# Parallel Forms Reliability and Measurement Accuracy Comparison of Adaptive and Conventional Testing Strategies

Marilyn F. Johnson and David J. Weiss
University of Minnesota

Prior research at the University of Minnesota has compared the parallel forms reliabilities of adaptive and conventional vocabulary tests as a function of test length. The results are shown in Figure 1, which displays alternate forms reliabilities of Owen´s Bayesian adaptive test and a conventional test as a function of number of items administered. The conventional test was peaked in information at $\theta = 0.0$; and test items were administered in order of information, from high to low values. The Bayesian adaptive test was scored by Bayesian methods; whereas the conventional test was scored by both proportion-correct and Bayesian methods. Both tests consisted of five-alternative multiple-choice vocabulary items.

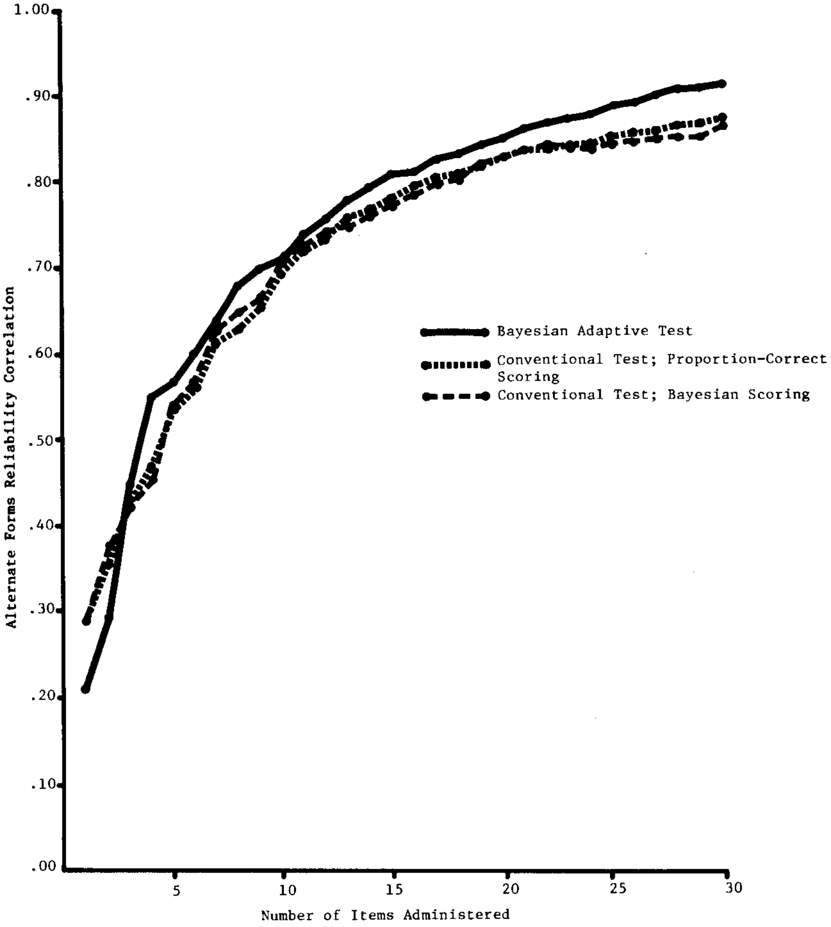
As expected, the plots in Figure 1 show an increase in reliability as test length increased for both testing strategies. However, rather than the expected asymptote of reliabilities for both strategies as test length increased, the reliability of the Bayesian adaptive test surpassed that of the conventional test. The approximate difference in reliabilities at test termination was $\underline{r}$ = .05, with a 30-item reliability of .92 for the Bayesian test and .87 for the conventional test scored by the Bayesian method. The difference in reliabilities between Bayesian and proportion-correct-scored conventional tests was .04 at the 30-item test length.

The analysis also included a comparison of concurrent validity obtained by correlating the ability estimates with number-correct scores on a 120-item vocabulary criterion test also composed of five-alternative multiple-choice questions. These results (see Figure 2) indicated that although the Bayesian adaptive test was more reliable than the conventional tests, the conventional tests yielded higher validities when correlated with the criterion test. Figure 2 shows that the validities, similar to the reliabilities, increased as a function of test length, with the conventional test yielding higher validities after four items. The validity of the Bayesian test at 30 items was .797; that of the Bayesian-scored conventional test was .834; and the proportion-correct-scored conventional test obtained a validity of .841.

## Purpose

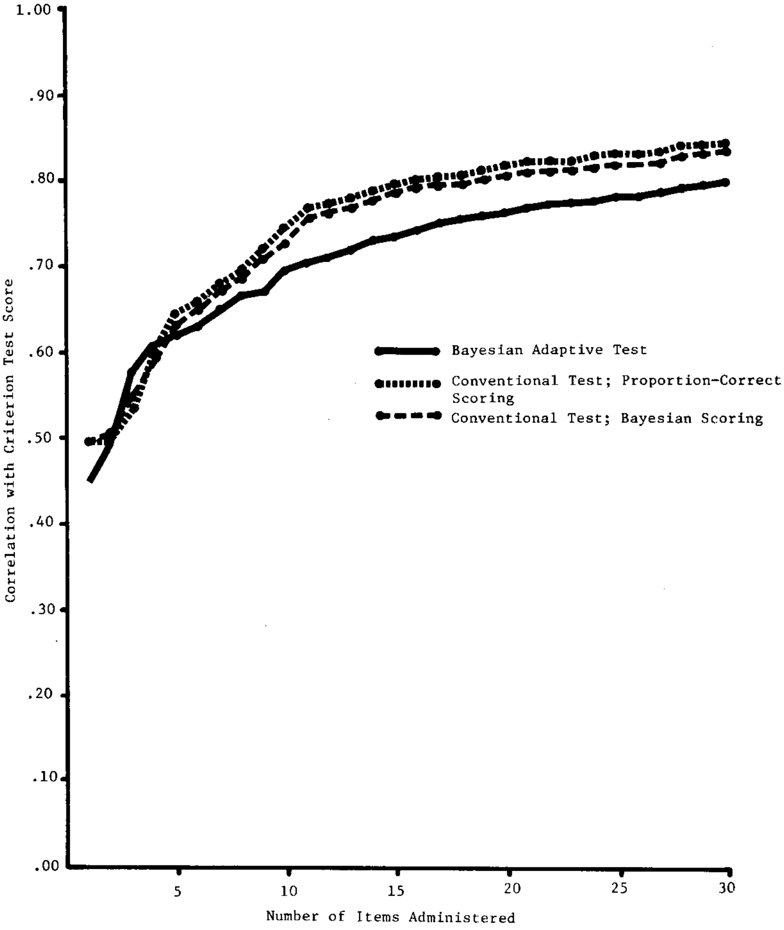
Due to the apparently contradictory nature of these findings, the present research was designed to replicate them. There were, however, some modifica-

Figure 1
Alternate Forms Reliabilities of Ability Level Estimates from a Bayesian Adaptive Test and a Conventional Test Scored by Proportion-Correct and Bayesian Scoring, as a Function of the Number of Items Administered



tions to the basic design of the comparison study, and an additional dependent variable, measurement accuracy, was used to compare the testing strategies. In addition, the present study compared peaked conventional, Bayesian adaptive, and maximum information adaptive testing strategies. The conventional test was also peaked in information evaluated at $\theta = 0.0$. Items on the conventional test were administered in order of item information but, for purposes of analysis, were arranged in random order. The item pool was composed of the same items that were used in the original study, but they were reparameterized after the original study and prior to the present investigation (Prestwood & Weiss, 1977). Comparisons of the three testing strategies were made in terms of parallel forms reliability as a function of test length and in terms of measurement accuracy as a function of $\theta$ level. Accuracy of measurement was operationalized as the posterior variance of the Bayesian-scored testing strategies and as standard errors of measurement for the maximum likelihood-scored testing strategies. Compari-

Figure 2
Correlations of Ability Level Estimates
from a Bayesian Adaptive Test and a Conventional Test
Scored by Proportion-Correct and Bayesian Scoring
with Criterion Test Score,
as a Function of the Number of Items Administered
(Averaged Across Two Test Forms)



sons of scoring strategies, including Bayesian, maximum likelihood, and proportion-correct scoring, were made on the basis of parallel forms reliability.

## Method

### Subjects

Undergraduate and graduate students from the University of Minnesota volunteered to participate in the fall 1978 and winter 1979 quarters. These students were recruited from Introductory Biology 1-011, Introductory Psychology 1-001, and a measurement course, Psychology 5-862. Students from the introductory psychology and biology courses participated in the study in order to obtain experimental points, which counted toward their final grade. Volunteers from the measurement course, both graduate and undergraduate students, participated at the request of the instructor.

There were 373 students in the conventional testing condition, 390 in the Bayesian testing condition, and 233 in the maximum information testing condition. Testing spanned two quarters in order to obtain an adequate number of students; a total of 996 students were tested during this period. Although students were recruited from varying subject pools, no difference in population was suggested because the undergraduate students were all from the College of Liberal Arts. In addition, students were sequentially assigned to one of the three testing strategies. The introductory biology and psychology students also participated in other studies during their experimental hour. In the case of the biology students, the experimental tests for this study were administered after a biology test. The fall 1978 introductory psychology students participated solely in this experiment, whereas the winter 1979 introductory psychology students first took the experimental test for this study, and then took another test. In each case, only data from the alternate forms verbal ability tests were analyzed.

## Procedure

All students took the tests at an individual cathode-ray terminal (CRT) connected to a Hewlett-Packard real-time computer system. A test proctor was present during testing to provide assistance to the examinees. The students were assured that they could take as much time as necessary to complete the tests. Prior to administration of items on the first test, however, instructional screens explaining the operation of the CRTs were displayed. After students reviewed the test instructions and responded to a number of identification and demographic questions, the experimental tests were administered. Students responded to the five-alternative multiple-choice vocabulary questions by typing a number into the CRT corresponding to the chosen alternative.

## Item Pools

Adaptive test. The Bayesian and maximum information tests used the same item pool from which to select items. The pool was composed of 256 items selected for the purposes of this study from the total vocabulary pool, which contained 358 items. The 358 items were newly parameterized items, based on combined data sources from conventional tests administered between fall 1969 and winter 1978. The items were parameterized with Urry's (1977) ESTEM program using a 3-parameter logistic ICC model. All items were assumed to have a guessing parameter of $\underline{c}$ = .20. (Details regarding the parameterization procedure can be found in Prestwood & Weiss, 1977.) Selection of items from the larger pool was based on several criteria, which varied by difficulty levels of the items. Because there were few very difficult or very easy items, fewer items at these extremes on the difficulty continuum were eliminated. Items with discrimination parameters of $\underline{a}$ = 3.00 were routinely rejected because this value was identified as a statistical artifact of the parameterization program and not as a true reflection of the item's discrimination value.

Based on a stratification of the items into difficulty levels, items were eliminated if their discriminations were low. This criterion, however, varied by difficulty level. In Levels 6 and 7, items were omitted if the discrimination parameter fell below $\underline{a}$ = .30. In Levels 3, 4, and 5, where there were more items, the culling criterion was set at _a_ = .35. In these levels, also, items were omitted if the sample size on which the parameters were calibrated was less than 100. In many cases the items rejected on the basis of sample size were also of low discrimination.

Conventional test. The alternate forms of the conventional test were each composed of 30 vocabulary items arranged in descending order of item information evaluated at $\theta = 0.0$. The 60 most informative items at $\theta = 0.0$ were selected from the vocabulary pool composed of 256 items. By this procedure, items with relatively higher discrimination levels and difficulties of about _b_ = 0.0 were selected. Each test was thus peaked with respect to item information. Items were ordered by information at $\theta = 0.0$, and the 60 items were divided into Test Form A and Test Form B according to an ABBABAAB selection scheme. This procedure was used to insure that the alternate forms did not systematically differ in item information. The items were administered in order of descending item information. However, for purposes of analysis, pairs of items from the two test forms were randomly formed to simulate conventional paper-and-pencil testing conditions. The conventional test items were selected from the adaptive test pool so that it was possible that adaptive test items could also be used in the conventional test, since an independent groups design was being used.

## Adaptive Testing and Scoring Strategies

Alternate forms of the adaptive tests were dynamically selected from the item pool by a special algorithm. Using an ABBABAAB rotational scheme, Form A of the adaptive test was given an opportunity to select an item from the pool of unadministered items, based on the item selection algorithm (Bayesian, maximum information) in use; and the ability estimate for that form of the adaptive test was updated. For administration of the next item to a testee, Form B then selected an item from the current pool of unadministered items; and the ability estimate for that form was updated. This procedure continued, using the ABBA-BAAB rotation, until 30 items were administered for each of the alternate forms--Form A and Form B--and the ability estimates for each form were saved after each item was administered.

Bayesian adaptive testing strategy. Items were selected and scored during the adaptive procedure according to Owen's (1975) Bayesian model. The prior distribution of ability was assumed to be normal, with a mean of 0.0 and a variance of 1.00. These values served as initial estimates of ability at the start of testing for each of the two forms for each individual. Testing was terminated after 30 items had been administered for each of the two forms. (Details concerning the Bayesian scoring algorithm can be found in McBride & Weiss, 1976.)

Maximum information adaptive testing strategy. Items were selected according to a maximum information item selection routine, and ability estimates were updated by scoring the responses by maximum likelihood methods (Bejar & Weiss, 1979). The initial estimate of ability was 0.0 for each form. Testing was terminated after 30 items had been adminstered for each of the two alternate forms.

The adaptive tests were scored after testing by a scoring strategy other than the one used during testing. The Bayesian test protocols were scored by maximum likelihood methods, and the maximum information test protocols were scored by Bayesian methods. Scores were calculated after each of the 30 items in both parallel tests. Responses to the two alternate forms of the conventional test were also rescored by Bayesian and maximum likelihood scoring methods at each test length from 1 to 30 items.

## Independent Variables

Testing strategy was the major independent variable of interest. The strategies compared were the conventional, Bayesian, and maximum information testing strategies. Methods of scoring were also compared. These included logistic maximum likelihood scoring, Bayesian scoring, and (for the conventional test) proportion-correct scoring. Test length was a third independent variable of interest. Thirty test lengths were obtained by scoring each 30-item test 30 times. That is, a test was scored after the first item, after the first two items, after the first three items, and so on until 30 scores were obtained. In this way, 30 test lengths, varying from 1 to 30 items, were generated for each of the alternate forms.

## Dependent Variables

Parallel forms reliabilities. Testing strategies were compared on the basis of parallel forms reliability by correlating corresponding ability estimates obtained from Forms A and B for a given testing strategy. Since the test protocols were scored in at least two ways, Bayesian and maximum likelihood, a total of seven testing-scoring conditions were compared on the basis of parallel forms reliability. Scoring strategy was compared on the basis of parallel forms reliability by comparing reliabilities of a single testing strategy scored by more than one method. Three of the parallel forms reliabilities paired the appropriate scoring method with each of the three testing strategies. These were proportion-correct scoring of conventional tests, maximum likelihood scoring of maximum information tests, and Bayesian scoring of Bayesian-administered tests.

The remaining four parallel forms reliabilities were obtained by scoring the test protocols by a scoring routine other than the appropriate one. In this way, reliabilities were obtained for the Bayesian-scored maximum information test, the maximum-likelihood-scored Bayesian test, the Bayesian-scored conventional test, and the maximum-likelihood-scored conventional test. Proportion-correct scores were not obtained for adaptive tests. Reliabilities were calculated as a function of test length. That is, reliability was calculated not only from end-of-test ability estimates but also for each of the 30 test lengths. Scoring method correlations were obtained by correlating estimates obtained from different scorings of the same testing strategy. These correlations were used to analyze the similarity of ability estimates obtained from different scoring techniques applied to a single set of data.

Errors of measurement. The three testing strategies were compared on the basis of their errors of measurement. This was assessed by two methods--one method estimated errors of measurement on the basis of maximum likelihood scoring methods; and the other, by Bayesian scoring methods. In the first method, test protocols were scored by maximum likelihood methods, and the standard errors of measurement (SEM) associated with each ability estimate was calculated. These values are the reciprocal of the square root of test information at a given $\theta$ level. They indicate how accurate the estimate is and how much it is likely to vary from the true $\theta$ value; the larger the standard error, the more likely the estimate will be inaccurate.

The SEM values were averaged within each of 20 $\hat{\theta}$ intervals ranging from approximately -3.0 to +2.0, and the mean SEM values were then plotted as a function of $\hat{\theta}$. This was done on a single randomly chosen parallel form for each of the three testing strategies.

The posterior variance of the Bayesian ability estimate was also used to compare the testing strategies on the basis of measurement accuracy. Posterior variances were averaged within each of 20 $\hat{\theta}$ intervals ranging from -2.0 to +2.0. These mean values were plotted at the midpoint of the $\hat{\theta}$ intervals and the points were connected to yield a continuous line. The posterior variance is analogous in meaning and interpretation to the standard errors of measurement.

Although one or the other of these measurement accuracy indices might have been adequate in comparing the testing strategies, both were included to minimize any biased conclusions regarding measurement accuracy of the adaptive tests. In general, posterior variance of Bayesian ability estimates will be less when items are selected according to a Bayesian testing strategy than when items are selected by any other adaptive procedure. Use of the posterior variance alone in the comparison of the adaptive testing strategies may bias conclusions toward the Bayesian testing strategy. For this reason the standard errors of measurement was also used as an index of measurement accuracy. This index, in general, will favor the maximum information testing strategy because items were selected and scored according to a maximum likelihood testing procedure.

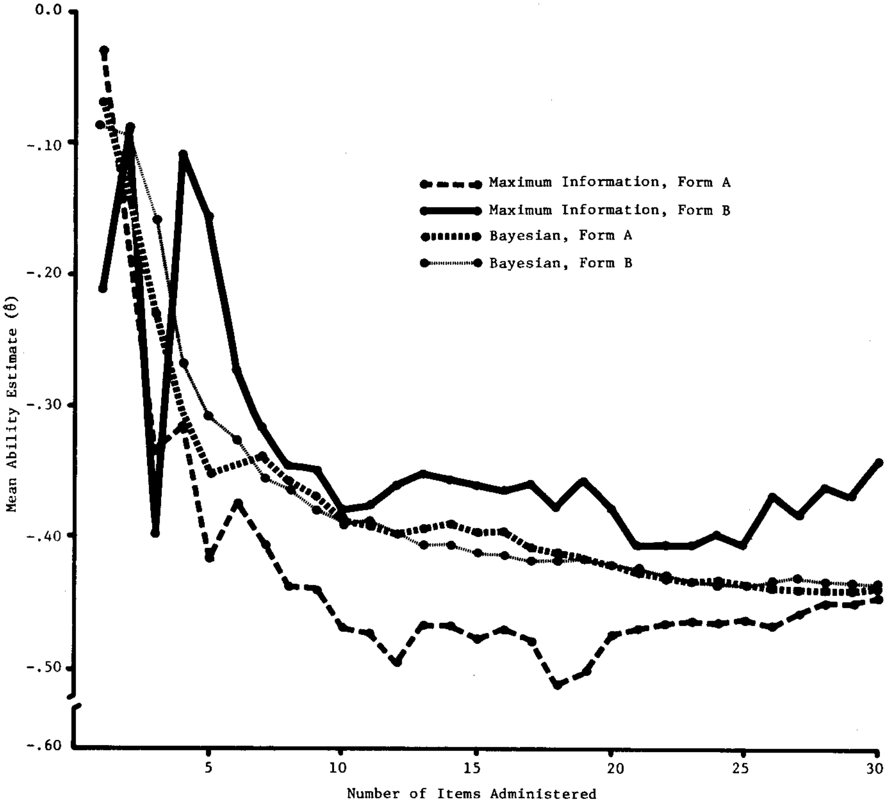
## Results

### Were the Tests Parallel?

Several analyses were performed to determine whether the alternate forms were functioning as parallel forms. These included comparisons of the means and variances of the ability estimates as a function of test length for the alternate forms of each testing strategy.

Score means. In general, the score means of the three testing strategies--conventional, Bayesian, and maximum information--showed an adequate level of parallel relationship between Forms A and B. Because the proportion correct score metric differs from the $\theta$ metric, the adaptive and conventional mean ability estimates are not directly comparable. Adaptive test comparisons of the means (Figure 3) show that there were greater differences between mean ability estimates for the alternate forms of the maximum information testing strategy than for the Bayesian testing strategy; this was because of the tendency of the Bayesian item selection and scoring routine to yield conservative estimates of ability. As testing progressed, however, differences between the ability estimates for the two alternate forms of each test decreased for both adaptive tests. Figure 3 also shows that the Bayesian mean ability estimates fell between the Form A and Form B means from the maximum information testing strategy. Thus, both adaptive procedures yielded about the same average ability estimates for the students selected from a common population.

Figure 3
Mean Ability Estimates from Parallel Forms A and B of Maximum Information and Bayesian Adaptive Tests, as a Function of Number of Items Administered

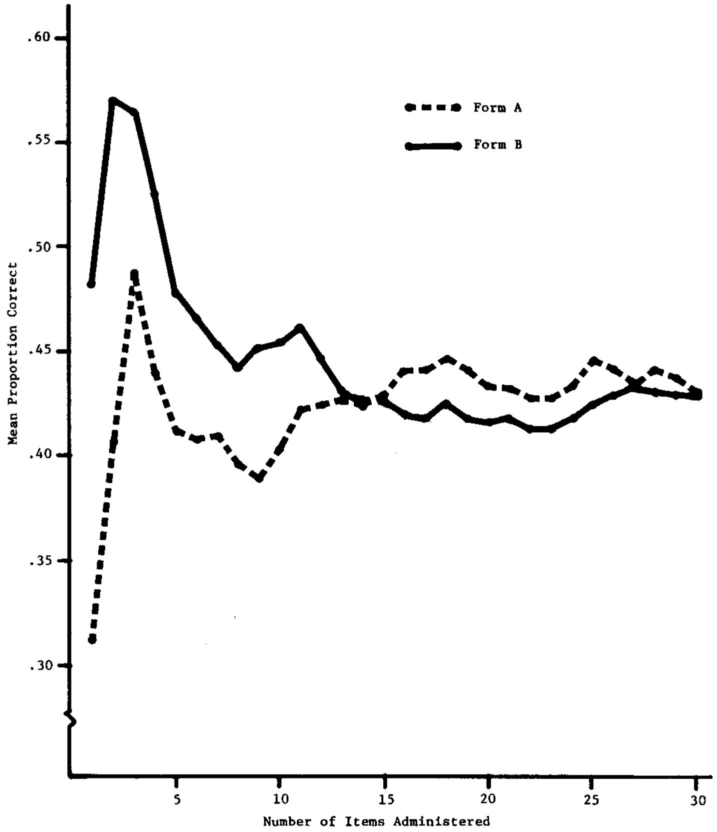


Means of the conventional parallel forms were obtained by averaging proportion-correct scores at each of 30 test lengths, based on randomly ordered items. Figure 4 shows that mean proportion-correct scores stabilized to a final value of .43.

Score variances. Variances of the ability estimates from the maximum information testing strategy (Figure 5) were relatively high up to 3 items, and then decreased steadily. The greatest difference in variance between the two alternate forms was at 3 items (1.25); whereas at 30 items the difference was only half (.75). Figure 5 also shows that ability score variances decreased from the beginning to the end of the test. Thus, score variances from the maximum information tests showed both a decrease in difference between alternate forms and a decrease in amount of variance as testing proceeded.

In comparison to the ability scores from the maximum information test, variance in Bayesian ability scores showed a similar maximum difference in vari-

Figure 4
Mean Proportion-Correct Score of the Conventional Test
for Alternate Forms A and B,
as a Function of Number of Items Administered

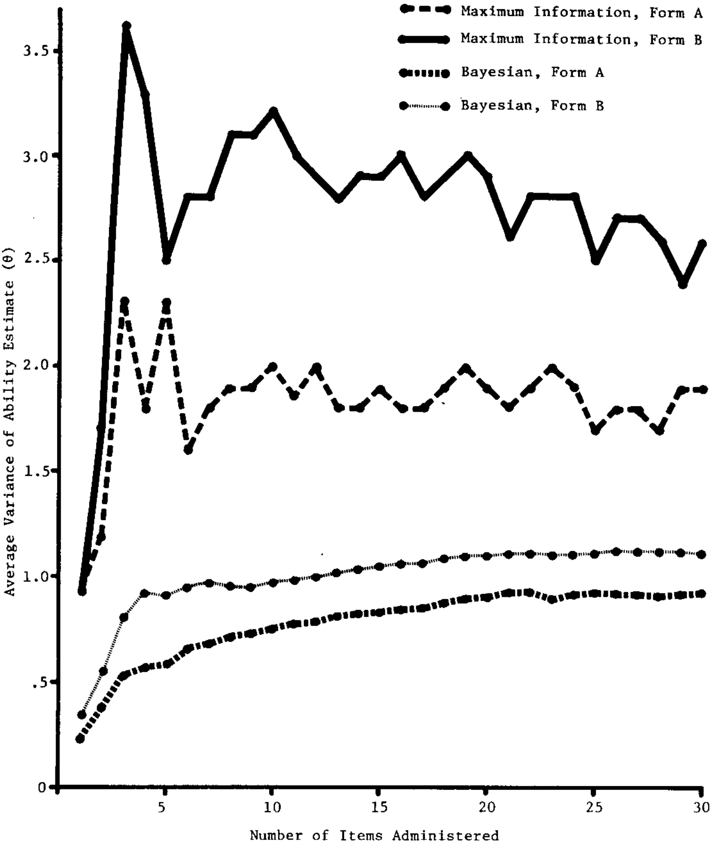


ance for tests of about 5 items in length, followed by decreased differences, as shown in Figure 5. Level of variance increased, however, as testing proceeded, reflecting the reduced dependence of the Bayesian ability estimates on the prior ability estimate. The restriction in Bayesian ability estimates due to the regression effect was still evident even at 30-item test lengths, since the ability estimate variances for the Bayesian tests were substantially lower than those of the maximum information tests.

Proportion correct score variance of both parallel forms of the conventional test decreased rapidly, from a possible maximum of .25 at 1 item to .06 at 30 items, as shown in Figure 6. Based on both the score means and score variances, the alternate forms of the conventional test were closer to being parallel than the alternate forms of either of the adaptive tests.

Errors of measurement as a function of test length. Samejima (1977) defines weakly parallel tests as tests that yield the same information functions. Thus, evidence for the parallel relationship between the adaptive forms included examination of their errors of measurement as a function of number of items administered. Average standard error of measurement, the reciprocal of the square root of theoretical test information, was used to compare alternate forms of the maximum information testing strategy. The error of measurement curves for the maximum information tests (Figure 7) showed the same form with variance decreasing rapidly to a final value of .40.

Figure 5
Average Variances of Ability Estimates for Forms A and B of Maximum Information Adaptive Tests and Bayesian Adaptive Tests, as a Function of Number of Items Administered

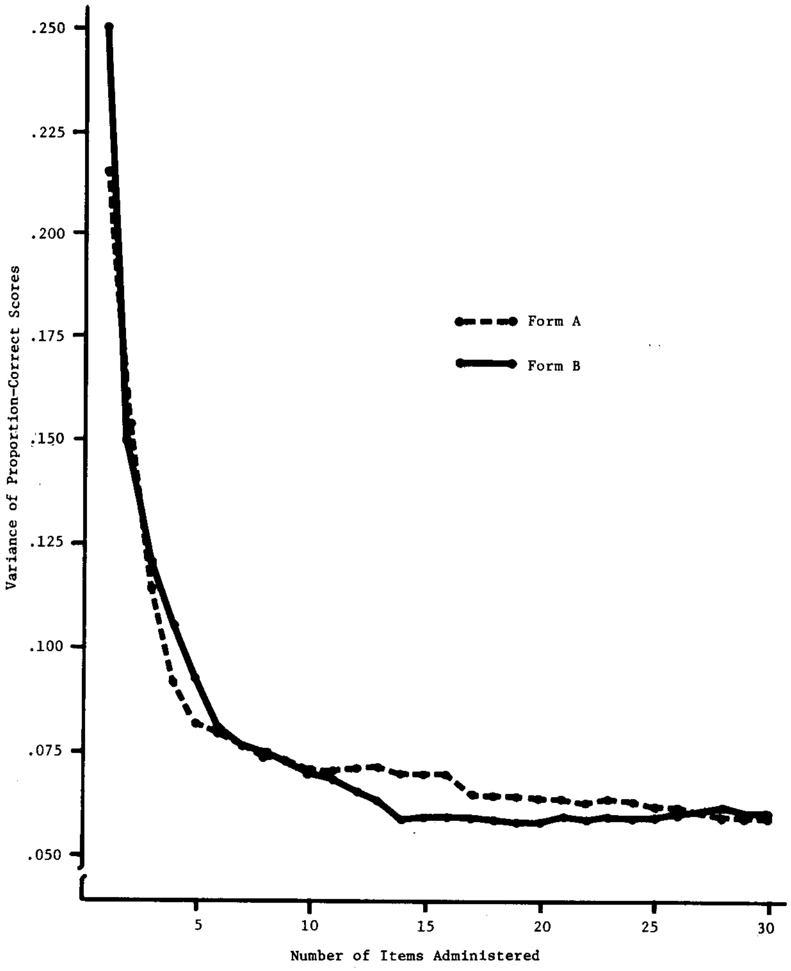


The error of measurement index for the Bayesian testing strategy was the posterior variance of the ability estimates. These data are also shown as a function of test length in Figure 7. Means of the Bayesian posterior variances for the two alternate forms were almost identical, decreasing from an initial value of .68, after 1 item was administered, to a final variance of .10, after 30 items were administered. As Figure 7 shows, there was less variance in Bayesian ability estimates than in the maximum likelihood ability estimates; but the data show that both the Bayesian and maximum information adaptive tests yielded parallel forms in terms of their mean errors of measurement, at almost all test lengths.

## Parallel Forms Reliability

Optimal scoring method. The optimal scoring method was maximum likelihood

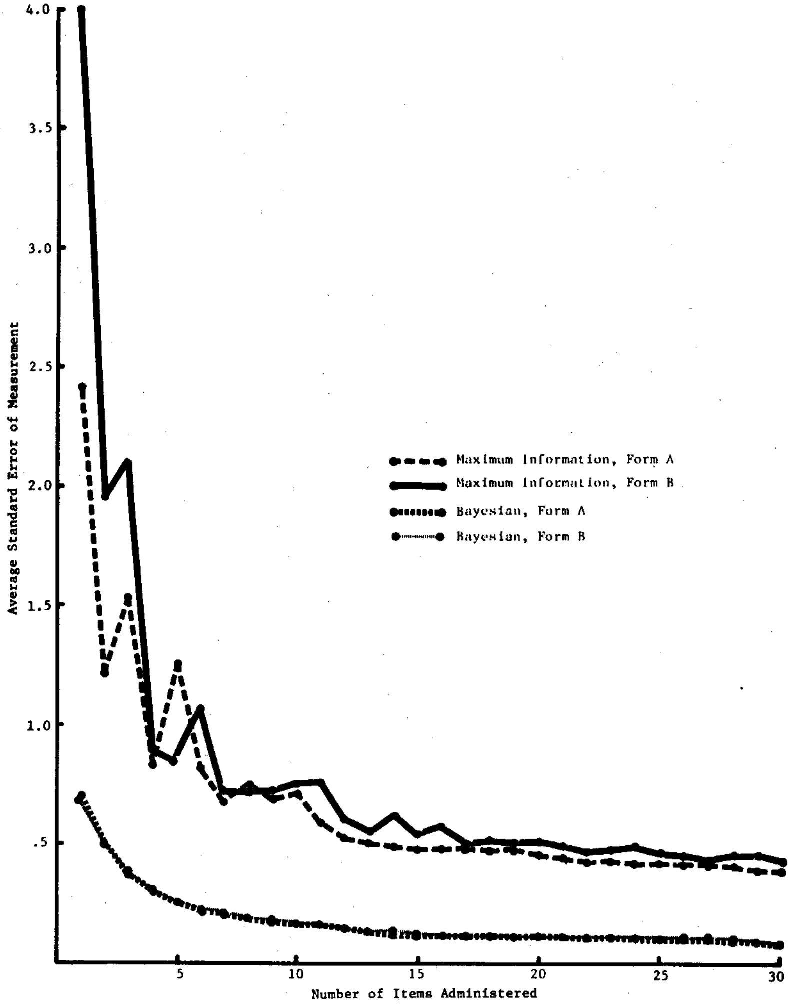
Figure 6
Variances of Proportion-Correct Scores from Alternate Forms A and B of the Conventional Test, as a Function of Number of Items Administered



for the maximum information testing strategy, Bayesian for the Bayesian testing strategy, and proportion correct for the conventional test. Alternate forms reliability correlations were computed at each test length for each testing strategy using these optimal scores.

Reliabilities of the three testing strategies as a function of test length are shown in Figure 8. The peaked conventional test yielded substantially higher reliabilities after 11 items than either of the adaptive tests. The greatest difference between reliabilities was $\underline{r}$ = .09 between the adaptive and conventional tests at the 30-item test length; the reliabilities of the adaptive tests were $\underline{r}$ = .81, compared with the final reliability of $\underline{r}$ = .90 for the conventional test. The data in Figure 8 show essentially the same level and shape in reliabilities for the adaptive tests, although there was greater fluctuation in reliabilities for the maximum information test. The conventional test reliability was nearly identical to that of the Bayesian test up to the 10-item test

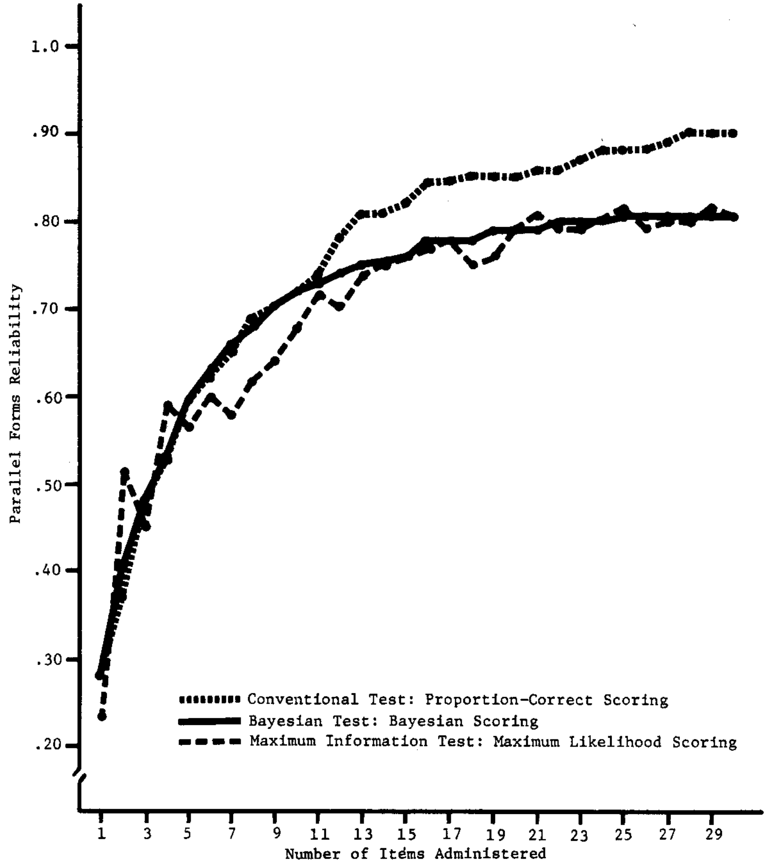
Figure 7
Means of Standard Error of Measurement from Parallel Forms A and B of Maximum Information Adaptive Tests and Mean Posterior Variance of Parallel Forms A and B of the Bayesian Adaptive Tests, as a Function of Number of Items Administered



length, but after that point the conventional test reliability increased more quickly than that of the adaptive tests. Although adaptive test reliabilities showed signs of leveling off toward the end of the test, the reliability of the conventional test seemed to increase steadily.

Other scoring strategy. Reliabilities were also obtained from testing strategies scored by other than optimal scoring strategies. Four testing-scoring combinations were of interest: Bayesian-scored maximum information tests, maximum-likelihood-scored Bayesian tests, Bayesian-scored conventional tests, and maximum-likelihood-scored conventional tests. These reliability results are shown in Figure 9 as a function of test length.

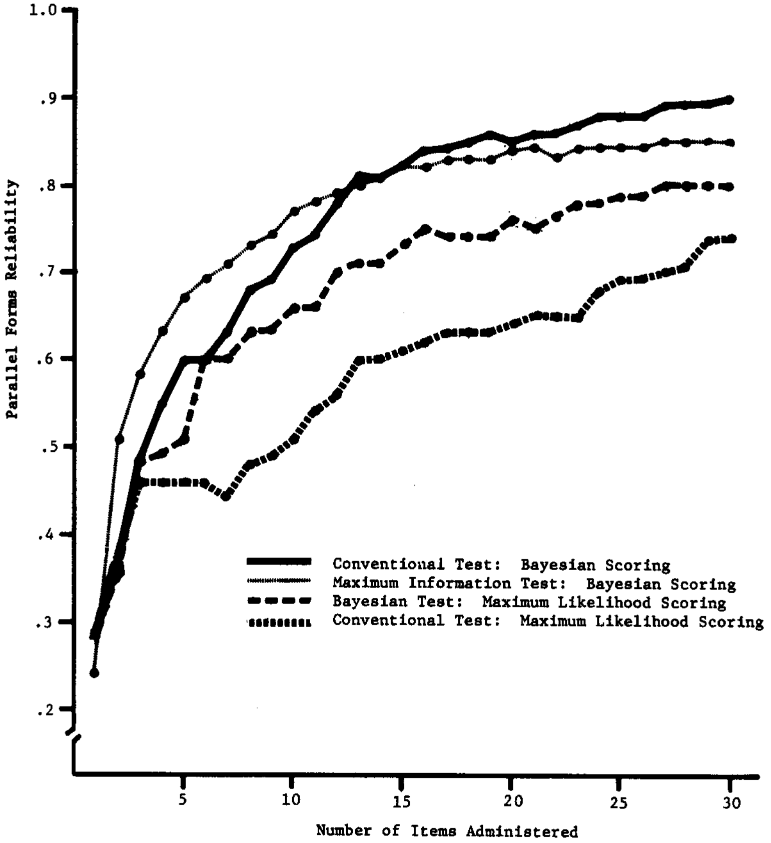
Figure 8
Parallel Forms Reliabilities of Optimally Scored Conventional, Bayesian, and Maximum Information Testing Strategies, as a Function of Number of Items Administered



In general, Figure 9 shows that the Bayesian scoring procedure yielded higher reliabilities under nonoptimal conditions than the maximum likelihood scoring procedure. Bayesian scoring of the conventional test yielded essentially equivalent reliabilities at every test length, as did proportion-correct scoring of the conventional test. Bayesian scoring of the maximum information tests yielded higher reliabilities at most test lengths beyond about 12 items than the optimal scoring strategy for that test. In addition, Bayesian scoring of the maximum information test tended to decrease substantially the differences in reliabilities observed between the conventional and adaptive tests. Figure 9 shows that the reliability for the Bayesian-scored maximum information test was higher than that of the conventional test for test lengths from 3 to 12 items. The maximum difference between these two reliabilities was $\underline{r}$ = .05 at 30 items, as compared to $\underline{r}$ = .09 for the data in Figure 8. These data indicate that Bayesian scoring of an adaptive test may yield more stable estimates of ability than maximum likelihood scoring.

The data also illustrate the inappropriateness of scoring conventional

Figure 9
Parallel Forms Reliabilities of Non-Optimally Scored Testing-Scoring Strategies, as a Function of Number of Items Administered



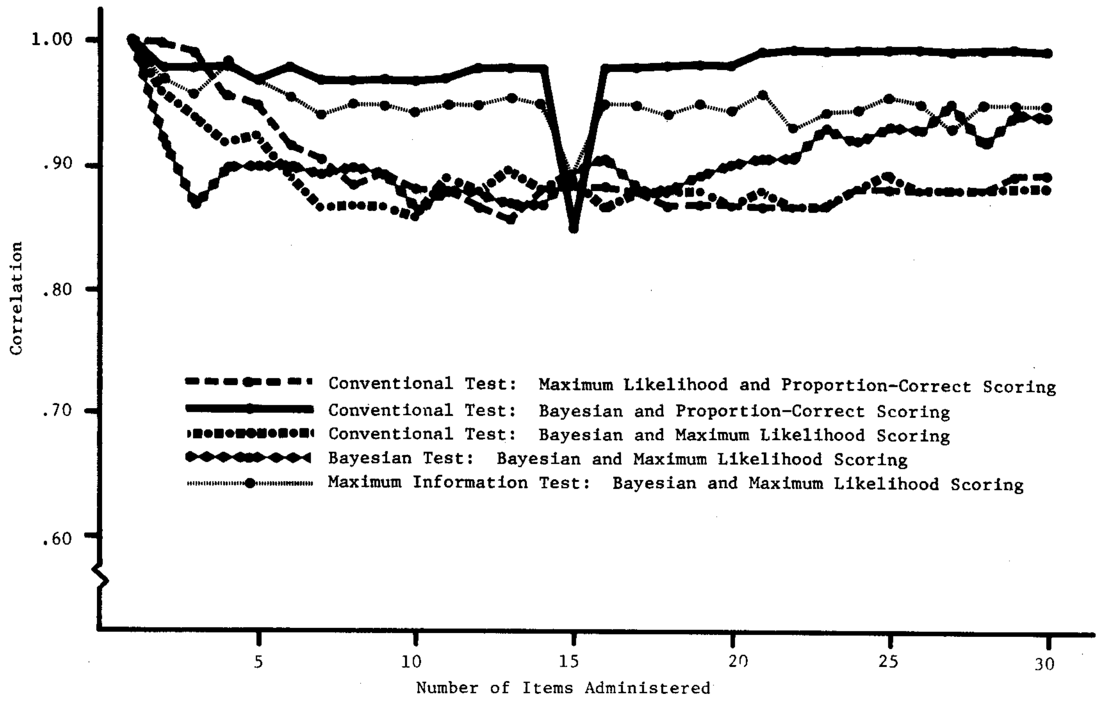
tests with maximum likelihood scoring methods. As Figure 9 shows, maximum likelihood scoring of the conventional test resulted in extremely low reliabilities at all test lengths, reaching a maximum of only .74 at 30 items.

## Scoring Method Correlations

To study the generality of the findings of Kingsbury and Weiss (1979), in their study of correlations among latent-trait scoring methods in achievement test data, comparisons of the ability estimates from the various scoring methods were made by correlating scores obtained from different ways of scoring the same testing strategy. For both adaptive testing strategies, Bayesian scores were correlated with maximum likelihood scores. Conventional test comparisons were made by correlating proportion-correct scores with Bayesian scores, proportion-correct scores with maximum likelihood scores, and Bayesian scores with maximum likelihood scores. For each testing strategy, one of the two alternate forms was randomly chosen for these analyses. These five scoring combinations are shown in Figure 10 as a function of test length.

As Figure 10 shows, the highest correlations were between Bayesian and proportion-correct scores of the conventional test. These correlations varied in

Figure 10
Correlations Between Scoring Methods for the Same Alternate Form, as a Function of Number of Items Administered



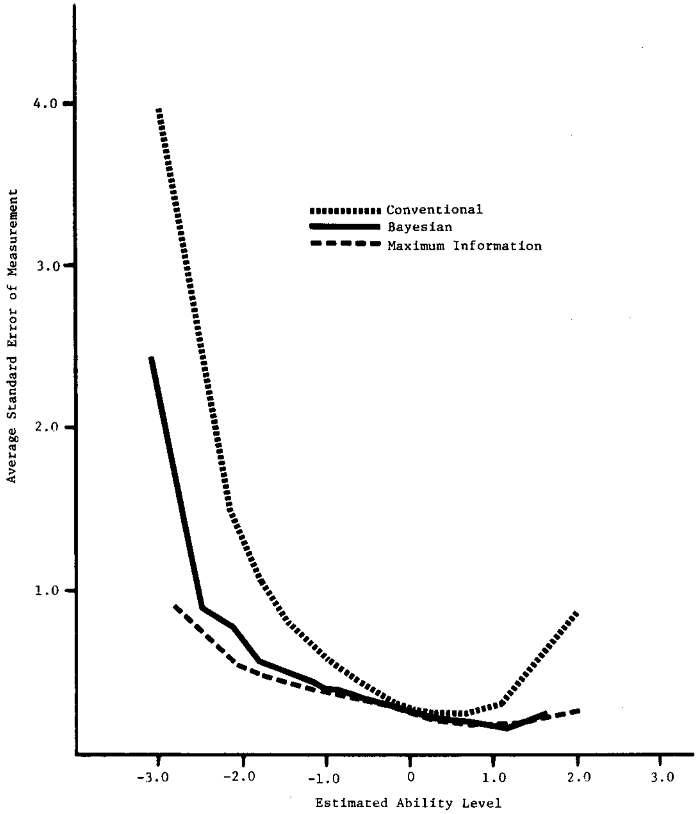
value between 1.00 for a 1-item test to .85 for a 15-item test, with most correlations between .97 to .99. The second highest level of correlation was between the Bayesian- and maximum-likelihood-scored maximum information test, with most correlations between .93 and .95. With the exception of the latter half of the correlations between Bayesian and maximum likelihood scores from the Bayesian test, there were few differences among the other three sets of correlations; the modal correlation for these three plots was .88. The correlations between Bayesian and maximum likelihood scores from the Bayesian test increased steadily after the 15-item test length to a final value of $\underline{r}$ = .94.

## Measurement Precision as a Function of Ability Level

Figure 11 shows plots of the average standard errors of measurement as a function of the maximum-likelihood-derived ability distribution. These data are the reciprocal of the square root of the test information function for each test. The distribution obtained from this sample varied from about -3.00 to +2.00 and was divided into equal frequency intervals ($N \geq 20$), separately for each testing strategy.

The data indicate that at no point on the ability continuum were the standard errors of measurement smaller in the conventional test than in the adaptive tests. In general, the maximum information testing strategy yielded smallest standard errors or greatest measurement precision. The Bayesian test, when scored by maximum likelihood, had poorer measurement precision at the lower ex-

Figure 11
Average Standard Error of Measurement as a Function of Ability Level for Conventional, Bayesian, and Maximum Information Testing Strategies (Non-Converging Values Eliminated)

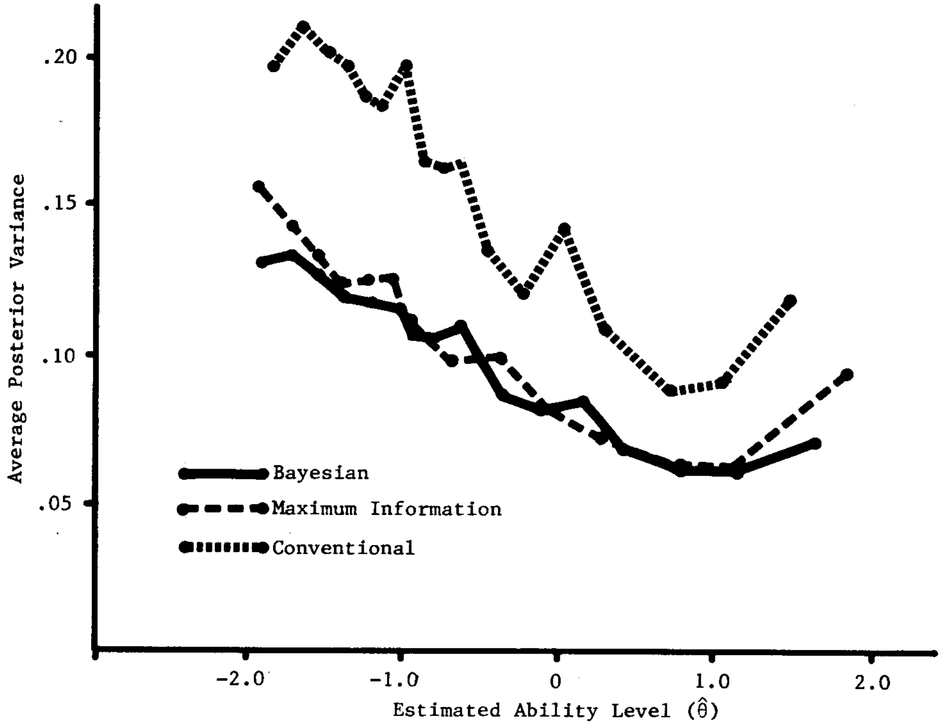


treme of the ability continuum than did the maximum information test. Precision of measurement for all the testing strategies was greatest at the central portion of the ability distribution than at the extremes.

Bayesian posterior variance comparisons are shown in Figure 12 as a function of the Bayesian-derived ability distribution. The distribution varied from about -2.00 to +2.00. The average posterior variance was greater at all points along the ability continuum for the conventional strategy than for either of the adaptive tests. The Bayesian and maximum information testing strategies had about the same level of measurement accuracy in the center of the ability distribution. At the extremes of the ability continuum, the Bayesian testing strategy resulted in slightly better measurement precision than did the maximum information testing strategy.

In both error of measurement comparisons, there was poorer measurement at the low end of the ability distribution, although the extremes--both positive and negative--were less precisely measured than the center of the ability continuum. The results indicate that the adaptive tests yield about the same level of measurement precision and that these levels were greater than those obtained from the conventional test at all levels of ability.

Figure 12
Average Bayesian Posterior Variance of Ability Estimates as a Function of Ability Level for Conventional, Bayesian, and Maximum Information Testing Strategies



## Discussion

The major finding in this study was that the conventional test yielded higher alternate forms reliability than did the adaptive tests. However, when the maximum information adaptive test was scored by the Bayesian scoring algorithm, reliabilities of short adaptive tests were higher than those of the conventional test, and differences in reliabilities were smaller at longer test lengths. Limitations of the item pool might account in part for the lowered reliability of the adaptive tests in comparison to the conventional test, since adaptive tests depend heavily on the quality of the items in the item pool. When an item pool consists of highly discriminating items, every ability level along the latent trait continuum can be measured with a high degree of precision using adaptive tests (McBride & Weiss, 1976). When there are few items to measure abilities at the extremes and/or the available items are of low discrimination, abilities at the extremes cannot be measured accurately.

The item pool used for the two adaptive tests had fewer items at the extremes of the ability range and these items had relatively lower discrimination parameters. It is likely that, especially at abilities where there were fewer items, the correlations between ability estimates would be attenuated and the adaptive process would be at a disadvantage as testing progressed. The result would be that toward the end of testing there would be fewer and fewer items available at a given ability level.

The adaptive test scoring process also depends on accurate parameterization of items and on testees responding according to a single latent trait. Experimental subjects taking a test that does not relate to any course they are taking and that does not count for a grade may respond carelessly, with less than full attention. It is unknown to what extent the item parameters are inaccurate. An optimal research strategy for comparison of conventional, Bayesian, and maximum information testing strategies on the basis of parallel forms reliability is through simulated testing. The disadvantage of inaccurate item parameters, non-optimal item pool characteristics, and the possibility that students did not respond exclusively in accordance with their ability level can be alleviated in simulation.

One additional factor that limits the comparison of the testing strategies in terms of alternate forms reliability correlations is the distribution of ability in the population. Since values of the Pearson product-moment correlations depend on the distributions of the ability estimates involved, different ability distributions can result in different levels of correlation. Thus, the reliability correlations confound the distribution of the ability estimates with the measurement precision of the testing strategies. Information is a measure of precision of measurement, yielding comparisons of testing strategies that are unconfounded by the distribution of the ability estimates. As Figure 11 shows, both adaptive testing strategies yielded scores with greater precision/information (lower errors of measurement) than did the conventional testing strategy.

On the basis of the reliability data, few conclusions can be drawn about the relative merits of the adaptive testing procedures. Bayesian scoring of the Bayesian test showed higher reliability than the maximum-likelihood-scored maximum information test. Bayesian scoring of the conventional and maximum information testing strategies yielded higher reliabilities than maximum likelihood scoring of the conventional and Bayesian testing strategies. This might indicate either that the Bayesian scoring algorithm yields more reliable estimates of ability or that it yields the same regressed or biased estimate of ability. The Bayesian test would tend to yield higher parallel forms reliabilities than the maximum information testing strategy in the case where most items measuring abilities at the extremes of the distribution are of lower discrimination. Because the Bayesian adaptive test yields regressed estimates of ability and requires fewer items measuring abilities at extreme $\theta$ values, the Bayesian ability estimates obtained, although biased, would be more stable than ability estimates from the maximum information testing strategy.

## REFERENCES

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1980. (NTIS No. AD A067752)

Kingsbury, G. G., & Weiss, D. J. A comparison of a Bayesian adaptive testing strategy to a conventional testing strategy: Alternate-forms reliability and criterion validity. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, in preparation.

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

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

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

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.

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

## ACKNOWLEDGMENTS

This study was supported by funds from Army Research Institute, Air Force Human Resources Laboratory and Office of Naval Research, under Contract N00014-79-0324 NR 150-431 with the Personnel and Training Research Programs of the Office of Naval Research.