DOCUMENT RESUME

ED 205 545 TM 810 423

AUTHOR Rosso, Martin A.; Reckase, Mark D.
TITLE A Comparison of a Maximum Likelihood and a Bayesian Ability Estimation Procedure for Tailored Testing.
INSTITUTION Missouri Univ., Columbia.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Apr 81
CONTRACT N00014-77-C0097
NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 14-16, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Comparative Analysis; *Computer Assisted Testing; Higher Education; *Maximum Likelihood Statistics; Test Format; Test Reliability
IDENTIFIERS Estimation (Mathematics); *Tailored Testing

ABSTRACT

The overall purpose of this research was to compare a maximum likelihood based tailored testing procedure to a Bayesian tailored testing procedure. The results indicated that both tailored testing procedures produced equally reliable ability estimates. Also an analysis of test length indicated that reasonable ability estimates could be obtained using 12 to 14 items. It was also seen in the results that the maximum likelihood tailored testing procedure yielded significantly less total test information than did the Bayesian tailored testing procedure. The major difference between the two procedures seems to be in the significantly different ability estimates that they yielded. The maximum likelihood procedure is the procedure of choice if an adequate prior distribution is not available. (Author/BW)

A Comparison of a Maximum Likelihood and a Bayesian Ability Estimation Procedure for Tailored Testing

Martin A. Rosso
Mark D. Reckase
University of Missouri-Columbia

Objectives of Inquiry

Within the last decade, tailored testing has become one of the motivating forces behind the application of latent trait theory to achievement and ability measurement. This growing attractiveness of tailored testing is the result of the problems inherent in conventional paper-pencil testing procedures and the recent availability of adequate computer technology. In the conventional testing situation items of inappropriate difficulty are administered to some of the examinees. For example, examinees of low ability often receive items that are too difficult for them, and subsequently they may become frustrated. Obversely examinees of high ability levels may receive items that are not challenging and as a consequence they may become bored by the testing procedure. Ideally everyone should receive items appropriate to his or her level of ability. Conventional tests are most appropriate and most accurate for examinees of average ability. Therefore the standard error of measurement is ordinarily higher at the extremes of the ability range than it is at the middle of the ability range (Koch and Reckase, 1979).

Tailored testing is designed to circumvent these problems by attempting to administer to each examinee only items of appropriate difficulty. Matching item difficulty to ability level should reduce the errors of measurement at the extremes of the ability range thus reducing one of the problems of conventional paper-pencil testing. In order for tailored tests to select items of more appropriate difficulty, the selection of an item is based upon the ability estimate obtained from the previously administered items. Because of the advantages accrued by this procedure and the growing availability of computer technology, tailored testing systems will most certainly proliferate in the future.

In tailored testing there are two commonly used methods of operation. These two methods are based on a maximum likelihood ability estimation procedure and a Bayesian ability estimation procedure (Owen, 1975). The first procedure estimates a subject's ability after each item using an empirical maximum likelihood technique. The ability estimate is then used to select the next item in such a way that the item information is maximized at that ability level (Birnbaum, 1968). With the second procedure, ability is estimated as the mean of the posterior ability distribution and items are selected to minimize the posterior variance of the ability estimate distribution, while assuming a normal prior distribution of ability. As these two methods of operation are substantially different, it is important to examine the quality of results from these two tailored testing procedures in order to make an educated decision in choosing which procedure to implement.

This research will therefore compare the two procedures on the basis of obtained ability estimates, obtained total test information, and reliability. However, since tailored tests need not be fixed in length, the first step in

---

Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles, April, 1981. This research was supported by Contract Number N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

this research is to determine the optimal test length for each procedure. This was done since Reckase (1974) has found that continued testing beyond the point at which the ability estimate stabilizes may introduce bias in the ability estimate. This is consequence of the fact that most of the appropriate items from the item pool for that ability level have been used and only inappropriate items are available. The determination of the appropriate test length will be accomplished by using information and posterior variance. Once the test lengths for the two procedures have been obtained the ability estimates yielded by the two procedures at those lengths will be compared. Also the total test information yielded and the reliability coefficients yielded by the two procedures will be compared. It is hoped that by making these comparisons a clearly preferred procedure will emerge thus making the selection of a tailored testing system that much easier.

Instruments

The major instruments used in this research were the tailored test based on the maximum likelihood ability estimation procedure, and the tailored test based on the Bayesian ability estimation procedure. The items used in the study were 137 items from the School and College Ability Test (SCAT), Forms 2A and 3A (Educational Testing Service, 1975). These items measured vocabulary knowledge using two different item formats, but all items were of the five choice multiple-choice form. The tailored tests were administered on an Amdahl 470/V7 computer via the IBM Time Sharing Option. The subjects received the items on an ADDS Consul 980 terminal.

Method

The experiment extended over the winter semester and summer session, of 1980, with a different group of subjects each time period. (The summer session will hereafter be referred to as the summer semester in order to avoid confusion). The subjects who participated in the study were graduate and undergraduate students enrolled in measurement courses at the University of Missouri-Columbia. During the winter semester the subjects were enrolled in a graduate/undergraduate course entitled, "Group Intelligence Testing", and an undergraduate course entitled, "Introduction to Educational Measurement and Evaluation". To recruit volunteers for the experiment, students in the classes were advised at the beginning of each semester that those who volunteered to participate in the study would receive extra credit towards their course grade. Each subject was required to participate in two sessions which were one week apart.

During both the winter and summer semester, the students who volunteered to participate in the experiment were randomly assigned to either the maximum likelihood tailored test procedure or the Bayesian tailored test procedure. During the winter semester there were 19 subjects assigned to the Bayesian tailored testing procedure and 18 subjects were assigned to the maximum likelihood procedure. Because three subjects failed to complete the second session only 16 subjects were included in the Bayesian ability estimation procedure and 18 in the maximum likelihood ability estimation procedure. During the summer semester there were 13 subjects included in the Bayesian procedure and 23 in the maximum likelihood procedure. In total there were 70 subjects who completed the ex-

periment.

Each subject who participated in the experiment received either a test administered under the maximum likelihood condition for both testing sessions or a test administered under the Bayesian condition for both sessions. The tests were administered one week apart for each subject. The two different testing sessions were started using two different ability estimates, either -.100 or .150, so that the two different testing sessions would not be identical.

Analyses

The first analysis performed was a comparison of ability estimates from the two semesters using analysis of variance techniques. Next a determination of the optimal test lengths was made by subjectively evaluating plots of the convergence of the ability estimates. The reliabilities were then computed across sessions and were compared using chi-square analyses. The total test information yielded by the two different procedures was compared using analyses of variance, and then the ability estimates yielded by the two procedures were compared using a 4-way analysis of variance. Test, session, semester, and length were the independent variables for the analysis.

Results

Before the data could be analyzed a determination had to be made whether the ability estimates from both the winter and summer semesters could be pooled. Because there were graduate students included in the summer semester group of subjects there was some reason to suspect that the mean ability estimates obtained from two semesters were different, and possibly the two groups of data should not be combined. To discern if there were differences in the ability estimates from the two semesters a three-way analysis of variance (ANOVA) was performed on the ability estimates obtained at the 20 item level. Since there was a potential difference between the ability estimate scales yielded by the two procedures, the ability estimates obtained within each tailored testing procedure were converted to T-scores to eliminate the test effect. In the ANOVA the independent variables were test (maximum likelihood vs. Bayesian), semester (winter vs. summer), and session, with session being a repeated measure. The results of this ANOVA are summarized in Table 1. As seen in Table 1 the semester main effect was significant ($F = 7.89$, $p < .05$). Subsequently the decision was made to analyze the data using semester as an independent variable. It can be seen in Table 2 that the ability estimates from the summer semester were greater than the ability estimates from the winter semester for both the maximum likelihood tailored testing procedure and for the Bayesian tailored testing procedure.

Table 1

Results of Three-Way ANOVA on the Ability Estimates Yielded at the 20 Item Level With Test, Semester, and Session as Independent Variables

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		

It can also be seen in Table 1 that the sessions main effect was significant ($F = 6.63$, $p < .05$). From Table 2 it can be seen that the mean ability estimates were greater in the second session than in the first session for both the maximum likelihood procedure and the Bayesian procedure.

Table 2

Mean Ability Estimates in T-Score Form at the 20 Item Level for the Winter and Summer Semester and for the Maximum Likelihood and Bayesian Tailored Testing Procedure

Test	Winter		Summer	
	Session 1	Session 2	Session 1	Session 2
Maximum Likelihood	46.35	47.24	51.96	52.46
Bayesian	45.76	47.82	53.53	55.00

After the decision was made not to combine the data from the two semesters a determination of the optimal test length for the two tailored testing procedures had to be made. For the maximum likelihood procedure the values of the ability estimates obtained after each item and the item information estimates at the ability estimates were plotted. For the Bayesian procedure the values of the ability estimates after each item and the new standard error of estimate were plotted. A visual evaluation of the plots from both semesters suggested that the point for which the curves flattened was at the 12 item level for the maximum likelihood procedure and at the 14 item level for the Bayesian procedure (See Figure 1 and Figure 2 for examples of these plots). The flattening of the curves indicated convergence to an ability estimate. Thus the decision was made to analyze the data from both semesters at these levels as well as the 20 item level.

After making this decision the next analysis to be performed was the comparisons of the reliabilities. The reliabilities for each test were computed across sessions at the 12, 14, and 20 item levels within each semester. The reliabilities were computed for both ability estimates and estimated true scores (Lord, 1968) and are shown in Table 3. The first comparison was a chi-square on the estimated true score reliabilities in order to determine if the reliabilities were estimates of the same correlation (Snedecor and Cochran, 1980). It was not significant. The second comparison was on the reliabilities for the ability estimates. It also was not significant. Although it appears as if there is not a significant difference between the reliabilities of the two different testing procedures, nor between the various test lengths, it must be remembered that these reliabilities were obtained using relatively small samples and thus it would take a large difference to be significant.

Table 3

Bayesian vs. Maximum Likelihood Tailored Test Reliabilities for Winter and Summer Using Abilities 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
Bayesian	True Score	.885	.900	.830	.946	.881	.855
Max. Like.	Ability	.925	.865	.943	.908	.748	.777
Max. Like.	True Score	.899	.820	.936	.921	.875	.839

The next analysis to be performed was the comparison of the test information yielded by the two procedures at the 20 item level. Using the 20 item level was deemed appropriate since the reliabilities for the different test lengths were not significantly different, and as indicated before both tests appeared to be yielding consistent ability estimates by the 14 item level. A three-way ANOVA was performed over the data using as independent variables test (maximum likelihood vs. Bayesian), semester (winter vs. summer), and session; with session being a repeated measure. The dependent variable was total test information at the final ability estimate for 20-item level. The results of the ANOVA on the total test information at the 20 item level are shown in Table 4.

Table 4

Results of the Three-Way ANOVA on the Total Test Information Yielded at the 20 Item Level Using Semester Test and Session as Independent Variables

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		

As seen in the table the test main effect was significant ($F = 6.63$, $p < .05$) indicating that the two procedures were significantly different for the average total test information. The mean information values presented in Table 5 show that the Bayesian procedure yielded more total test information than did the maximum likelihood procedure. The only other significant effect was the semester main effect ($F = 6.11$, $p < .05$). This was not surprising, as earlier results showed that the ability estimates from the two procedures were different for the two semesters. Since the summer semester yielded higher ability estimates, this would have resulted in items with greater b-values being selected for the subjects during the summer session. Because there are fewer optimal items available at the extremes of the item pool this resulted in items being selected that yielded less than optimal item information. Since the total test information is contingent upon item information, this would result in lower test information during the summer. This can be seen in Table 5.

Table 5

Mean Total Test Information for the Bayesian and Maximum Likelihood Tailored Test Procedures for the Winter and Summer Semester

Semester	Session	Bayesian			Maximum Likelihood		
		20 Item	14 Item	12 Item	20 Item	14 Item	12 Item
Winter	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
Summer	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

After comparing the total test information yielded by the maximum likelihood tailored testing procedure and the Bayesian tailored testing procedure, the next analysis was the comparison of the ability estimates yielded by the two procedures. To make this comparison a four-way ANOVA was used in order to examine the effect of the test length on the ability estimate as well as the effect of the two different tailored testing procedures. The independent variables were test (maximum likelihood vs. Bayesian), length (20 items, 14 items, and 12 items), semester (winter vs. summer), and session, with session and length being repeated measures. The dependent variable was the ability estimates from the 12 item, 14 item, and 20 item levels. It was expected from prior results that there would be a semester main effect as well as a session main effect. The results shown in Table 6 indicate that both the semester main effect ($F = 8.33$, $p < .05$) and the session main effect ($F = 7.50$, $p < .05$) were significant. The session main effect was probably due to practice.

It is also seen from Table 6 that the test main effect was significant ($F = 15.43$, $p < .05$), indicating a significant difference between the ability estimates yielded by the two different tailored testing procedures. An examination of Table 7 indicates that the maximum likelihood procedure yielded greater ability estimates for both semesters, across sessions, and for all three different test lengths. Also important to be noted is the lack of a main effect for test length ($F = 0.61$, $p > .05$). This indicates that the mean ability estimates at the different lengths were not significantly different from one another. There was an interaction of test length and test ($F = 6.39$, $p < .05$). This interaction is explainable by the ability estimates from the maximum likelihood tailored testing procedure staying relatively stable while the ability estimates from the Bayesian tailored testing procedure changing with test length. (See Table 8).

Table 6

Results of the Four-Way ANOVA on the Ability Estimates From the Maximum Likelihood and Bayesian Tailored Test Procedures at the Three Different Test Lengths

Source	SS	df	MS	F	p
Semester	12.69	1	12.69	8.33	0.005
Test	23.49	1	23.49	15.43	0.000
Semester x Test	0.27	1	0.27	0.18	0.675
Error	100.50	66	1.52		
Session	1.01	1	1.01	7.50	0.008
Session x Semester	0.00	1	0.00	0.00	0.992
Session x Test	0.02	1	0.02	0.15	0.700
Session x Semester x Test	0.13	1	0.13	0.97	0.328
Error	8.92	66	0.14		
Length	0.04	2	0.02	0.61	0.546
Length x Semester	0.01	2	0.00	0.11	0.89
Length x Test	0.37	2	0.18	6.39	0.002
Length x Semester x Test	0.04	2	0.02	0.72	0.488
Error	3.83	132	0.03		
Session x Length	0.06	2	0.03	1.63	0.199
Session x Length x Semester	0.01	2	0.01	0.32	0.728
Session x Length x Test	0.02	2	0.01	0.46	0.630
Session x Length x Semester x Test	0.06	2	0.03	1.43	0.242
Error	2.54	132	0.02		

Table 7

Mean Ability Estimates for the Maximum Likelihood and Bayesian Ability Estimation Procedures at the 12, 14, and 20 Item Levels

		Mean Ability Estimates			
		Maximum Likelihood		Bayesian	
Semester	Item	Session 1	Session 2	Session 1	Session 2
Winter	12	1.32	1.36	0.67	0.85
	14	1.28	1.33	0.69	0.98
	20	1.25	1.30	0.78	0.89
Summer	12	1.53	1.68	1.07	1.15
	14	1.50	1.68	1.12	1.78
	20	1.53	1.55	1.18	1.26

Since the ability estimates yielded by the two different procedures were significantly different, it seems likely that the items selected by the procedures would be different. As inspection of a frequency count of item usage indicated that the maximum likelihood procedure was utilizing items with higher b-values than was the Bayesian procedure.

Table 8

Mean Ability Estimates for the Maximum Likelihood and Bayesian Ability Estimation Procedures Combined over Semesters and Sessions

Item	Ability Estimates	
	Maximum Likelihood	Bayesian
12	1.48	.906
14	1.47	.914
20	1.42	.992

Summary and Conclusions

The overall purpose of this research was to compare a maximum likelihood based tailored testing procedure to a Bayesian tailored testing procedure. The results indicated that both tailored testing procedures produced equally reliable ability estimates. Also an analysis of test length indicated that reasonable ability estimates could be obtained using 12 to 14 items.

It was also seen in the results that the maximum likelihood tailored testing procedure yielded significantly less total test information than did the Bayesian tailored testing procedure. This seemed to be a result of the fact that the maximum likelihood procedure yielded significantly higher ability estimates, thus utilizing from the item pool items with greater b-values. At the extremes of the item pool there were fewer optimal items from which to choose. This problem may have been alleviated had the item pool had more items with greater b-values.

The major difference between the two procedures seems to be in the significantly different ability estimates that they yielded. An examination of the ability estimation procedure used by the two procedures explains why this discrepancy exists. The Bayesian tailored testing procedure performs its ability estimation on the basis of the prior ability distribution. This results in a regression towards the prior mean for this procedure's ability estimates. Since

in this study the initial ability distribution had a mean well below the population ability level, the result was an inhibiting effect on the final ability estimates. It was predicted that had the prior ability distribution been greater than the population value the result would have been that the final ability estimates would have been greater. This result was borne out when the ability estimates for the Bayesian group were recalculated using a prior mean ability estimate of 2.00. The results were that the recalculated ability estimates were significantly higher ($\bar{x}$ = 1.06, t = 4.34, $p \leq .05$). This result points out the importance of the prior to the Bayesian procedure. An inaccurate prior can affect the ability estimates. Since knowledge of the prior is often not available this procedure could result in biased estimates of ability. It thus seems that the maximum likelihood procedure is the procedure of choice if an adequate prior distribution is not available.

FIGURE 1

ABILITY ESTIMATES AND
INFORMATION VALUES AFTER EACH
ITEM IN A MAXIMUM LIKELIHOOD
TAILORED TEST
WINTER SEMESTER
SESSION 1

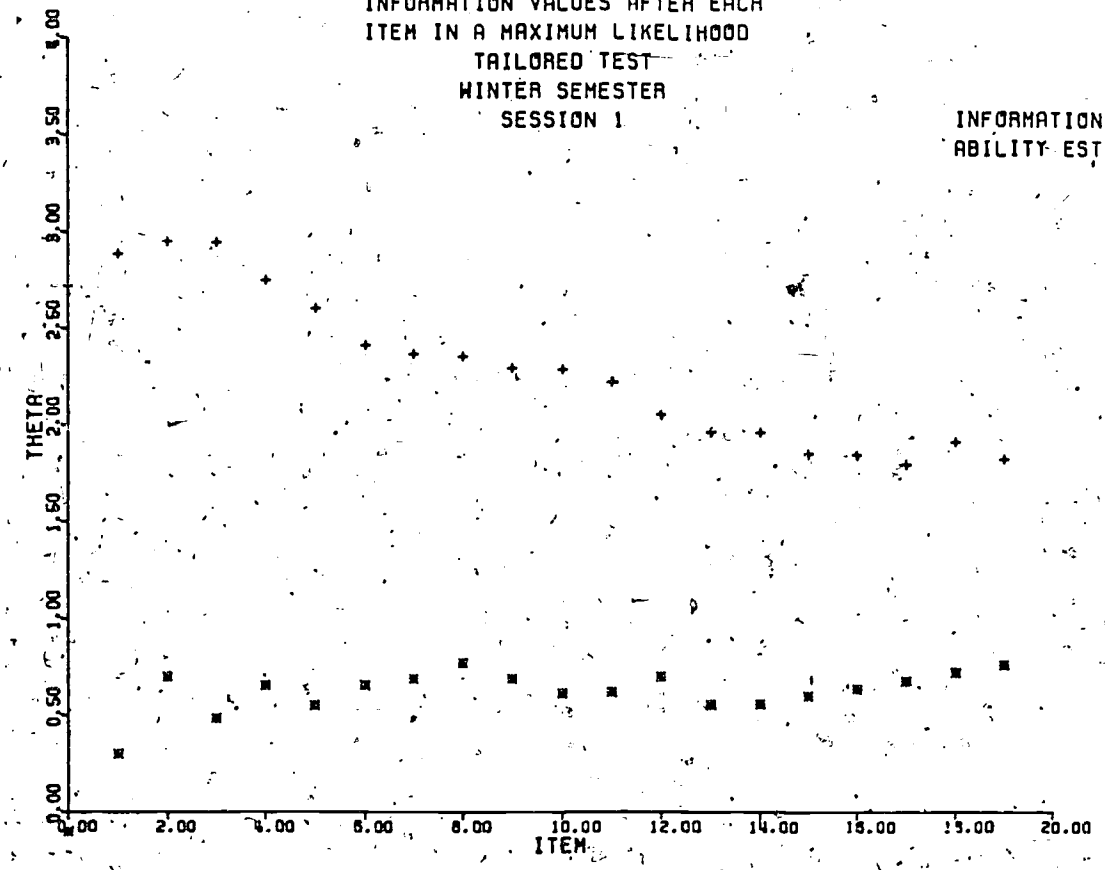
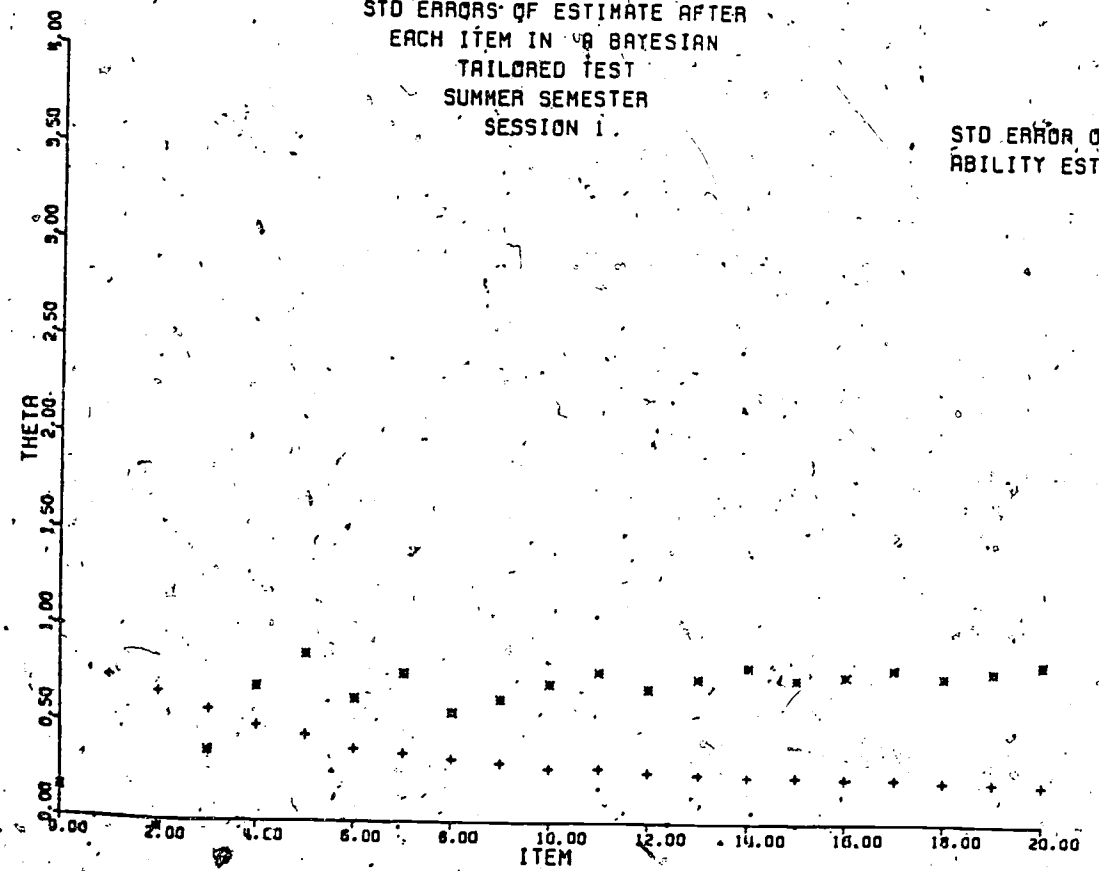


FIGURE 2

ABILITY ESTIMATES AND
STD ERRORS OF ESTIMATE AFTER
EACH ITEM IN A BAYESIAN
TAILORED TEST
SUMMER SEMESTER
SESSION 1



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, Massachusetts: Addison Wesley, 1968.

Educational Testing Service. Cooperative School and College Ability Tests. Princeton, New Jersey, 1955.

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

Lord, F. M. and Novick, M. R. Statistical theories of Mental Test Scores. Reading, Massachusetts: Addison Wesley, 1968.

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.

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.

Snedecor, G. W. and Cochran, W. G. Statistical Methods. Ames, Iowa: Iowa State University Press, 1980.