DOCUMENT RESUME

ED 206 733	TM 810 662
AUTHOR	McKinley, Robert L.; Reckase, Mark D.
TITLE	A Comparison of a Bayesian and a Maximum Likelihood Tailored Testing Procedure.
INSTITUTION	Missouri Univ., Columbia. Tailored Testing Research Lab.
SPONS AGENCY	Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
REPORT NO	TTRL-RR-81-2
PUB DATE	Jun 81
CONTRACT	N00014-77-C-0097
NOTE	65p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Academic Ability; *Bayesian Statistics; *Comparative Analysis; Higher Education; Latent Trait Theory; Mathematical Models; *Maximum Likelihood Statistics; *Test Construction; Test Items; Test Reliability; Verbal Ability
IDENTIFIERS	School and College Ability Tests; *Tailored Testing; Test Length

ABSTRACT

A study was conducted to compare tailored testing procedures based on a Bayesian ability estimation technique and on a maximum likelihood ability estimation technique. The Bayesian tailored testing procedure selected items so as to minimize the posterior variance of the ability estimate distribution, while the maximum likelihood tailored testing procedure selected items so as to maximize the item information for the current ability estimate. Results of the analyses for the two procedures indicated that the optimal test length of Bayesian procedure was 14 items, while the optimal test length of the maximum likelihood procedure was 12 items. No difference was found between the procedures in terms of the reliability of the ability estimates. The Bayesian procedure yielded greater mean total test information than the maximum likelihood. The goodness of fit comparison indicated that the Bayesian procedure yielded poorer fit of the 3PL model to the data than did the maximum likelihood procedure. It was concluded that for large scale tailored testing, a maximum likelihood tailored testing procedure with item selection based on information is the procedure of choice. (Author/GK)

A Comparison of a Bayesian and a Maximum Likelihood Tailored Testing Procedure

Robert L. McKinley
and
Mark D. Reckase

Research Report 81-2
June 1981

Tailored Testing Research Laboratory
Educational Psychology Department
University of Missouri
Columbia, MO 65211

Prepared under contract No. N00014-77-C-0097, NR 150-395
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1 REPORT NUMBER Research Report 81-2	2 GOVT ACCESSION NO
3 RECIPIENT'S CATALOG NUMBER	
4 TITLE (and Subtitle) A Comparison of a Bayesian and a Maximum Likelihood Tailored Testing Procedure	5 TYPE OF REPORT & PERIOD COVERED Technical Report
	6 PERFORMING ORG REPORT NUMBER
7 AUTHOR(s) Robert L. McKinley and Mark D. Reckase	8 CONTRACT OR GRANT NUMBER(s) N00014-77-6-0097
9 PERFORMING ORGANIZATION NAME AND ADDRESS Department of Educational Psychology University of Missouri Columbia, MO 65201	10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.: 61153N Proj:RR042-04 T.A.: 042-04-01 W.V.: NR150-395
11 CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217	12 REPORT DATE June, 1981
	13 NUMBER OF PAGES 50
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)

Approval 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

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

Latent Trait Theory
Owen's Bayesian Ability Estimation
Maximum Likelihood Ability Estimation
Tailored Testing

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

A study was conducted to compare tailored testing procedures based on Owen's Bayesian ability estimation technique and on a maximum likelihood ability estimation technique. The Bayesian tailored testing procedure selected items so as to minimize the posterior variance of the ability estimate distribution, while the maximum likelihood tailored testing procedure selected items so as to maximize the item information for the current ability estimate. The study was conducted over the winter semester and summer session of 1980 using as subjects volunteers from a undergraduate introductory course in measurement and a graduate/undergraduate course in group intelligence testing. Analyses for the two procedures included a determination of the optimal test length, a comparison of the test-retest reliability, a comparison of the total test information, a comparison of the obtained ability estimates, a comparison of the goodness of fit of the 3PL model to the test data, and the compiling of descriptive statistics including average testing time and average test difficulty. Results of the analyses indicated that the optimal test length of the Bayesian procedure was 14 items, while the optimal test length of the maximum likelihood procedure was 12 items. No difference was found between the procedures in terms of the reliability of the ability estimates. The Bayesian procedure yielded greater mean total test information than the maximum likelihood, but this was found to be due to the regression of the Bayesian ability estimates to the mean of the assumed prior distribution, where more information was available. In the range of ability where there were ability estimates for both procedures there was no difference in total test information. Further analyses showed that the assumption of different priors can significantly alter the ability estimates obtained from the Bayesian tailored test, as well as the total test information yielded by the test and the optimal length of the test. The goodness of fit comparison indicated that the Bayesian procedure yielded significantly poorer fit of the 3PL model to the data than did the maximum likelihood procedure. Based on the results of these analyses it was concluded that for large scale tailored testing a maximum likelihood tailored testing procedure with item selection based on information is the procedure of choice.

CONTENTS

Introduction 1

Comparison of Latent Trait Models 1

Comparison of Ability Estimation Procedures 2

Method 3

Item Pool 3
Tailored Testing Procedures 7
Ability Estimation and Item Selection Procedures 7
Design. 10
Sample. 10
Analyses. 10

Results. 17

Optimal Cutoffs 17
Reliabilities 25
Total Test Information. 26
Ability Estimates 29
Items Administered. 36
Goodness of Fit 37
Descriptive Statistics. 37
Latency 41

Discussion 43

Optimal Cutoffs 43
Reliabilities 43
Total Test Information. 43
Ability Estimates 44
Items Administered. 45
Goodness of Fit 46
Descriptive Statistics. 47
Latency 47
Nonconvergence. 47

Summary and Conclusions. 47

References 49

A Comparison of a Bayesian and a Maximum Likelihood Tailored Testing Procedure

It is possible that in the near future there will be a widespread usage of tailored testing as an alternative to paper-and-pencil tests. For example, the Armed Services plan to implement tailored testing procedures in the near future. The possibility of large-scale implementation of tailored testing has increased the need to identify the optimal tailored testing procedures among the many that are available.

When selecting a tailored testing procedure a decision must be made as to which of numerous available techniques should be used for the component parts of the tailored testing procedure. For instance, one requirement for tailored testing is the calibration of items. For the calibration of items a number of models are available (e.g., one-, two-, and three-parameter logistic models), and for each model there may be a number of calibration programs available (e.g., the ANCILLES, LOGIST, and OGIVIA procedures for the three-parameter logistic model). Two other important components of tailored testing are the item selection procedure and the ability estimation procedure. While there are several ability estimation procedures available, the most common procedures are Owen's Bayesian and maximum likelihood estimation procedures. For selecting items the two most frequently used procedures are either to select items to maximize information at a given ability level or to select items to minimize the posterior variance of the ability estimates. While a number of studies have been done comparing various procedures available for a number of components, little has been done to directly compare tailored testing procedures employing these different ability estimation and item selection procedures. The purpose of the current study, then, is to compare in a live testing setting tailored testing procedures based on maximum likelihood ability estimation and maximum information item selection, and on Owen's Bayesian ability estimation (Owen, 1975) and minimum posterior variance item selection. Before proceeding with a presentation of the current study, however, previous studies investigating different procedures for tailored testing will be discussed.

Comparison of Latent Trait Models

Several studies have been done to compare the use of different models for tailored testing. One such study, a direct comparison of the 1PL and 3PL models in a live tailored testing setting, was reported by Koch and Reckase (1978). The purpose of this study was to compare the 1PL and 3PL models in a tailored testing application to vocabulary ability measurement. Both procedures used maximum likelihood techniques for item and ability parameter estimation. In both procedures items were selected to maximize the information function at the current ability estimate. The results of this study indicated that both models could be successfully applied to vocabulary ability measurement. The reliabilities reported (a combination of test-retest and equivalent forms reliabilities) indicated that the 3PL procedure yielded a slightly higher reliability than the 1PL procedure ($\underline{r}$=.77 for the 3PL procedure and $\underline{r}$=.61 for the 1PL procedure). One important finding of this study was that, if careful attention is not paid to the operational characteristics of the procedures, nonconvergence of the maximum likelihood ability estimation procedure can be a serious problem. In this study the 3PL procedure failed to converge to ability estimates in nearly one-third of the cases. Nonconvergence was not a serious problem with the 1PL procedure.

In a second study, reported by Koch and Reckase (1979), in which tne 1PL and 3PL models were applied to a multidimensional achievement test, nonconvergence of the 3PL maximum likelihood ability estimation procedure was encountered in about eight percent of the cases. The substantial reduction in nonconvergence cases over the previous study was attrituted to use of an item pool of more appropriate difficulty. Despite the reduction of the number of cases of nonconvergence in this study, the results still indicated a number of problem areas. Reliabilities were quite low for both procedures, as was the information yielded by both procedures. A number of possible explanations were suggested for the inadequate performance of the procedures. Among these were unstable item parameter estimates due to small sample sizes, instability due to poor linking procedures, and poor selection of entry points into the item pool. These problems appeared to have equally serious effects on both the 1PL and 3PL procedures.

A study reported by McKinley and Reckase (1980a) attempted to correct the problems encountered in the Koch and Reckase studies. Close attention was paid to appropriate item parameter linking and entry points for the 1PL and 3PL procedures. The results of this study indicated that both models could be quite successfully applied to tailored testing if correctly implemented. Both 1PL and 3PL reliabilities were higher than the reliability of a classroom test over the same material. A comparison of the 1PL and 3PL procedures indicated that the 3PL procedures yielded more information than the 1PL procedure or the classroom test. The 3PL procedure also fit the data better than the 1PL model. This study concluded that for tailored testing applications the 3PL model was the model of choice.

A similar conclusion was reached by Urry (1970, 1977b). Through a series of simulation studies, Urry found that tailored testing becomes less effective when a model with insufficient parameters is used. He concluded that construct validity decreases as a function of the degree of degeneracy of the model, and the 1PL model was particularly inappropriate for use with multiple-choice items because it did not portray multiple-choice response data with fidelity (Urry, 1977b).

Comparisons of Ability Estimation Procedures

It would appear to be clearly established in the literature that for tailored testing applications the 3PL model is more appropriate than the 1PL model. However, very little appears in the literature concerning the optimal procedures for ability estimation and item selection to be used in the 3PL tailored testing procedure.

One study that did compare different ability estimation procedures was conducted by Maurelli (1978). In this study a comparison was made of maximum likelihood and Bayesian ability estimation procedures in a simulated stradaptive testing application. The Bayesian ability estimation procedure was a modification of the procedure proposed by Owen (1975), and the maximum likelihood estimation procedure was similar to the one proposed by Lord (1975). Modifications included altering item selection procedures to make them compatible with the branching scheme of the stradaptive model. The variables investigated included the ability estimation procedures, test lengths (15, 30, and 45 items), and the use or non-use of prior information to determine entry level (variable entry point). The conclusions reached in this study included the finding that the maximum likelihood procedure performed best overall when quality of performance was measured in terms of bias (mean error of estimate), linearity of the regression of $\hat{\theta}$ on θ, average information, and fidelity (correlation of $\hat{\theta}$ with θ). The Bayesian procedure showed acceptable performance only at the longest test length when using prior information to determine the entry point. This procedure was found to be most deficient in the lower third of the ability scale. Maurelli also concluded that assuming a normal prior assures a regression of the estimates towards the mean of that prior. Unfortunately, since this study was conducted using the classic stradaptive item selection procedure for both ability estimation procedures, no comparison of item selection procedures was made.

From a review of the literature it is apparent that there is little evidence for determining whether maximum likelihood or Bayesian estimation is better, in general, for any application. Virtually no attempt has been made to directly compare the two procedures in a live tailored testing application. Nor has there been any comparison of the two most common item selection procedures used with these estimation procedures. The purpose of the present study, then, is to compare,in a live tailored testing application, maximum likelihood ability estimation and maximum information item selection with minimum posterior variance item selection and Bayesian ability estimation.

Method

Item Pool

Both the Bayesian and the maximum likelihood tailored tests used the same pool of 137 items. Items used for this study were selected from the first and third subtests of the School and College Ability Tests (SCAT), forms 2A and 3A. Estimates of the 3PL item parameters were obtained from the Educational Testing Service (ETS). The distributions of the item parameter estimates are shown in Figures 1-A, 1-B, and 1-C, and a summary of the descriptive statistics for these distributions is presented in Table 1. As can be seen in Figure 1-A the item discriminations (a-values) were fairly evenly distributed, with most of the items having a-values greater than .75. The item difficulties (b-values), shown in Figure 1-B, were approximately normally distributed between -2.0 and +2.0, with a slight tail at the negative end. The item guessing values (c-values) were very tightly clustered around the mean of .14. In terms of these item parameter estimates the item pool was of high quality. It came very close to satisfying the requirements for tailored testing item pools set out by Urry (1977a).

Table 1

Summary of Descriptive Statistics of Item Parameter Estimates for Tailored Testing Item Pool

Statistic	a_i	b_i	c_i
Mean	1.12	-.05	.14
Median	1.04	.06	.14
St. Dev.	.46	1.17	.05
Skewness	.27	-.22	1.61
Kurtosis	-.90	2.14	6.15
Minimum	.14	-4.25	.06
Maximum	1.94	4.56	.39

Note: The item pool contained 137 items.

Figure 2 shows the total information curve for the item pool. As can be seen in the figure the curve is slightly negatively skewed. The curve is very high near the center of the ability estimate range and drops off rather sharply toward the extremes of the ability estimate range. Information was high between -1.5 and +2.5, but outside the range not much information was available.

FIGURE 1

ITEM PARAMETER ESTIMATE DISTRIBUTIONS FOR TAILORED TESTING ITEM POOL

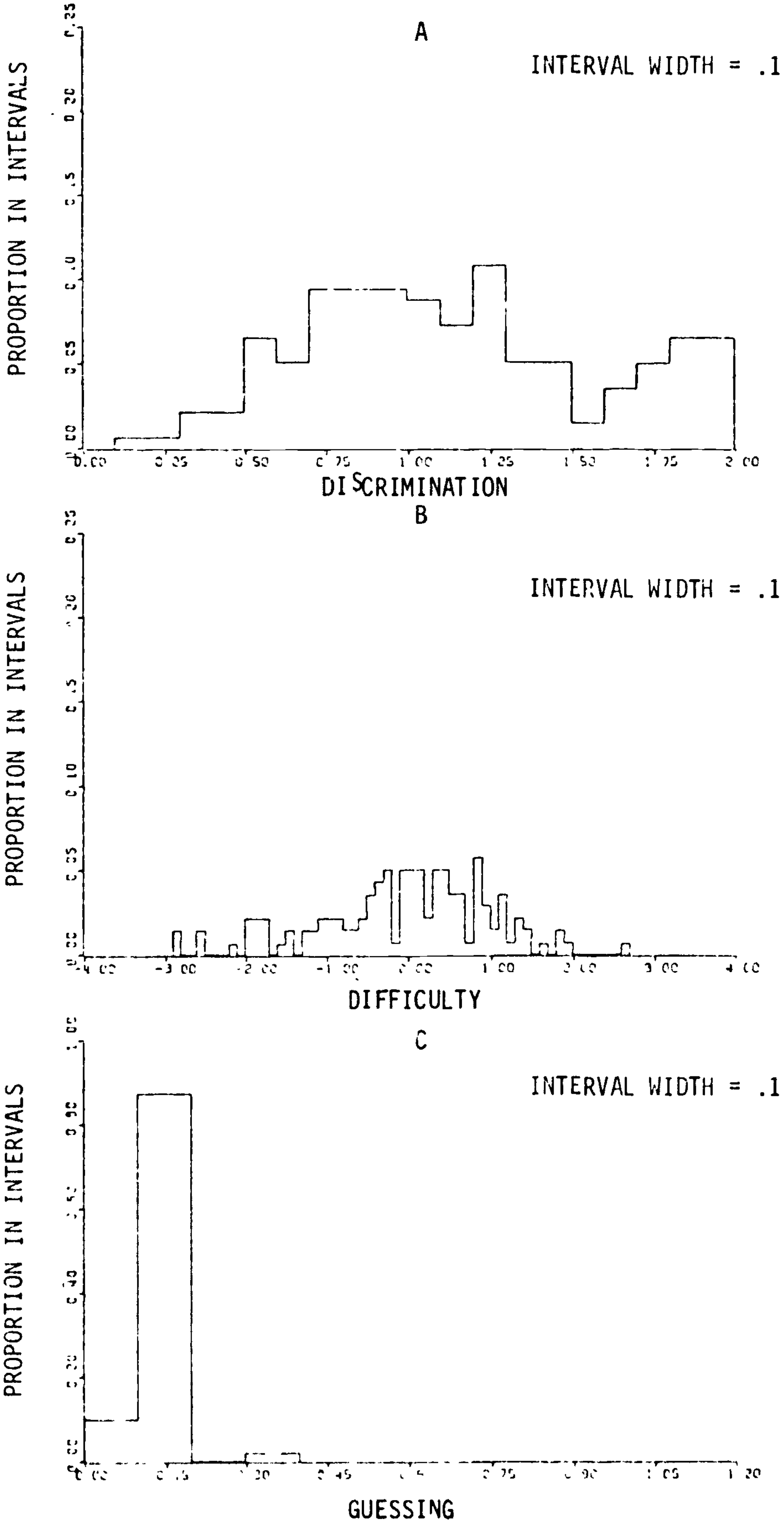
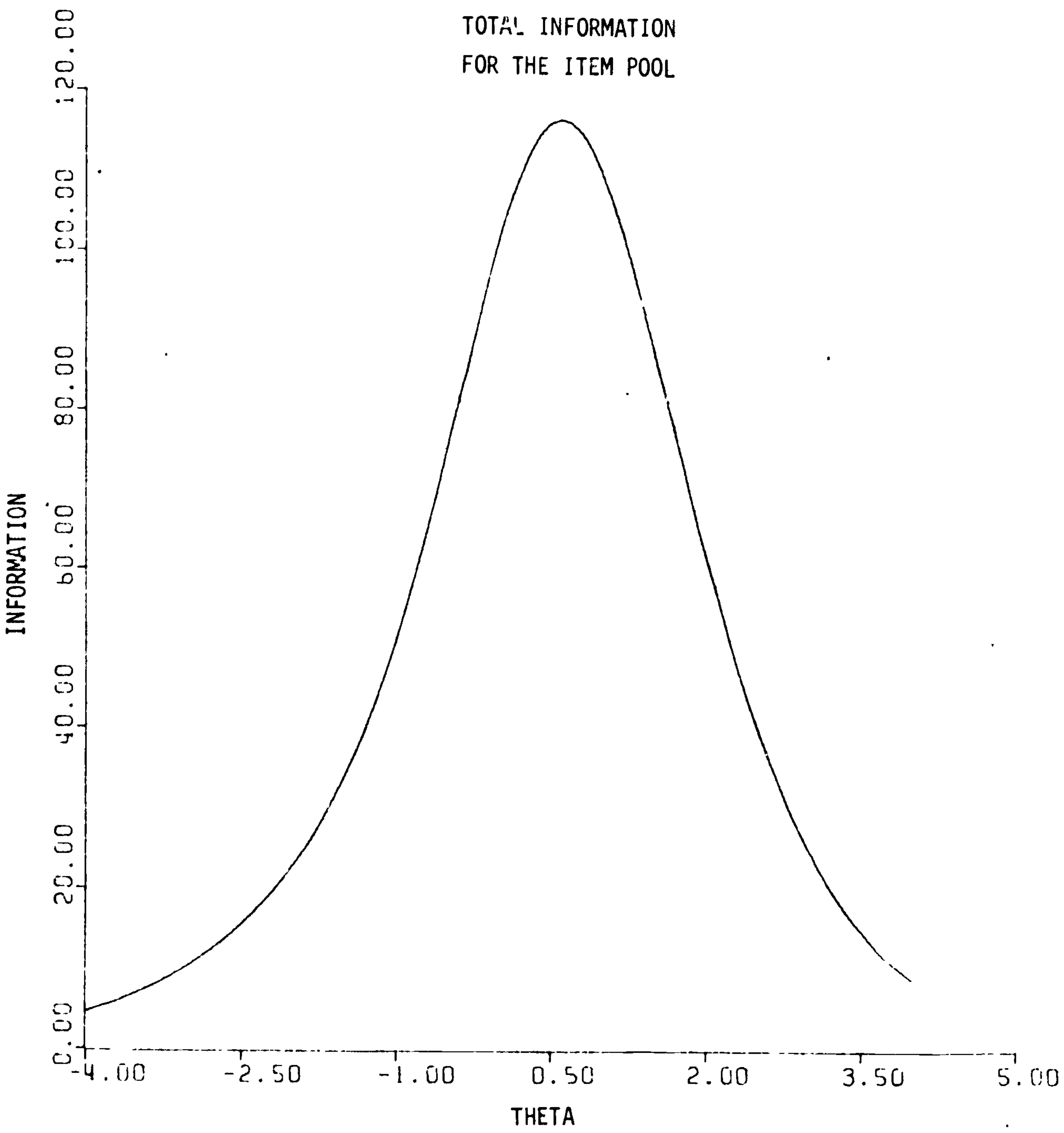


FIGURE 2
TOTAL INFORMATION
FOR THE ITEM POOL



Tailored Testing Procedures

In general, tailored testing procedures have three main components: an item selection routine, an ability estimation procedure, and a stopping rule. In this study two combinations of item selection and ability estimation procedures were used. For one group of subjects, items that had the most item information (Birnbaum, 1968) at the most recent ability estimate were selected and a maximum likelihood estimation procedure using all previous responses was used for ability estimates. For the rest of the subjects items were selected to minimize the posterior variance of the ability estimate distribution and Owen's Bayesian ability estimation procedure was used for ability estimates. These procedures will be described in greater detail shortly.

Before testing began no ability estimates were available for the subjects, so initial estimates were assigned to determine the starting points in the item pool. For both procedures the initial ability estimates were randomly assigned to each subject to be either +.150 or -.100. These values represent difficulty values near the center of the item pool difficulty distribution with one starting point on either side of the median. For the second session subjects were assigned the alternative initial ability estimate in order to provide different initial items from one session to the next. Both procedures used the same stopping rule. The tailored tests continued until 20 items had been administered.

Ability Estimation and Item Selection Procedures

For the maximum likelihood tailored tests, items were selected for administration that yielded the maximum item information at the most recent ability estimate. For the 3PL model the formula for item information is given by

$$I_i(\theta_j) = D^2 a_i^2 \psi[DL_i(\theta_j)] - D^2 a_i P_i(\theta_j) \psi[DL_i(\theta_j) - \log c_i] \qquad (1)$$

where $I_i(\theta_j)$ is the value of the item information at Ability θ_j, $L_i(\theta_j) = a_i(\theta_j - b_i)$, $P_i(\theta_j)$ is the probability of a correct response to Item i given Ability θ_j, and $\psi(x)$ is the logistic probability density function. Total test information is the sum of the item information values:

$$I(\theta_j) = \sum_{i=1}^{n} I_i(\theta_j). \qquad (2)$$

Formula 1 was used in the tailored testing procedure to compute the information for each item in the item pool at the examinee's current ability estimate. The item with the greatest information at the ability estimate was then administered to the examinee. The first item was selected to maximize information at the initial, randomly assigned ability estimate. If that item were correctly answered the ability estimate was increased by a fixed stepsize of .4, and if it were incorrectly answered the ability estimate was decreased by the fixed stepsize. The .4 stepsize was selected on the basis of previous research as giving the best combination of minimum error and least statistical bias (Patience and Reckase, 1980). This fixed stepsize procedure was used until a maximum likelihood ability estimate, the mode of the likelihood distribution, could be obtained (i.e., when both correct and incorrect responses were obtained). Each new item was then selected to maximize the information at the new ability estimate, with the restriction that no item could be used more than once.

For the Bayesian tailored tests items were selected to minimize the posterior variance of the ability estimate distribution. Owen's procedure assumes a normal distribution of ability as a prior. In this study the mean of that prior distribution was set equal to the initial, randomly assigned ability estimate, and the standard deviation of the prior was set equal to one. The first item was then selected so as to result in the greatest possible reduction in the standard error of estimate (Urry, 1977a). This was accomplished in the following manner. Rather than computing the actual value of the standard error of estimate, a statistic labelled α by Jensema (1974),was used for efficiency of computation. The expected standard error is a function of α. The value α_i was calculated for each item in the item pool according to the formula.

$$\alpha_i = \frac{W_i(2-u_i)\exp(2D_i{}^2)}{2(1-c_i)t_i} \qquad (3)$$

where c_i is the item guessing value and the following relationships hold:

$$D_i = \frac{b_i - \theta_j}{[2(a_i{}^{-2} + \sigma_j{}^2)]^{1/2}}\ , \qquad (4)$$

$$\text{erf } D_i = \sqrt{\pi} \int_0^{D_i} \exp(-x^2)dx, \qquad (5)$$

$$u_i = 1 - \text{erf } D_i, \qquad (6)$$

$$W_i = c_i + \frac{(1 - c_i)u_i}{2}\ , \qquad (7)$$

and

$$t_i = \frac{a_i{}^2 \; \sigma_j{}^2}{1 + a_j{}^2 \; \sigma_j{}^2} \quad . \tag{8}$$

In the above equations a_i is the discrimination of Item i, b_i is the difficulty of Item i, θ_j is the jth ability estimate, and $\sigma_j{}^2$ is the variance of the j ability estimates.

The item with the smallest α_i value was administered to the examinee. Then new estimates of θ and σ^2 were made based on the examinee's response. If the response were correct, new estimates of ability and variance were computed as

$$\theta_{j+1} = \theta_j + \frac{S_i(1-c_i)}{\sqrt{2\pi}\; W_i \exp(D_i{}^2)} \tag{9}$$

and

$$\sigma^2{}_{j+1} = \sigma_j{}^2[1 - \frac{t_i(1-c_i)[W_i-c_i-D_i W_i u_i \sqrt{\pi}\;\exp(d_i{}^2)]}{W_i{}^2 u_i \pi \exp(2D_i{}^2)}] \tag{10}$$

where

$$S_i = \frac{\sigma_j{}^2}{(a_i{}^{-2} + \sigma_j{}^2)^{1/2}} \quad . \tag{11}$$

and the other parameters are as previously defined. If an incorrect response were made the new ability and variance estimates were computed as

$$\theta_{j+1} = \theta_j - \frac{S_i\sqrt{2}}{\sqrt{\pi}\; V_i \exp(D_i{}^2)} \tag{12}$$

and

$$\sigma_{j+1}{}^2 = \sigma_j{}^2[1 - \frac{2t_i[1 + \sqrt{\pi}\; D_i V_i \exp(D_i{}^2)]}{\pi[V_i \exp(D_i{}^2)]^2}] \; , \tag{13}$$

where

$$V_i = 1 + \text{erf } D_i. \tag{14}$$

Once new estimates of ability and variance were calculated new α_i values were computed for all unused items in the item pool. The above sequence was then repeated. For further discussion of the mechanics of this procedure see Owen (1975), Urry (1971), Jensema (1972) or Jensema (1974). One further point about this procedure that should be noted is that the prior assumption of normality is maintained throughout the procedure. That is, the distribution of ability is not recomputed after each item. Rather, it is assumed to remain normal with a mean equal to the current ability estimate and variance equal to the current estimate of variance.

Design

This study employed a test-retest design, with two sessions one week apart. Subjects were randomly assigned to take either a maximum likelihood or a Bayesian tailored test, and subjects were randomly assigned an initial ability estimate of -.100 or +.150 for the first session. For the second session subjects were assigned to the alternative initial ability estimate from the one assigned the first session. Subjects received the same type of test, Bayesian or maximum likelihood, for both sessions in order to make test-retest reliability comparisons possible. The tailored tests were administered on Applied Digital Data Systems (ADDS) Consul 980 cathode ray tube terminals connected to an Amdahl 470/V7 via Time Sharing Option facilities.

Sample

This study was conducted over the winter semester and summer session of 1980. For the rest of this paper both will be referred to as semesters. The winter semester study was conducted using 34 volunteers from an introductory course in measurement. Of these 34 volunteers, 31 were female and three were male, 33 were seniors and one was a graduate student. The second semester subjects included volunteers from two courses, the introductory course in measurement mentioned above, and a graduate/undergraduate course in group intelligence testing. There were 36 volunteers, of whom 14 were in the introductory course in measurement. The remaining 22 were from the group intelligence tasting course. There were 25 females in this second semester group, and 11 males. There were 14 graduate students, 15 seniors, six juniors, and one sophomore.

Analyses

Before any of the planned analyses were performed, preliminary analyses were performed to determine whether data from the two semesters should be combined. These analyses included the plotting and visual comparison of the ability estimate distributions obtained for the two semesters and a comparison of the ability estimate means from the two semesters using analysis of variance (ANOVA) procedures. Because the second semester study included students enrolled in graduate school, and because it occurred during a summer semester, there was some reason to suspect the two groups were not comparable and should not be combined.

Figures 3 and 4 show the ability estimate distributions for the Bayesian tailored tests for the winter and summer semesters, respectively. Figures 5 and 6 show the maximum likelihood tailored test ability estimate distributions for the winter and summer semesters, respectively. A visual comparison of these plots indicated that the ability estimates from the summer semester tended to be higher than the winter semester ability estimates for both the Bayesian and the maximum likelihood tailored tests. This indicated that the subjects in the summer semester may have had higher vocabulary ability than the winter semester subjects, since both groups took tests using the same item pool. The means of these distributions are shown in Table 2.

FIGURE 3

ABILITY ESTIMATE FREQUENCY DISTRIBUTION FOR BAYESIAN TAILORED TESTS FOR WINTER COMBINED OVER SESSIONS

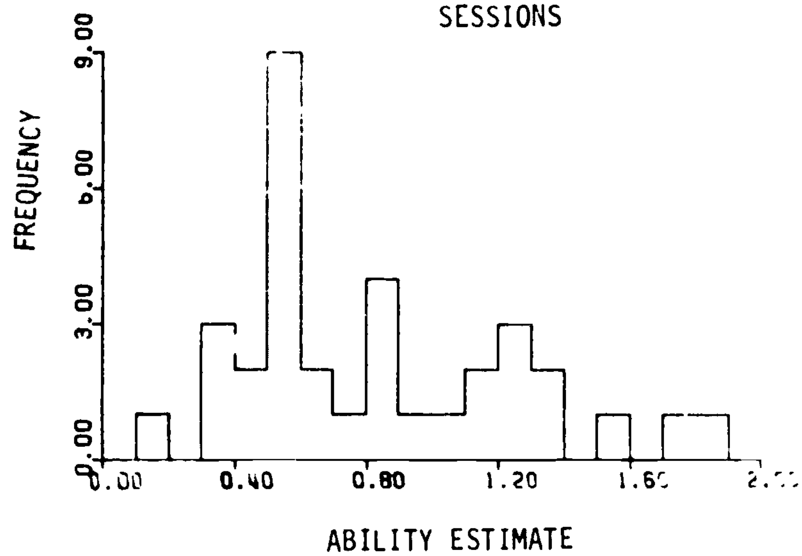


FIGURE 4
ABILITY ESTIMATE FREQUENCY
DISTRIBUTION FOR BAYESIAN
TAILORED TESTS FOR SUMMER
COMBINED OVER SESSIONS

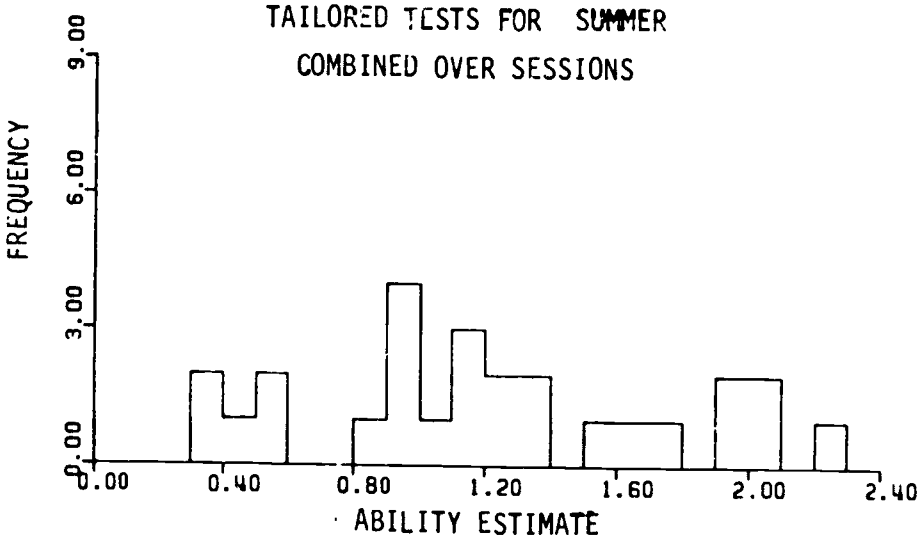


FIGURE 5
ABILITY ESTIMATE FREQUENCY
DISTRIBUTION FOR MAXIMUM
LIKELIHOOD TAILORED TESTS FOR
WINTER COMBINED OVER SESSIONS

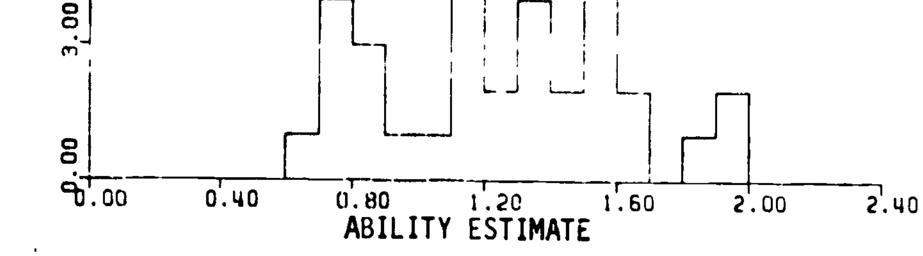


FIGURE 6

ABILITY ESTIMATE FREQUENCY DISTRIBUTION FOR MAXIMUM LIKELIHOOD TAILORED TESTS FOR SUMMER COMBINED OVER SESSION

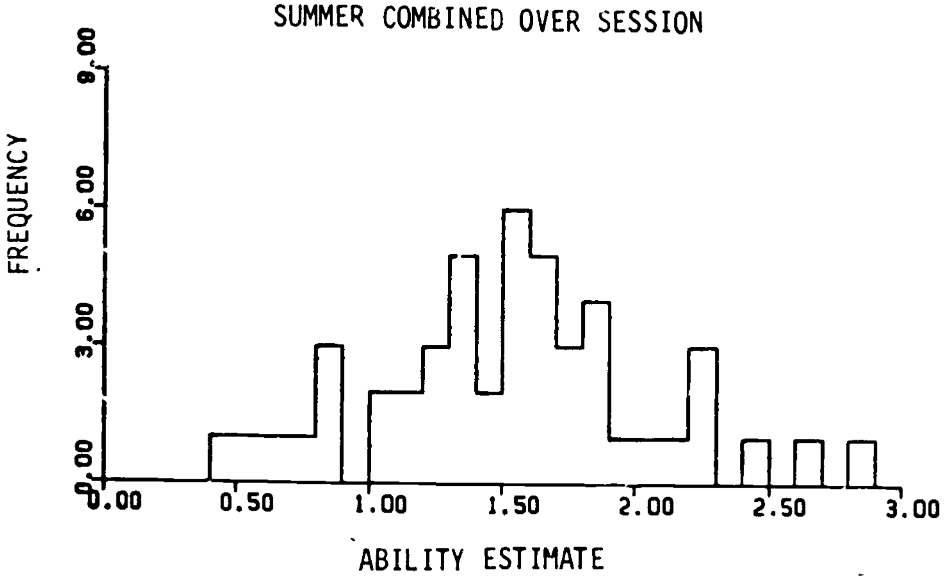


Table 2

Mean Ability Estimates for Bayesian and Maximum Likelihood Tailored Tests for Winter and Summer

Test	Spring		Summer	
	Session 1	Session 2	Session 1	Session 2
Bayesian	.75	.86	1.18	1.27
Maximum Likelihood	1.25	1.30	1.53	1.56

In an attempt to confirm that there was a real difference between the groups a three-way analysis of variance with repeated measures on one factor was performed. The independent variables were test (Bayesian or maximum likelihood), semester (winter or summer), and session. The repeated measures were over the sessions. In order to facilitate the interpretation of the results of this analysis the Bayesian and maximum likelihood ability estimates were put on the same scale by converting them to _Z_-scores. The _Z_-scores were computed for each procedure using the within procedure means and standard deviations. This put both sets of ability estimates on the same scale, thus eliminating any differences due to different scales. This was done because at this stage the differences in the procedures were not an issue. The ANOVA was then run using the _Z_-scores as the dependent variable. The results of the ANOVA are shown in Table 3.

Table 3

Analysis of Variance Table

for Preliminary Comparison of Winter

and Summer Ability Estimate Distributions

Source	_SS_	_df_	_MS_	_F_	_p_
Test	35.96	1	35.96	0.20	.655
Semester	1406.51	1	1406.51	7.89	.006
Test x semester	35.96	1	35.96	0.20	.655
Error	11942.65	67	178.25		
Session	50.87	1	50.87	6.63	.012
Test x Session	9.67	1	9.67	1.26	.265
Semester x Session	2.03	1	2.03	0.26	.608
Test x Semester x Session	0.10	1	0.10	0.01	.910
Error	513.97	67	7.67		

Because _Z_-scores were used the presence of different scales for the two test procedures did not result in a problem in interpretation. Also, because the means for the two tests were set to 50 the test main effect was eliminated. The semester main effect, however, was significant ($\underline{F}$=7.89, $\underline{p}$<.01), indicating that the examinees in the summer study had significantly higher vocabulary ability estimates. Because of this the decision was made not to combine the two groups, but rather to treat semester as an independent variable with two levels.

Once the determination was made that the data from the two semesters should not be combined, a number of analyses were performed separately on the two sets of data. The first analysis was the determination of optimal test lengths for the two procedures. This analysis was performed since the administration of inappropriate items may induce ability estimate bias. It is important to not allow the tailored tests to continue beyond the optimal length. As the items appropriate for an ability are used up, bias can be introduced into the ability estimates if the test is continued, since the procedure may begin to administer less appropriate items (Reckase, 1974). Therefore, it is important that the tailored tests do not continue beyond the optimal length. The test length analysis was accomplished by plotting the convergence of the procedures to ability estimates for each tailored test. That is, for a given tailored test the ability estimate obtained after each item was plotted against the item number. Then a second plot was done as an overlay on the same set of axes. For the maximum likelihood tests the overlay was the plot of the item information that was obtained for each item at the previous ability estimate against the item number. For the Bayesian tests the overlay was the plot of the standard error of estimate obtained after each item against the item number. The purpose of these plots was to graphically represent the interrelationships of test length, ability estimate, and item information or standard error of estimate, so that a determination could be made as to what test length and item information or standard error of estimate would be optimal as cutoff values for terminating the tailored tests.

Other analyses performed included comparisons of the Bayesian and maximum likelihood test-retest reliabilities, the total test information yielded by the two procedures, and the ability estimates yielded by the two procedures. All of these comparisons were made using the 20 item level as well as at the various test lengths determined by the optimal cutoff analyses. All correlations used in the reliability analyses were computed using both ability estimates and estimated true scores (Lord, 1979). The computation of the estimated true scores was accomplished by summing the probabilities of correct responses at the examinee's final ability estimate for all the items in the item pool. The formula for estimated true scores is as follows:

$$t(\theta_j) = \sum_{i=1}^{n} P_i(\theta_j) , \qquad (15)$$

where $t(\theta_j)$ is the estimated true score for Examinee j.

The reliabilities computed for this study were not strictly test-retest reliabilities, but rather a mixture of test-retest and equivalent forms reliabilities, since the tests in one session were not identical to tests taken in the other session. The hypothesis that all of the reliabilities were estimates of the same reliability was tested using a chi-square test given in Snedecor and Cochran (1980).

The total test information analyses were done to compare the amount of information yielded at the final ability estimate by the two procedures. Total test information was computed using Equations 1 and 2, where the summation in Equation 2 is over the items of each tailored test. Total test informations were compared using both plots and analysis of variance procedures.

Comparisons of the ability estimates included a number of analyses. One analysis was the comparison of the distributions of ability estimates yielded by the two procedures using plots of the distributions. Also, analysis of variance procedures were used to compare the mean ability estimates. Another comparison involved the use of the Bayesian and maximum likelihood ability estimation procedures with the item selection procedures switched. That is, Bayesian ability estimates were obtained for the items selected by the maximum likelihood tailored test procedure, and maximum likelihood ability estimates were obtained for the items selected by the Bayesian tailored test procedure. The purpose of this analysis was to determine whether the differences found between the two procedures were due solely to the ability estimation procedures, or whether the item selection procedures also had an effect.

Another set of analyses performed was the comparison of the items that were administered by the tailored tests. Included in these analyses were a comparison of the items administered by the two procedures and a comparison for each procedure of the items administered for the two sessions.

The goodness of fit of the 3PL model to the test data for the two procedures were also compared. The goodness of fit statistic used in this study was the mean square deviation (MSD), calculated by summing over examinees the squared differences between the actual responses to the items and the expected responses to the items (probability of a correct response) as predicted by the model (Reckase, 1977). The formula for the MSD statistic is

$$MSD_j = \sum_{i=j}^{n} \frac{(u_{ij} - P_i(\theta_j))^2}{n_j} , \qquad (16)$$

where MSD_j is the mean squared deviation for Examinee j, u_{ij} is the actual response to Item i by Examinee j, $P_i(\theta_j)$is the probability of a correct response to Item i by Examinee j determined from the model using the final ability estimate and the item parameter estimates, and n_j is the number of items in the tailored test for Examinee j. The goodness of fit of the two procedures was compared using the MSD statistic as the dependent variable in a dependent *t*-test.

Other analyses run on the data included two correlational analyses. One such analysis performed was the correlation of item response latency times with the ability estimates. Correlations were obtained between mean item response latency times and final ability estimates. Also the mean item response latency times for correct responses and incorrect responses were compared using an analysis of variance procedure.

A final set of analyses performed was the compilation of descriptive statistics for the two procedures for both sessions. Descriptive statistics included average testing time and average test difficulty.

Results

Optimal Cutoffs

Figure 7 shows typical convergence plots that were obtained for one person using the winter data for the maximum likelihood tailored tests and Figure 8 shows typical convergence plots obtained for one person using the winter data for the Bayesian tailored tests. The values of the ability estimates and the item information estimates at the estimated ability that were plotted in Figure 8 are shown in Table 5. These figures and tables and others like these were examined in order to determine the minmum test length at which the ability estimates obtained from the two procedures appeared to be stable. In the plots of the ability estimates obtained from the maximum likelihood tailored test procedure the curve appeared to flatten out at about 12 items, indicating that 12 items was a sufficient length for the tailored tests. For the Bayesian procedure the curves flattened out around the 14 item level. In terms of item information the 12 item cutoff for the maximum likelihood procedure would represent an information cutoff of approximately 1.64. That is, the average item information for Item 12, using the 12 item ability estimates, was 1.64 (n=18). The Bayesian cutoff of 14 items would represent a standard error of estimate cutoff of .25, which was the average standard error of estimate of the 14 item level (n=16).

FIGURE 7

ABILITY ESTIMATES AND INFORMATION VALUES AFTER EACH ITEM IN A MAXIMUM LIKELIHOOD TAILORED TEST FOR THE WINTER SEMESTER

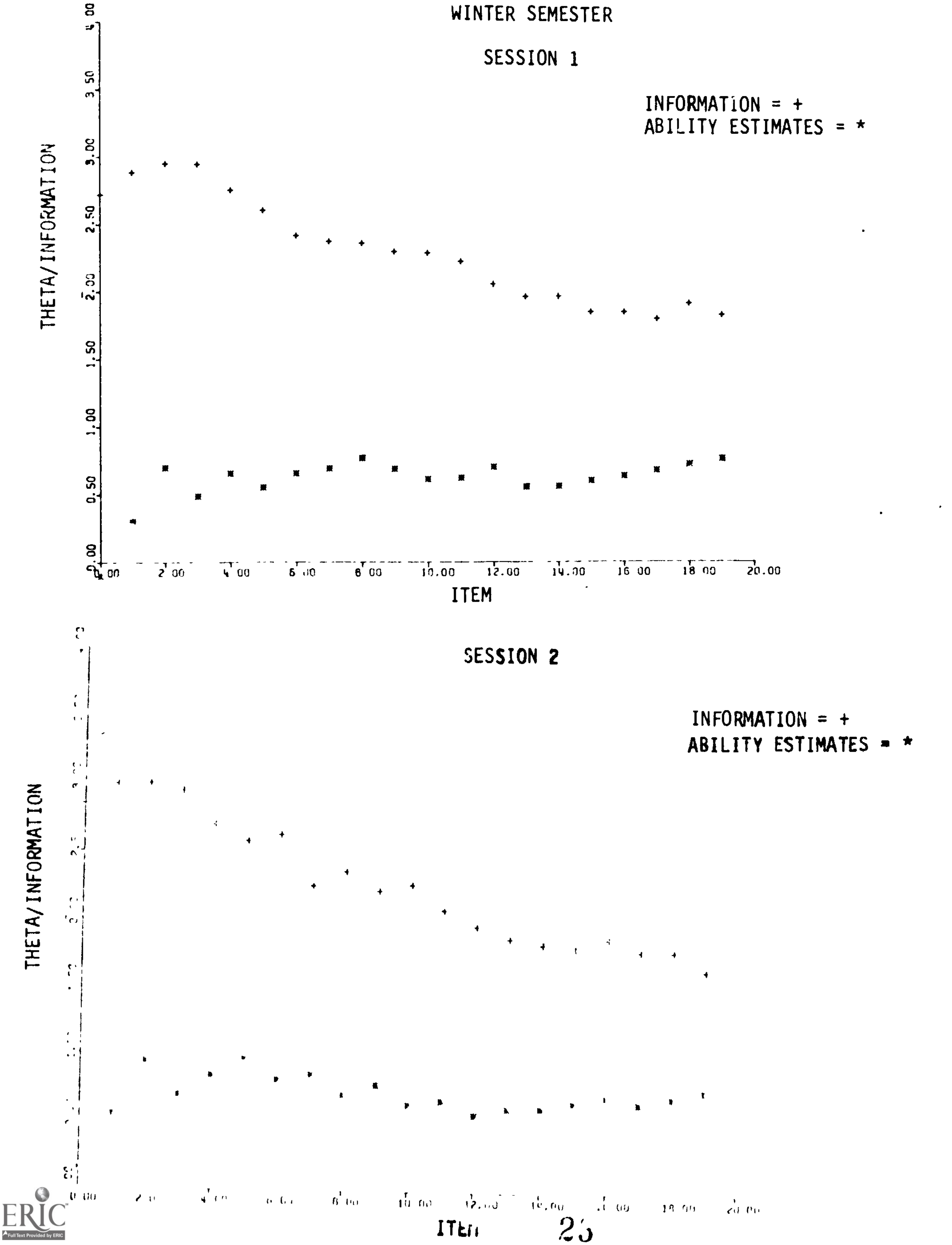
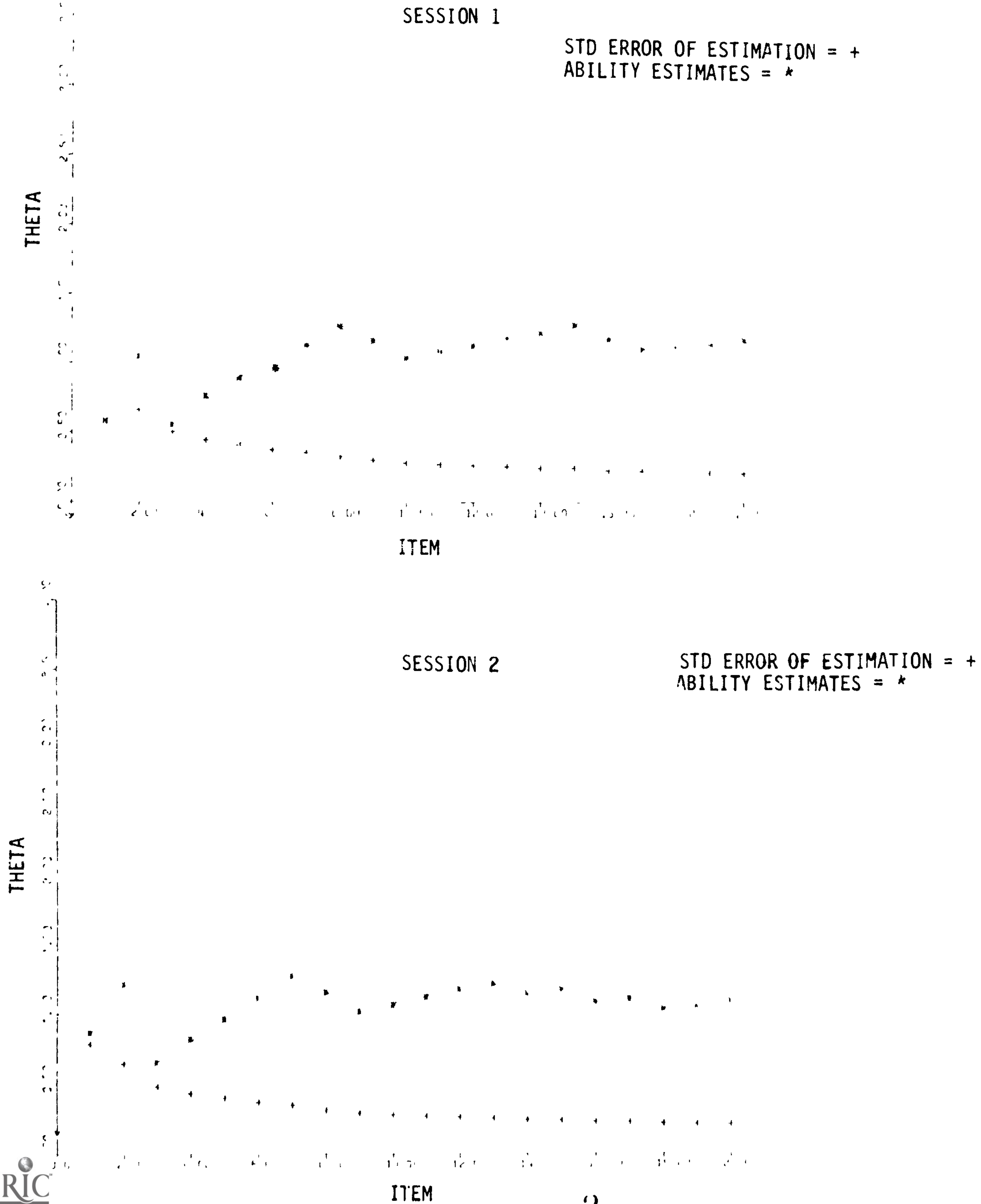


FIGURE 8

ABILITY ESTIMATES AND STD ERRORS OF ESTIMATE AFTER EACH ITEM IN A BAYESIAN TAILORED TEST FOR THE WINTER SEMESTER



SESSION 1

SESSION 2

Table 4

Ability Estimates and Item Information for
Both Sessions of a Maximum Likelihood Tailored Test
for the Winter Semester

Item	Session 1		Session 2	
	Ability Estimate	Information	Ability Estimate	Information
0	-.100	2.721	.150	2.850
1	.300	2.884	.550	2.994
2	.700	2.952	.950	3.003
3	.487	2.943	.701	2.952
4	.654	2.752	.847	2.705
5	.548	2.607	.964	2.584
6	.654	2.413	.817	2.637
7	.692	2.371	.855	2.259
8	.768	2.361	.708	2.363
9	.690	2.301	.786	2.224
10	.614	2.288	.639	2.276
11	.621	2.225	.677	2.088
12	.699	2.055	.571	1.969
13	.553	1.966	.619	1.880
14	.560	1.964	.626	1.838
15	.598	1.851	.664	1.814
16	.635	1.846	.712	1.887
17	.673	1.793	.664	1.795
18	.721	1.911	.712	1.796
19	.759	1.826	.760	1.657

Table 5

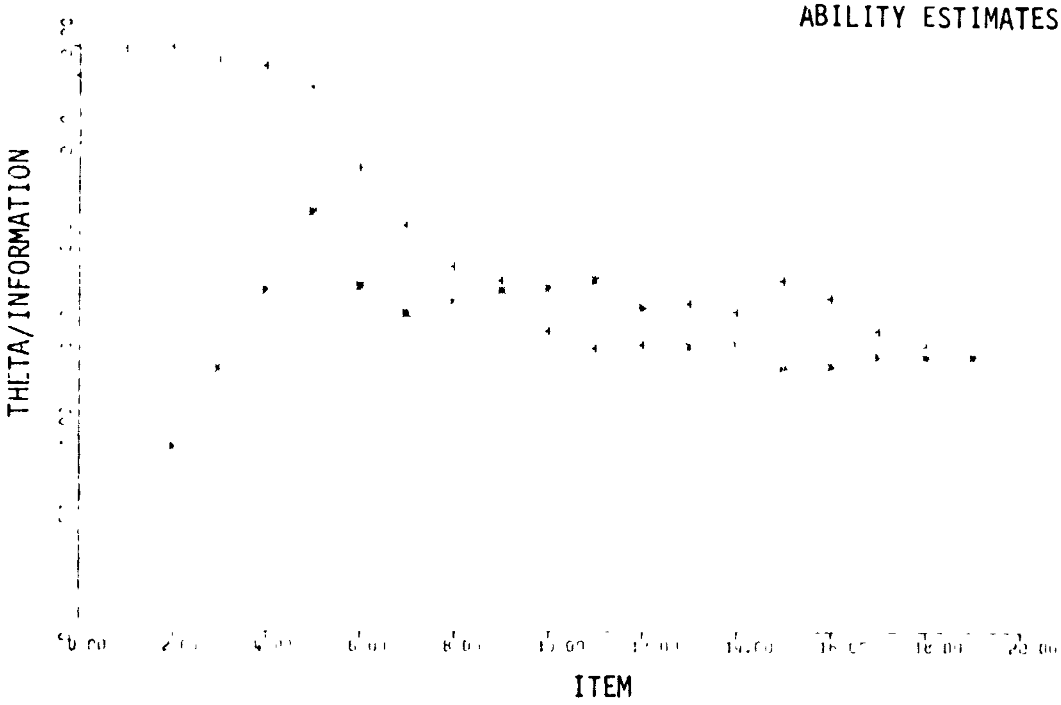
Ability Estimates and Standard Errors of Estimate
for Both Sessions of a Bayesian Tailored Test
for the Winter Semester

Item	Session 1		Session 2	
	Ability Estimate	Standard Error of Estimate	Ability Estimate	Standard Error of Estimate
0	-.100	1.000	.150	1.000
1	.579	.818	.872	.790
2	1.043	.660	1.216	.650
3	.551	.502	.660	.485
4	.760	.440	.827	.435
5	.884	.401	.964	.399
6	.995	.371	1.113	.371
7	1.124	.348	1.270	.348
8	1.264	.327	1.155	.316
9	1.161	.300	1.019	.292
10	1.037	.279	1.065	.281
11	1.078	.268	1.119	.272
12	1.126	.261	1.173	.263
13	1.177	.253	1.214	.257
14	1.214	.247	1.146	.243
15	1.275	.243	1.168	.239
16	1.176	.232	1.082	.228
17	1.101	.223	1.102	.224
18	1.121	.219	1.037	.217
19	1.138	.217	1.057	.212
20	1.165	.212	1.086	.209

FIGURE 9

ABILITY ESTIMATES AND INFORMATION VALUES AFTER EACH ITEM IN A MAXIMUM LIKELIHOOD TAILORED TEST FOR THE SUMMER SEMESTER

SESSION 1 INFORMATION = +
ABILITY ESTIMATES = *



SESSION 2 INFORMATION = +
ABILITY ESTIMATES = *

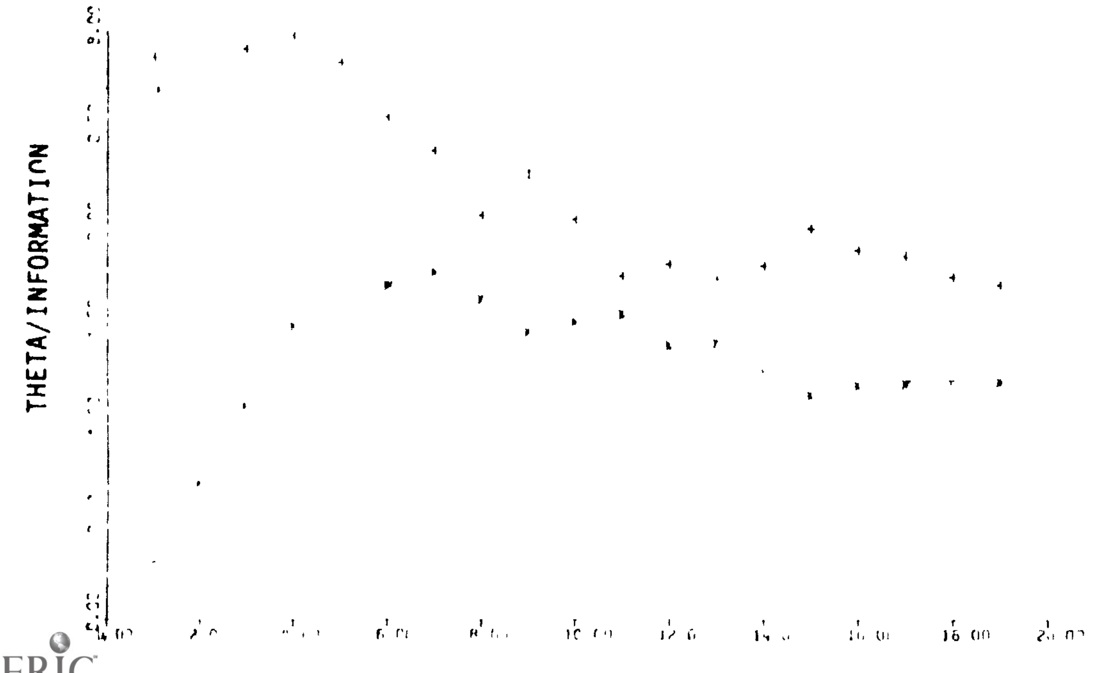
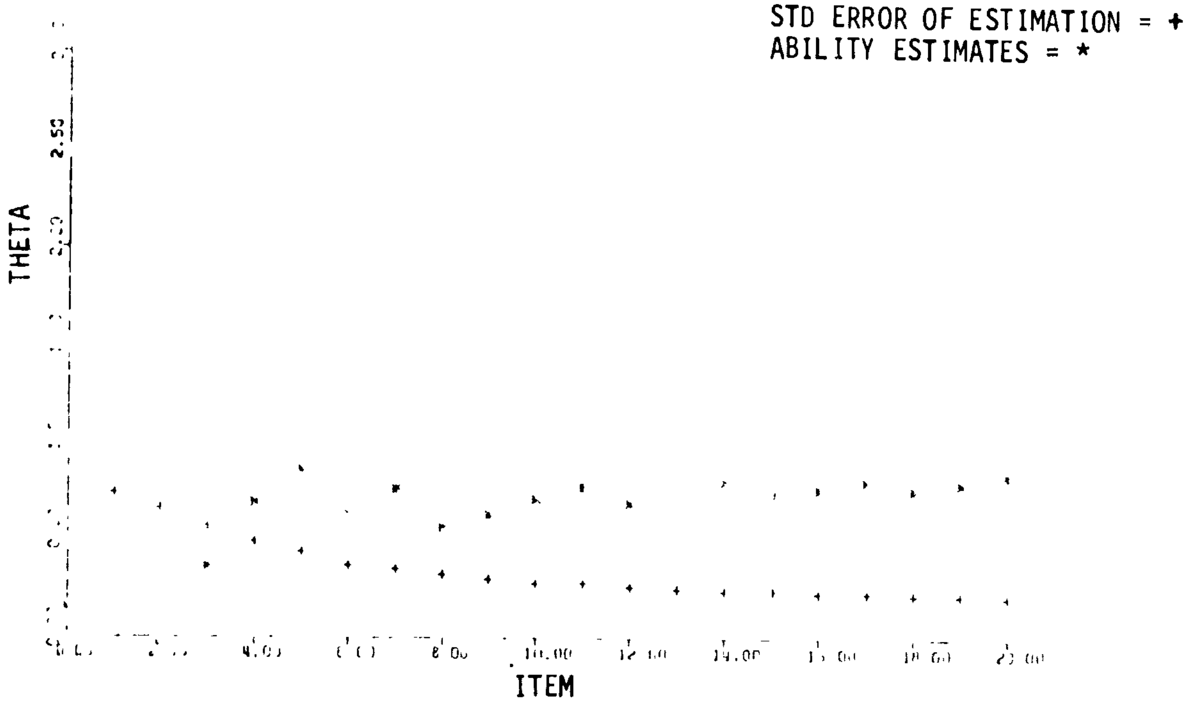


FIGURE 10

ABILITY ESTIMATES AND STD ERRORS OF ESTIMATE AFTER EACH ITEM IN A BAYESIAN TAILORED TEST FOR THE SUMMER SEMESTER

SESSION 1



SESSION 2

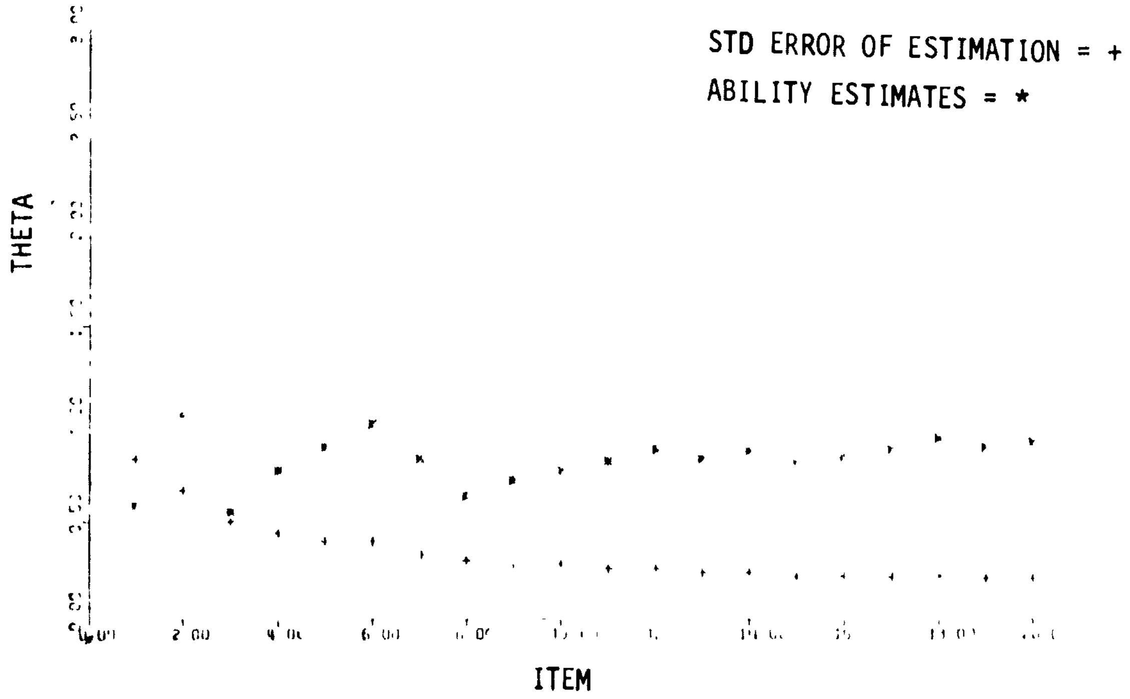


Figure 9 shows typical convergence plots for the maximum likelihood tailored tests for the summer semester and Figure 10 shows typical Bayesian convergence plots for the summer semester. Table 6 shows the ability estimates and item information estimates at the estimated ability used for Figure 9 and Table 7 shows the ability estimates and standard errors of estimate used in Figure 10. A visual examination of these plots indicates that the optimal cutoffs for the summer semester data were roughly the same as those determined for the winter data.

The mean item informations and mean standard errors of estimate at the 12 and 14 item levels, respectively, were not significantly different from the values obtained for the winter semester data. Because of this, all of the analyses were performed using ability estimates based on tailored tests that were 12 items long, 14 items long, and 20 items long, regardless of the semester or whether the testing procedure was Bayesian or maximum likelihood. In this way the Bayesian and maximum likelihood tailored test procedures could be compared at their respective optimal cutoffs and at various equal test lengths.

Table 6

Ability Estimates and Item Information for Both Sessions of a Maximum Likelihood Tailored Test for the Summer Semester

Item	Session 1		Session 2	
	Ability Estimate	Information	Ability Estimate	Information
0	.150	2.850	-.100	2.721
1	.550	2.994	.300	2.884
2	.950	3.003	.700	2.952
3	1.350	2.942	1.100	2.922
4	1.750	2.898	1.500	2.990
5	2.150	2.794	1.900	2.853
6	1.772	2.383	1.708	2.571
7	1.626	2.077	1.783	2.396
8	1.701	1.869	1.637	2.072
9	1.749	1.800	1.470	2.285
10	1.756	1.541	1.518	2.049
11	1.804	1.452	1.565	1.764
12	1.658	1.473	1.398	1 823
13	1.465	1.684	1.406	1.742
14	1.473	1.644	1.259	1.812
15	1.356	1.804	1.143	2.002
16	1.363	1.708	1.190	1.889
17	1.411	1.542	1.198	1.857
18	1.404	1.464	1.205	1.750
19	1.411	1.447	1.212	1.713

Table 7

Ability Estimates and Standard Errors of Estimate for Both Sessions of a Bayesian Tailored Test for the Summer Semester

Item	Session 1		Session 2	
	Ability Estimate	Standard Error of Estimate	Ability Estimate	Standard Error of Estimate
0	.150	1.000	-.100	1.000
1	-.429	.741	.579	.818
2	-.036	.660	1.043	.660
3	.358	.569	.551	.502
4	.694	.491	.760	.440
5	.857	.438	.884	.401
6	.630	.375	.995	.371
7	.761	.352	.824	.332
8	.561	.316	.632	.298
9	.632	.302	.711	.286
10	.713	.290	.760	.276
11	.770	.279	.809	.268
12	.694	.263	.874	.259
13	.743	.255	.824	.247
14	.805	.249	.860	.241
15	.742	.239	.797	.232
16	.765	.232	.827	.226
17	.796	.228	.872	.221
18	.763	.221	.930	.219
19	.787	.217	.878	.212
20	.830	.212	.911	.208

Reliabilities

Table 8 shows the test-retest reliabilities that were obtained for this study. It includes reliabilities at the 12, 14, and 20 item levels, for both the Bayesian and maximum likelihood tailored tests, for both the winter and summer session data. The reliabilities in Table 8 were computed using both ability estimates and estimated true scores. Fisher's _r_ to _z_ transformation was applied, and then a chi-square test (Snedecor and Cochran, 1980) was performed to determine whether all the reliabilities were estimates of the same reliability. The obtained chi-square statistic was found to be not significant. Thus, the 12 item test length was not significantly different from the 20 item length in terms of reliability. Moreover, based on these results it would appear that there were no significant differences between the reliabilities of the maximum likelihood and Bayesian procedures, regardless of test length. That is, the reliability at a test length of 12 items was approximately the same for the Bayesian tailored tests as it was for the maximum likelihood tailored tests, even though 12 items was selected as the optimal cutoff for the maximum likelihood tailored tests. Also, although the 12 and 14 item cutoffs were determined using the winter semester data, the reliabilities obtained at those test lengths for the summer data were about the same. However, it should be remembered that these reliabilities were obtained using small sample sizes, so a large difference was needed for significance.

Table 8

Bayesian and Maximum Likelihood Tailored Test Reliabilities for Winter and Summer Using Ability Estimates and Estimated True Scores

Test	Estimate	Winter			Summer		
		20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
Bayesian	Ability	.914	.919	.866	.963	.929	.905
	True Score	.885	.900	.830	.946	.881	.855
Max. Like.	Ability	.925	.865	.943	.908	.748	.777
	True Score	.899	.820	.936	.921	.875	.839

Note. Sample sizes for computation of reliabilities were n=16 for the winter semester Bayesian reliabilities, n=13 for the summer semester Bayesian reliabilities, n=18 for the winter semester maximum likelihood reliabilities, and n=23 for the summer semester maximum likelihood reliabilities.

Total Test Information

The mean total test information at the ability estimates obtained for the Bayesian and maximum likelihood tailored tests at the 12, 14, and 20 item test lengths for both the winter and summer semesters are shown in Table 9. It was expected that the mean total test informations for both the Bayesian and the maximum likelihood tests would be greater for the winter semester than for the summer semester. This was expected because it had already been determined that the ability estimates for the summer semester were significantly higher than the ability estimates for the winter semester. This would have resulted in items with greater b-values being selected during the summer tests. Since fewer items were available farther away from the center of the item pool, the total information in that region of the pool would be lower. In addition, the fewer items available would result in greater mismatching of ability estimates and item difficulty, which would also lower total test information. In order to confirm this, a three-way ANOVA was run using the 20 item mean total test information as the dependent measure, with semester, session, and test type (Bayesian or maximum likelihood) as independent variables. The session variable was a repeated measure. The results of this ANOVA are summarized in Table 10. As indicated in Table 10, an $F=6.11$ ($p<.05$) was obtained for the semester main effect, indicating that the mean total test informations for the winter semester were higher than the mean total test informations for the summer. Thus,the results of the ANOVA on the mean total test informations are consistent with the prediction based on the finding that the ability estimates were higher for the summer than for the winter.

Table 9

Mean Total Test Information for Bayesian and

Maximum Likelihood Tailored Tests for Winter and Summer

Semester	Session	Bayesian			Maximum Likelihood		
		20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
	1	40.89	30.83	26.62	38.20	27.89	24.64
Winter	2	41.33	31.61	27.61	36.98	27.60	23.98
	combined	41.11			37.59		
	1	38.00	29.35	26.13	33.95	25.84	22.56
Summer	2	37.49	29.09	25.67	33.29	24.79	21.62
	combined	37.74			33.62		

Table 10

Results of Three-Way ANOVA on 20 Item Mean Total Test

Informations Using Semester and Test as Independent

Variables with Repeated Measures over Sessions

Source	SS	df	MS	F	p
Test	494.71	1	494.71	6.63	0.012
Semester	455.64	1	455.64	6.11	0.016
Test x Semester	3.14	1	3.14	0.04	0.838
Error	4997.47	67	74.59		
Session	8.04	1	8.04	2.22	0.141
Session x Test	6.90	1	6.90	1.90	0.172
Session x Semester	0.33	1	0.33	0.09	0.764
Session x Test x Semester	4.77	1	4.77	1.32	0.255
Error	243.05	67	3.63		

Table 9 shows that the mean total test informations from the Bayesian tests were higher than the mean total test informations from the maximum likelihood tests at all test lengths. The ANOVA summarized in Table 10 indicates that this difference is significant, since an F=6.63 ($p<.05$) was obtained for the test main effect. This result may indicate that the mean Bayesian ability estimate was less than the mean maximum likelihood ability estimate since, as was previously pointed out, the information per item was lower for items farther away from the center of the item pool. In order to further compare the total test information yielded by the two procedures, the total test information for the two procedures were plotted on the same set of axes. These plots are shown in Figure 11.

FIGURE 11

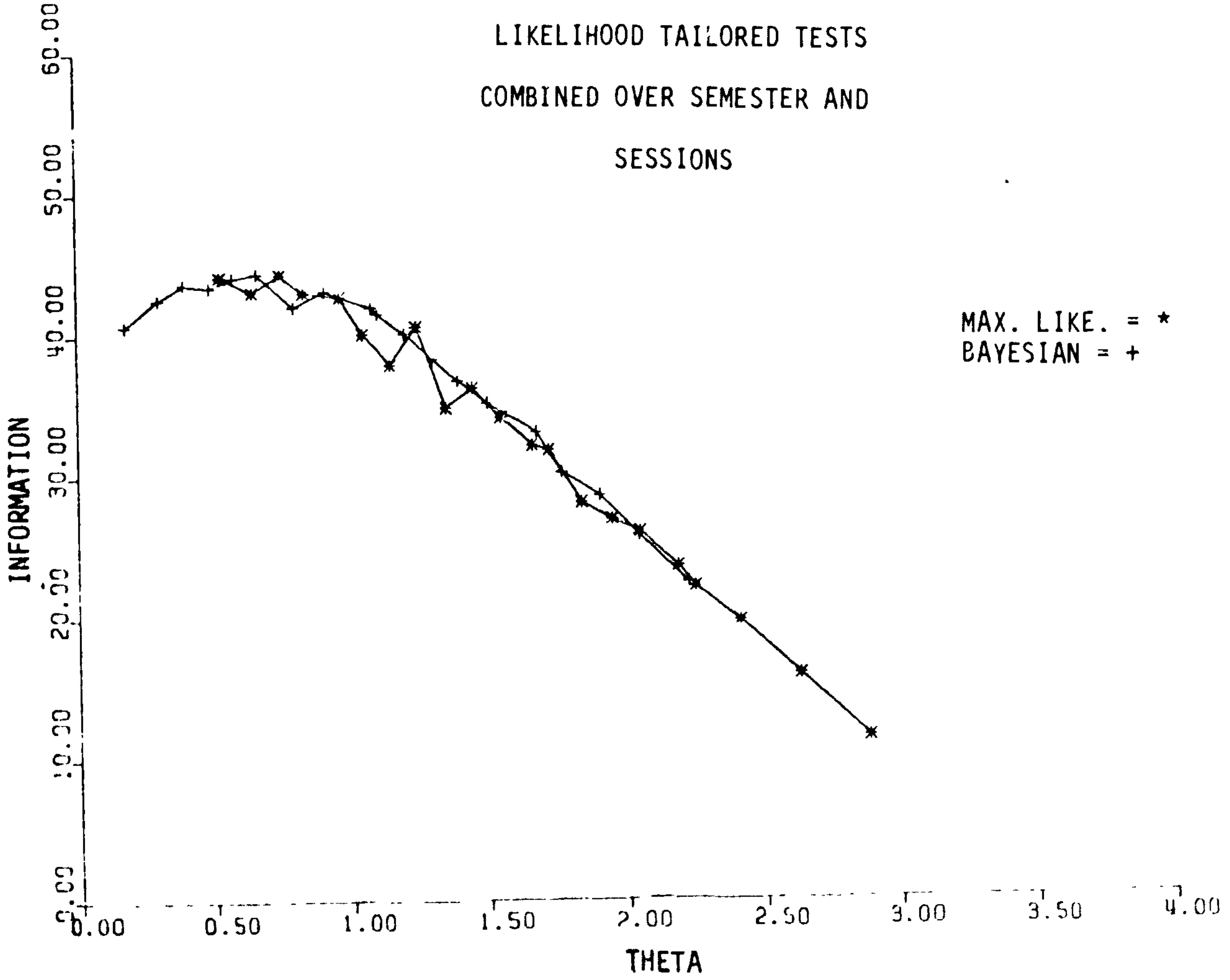
TOTAL TEST INFORMATION

FOR THE BAYESIAN AND MAXIMUM

LIKELIHOOD TAILORED TESTS

COMBINED OVER SEMESTER AND

SESSIONS



As can be seen in Figure 11, the test information functions for the Bayesian and maximum likelihood procedures were, for all practical purposes, the same. The Bayesian curve is shifted toward the lower end of the ability scale relative to the maximum likelihood curve, however, indicating that the Bayesian ability estimates fall in a slightly lower region of the scale. This result will be amplified in the next section.

Ability Estimates

A summary of the descriptive statistics for the ability estimate distributions for the Bayesian tailored tests at the 12, 14, and 20 item test lengths for both sessions of the winter semester study is shown in Table 11. Table 12 contains the same data for the summer study. The summary statistics for the maximum likelihood tailored test ability estimate distributions at the 12, 14, and 20 item test lengths for the winter and summer semesters are shown in Table 13 and 14, respectively. Plots of the 20 item ability estimate distributions were shown earlier in Figures 3 through 6. An ANOVA previously discussed in conjunction with these figures indicated that the summer semester ability estimates were significantly higher than the winter semester ability estimates. A comparison of the means presented in Tables 11 through 14 also indicates that the maximum likelihood ability estimates were higher than the Bayesian ability estimates. In order to confirm this fact, a four-way ANOVA was run on the ability estimates using semester, test, session and test length as independent variables, with repeated measures over sessions and test lengths. Recall that the plot of total test information indicated that there might be a significant difference in the two sets of ability estimates.

Table 11

Descriptive Statistics for the Bayesian

Ability Estimate Distributions for the Winter Semester

Statistic	Session 1			Session 2		
	20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
N	16	16	16	16	16	16
Mean	.746	.650	.643	.859	.800	.833
St. Dev.	.452	.455	.503	.419	.401	.394
Skewness	.969	.942	.616	1.021	1.084	1.004
Kurtosis	.416	.916	-.147	-.135	.293	.030
Minimum	.181	.003	-.141	.497	.391	.388
Maximum	1.818	1.764	1.723	1.775	1.740	1.693

Table 12

Descriptive Statistics for the Bayesian
Ability Estimate Distributions for the Summer Semester

Statistic	Session 1			Session 2		
	20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
N	13	13	13	13	13	13
Mean	1.183	1.115	1.070	1.178	1.178	1.156
St. Dev.	.554	.507	.477	.584	.584	.597
Skewness	.037	.230	.466	.425	.425	.109
Kurtosis	-.933	-.643	-.272	-.216	-.216	-.263
Minimum	.387	.316	.330	.191	.191	.071
Maximum	2.049	1.967	1.960	2.256	2.256	2.183

Table 13

Descriptive Statistics for the Maximum Likelihood
Ability Estimate Distributions for the Winter Semester

Statistic	Session 1			Session 2		
	20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
N	18	18	18	18	18	18
Mean	1.255	1.280	1.314	1.296	1.312	1.336
St. Dev.	.332	.427	.408	.378	.435	.470
Skewness	-.184	-.227	.037	-.029	.003	.243
Kurtosis	-.349	-.969	-1.190	-1.035	-1.191	-.940
Minimum	.649	.539	.699	.750	.626	.571
Maximum	1.913	1.849	2.011	1.954	2.052	2.112

Table 14

Descriptive Statistics for the Maximum Likelihood Ability Estimate Distributions for the Summer Semester

Statistic	Session 1			Session 2		
	20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
N	23	23	23	23	23	23
Mean	1.535	1.517	1.545	1.563	1.698	1.701
St. Dev.	.511	.554	.525	.580	.952	.789
Skewness	-.149	-.103	-.166	.283	3.054	2.610
Kurtosis	.343	.070	-4.260	-.075	12.050	9.398
Minimum	.457	.483	.665	.532	.576	.727
Maximum	2.634	2.592	2.537	2.887	5.500	4.700

The results of the four-way ANOVA on ability estimates are summarized in Table 15. As can be seen in Table 15, the test main effect was significant (F=15.43, p<.01), indicating that the maximum likelihood ability estimates were significantly higher than the Bayesian ability estimates. Thus, the hypothesis formulated on the basis of the information analyses was confirmed. The significance of the semester main effect (F=8.33, p<.01) is further evidence supporting the conclusion that the summer study ability estimates were significantly higher than the winter study ability estimates. This was true for both procedures, as indicated by the nonsignificance of the semester x test interaction.

The significance of the session main effect (F=7.50, p<.01) indicates that the second session ability estimates were significantly higher than the first session ability estimates. The lack of significance of the interaction of session with test indicates that the second session ability estimates were significantly higher than the first session ability estimates for both procedures.

Table 15

Results of Four-Way ANOVA on Ability Estimates

Using Semester, Test, Session, and Test Length as Independent Variables with Repeated Measures over Sessions and Test Length

Source	SS	df	MS	F	p
Semester	12.69	1	12.69	8.33	.005
Test	23.49	1	23.49	15.43	.000
Semester x Test	.27	1	.27	.18	.675
Error	100.50	66	1.52		
Session	1.01	1	1.01	7.50	.008
Session x Semester	.00	1	.00	.00	.992
Session x Test	.02	1	.02	.15	.699
Session x Semester x Test	.13	1	.13	.97	.328
Error	8.92	66	.14		
Length	.04	2	.02	.61	.546
Length x Semester	.01	2	.00	.11	.895
Length x Test	.37	2	.19	6.39	.002
Length x Semester x Test	.04	2	.02	.72	.488
Error	3.83	132	.03		
Session x Length	.06	2	.03	1.63	.199
Session x Length x Semester	.01	2	.01	.32	.728
Session x Length x Test	.02	2	.01	.46	.630
Session x Length x Semester x Test	.06	2	.03	1.43	.242
Error	2.54	132	.02		

The test length main effect was not significant. However, the interaction of test type with test length was significant ($F=6.39$, $p<.01$). In order to explore this effect Fisher's LSD test was applied to the mean ability estimates at the different test lengths. When the LSD test was applied to the maximum likelihood mean ability estimates no significant differences were found. For the winter semester Bayesian mean ability estimates, a value of LSD=.073 at $\alpha=.05$ was obtained. Comparisons of this value with the differences in means indicated that the 12 item and 20 item mean ability estimates were significantly different, while the other pairings, 12 item with 14 item and 14 item with 20 item, were not significantly different.

The results of these LSD tests are consistent with previously reported results. The mean ability of the winter group was closer to the mean of the assumed prior distribution than was the mean ability of the summer group. As a result, the effect of the low prior may have been overcome by the 12 item level for the winter semester. The 12 item and 14 item mean ability estimates were not significantly different, nor were the 12 item and 20 item mean ability estimates. This finding is an anomaly for which no explanation could be found. Reckase (1974) found that continuation of a tailored test beyond the optimal test length introduces bias into the ability estimates. From the convergence plots it appears that this was the case here. For the summer semester the significance of the difference between the 12 item and 20 item mean ability estimates perhaps indicates that the effect of the prior distribution was not overcome by 12 items, but rather that the ability estimates continued to increase beyond the twelfth item. The lack of significance of the difference between the 12 item and 14 item mean ability estimates may just be an indication that the increase in ability estimates was too gradual for two items to make a significant difference. This explanation is, of course, only one possibility. Other reasonable explanations may be found.

The final set of analyses run on the ability estimates involved an investigation of the interaction of the ability estimation procedures and item selection procedures. Because the two tailored testing techniques utilized different item selection procedures, any differences in the ability estimates obtained from the techniques could have been due to differences in the ability estimation procedures, differences in the item selection procedures, or both. In order to determine the source of the differences in ability estimates, maximum likelihood ability estimates were obtained using the items selected by the Bayesian tailored testing procedure, and the Bayesian ability estimates were obtained using the items selected by the maximum likelihood tailored testing procedure. These new ability estimates were analyzed with a three-way ANOVA using ability estimation procedure, item selection procedure, and session as independent variables, with repeated measures over sessions. The results of this analysis are reported in Table 16 and the mean ability estimates obtained are reported in Table 17.

Table 16

Three-Way ANOVA on Recalculated Ability Estimates Using Item Selection Procedure, Estimation Procedure, and Session as Independent Variables, with Repeated Measures over Sessions

Source	SS	df	MS	F	p
Selection	10.20	1	10.20	11.85	.001
Error	59.40	69	.86		
Session	.80	1	.80	7.24	.009
Session x Selection	.01	1	.01	.10	.758
Error	7.57	69	.11		
Estimation	.17	1	.17	29.10	.000
Estimation x Selection	.03	1	.03	4.51	.037
Error	.41	69	.01		
Session x Estimation	.01	1	.01	5.46	.022
Session x Estimation x Selection	.01	1	.01	5.08	.027
Error	.13	69	.00		

Table 17

Means and Standard Deviations Associated with the Three-Way ANOVA on Recalculated Ability Estimates

Item Selection Procedure	Statistic	Ability Estimation Procedure			
		Bayesian		Maximum Likelihood	
		Session 1	Session 2	Session 1	Session 2
Bayesian	Mean	.954	1.048	.984	1.079
	Std. Dev.	.533	.516	.579	.563
Max. Like.	Mean	1.317	1.413	1.363	1.507
	Std. Dev.	.409	.433	.456	.504

As can be seen in these tables, regardless of which procedure selected the items, the maximum likelihood ability estimates were significantly greater than the Bayesian ability estimates ($\underline{F}$=29.10, $\underline{p}$<.01). It is clear from these results that at least part of the differences found between the Bayesian and maximum likelihood tailored test ability estimate distributions was due to differences in the ability estimation procedures. However, it should also be noted from Table 16 and 17 that for both procedures the ability estimates were higher when based on the maximum likelihood items than when based on the Bayesian items ($\underline{F}$=11.85, $\underline{p}$<.01). It is also clear, then, that part of the differences found between the ability estimates obtained from the two tailored test procedures was due to the difference in items selected for administration. One possible explanation for these differences is that the assumption of a prior distribution of ability made by the Bayesian procedure imposed a restriction on the range of ability estimates obtained from that procedure, which in turn would restrict the range of the b-values of the items selected. The restriction of the range of b-values would have further limited the range of ability estimates. Thus, there may have been an interaction of item selection and ability estimation procedures that, due to an inappriately low prior, limited the magnitude of the resulting ability estimates. This is supported by the finding that the estimation procedure x item selection procedure interaction was significant ($\underline{F}$=4.51, p .05).

The session main effect reported in Table 16 was also significant ($\underline{F}$=7.24, $\underline{p}$<.01), as was the session x estimation procedure interaction ($\underline{F}$=5.46, $\underline{p}$<.05). The significante of the session main effect was consistent with previously reported findings. The significance of the session x estimation procedure interaction was probably due to the restriction in the range of the Bayesian ability estimates. The three-way interactton among session, estimation procedure, and item selection procedure ($\underline{F}$=5.08, $\underline{p}$<.05) is difficult to interpret.

In order to further investigate the effect of the prior distribution on the obtained Bayesian ability estimates an additional analysis was performed. This analysis involved obtaining Bayesian ability estimates using both sets of tailored test items, but using a prior with a mean of 2.0, as opposed to the mean of -.100 or .150 originally employed. These new ability estimates were also analyzed using a three-way ANOVA, but this time with prior distribution in place of ability estimation procedure as the third independent variable. The results of this analysis are reported in Table 18 and the obtained means are reported in Table 19. As can be seen in Tables 18 and 19 the new Bayesian ability estimates were significantly higher using the high prior than when using the low prior for both item selection procedures. The prior distribution main effect had an _F_=91.84, _p_<.01, while the prior x selection interaction was not significant. As can be seen from the means reported in Table 19, use of the high prior increased the Bayesian ability estimates using the Bayesian tailored test items, but not to the level of the Bayesian ability estimates using the maximum likelihood tailored test items. This is supported by the significance of the selection main effect (_F_=10.04, _p_<.01). It should be remembered that for the Bayesian tests using the high prior the items were still those selected when the low prior was being used. As a result, all of the items were too easy for the ability estimates obtained using a high prior. Thus, when an item was correctly answered the ability estimate would have increased only minimally. When an item was answered incorrectly, on the other hand, the low _b_-values would have resulted in a large decrease in the ability estimate. That is, the _b_-values pulled the ability estimates down close to that level for which the items had been selected. The effect of the high prior, then, was to increase the ability estimates only a small, though statistically significant, amount.

Table 18

Three-Way ANOVA on Recalculated Bayesian Ability Estimates Using Item Selection Procedure, Prior Distribution, and Session as Independent Variables, with Repeated Measures over Sessions

Source	_SS_	_df_	_MS_	_F_	_p_
Selection	9.34	1	9.34	10.04	.002
Error	64.22	69	.93		
Session	.15	1	.15	3.78	.056
Session x Selection	.13	1	.13	3.18	.079
Error	2.71	69	.04		
Prior	.38	1	.35	91.84	.000
Prior x Selection	.00	1	.00	.43	.513
Error	.23	69	.00		
Session x Prior	.00	1	.00	.49	.484
Session x Prior x Selection	.00	1	.00	.03	.858
Error	.48	69	.01		

Table 19

Means and Standard Deviations Associated with the Three-Way ANOVA on Recalculated Bayesian Ability Estimates

Item Selection Procedure	Statistic	Prior Distribution			
		Low		High	
		Session 1	Session 2	Session 1	Session 2
Bayesian	Mean	.954	1.037	1.017	1.111
	Std. Dev.	.533	.521	.546	.582
Max. Like.	Mean	1.360	1.355	1.430	1.443
	Std. Dev.	.408	.469	.454	.485

This result serves to point out the serious effect of an inappropriate prior. Selecting a prior and selecting items on the basis of that prior forces item b-values to remain in the region of the prior mean. Because only items with b-values in that region are administered, subsequent ability estimates are also forced to remain in the region of the prior mean. The results of the ANOVA and LSD tests on the ability estimates indicate than an inappropriate prior may eventually be overcome, but it may significantly increase the length of the tailored test that is required. Moreover, the different results of the LSD test for the two semesters point out that the appropriateness of the prior must be determined for every distinct group of examinees.

It may be true that the Bayesian ability estimates obtained using the high prior were still smaller than the maximum likelihood ability estimates because the estimates were obtained using items with inappropriately low b-values. If so, the fact that the Bayesian ability estimates obtained using the maximum likelihood items were higher than when the Bayesian items were used would indicate that the maximum likelihood procedure administered items with higher b-values than did the Bayesian procedure. A comparison of the mean b-value for the items administered by the Bayesian procedure (.678) with the maximum likelihood mean b-value (.903) yielded a $t=11.45$, $p<.001$. Clearly, then, the maximum likelihood procedure administered items with greater b-values than did the Bayesian procedure. This result supports the hypotheses set out above.

Items Administered

As was discussed in the previous section, the maximum likelihood procedure tended to administer items with higher b-values than did the Bayesian procedure. Further comparisons indicated that the mean a-value of the items administered by the maximum likelihood procedure (1.766) was significantly greater than the Bayesian mean a-value (1.749), yielding a $t=2.183$, $p<.05$. The mean c-values were not significantly different.

It was also found that there were operational differences in the two item selection procedures. It was found that the b-value of an item selected by the Bayesian tailored test procedure was more closely related to the current ability estimate than was the b-value of an item selected by the maximum likelihood procedure. The correlation of the item b-values with the ability estimates used to select the items was $r=.77$ for the Bayesian procedure, $r=.61$ for the maximum likelihood procedure. A comparison of these correlations using Fisher's r to z transformation resulted in a $z=3.31$, $p<.01$. There were no significant differences in the correlations of the a-values and c-values with the ability estimates.

In order to determine whether the difference in entry point into the item pool between sessions affected one procedure more than another, an analysis was performed to discover whether one procedure had more items in common between sessions than the other procedure. This analysis was also used to compare the similarity of items over sessions of the two procedures.

For the Bayesian tailored tests the proportion of items administered in the first session that were repeated in the second session was $p=.827$. For the maximum likelihood procedure the proportion of repeated items was $p=.848$. A comparison of these two proportions to determine whether the difference was significant yielded a $z=1.07$, which was not significant. It is seen from this result that the two procedures were equally consistent in the items that were administered across sessions. Also, both procedures tended to use only about a third of the items in the item pool.

Goodness of Fit

From the analyses reported previously the conclusion was reached that the Bayesian procedure was producing ability estimates that were perhaps too low. If that were true then the probability of a correct response to a given item computed from the 3PL model using those ability estimates would also be too low. This would be reflected in poorer fit of the model to the data when using the Bayesian procedure, which should have been detected by the comparison of the MSD statistic obtained for the two procedures. This was the case. The MSD statistic obtained for the maximum likelihood procedure was MSD=.244, and the value obtained for the Bayesian procedure was MSD=.266. A comparison of these two values yielded a $t=5.64$, $p<.01$, indicating that the Bayesian procedure yielded significantly poorer fit than the maximum likelihood procedure.

Descriptive Statistics

The first descriptive statistic compiled for the procedures was the average test difficulty measured as the proportion of items answered correctly. To analyze these proportion correct values a three-way ANOVA was run using semester, test, and session as independent variables, with repeated measures over sessions. In order to meet the assumption of normality made by the ANOVA, the proportion correct values were first transformed using the arc sine transformation. The results of this ANOVA are summarized in Table 20. As can be seen in the table, both the semester and test main effects were significant ($F=6.55$, $p<.05$ for semester; $F=5.53$, $p<.05$ for test). The means and standard deviations for this

ANOVA are reported in Table 21. An examination of Table 21 shows that the summer tests were easier for the examinees than were the winter tests. The nonsignificance of the semester x test interaction indicates that this was true for both procedures. The session main effect was not significant, indicating that neither session was easier than the other for either procedure or semester. However, these tables do indicate that the maximum likelihood tests were easier for the examinees than were the Bayesian tests. This finding appears to be contrary to the expected result. If the Bayesian procedure were administering items with inappropriately low b-values, as was previously hypothesized, the Bayesian tests would have been easier for the examinees than were the maximum likelihood tests. The finding that the maximum likelihood tests were easier indicates that at least some part of the difference in ability estimates obtained from the two procedures was due to actual differences in vocabulary ability. However, the ability estimate analysis investigating the effect of the prior on subsequent ability estimation and item selection clearly demonstrates that not all of the differences in ability estimates were due to differences in group ability.

An analysis of the test difficulty separately for each semester is revealing. It was hypothesized previously that the Bayesian procedure actually overcame the inappropriate prior for the winter semester examinees, and that the ability estimates leveled off somewhat at a level substantially below the level of the maximum likelihood ability estimates. This indicates an actual difference in ability. This is supported by a comparison of the mean test difficulties for the two procedures for the winter semester, which yields a $t=2.47$, $p<.05$. For the summer semester it was suggested that the inappropriate prior may not have been overcome. The summer Bayesian ability estimates were significantly lower than the maximum likelihood ability estimates, but were increasing with increased test length. Had the Bayesian tailored tests been sufficiently long to overcome the effect of the prior it is possible that the Bayesian abllity estimates would have approached the level of the maximum likelihood ability estimates. It is likely, then, that there was considerably less difference between the group vocabulary abilities for the summer semester than for the winter semester. This is supported by the finding that the difference in mean proportion correct values for the two procedures was not significant for the summer semester. Had the examinees taking the Bayesian tests been of the same ability as the maximum likelihood examinees, the Bayesian tailored tests would have been easier than the maximum likelihood tests.

Table 20

Three-Way ANOVA on Test Difficulties Using Semester, Test, and Session as Independent Variables, with Repeated Measures over Sessions

Source	SS	df	MS	F	p
Semester	440.14	1	440.14	7.10	.001
Test	300.08	1	300.08	4.84	.031
Semester x Test	35.72	1	35.72	.58	.450
Error	4089.04	66	61.96		
Session	3.80	1	3.80	.66	.419
Session x Semester	3.57	1	3.57	.62	.433
Session x Test	2.21	1	2.21	.39	.537
Session x Semester x Test	5.31	1	5.31	.93	.339
Error	378.29	66	5.73		

Table 21

Means and Standard Deviations of Test Difficulties for Both Sessions of the Bayesian and Maximum Likelihood Tailored Tests for Both Semesters

Semester	Statistic	Bayesian		Maximum Likelihood	
		Session 1	Session 2	Session 1	Session 2
Winter	Mean	.69(56.24)	.70(56.76)	.75(60.93)	.76(60.93)
	Std. Dev.	.09(5.59)	.07(4.67)	.05(3.60)	.07(4.35)
Summer	Mean	.75(60.83)	.77(61.49)	.79(63.44)	.78(62.79)
	Std. Dev.	.10(7.46)	.09(6.55)	.08(6.11)	.10(7.15)

Note: Values in parentheses represent the results of the arc sine transformations.

Another statistic compiled for the two procedures was the mean testing time for the 20 item test, measured in seconds. The means and standard deviations for both sessions of the two procedures for both semesters are shown in Table 22. Table 23 summarizes the results of a three-way ANOVA on the testing times using semester, test, and session as independent variables, with repeated measures over sessions. As can be seen in these tables, the tailored tests in the summer study took significantly longer than did the winter semester tailored tests (F=11.71, p<.01). This was true for both procedures, since the semester x test interaction was not significant. The test main effect was not significant, indicating that there were no significant differences in the amount of time the two types of tailored tests lasted. The session main effect was significant (F=5.96, p<.05), with the second session tests ending more quickly than the first session tests in all cases except the summer maximum likelihood condition. The session x semester interaction was significant (F=5.97, p<.05), with the difference between the two sessions being larger for the winter semester.

Table 22

Means and Standard Deviations of Testing Time in Seconds for Both Sessions of the Bayesian and Maximum Likelihood Tailored Tests for Both Semesters

Semester	Statistic	Bayesian		Maximum Likelihood	
		Session 1	Session 2	Session 1	Session 2
Winter	Mean	607.88	498.12	566.72	480.78
	Std. Dev.	176.02	102.69	94.02	106.13
Summer	Mean	715.92	707.62	601.39	609.78
	Std. Dev.	256.59	222.57	154.37	181.94

Table 23

Three-Way ANOVA on Testing Time Using Semester, Test, and Session as Independent Variables. with Repeated Measures over Sessions

Source	SS	df	MS	F	p
Semester	483310.04	1	483310.04	11.71	.001
Test	152578.77	1	152578.77	3.70	.059
Semester x Test	50360.16	1	50360.16	1.22	.273
Error	2723098.41	66	41259.07		
Session	79322.81	1	79322.81	5.96	.017
Session x Semester	79459.27	1	79459.27	5.97	.017
Session x Test	3631.40	1	3631.40	.27	.603
Session x Semester x Test	141.60	1	141.60	.01	.918
Error	878721.87	66	13313.97		

Latency

The correlations obtained between the mean item response latencies for a person and their ability estimates for both sessions of both semesters are shown in Table 24. As can be seen in the table, none of the correlations were significant for the winter semester. For the summer semester the correlations were significant for the first session Bayesian tests (r=-.57, p<.05)and for both sessions of the maximum likelihood tests (r=.51, p<.05 for the first session; r=-.47, p<.05 for the second session).

Table 24

Correlations of Ability Estimates and Mean Latencies for Both Sessions of the Bayesian and Maximum Likelihood Tailored Tests for the Winter and Summer Semesters

Semester	Bayesian		Maximum Likelihood	
	Session 1	Session 2	Session 1	Session 2
Winter	-.17	-.02	.06	-.18
Summer	-.57*	-.04	-.51*	-.47*

* p<.05.

The final analysis performed was the comparison of mean latencies for correct and incorrect responses. The results of a four-way ANOVA on the mean latencies are summarized in Table 25. The means and standard deviations for this analysis appear in Table 26. For this analysis the independent variables were semester, test, session, and response (correct or incorrect). Session and response were repeated measures. As can be seen in Table 25 the session and response main effects were significant. The semester and test main effects were not significant, nor were any of the interactions. From Table 26 it can be seen that the first session response latencies were greater than the second session response latencies. Also, response latencies for the incorrect responses were greater than the latencies for correct responses. No differences were found between the two test procedures.

Table 25

Results of Four-Way ANOVA on Mean Response Latencies Using Semester, Test, Session, and Response as Independent Variables with Repeated Measures over Session and Response

Source	SS	df	MS	F	p
Semester	527.87	1	527.87	3.35	.072
Test	315.29	1	315.29	2.00	.162
Semester x Test	75.61	1	75.61	.46	.474
Error	10570.06	67	157.76		
Session	2000.51	1	2000.51	52.42	.000
Session x Semester	4.42	1	4.42	.12	.735
Session x Test	42.52	1	42.52	1.11	.295
Session x Semester x Test	17.61	1	17.61	.46	.499
Error	2556.81	67	38.16		
Response	2794.22	1	2794.22	60.72	.000
Response x Semester	45.40	1	45.40	.99	.324
Response x Test	12.99	1	12.99	.28	.597
Response x Semester x Test	3.26	1	3.26	.07	.791
Error	3083.28	67	46.02		
Session x Response	93.99	1	93.99	3.54	.064
Session x Response x Semester	4.34	1	4.34	.16	.687
Session x Response x Test	3.83	1	3.83	.14	.705
Session x Response x Semester x Test	3.15	1	3.15	.12	.732
Error	1777.85	67	26.54		

Table 26

Means and Standard Deviations of Response Latencies for Correct and Incorrect Responses for Both Sessions of the Bayesian and Maximum Likelihood Tailored Tests for Both Semesters

Semester	Response	Statistic	Bayesian		Maximum Likelihood	
			Session 1	Session 2	Session 1	Session 2
Winter	Correct	Mean	13.86	8.86	11.80	8.27
		Std. Dev.	7.32	4.65	4.15	3.71
	Incorrect	Mean	20.23	13.27	19.51	12.22
		Std. Dev.	11.79	6.10	11.02	5.75
Summer	Correct	Mean	17.40	11.82	12.20	9.27
		Std. Dev.	6.36	4.32	4.99	5.92
	Incorrect	Mean	24.89	17.50	21.05	16.23
		Std. Dev.	13.45	8.28	12.72	9.91

Discussion

In order to put this study in the proper perspective, it is necessary to view it as one of a series of studies designed to evaluate alternative components for tailored testing. The series began with several studies designed to determine which of the available latent trait models was optimal (Koch and Reckase, 1978, 1979; McKinley and Reckase, 1980a; Reckase, 1977). Once a model was selected (the 3PL model), a study was done to identify the optimal item calibration procedure to be used with the model (McKinley and Reckase, 1980b). After an item calibration procedure was selected for the model (LOGIST), a set of studies was begun to determine what the optimal operating characteristics of the tailored testing procedure should be. These characteristics included item selection and ability estimation procedures, which are the topics of the current study.

Optimal Cutoffs

The convergence plot analyses performed indicated that the optimal test length for the Bayesian procedure was 14 items. This result was consistent across sessions and semesters. For the maximum likelihood procedure the optimal test length was 12 items. If cutoffs are expressed in terms of item information and standard error of estimate, the optimal cutoff value for the Bayesian procedure was a standard error of estimate of .25, and for the maximum likelihood procedure the optimal cutoff value was an item information of 1.64. A comparison of the optimal test lengths for the two procedures indicates that the Bayesian procedure requires more items to obtain stable ability estimates. This conclusion is supported by the finding that the test length main effect was not significant for the maximum likelihood ability estimates but was significant for the Bayesian ability estimates. This finding was also consistent across semesters.

Reliabilities

In terms of reliability, no significant differences were found between the two procedures. Moreover, for neither procedure was there any significant differences in reliability across the different test lengths. The results were the same when reliabilities were computed using estimated true scores. It should be pointed out here that the relative instability of the Bayesian ability estimates did not lower the reliability of the Bayesian tailored tests at the shorter test lengths. As was stated earlier, these reliabilities were obtained with relatively small sample sizes, so large differences were necessary for significance.

Total Test Information

At the 20 item level the mean total test information yielded by the Bayesian procedure was significantly greater than the mean total test information yielded by the maximum likelihood procedure. It is apparent from these findings that the Bayesian procedure was yielding ability estimates in a range where more items with high information at those ability estimates were available. Since the mean of the assumed prior distribution of ability was in that region of the ability scale for which the item pool would yield the greatest information, this result indicates that the ability estimates yielded by the Bayesian procedure tended to be relatively close to the mean of the prior distribution. In that region of the ability scale where both procedures yielded ability estimates the Bayesian procedure did not appear to yield more total test information. It should be pointed out that the prior of the Bayesian procedure held ability estimates in that range where there was high information because the mean of the prior was selected as an ability near the mode of the total information curve of the pool. Had the prior been set higher the mean total test information for the Bayesian procedure would have decreased. Thus, the high information of the Bayesian procedure was due to the selection of the prior and the structure of the item pool.

Ability Estimates

The four-way ANOVA on the ability estimates confirmed the hypothesis that the maximum likelihood ability estimates were significantly greater than the Bayesian ability estimates. On the basis of this finding the hypothesis was formulated that the Bayesian ability estimates were smaller because the mean of the assumed prior distribution was too low. That is, the effect of the prior distribution was to lower the ability estimates obtained from the Bayesian procedure. If these hypotheses were true, then the Bayesian ability estimates should have increased as test length increased, since additional items would give the procedure opportunity to overcome the effect of the inappropriate prior. The increase should have continued until the prior was overcame, and then the ability estimates should have begun to stabilize. Evidence supporting this prediction was obtained from the test length analyses. The maximum likelihood ability estimates did not change significantly after 12 items, while for the Bayesian ability estimates the test length effect was significant. For the summer semester the Bayesian mean ability estimates continued to increase across the different test lengths. This is an indication that the mean ability of the summer examinees was sufficiently higher than the mean of the assumed prior distribution of ability that the retarding effect of the prior was never completely overcome, even after 20 items. The mean ability of the winter semester examinees was lower than the mean ability of the summer group, and as a result the prior was more appropriate, yielding stable estimates by the twelfth item. From these results it appears that use of an inappropriate prior distribution of ability may have a serious effect on the ability estimates obtained from the Bayesian procedure, thus affecting the length of test required to obtain accurate estimates.

The investigation into the interaction of the ability estimation procedures and the item selection procedures yielded further evidence as to the restricting effect of the assumed prior distribution. The Bayesian ability estimation procedure consistently yielded ability estimates that were lower (closer to the mean of the assumed prior distribution) than the maximum likelihood ability estimates, even when ability estimates were obtained from the two procedures using the same set of items. When ability estimates were obtained from the Bayesian procedure on the maximum likelihood items using a higher prior mean, the ability estimates increased to the same level as the maximum likelihood ability estimates. However, using a higher prior did not significantly increase the Bayesian ability estimates on the Bayesian items. Because the maximum likelihood items had significantly higher b-values, it was hypothesized that raising the prior would affect ability estimates only if items were selected on the basis of the new prior (i.e., with higher b-values). These results indicate that, had a higher prior distribution of ability been assumed for the summer

Bayesian tailored tests, the procedure would have administered items witn greater b-values and would have yielded ability estimates close to the magnitude of the maximum likelihood ability estimates. For the winter semester, due to the actual differences in ability between the two groups, a higher prior probably would have significantly increased the Bayesian ability estimates, but not to the level of the maximum likelihood ability estimates.

The results of the ability estimate analyses lead to two general conclusions. The first conclusion is that use of an inappropriate prior distribution of ability in the estimation of ability may significantly increase the test length required to obtain accurate ability estimates. The greater the degree of inappropriateness of the assumed prior, the longer the tailored test will have to be to obtain good ability estimates. The second conclusion is that the commonly assumed prior distribution of ability will not be appropriate for a heterogeneous group. The same prior was used for the winter and summer examinees, two groups clearly different in ability. For the winter semester the effect of the inappropriate prior was **not** as pronounced as it was for the summer session.

These conclusions have special significance for criterion referenced type testing, where some absolute level of performance is sought. An inappropriate prior could prevent an examinee's ability estimate from reaching the criterion, or could artificially elevate the ability estimates to a level above the criterion. Making valid decisions in such situations would be quite difficult.

Items Administered

Analysis of the items administered by the two procedures indicated that the maximum likelihood procedure administered items with higher b-values than the Bayesian procedure. There appeared to be two reasons for this. First, the Bayesian ability estimates were lower than the maximum likelihood ability estimates, and therefore the administration of items appropriate for the current ability estimate resulted in the selection of easier items. Second, the item selection procedure used by the Bayesian tailored test procedure selected items with b-values more highly correlated with the ability estimates than did the item selection procedure employed by the maximum likelihood tailored test procedure. The result of this was to strengthen the effect of the lower ability estimates yielded by the Bayesian procedure. The effect of the inappropriate prior might have been less had the Bayesian procedure selected items on the basis of information.

Further comparisons indicated that the maximum likelihood procedure administered items with higher a-values than the Bayesian procedure. This was probably due to the fact that selection using the information function more heavily weighted tne a-values in the selection of items than did the Bayesian procedure. No differences were found in the c-values of the items administered by the two procedures.

Goodness of Fit

The results of the ability estimate analyses indicated that the Bayesian procedure may have underestimated ability. As a result, an examinee would have had a higher probability of correctly responding to items than would have been predicted by the model on the basis of their Bayesian ability estimate. This would result in poorer fit of tne 3PL model to the data when using Bayesian ability estimates than when using the maximum likelihood ability estimates. This was found to be the case. The MSD value obtained for the Bayesian procedure was significantly greater than the MSD value obtained for the maximum likelihood procedure.

Descriptive Statistics

The results of the analyses of the mean proportion correct for each test at first appeared inconsistent with the results of other analyses. The maximum likelihood tailored tests were found to be significantly less difficult than the Bayesian tailored tests. Since the maximum likelihood procedure administered items with greater b-values, it was expected that the maximum likelihood tests would be found to be more difficult than the Bayesian tests. Further analyses apparently resolved this conflict. Since the winter Bayesian ability estimates stabilized at a lower level than the maximum likelihood ability estimates, it was hypothesized that the examinees taking the Bayesian tests were of lower vocabulary ability than the examinees taking the maximum likelihood tests. This would explain why the examinees taking the Bayesian tests received easier items than the examinees taking the maximum likelihood tests but missed more items. For the summer semester the Bayesian ability estimates did not stabilize, indicating that the examinees taking the Bayesian tests were of vocabulary ability closer to the ability of the examinees taking the maximum likelihood tests than was the case with the winter group. In support of this interpretation was the finding that for the summer semester the mean proportion correct for the two procedures were not significantly different.

The results of these analyses leave unclear the degree to which the difference in ability estimates obtained from the two procedures was due to actual differences in vocabulary ability and how much of the difference was due to differences in the ability estimation and item selection procedures. However, on the basis of the recalculated ability estimates discussed in conjunction with the ability estimate analyses, it would appear that a substantial part of the difference was due to the interaction of the ability estimation procedure, including the prior, with the item selection procedure.

The other statistic compiled for the two procedures was mean testing time. An ANOVA on mean testing times indicated that there was not a significant difference in the amount of time the two types of tailored tests required when the number of items administered was the same.

Latency

The correlations obtained between mean latencies and ability estimates followed no meaningful pattern. For the winter semester the mean latencies were not significantly correlated with the ability estimates. That is, the magnitude of the ability estimates apparently had no bearing on the mean amount of time to respond. For the summer semester the correlations were significant, but there were no differences indicated between the two procedures. The significant correlations for the summer indicated that the brighter students took less time to respond to the items. No hypothesis could be produced to explain why the summer and winter groups behaved differently.

The final analysis performed was the comparison of mean latencies for correct and incorrect responses. The ANOVA on the response latencies indicated that response time was greater for incorrect responses tnan for correct responses, but no differences between the two procedures were indicated.

Nonconvergence

Nonconvergence was not actually a research question in this study, but whenever maximum likelihood estimation procedures are employed it is an important issue. Earlier studies using maximum likelihood estimation in tailored test procedures (Koch and Reckase, 1978, 1979; McKinley and Reckase, 1980a), found nonconvergence to be a serious problem for the 3PL model. The incidence of nonconvergence was reduced by properly selecting entry points into the item pool and more accurately linking the item calibrations used in the tailored testing procedure, but nonconvergence was not completely eliminated. An important observation concerning the current study is that there were no cases of nonconvergence.

Summary and Conclusions

Previous studies investigating alternatives for the various components of tailored testing indicated that 3PL model was preferred to the 1PL model. It was also found that the LOGIST calibration program was better than the ANCILLES procedure for calibrating the item pool. Once these components had been selected several studies were undertaken to determine the optimal operational characteristics of a tailored testing procedure using these components. The present study was designed to compare alternative ability estimation and item selection procedures.

This study involved a live tailored testing comparison of a tailored testing procedure based on a Bayesian ability estimation procedure and a tailored testing procedure based on maximum likelihood ability estimation. The Bayesian tailored testing procedure selected items so as to minimize the posterior variance of the ability estimate distribution, while the maximum likelihood tailored testing procedure selected items so as to maximize the item information for the current ability estimate. Attempts were made to first determine the optimal test length for the two procedures, and then to compare the procedures at those test lengths, as well as at a 20 item test length.

Analyses indicated that the optimal test length of the maximum likelihood procedure was about 12 items, while the optimal length of the Bayesian test was 14 items. Comparisons of the two procedures at these test lengths and at the 20 item length yielded the following results. There was no difference at any test length between the two procedures in terms of reliability. The Bayesian procedure did yield greater mean total test information than did the maximum likelihood procedure. However, it was found that the higher information of the Bayesian procedure was due to the regression of the ability estimates to the mean of the assumed prior distribution of ability. In the range of ability where there were ability estimates for both procedures the difference in total test information was negligible. Further analyses showed that the assumption of different priors can significantly alter the ability estimates obtained from a Bayesian tailored test, as well as the total test information yielded by the tailored test. It was found that the more inappropriate the prior the longer the Bayesian tailored test had to be to obtain accurate ability estimates. Thus, the winter semester Bayesian tests yielded stable ability estimates by the twelfth item, on the average, while the summer semester Bayesian ability estimates generally did not converge to a stable value. This was consistent with the finding that the subjects in the summer semester were of higher vocabulary ability than were the winter semester subjects. Analyses of the item selection procedures indicated that selection of items to minimize the posterior variance of the ability estimates magnified the effect of the inappropriate prior. The goodness of fit comparison indicated that the Bayesian procedure yielded significantly poorer fit of the 3PL model to the data than did the maximum likelihood procedure, which was consistent with the finding that the Bayesian ability estimates were too low.

Based on the results reported above it was concluded that selection of an inappropriate prior significantly increased the test length required for accurate estimation using a Bayesian tailored test. At any length less than the optimal test length, Bayesian ability estimates are biased in the direction of the mean of the prior distribution. If testing continues beyond the optimal test length, bias is again introduced into the ability estimates if inappropriate items are administered. Because the optimal test length varies depending on the appropriateness of the prior, in order to avoid bias in the ability estimates it is essential to determine an appropriate prior. Also, it is clear from this study that the N(0,1) prior can be appropriate for only a relatively homogeneous group. For large heterogeneous groups determination of an appropriate prior is much more difficult, and bias in the ability estimates can often result. Therefore,the Bayesian tailored testing procedure seems appropriate only when good prior information can be obtained. For large scale tailored testing a maximum likelihood tailored testing procedure with item selection based on information is the procedure of choice.

REFERENCES

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

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation, University of Washington, 1972.

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

Koch, W. R. and Reckase, M. D. A live tailored testing comparison study of the one- and three-parameter logistic models (Research Report 78-1). Columbia: University of Missouri, Department of Educational Psychology, June 1978.

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

Lord, F. M. A broad range test of verbal ability. Proceedings of the First Conference on Computerized Adaptive Testing, June, 1975, United States Civil Service Commission, 1975.

Lord, F. M. Personal communication, June, 1979.

Maurelli, V. A. A comparison of Bayesian and maximum likelihood scoring in a cumulated stradaptive test. Unpublished Master's Thesis, St. Mary's University, 1978.

McKinley, R. L. and Reckase, M. D. A successful application of latent trait theory to tailored achievement testing (Research Report 80-1). Columbia: University of Missouri, Department of Educational Psychology, February 1980. (a)

McKinley, R. L. and Reckase, M. D. A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness of fit as a criterion (Research Report 80-2). Columbia: University of Missouri, Department of Educational Psychology, December 1980. (b)

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.

Patience, W. M. and Reckase, M. D. Effects of program parameters and item pool characteristics on the bias of a three-parameter tailored testing procedure. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April, 1980.

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, 1974, 6, 208-212.

Reckase, M. D. Ability estimation and item calibration using the one- and three-parameter logistic models: A comparative study (Research Report 77-1). Columbia: University of Missouri, Department of Educational Psychology, November 1977.

Snedecor, G. W. and Cochran, W. G. Statistical Methods (7th ed.). Ames, IA: Iowa State University Press, 1980.

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

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

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14(2), 181-196. (a)

Urry, V. W. Tailored testing: A spectacular success for latent trait theory. Springfield, VA: National Technical Information Service, 1977. (b)

Urry, V. W. ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, D.C.: U. S. Civil Service Commission, Personnel Research and Development Center, 1978.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (ETS Research Memorandum RM-76-6). Princeton, NJ: Educational Testing Service, June, 1976.

Navy

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-711
NAVTRAEQUIPCEN
Orlando, FL 32813

1 Chief of Naval Education and Training
Liason Office
Air Force Human Resource Laboratory
Flying Training Division
WILLIAMS AFB, AZ 85224

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Dr. Richard Elster
Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93940

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Henry M. Halff
Department of Psychology,C-009
University of California at San Diego
La Jolla, CA 92093

Navy

1 Dr. Patrick R. Harrison
Psychology Course Director
LEADERSHIP & LAW DEPT. (7b)
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 CDR Charles W. Hutchins
Naval Air Systems Command Hq
AIR-340F
Navy Department
Washington, DC 20361

1 CDR Robert S. Kennedy
Head, Human Performance Sciences
Naval Aerospace Medical Research Lab
Box 29407
New Orleans, LA 70189

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

Navy

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
(OP-115)
Washington, DC 20350

1 LT Frank C. Petho, MSC, USN (Ph.D)
Selection and Training Research Division
Human Performance Sciences Dept.
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 Dr. Bernard Rimland (03B)
Navy Personnel R&D Center
San Diego, CA 92152

Navy

1 Dr. Worth Scanland, Director
Research, Development, Test & Evaluation
N-5
Naval Education and Training Command
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy,
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Wisher
Code 309
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Dexter Fletcher
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria,VA 22333

1 Dr. Michael Kaplan
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 DR. JAMES L. RANEY
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Army

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commandant
US Army Institute of Administration
Attn: Dr. Sherrill
FT Benjamin Harrison, IN 46256

1 Dr. Frederick Steinheiser
Dept. of Navy
Chief of Naval Operations
OP-113
Washington, DC 20350

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Research and Measurment Division
Research Branch, AFMPC/MPCYPR
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Dr. Marty Rockway
Technical Director
AFHRL(OT)
Williams AFB, AZ 58224

Marines

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Major Michael L. Patrow, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. William Graham
Testing Directorate
MEPCOM/MEPCT-P
Ft. Sheridan, IL 60037

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Dr. Wayne Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

1 DARPA
1400 Wilson Blvd.
Arlington, VA 22209

Civil Govt

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Capt. J. Jean Belanger
Training Development Division
Canadian Forces Training System
CFTSHQ, CFB Trenton
Astra, Ontario KOK 1B0

1 CDR Robert J. Biersner
Program Manager
Human Performance
Navy Medical R&D Command
Bethesda, MD 20014

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Liaison Scientists
Office of Naval Research,
Branch Office , London
Box 39 FPO New York 09510

1 Col Ray Bowles
800 N. Quincy St.
Room 804
Arlington, VA 22217

Non Govt

1 Dr. Robert Brennan
American College Testing Pr.grams
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. William E. Coffman
Director, Iowa Testing Programs
334 Lindquist Center
University of Iowa
Iowa City, IA 52242

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

Non Govt

1 Dr.,Fritz Drasgow
Yale School of Organization and Manageme
Yale University
Box 1A
New Haven, CT 06520

1 Dr. Mavin D. Dunnette
Personnel Decisions Research Institute
2415 Foshay Tower
821 Marguette Avenue
Mineapolis, MN 55402

1 Mike Durmeyer
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, Texas 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Non Govt

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organ.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

Non Govt

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, [illegible] 29208

1 Professor John A. Keats
University of Newcastle
AUSTRALIA 2308

1 Mr. Marlin Kroger
1117 Via Goleta
Palos Verdes Estates, CA 90274

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

Non Govt

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Bill Nordbrock
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

Non Govt

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

Non Govt

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. Robert Smith
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

Non Govt

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. Phyllis Weaver
Graduate School of Education
Harvard University
200 Larsen Hall, Appian Way
Cambridge, MA 02138

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2
WEST GERMANY