1.438 | .701 |
| 9 | A | 1.061 | -.393 | 39 | B | .771 | -.409 |
| 10 | B | 1.241 | -.763 | 40 | A | .742 | -.179 |
| 11 | B | 1.594 | .544 | 41 | A | 1.057 | .678 |
| 12 | A | .972 | -.396 | 42 | B | .758 | -.677 |
| 13 | A | 3.000 | .486 | 43 | B | .728 | -.452 |
| 14 | B | 2.275 | .549 | 44 | A | .712 | -.527 |
| 15 | B | .943 | .050 | 45 | A | .730 | .218 |
| 16 | A | 1.180 | .518 | 46 | B | 1.264 | .786 |
| 17 | A | .922 | -.524 | 47 | B | .701 | -.544 |
| 18 | B | .876 | -.105 | 48 | A | .814 | .579 |
| 19 | B | 1.107 | -.861 | 49 | A | 3.000 | .572 |
| 20 | A | .856 | -.198 | 50 | B | .680 | -.690 |
| 21 | A | .977 | -.754 | 51 | B | .658 | .011 |
| 22 | B | 1.790 | -.959 | 52 | A | .649 | -.131 |
| 23 | B | .856 | -.010 | 53 | A | .652 | -.499 |
| 24 | A | .853 | -.380 | 54 | B | .722 | .515 |
| 25 | A | .841 | -.166 | 55 | B | .637 | -.478 |
| 26 | B | .872 | .176 | 56 | A | 1.002 | .850 |
| 27 | B | .840 | -.364 | 57 | A | .623 | .000 |
| 28 | A | .983 | .478 | 58 | B | 1.087 | .885 |
| 29 | A | .939 | .413 | 59 | B | .620 | .058 |
| 30 | B | .820 | -.384 | 60 | A | .603 | -.385 |

Table C
Item Parameter Estimates for Items in the Bayesian Adaptive Testing Item Pool
($c$=.20 for All Items)

| Item | $a$ | $b$ | Item | $a$ | $b$ | Item | $a$ | $b$ | Item | $a$ | $b$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.960 | .223 | 46 | .745 | .311 | 91 | 3.000 | 1.381 | 136 | 1.075 | -1.345 |
| 2 | 1.529 | -.146 | 47 | .689 | -.050 | 92 | 3.000 | 1.374 | 137 | 1.067 | -1.335 |
| 3 | 1.424 | .176 | 48 | .678 | -.257 | 93 | 3.000 | 1.860 | 138 | .943 | -1.313 |
| 4 | 1.384 | .131 | 49 | .681 | -.684 | 94 | 2.321 | 1.442 | 139 | .875 | -1.448 |
| 5 | 1.202 | -.550 | 50 | .669 | -.567 | 95 | 2.111 | 1.518 | 140 | .887 | -1.189 |
| 6 | 1.109 | .135 | 51 | .651 | -.173 | 96 | 3.000 | 1.945 | 141 | 2.128 | -1.790 |
| 7 | 1.073 | -.355 | 52 | .693 | .321 | 97 | 1.716 | 1.420 | 142 | 1.887 | -1.552 |
| 8 | 1.036 | -.152 | 53 | .674 | .242 | 98 | 1.618 | 1.506 | 143 | 1.701 | -1.640 |
| 9 | 1.200 | .351 | 54 | .712 | .470 | 99 | 1.380 | 1.515 | 144 | 1.728 | -2.022 |
| 10 | 1.375 | .468 | 55 | .664 | -.776 | 100 | 1.289 | 1.433 | 145 | 1.427 | -1.674 |
| 11 | 1.570 | .546 | 56 | .886 | .796 | 101 | 3.000 | .960 | 146 | 1.235 | -1.875 |
| 12 | 1.109 | -.701 | 57 | .959 | .858 | 102 | 3.000 | 1.000 | 147 | 1.200 | -1.970 |
| 13 | 3.000 | .486 | 58 | 1.210 | .875 | 103 | 3.000 | 1.017 | 148 | 1.128 | -1.722 |
| 14 | .939 | -.281 | 59 | .619 | -.655 | 104 | 3.000 | 1.064 | 149 | 1.083 | -1.996 |
| 15 | .949 | -.439 | 60 | .610 | .012 | 105 | 3.000 | .792 | 150 | 1.067 | -1.936 |
| 16 | 1.244 | .542 | 61 | 3.000 | 2.287 | 106 | 3.000 | 1.156 | 151 | .873 | -2.016 |
| 17 | .917 | .171 | 62 | 3.000 | 2.363 | 107 | 3.000 | 1.180 | 152 | .829 | -1.582 |
| 18 | 1.086 | .483 | 63 | 3.000 | 2.405 | 108 | 3.000 | .670 | 153 | .768 | -1.927 |
| 19 | .872 | -.124 | 64 | 3.000 | 2.138 | 109 | 2.778 | 1.171 | 154 | .745 | -2.158 |
| 20 | .860 | -.235 | 65 | 3.000 | 2.138 | 110 | 3.000 | 1.219 | 155 | .812 | -1.244 |
| 21 | .934 | -.670 | 66 | 3.000 | 2.138 | 111 | 2.291 | .765 | 156 | .722 | -2.141 |
| 22 | .870 | .067 | 67 | 2.935 | 2.411 | 112 | 3.000 | 1.244 | 157 | .692 | -2.144 |
| 23 | .910 | -.633 | 68 | 3.000 | 2.069 | 113 | 3.000 | 1.259 | 158 | .672 | -2.009 |
| 24 | .939 | -.709 | 69 | 3.000 | 2.066 | 114 | 1.843 | .780 | 159 | .757 | -1.191 |
| 25 | .910 | .286 | 70 | 3.000 | 2.066 | 115 | 1.765 | 1.161 | 160 | .663 | -1.781 |
| 26 | 1.069 | .536 | 71 | 3.000 | 2.066 | 116 | 1.314 | 1.097 | 161 | 3.000 | -2.363 |
| 27 | .872 | .195 | 72 | 3.000 | 2.504 | 117 | 1.267 | 1.113 | 162 | 3.000 | -2.363 |
| 28 | .822 | -.278 | 73 | 3.000 | 2.022 | 118 | 1.317 | 1.204 | 163 | 3.000 | -2.324 |
| 29 | .896 | .336 | 74 | 3.000 | 2.022 | 119 | 1.168 | .919 | 164 | 3.000 | -2.324 |
| 30 | 1.232 | .643 | 75 | 3.000 | 2.632 | 120 | 1.256 | 1.207 | 165 | 3.000 | -2.632 |
| 31 | .844 | .205 | 76 | 3.000 | 2.632 | 121 | 1.432 | -1.043 | 166 | 3.000 | -2.632 |
| 32 | .860 | .275 | 77 | 1.162 | 2.676 | 122 | 1.235 | -1.031 | 167 | 3.000 | -2.632 |
| 33 | .797 | -.257 | 78 | .632 | 2.153 | 123 | 1.093 | -1.093 | 168 | 2.208 | -2.461 |
| 34 | .876 | -.742 | 79 | .613 | 2.004 | 124 | .882 | -1.061 | 169 | 1.749 | -2.366 |
| 35 | .800 | -.390 | 80 | .556 | 1.991 | 125 | .835 | -1.022 | 170 | 1.753 | -2.580 |
| 36 | 1.058 | -.998 | 81 | 3.000 | 1.606 | 126 | .777 | -1.055 | 171 | 1.452 | -2.239 |
| 37 | .791 | .085 | 82 | 3.000 | 1.576 | 127 | .736 | -1.085 | 172 | 1.286 | -2.236 |
| 38 | .773 | -.235 | 83 | 3.000 | 1.709 | 128 | .672 | -1.091 | 173 | 1.241 | -2.670 |
| 39 | .767 | -.374 | 84 | 3.000 | 1.481 | 129 | .568 | -1.054 | 174 | 1.087 | -2.635 |
| 40 | .876 | -.924 | 85 | 3.000 | 1.472 | 130 | .564 | -1.023 | 175 | 1.104 | -2.187 |
| 41 | .779 | .246 | 86 | 3.000 | 1.758 | 131 | 1.817 | -1.439 | 176 | 1.020 | -2.584 |
| 42 | .788 | .295 | 87 | 3.000 | 1.464 | 132 | 1.749 | -1.256 | 177 | 1.014 | -2.479 |
| 43 | .745 | -.684 | 88 | 3.000 | 1.455 | 133 | 1.274 | -1.351 | 178 | .981 | -2.634 |
| 44 | .767 | -.803 | 89 | 3.000 | 1.801 | 134 | 1.165 | -1.395 | 179 | .956 | -2.266 |
| 45 | .699 | -.324 | 90 | 2.518 | 1.607 | 135 | 1.145 | -1.412 | 180 | .859 | -2.251 |

Table D
Correlations Between Scores from Alternate Forms for Three Combinations of Testing Strategy and Test Scoring, at Test Lengths from 1 to 30 Items

| Test Length | Bayesian Adaptive Test | Conventional Test | |
|---|---|---|---|
| | | Bayesian Scoring | Proportion-Correct Scoring |
| 1 | .211 | .288 | .288 |
| 2 | .293 | .374 | .352 |
| 3 | .446 | .422 | .419 |
| 4 | .551 | .454 | .467 |
| 5 | .568 | .536 | .534 |
| 6 | .599 | .566 | .562 |
| 7 | .638 | .624 | .613 |
| 8 | .678 | .649 | .626 |
| 9 | .698 | .662 | .652 |
| 10 | .706 | .703 | .696 |
| 11 | .738 | .724 | .723 |
| 12 | .759 | .737 | .734 |
| 13 | .780 | .754 | .757 |
| 14 | .791 | .763 | .764 |
| 15 | .810 | .774 | .780 |
| 16 | .812 | .790 | .795 |
| 17 | .830 | .801 | .808 |
| 18 | .835 | .807 | .812 |
| 19 | .844 | .823 | .822 |
| 20 | .851 | .831 | .831 |
| 21 | .864 | .837 | .837 |
| 22 | .872 | .840 | .838 |
| 23 | .877 | .841 | .842 |
| 24 | .885 | .842 | .845 |
| 25 | .892 | .850 | .857 |
| 26 | .896 | .854 | .861 |
| 27 | .906 | .856 | .864 |
| 28 | .911 | .860 | .869 |
| 29 | .915 | .861 | .871 |
| 30 | .920 | .868 | .879 |

Table E
Correlations Between Criterion Test Scores and Scores Obtained from Three Combinations of Testing Strategy and Scoring Method at Test Lengths from 1 to 30 Items, Averaged Across Test Forms

| Test Length | Bayesian Adaptive Test | Conventional Test | |
|---|---|---|---|
| | | Bayesian Scoring | Proportion-Correct Scoring |
| 1 | .445 | .492 | .492 |
| 2 | .490 | .501 | .493 |
| 3 | .576 | .543 | .536 |
| 4 | .610 | .590 | .597 |
| 5 | .621 | .635 | .644 |
| 6 | .630 | .653 | .657 |
| 7 | .650 | .676 | .680 |
| 8 | .665 | .688 | .693 |
| 9 | .671 | .710 | .720 |
| 10 | .691 | .729 | .741 |
| 11 | .702 | .758 | .767 |
| 12 | .712 | .764 | .772 |
| 13 | .720 | .769 | .781 |
| 14 | .729 | .776 | .787 |
| 15 | .735 | .782 | .792 |
| 16 | .741 | .791 | .801 |
| 17 | .750 | .795 | .805 |
| 18 | .755 | .797 | .807 |
| 19 | .758 | .803 | .812 |
| 20 | .763 | .808 | .818 |
| 21 | .768 | .810 | .820 |
| 22 | .771 | .813 | .824 |
| 23 | .775 | .814 | .823 |
| 24 | .776 | .818 | .828 |
| 25 | .779 | .820 | .832 |
| 26 | .783 | .820 | .830 |
| 27 | .786 | .822 | .833 |
| 28 | .790 | .827 | .840 |
| 29 | .795 | .833 | .840 |
| 30 | .797 | .834 | .841 |