DOCUMENT RESUME

ED 159 200 TM 007 467

AUTHOR Koch, Bill R.; Reckase, Mark D.
TITLE A Live Tailored Testing Comparison Study of the One and Three Parameter Logistic Models.
PUB DATE Mar 78
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Asscciation (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Comparative Analysis; Goodness of Fit; Higher Education; *Mathematical Models; Multiple Choice Tests; *Testing; *Test Reliability
IDENTIFIERS Latent Trait Models; *Rasch Model; Tailored Testing; *Three Parameter Model

ABSTRACT

A live tailored testing study was conducted to compare the results of using either the one-parameter logistic model or the three-parameter logistic model to measure the performance of college students on multiple choice vocabulary items. The results of the study showed the three-parameter tailored testing procedure to be superior to the one-parameter procedure on the basis of goodness of fit of observed to predicted item responses, test-retest reliability, convergence to stable ability estimates, and test information. No differences were found in the prediction of an outside criterion. However, implicit in these results was the assumption that the nonconvergence problem encountered in one-third of the cases for the three-parameter procedure could be solved. Thus, based on the data reported in this study, the three-parameter tailored testing method was deemed the technique of choice, at least for unidimensional tests consisting of multiple choice items where guessing is a factor. (Author)

A Live Tailored Testing Comparison Study of the One and Three Parameter Logistic Models

by

Bill R. Koch and Mark D. Reckase
University of Missouri-Columbia

Tailored testing derives its name from its primary aim and characteristic, which is to attempt to "tailor" a test for a given individual, often using computer capabilities. That is, rather than administering the same set of test items to all examinees, the tailored testing procedure presents a unique set of items that tries to match item difficulty levels to a person's ability. "An examinee is measured most effectively when the test items are neither too difficult nor too easy for him" (Lord, 1970). Thus, one goal of the tailored testing procedure is to select items from a precalibrated item pool stored in the computer so that the probability of a correct response by the examinee is .50 on each item. In general, tailored testing procedures require the three components of a pool of calibrated items, an item selection technique, and a scoring method (Patience, 1977).

Although several tailored testing procedures have been developed, most of the procedures employ either a one-parameter or a three-parameter logistic model for item calibration and ability estimation purposes. However, no empirical studies have been reported in the literature that directly compare these two tailored testing models on the basis of their relative performances and characteristics in actual live-testing settings. The primary purpose of the present study, therefore, was to deal with this issue and hopefully collect evidence for the recommendation of one model over the other in this specific situation. We begin with a brief discussion of the two latent trait models.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Toronto, 1978. This research was supported by contract number N00014-77-C-0097 from the Personnel and Training Research Programs of the Office of Naval Research.

The Rasch model (1960) or one-parameter logistic model, is thoroughly described in a recent article by Wright (1977). Here let it suffice to say that the one parameter model requires only one ability parameter θ_j for each person and one item difficulty parameter b_i to describe the interaction between an examinee and a test item. The exponential form of the simple logistic model is

$$P\{u_{ij}\} = \frac{e^{u_{ij}(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \qquad (1)$$

where u_{ij} is the score (0 or 1) on Item i by Person j, θ_j and b_i are as defined above, and $P\{u_{ij}\}$ is the probability of a correct or incorrect response.

In contrast, the three-parameter logistic model presented by Birnbaum (1968) requires the estimation of three item parameters to describe the interaction between test items and examinees. The model is given by

$$P_{ij} = P\{u_{ij} = 1\} = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (2)$$

where $P\{u_{ij} = 1\}$ is the probability of a correct response by Person j to Item i; c_i is the guessing parameter for Item i; D is a scaling constant equal to 1.7; a_i is the item discrimination parameter; b_i is the item difficulty parameter; and θ_j is the ability parameter for Person j. Q_{ij}, the probability of an incorrect response, is defined simply as $1 - P_{ij}$.

Both models have in common the assumptions that the items may be scored dichotomously, that the latent trait being measured by the items is unidimensional, that item parameters remain invariant across groups of examinees, and that local independence holds (Lord and Novick, 1968).

The bases for the comparisons of the two tailored testing procedures will be (a) the goodness of fit of the models using mean squared deviations of observed from predicted response data, (b) the reliabilities of the two methods, (c) the ability estimates yielded by the two procedures, (d) the correlation of the ability estimates with the same outside criterion, (e) descriptive statistics for each procedure, (f) the rates at which the two methods converge to ability estimates, and (g) the information functions for the two procedures.

Method

Item Calibrations

The source of items used for the tailored testing comparison study was the Syracuse Adult Development Study vocabulary tests, Forms C2, D2, and E(1972). All of the items were of the multiple choice form with five alternatives per item. A principal components factor analysis of the inter-item tetrachoric correlation coefficients conducted on form D-2 indicated that only one factor was present in the test, accounting for approximately 41% of the variance, with a sample size of 1,000 (Reckase, 1972).

Two identical pools of 72 vocabulary items were constructed, one for use with the one-parameter model and the other for the three-parameter model. The one-parameter pool was calibrated using a modified version of a program given in an article by Wright and Panchapakesan (1969). For the three parameter pool, the LOGIST program developed by Wood, Wingersky, and Lord (1976) was used. Table 1 presents the means, standard deviations, and ranges of the item parameter estimates resulting from the two calibration procedures, along with the sample sizes upon which they were based.

Insert Table 1 about here

Specific Tailored Testing Procedures

For the one-parameter procedure, items were selected for administration based on difficulty values (b_i). The procedure began with an ability estimate of $\pm.50$ for the examinees, depending on the experimental condition to which they had been assigned. Thus, the first item administered was the first one encountered in the pool that was equal to the initial ability estimate, within a $\pm.30$ acceptance range. If the examinee answered the first item correctly, the next item administered was the item in the pool at a fixed stepsize away (.693) in a positive direction, i.e. a more difficult item, still within the acceptance range. On the other hand, an incorrect response led to the next item that was -.693 away, i.e. an easier item. The .693 fixed stepsize value had been previously determined through an analysis of tailored testing operation (Reckase, 1976).

When at least one item had been answered correctly and one incorrectly, the ability level of the examinee was estimated using an empirical maximum likelihood procedure. The technique used was an iterative search to determine the mode of the likelihood distribution, which became the new ability estimate. The next item administered was one selected so that it had probability .50 of being answered correctly. For the one-parameter model this was an item with difficulty equal to the ability estimate, within the $\pm.30$ acceptance range of easiness. The tailored test was terminated when no items remained in the pool that fell within the $\pm.30$ range or when a maximum of 20 total items had been administered.

For the three-parameter procedure, items were selected for administration based on values of the information function. Actually, this was equivalent to the one-parameter item selection procedure since, for the one-parameter model, selecting items to maximize the information that an item provided about a person's ability

was the same as selecting items on the basis of appropriate easiness value. That is, the information function was maximal for the one-parameter model when the item administered equalled the ability estimate.

However, for the three-parameter model, the information function was more complex. In particular, the added discrimination and guessing parameters played a crucial role in determining the amplitude of the information curve. The formula used to compute item information for the three-parameter logistic model was given in Birnbaum (1968) as

$$I(\theta_j,u_i) = D^2a_i^2\psi[DL_i'(\theta_j)] - D^2a_iP_{ij}(\theta_j)\psi[DL_i(\theta_j) - \log c_i] \quad (3)$$

where $I(\theta_j,u_i)$ is the information of Item i at ability level θ for Person j, given item response u_i; $L_i(\theta_j) = a_i(\theta_j - b_i)$; $P_{ij}(\theta_j)$ is the probability of a correct response to Item i given ability level θ_j; $\psi(x)$ is the logistic probability density function; and the other parameters have their meanings mentioned previously. The total test information was then simply the sum of the item information (Birnbaum, 1968) given by

$$I(\theta) = \sum_{i=1}^{n} I(\theta_j,u_i). \quad (4)$$

The tailored testing procedure for the three-parameter model began the same way as described above. Namely, a fixed .693 stepsize was used to select items until at least one correct and incorrect response had been obtained. Ability estimates were again computed using the maximum likelihood technique. However, to select the next item to be administered, the item pool was searched for the item which had the most information (i.e. $I(\theta_j,u_i)$ was maximal) for that particular ability estimate. This process was repeated until either no item was available in the pool with $I(\theta_j,u_i) > .70$ or until a total of 20 items had been administered.

Design

The sutdy employed a counterbalanced design in which there wer two separate sessions one week apart for each examinee, with both the one- and three-parameter tests administered at each session. The order of test presentation was reversed from one session to the next for each examinee, but the test was arranged so that the examinees were not aware of receiving two tests. The second test was initiated immediately after a final ability estimate was obtained from the first test. The tests were all administered on ADDS Consul 980 cathode ray tube terminals connected to an IBM 370/168 computer through a timesharing system.

The subjects who participated in the study were undergraduate and graduate students enrolled in educational psychology and measurement courses at the University of Missouri-Columbia. A total of 142 students took part in the study, but 14 cases were deleted due to missing data, resulting in 128 net examinees. All students received extra credit for their participation.

Analyses

The measure used to determine the goodness of fit of the observed response data to the models was the mean squared deviation (MSD) statistic given by

$$MSD_j = \frac{\sum_{i=1}^{N} (u_{ij} - P_{ij})}{N} . \qquad (5)$$

where MSD_j was the mean squared deviation for Person j; u_{ij} was the actual response; P_{ij} was the predicted response from the model; and N was the number of items from the tailored test. Two MSD statistics were calculated for each examinee, one for each model from the first test session. A systematic sample of 22 examinees was taken to compare the two models using the MSD criterion in a t-test analysis, since

it was desired that MSD values be computed across the range of ability estimates yielded by the tailored tests.

The reliability comparison of the two models was not a true test-retest reliability, but rather was a hybrid of test-retest and equivalent forms reliability. It was impossible for an examinee to receive exactly the same tailored test twice due to differences in entry points into the item pool and to changes in response strings. However, numerous items were repeated over test sessions as a function of the consistency in ability estimation for a person since items were selected from the same pool. Several descriptive statistics were also computed for the two testing procedures such as average test length, average difficulty, and percentage of test items in common over the two sessions. Where differences were found, the effects on reliability were partialed out.

Correlation analyses were conducted between ability estimates yielded by the one- and three-parameter models over the two test sessions, as well as between the ability estimates and an outside criterion of performance, namely, traditional paper and pencil exam scores over course material. The purpose of these correlations was to determine the degree to which the two test procedures were measuring the same thing, and whether one model did better than the other in prediction of the criterion.

Information function analyses were performed to compare the two models in terms of relative efficiency, the ratio of tailored test information to total test information (Lord, 1970). A plot was constructed of the relative efficiency of both the one-parameter and the three-parameter tailored tests against the same 30-item traditional vocabulary test. Again, data for the plot were selected with a systematic rather than random sample to insure broad coverage over the range of ability estimates.

Convergence plots were drawn for the tailored tests taken by each examinee over both sessions. On one axis were plotted the ability estimates calculated after each item was administered, and on the other axis were plotted the items received. The purpose was to provide a graphic description of the rates at which the two models converged to stable ability estimates. Direct comparisons in this regard were not possible since the one- and three-parameter ability estimates were on different scales. However, representative plots were selected and subjective summary judgements were made.

Results

Goodness of Fit

The results of the MSD statistic to compare the goodness of fit of the one- and the three-parameter models are presented in Table 2. The MSD values are shown for 22 cases along with descriptive statistics and the results of a paired samples t-test analysis on the data. The t-test showed that the MSD statistic was significantly smaller ($p < .05$) for the three-parameter model, indicating better fit of the model to the observed response data.

Insert Table 2 about here

Reliability

The correlation matrix in Table 3 consists of the coefficients obtained from intercorrelating the various ability estimates yielded in the tailored tests from the two models. Of special interest is the correlation between the ability estimate from the first one-parameter logistic tailored test (1PL 1) and the second one-parameter logistic tailored test (1PL 2). The .61 value shown in Table 3 is the

reliability coefficient for the one-parameter logistic tailored test. This is significantly lower ($p < .05$) than the .77 reliability coefficient obtained by correlating the ability estimates from the first three-parameter logistic tailored test (3PL 1) and the second corresponding test (3PL 2).

Insert Table 3 about here

It is very important to note, however, that these reliabilities are based on only 89 rather than 128 cases. The difference is due to the failure of the three-parameter tailored test to converge at ability estimates for 39 cases. The non-convergence problem was common when using maximum likelihood ability estimation for the three-parameter model when very difficult items were encountered which substantially raised the lower asymptote of the logistic function, $\hat{c}_i$, the chance of obtaining a correct response by random guessing. In such cases, the mode of the likelihood distribution could not be found, and the estimation procedure did not yield an ability estimate. The values in parentheses in Table 3 indicate the reliability coefficients obtained when the 39 nonconvergence cases remain in the analyses. The three-parameter reliability now drops from .77 to only .36. The one-parameter reliability also drops slightly from .61 to .55. However, the difference between the reliabilities for 128 cases (.36 vs. .55) is not statistically significant.

Since it was common for each tailored test administered to an examinee to have different numbers of test items, and since test length often impacts on reliability, another comparison was undertaken in which ability estimates were equated for test length. The correlation between the first and second one-parameter tailored test ability estimates, .61, was compared to the correlation between the first and second three-parameter ability estimates for tests with an equal number of items

presented (3PLEQI 1 vs. 3PLEQI 2). The resulting difference between these correlations, .61 and .78, was still significant.

The number of test items in common from one test to another was also investigated for a possible effect on reliability, since the three-parameter tests had 85% of such items in common, compared to only 20% for the one-parameter tests. Partial correlation coefficients were computed to factor out the effects of repeated test items on the overall reliabilities, but the results showed this variable to have no effect.

Table 4 presents several additional descriptive statistics for the one- and the three-parameter tests. For example, the mean test difficulty for both procedures was about the same, close to .50. This indicated that, in general, items of appropriate difficulty were being administered. Also note that the three-parameter tests tended to be slightly longer than the one-parameter tests.

Insert Table 4 about here.

Other Correlation Analyses

Table 3 illustrates the degree of similarity among all the ability estimate intercorrelations, regardless of the procedure. The ability estimates yielded by the one-parameter tests and the three-parameter tests consistently fall in the range from .44 up to .70. Not shown in the table, but also computed, were the correlations between the ability estimates yielded by the tailored tests and the outside criterion of scores on traditional course exams. These correlations were consistently in the .30's for both procedures over both sessions, meaning that both the one-parameter and the three-parameter tests predicted the outside criterion equally well.

Information Function Analyses

The results of the relative efficiency comparison are shown in Figure 1. The horizontal dashed line indicates the information of the traditional 30-item vocabulary test as the reference position to compare the two types of tailored tests. However, the ability scales used for plotting the two relative efficiency curves are not the same. The plot shows that the three-parameter tailored test yielded substantially greater information than the traditional test, but only in a peaked fashion for ability estimate levels between -2.0 and +.50, falling off sharply outside this range. However, at no point did the one-parameter tailored test exceed the traditional test information, and its information curve was rectangular rather than peaked. Also shown in Figure 1 are the frequency distributions of ability estimates obtained from the two procedures. Note that the information from the three-parameter test is greatest ere most of the examinees were concentrated.

Insert Figure 1 about here

Convergence Plots

In Figure 2 are pictured four individual tailored testing convergence plots, including good and poor examples of convergence for each of the two types of tailored tests. Plot 2-A shows a case where neither procedure converged very well, 2-B a case where the one-parameter test did well but not the three-parameter test, 2-C a case in which the three-parameter test converged better than the one-parameter test, and 2-D where both procedures converged nicely. A subjective classification method applied to 44 separate cases resulted in the following breakdown: 2-A, 7 plots; 2-B, 5 plots; 2-C, 18 plots; and 2-D, 14 plots. However,

recall that in 39 cases, not included in the above categories, the three-parameter tailored testing procedure failed to converge at all.

Insert Figure 2 about here

Discussion

Theoretically the MSD statistic had a possible range in value from 0 to 1 -- 0 for perfect fit and 1 for perfect lack of fit. In actual practice, however, the value of the MSD for an examinee rarely exceeded .25 for either model. Although the sampling distribution of the MSD statistic was unknown, previous research had shown the distribution to be approximately normal (Reckase, 1977). Thus the t-test results may be interpreted for this data as evidence that the three-parameter tailored testing procedure did a significantly better job of fitting the response data than the one-parameter test. The result showed a closer match between the item responses predicted by the model and the actual observed responses for the three-parameter tailored test.

The reliability comparison also showed the three-parameter procedure to be superior, but only when about one-third of the nonconverging tests were removed from the data analysis. This superiority held even when the effects of test length and repeated items were controlled or equated for both procedures.

However, the consistent, moderately high degree of intercorrelation among the ability estimates yielded by both models over both sessions indicated that both procedures were measuring the same thing. Moreover, both of the tailored testing methods correlated equally well with the outside criterion measure. In this regard it should be noted that high correlations were not expected, since performance levels on a general vocabulary test would not necessarily lead to similar performances

on course achievement tests. However, the achievement test scores were the only outside criterion available for the examinees.

The descriptive statistics for the two tailored testing procedures showed the three-parameter tests to be slightly longer on the average, although test length differences would best be interpreted as being a function of the different item selection methods and stopping rules employed. Since the ±.30 acceptance range for the one-parameter method and the .70 information level cutoff for the three-parameter method were both somewhat arbitrary values derived from simulation and empirical studies, changes in these values would have changed the number of items administered. Both procedures functioned well on the average in administering items of appropriate difficulty (near .50) for the examinees.

The relative efficiency comparison of the two procedures based on their respective test information curves showed that neither type of tailored test provided as much information across the broad range of ability estimates as did the traditional test. However, the three-parameter procedure did exceed the traditional test information for a limited range of abilities, the range in which most persons were concentrated, while in no case did the one-parameter test information do so.

The subjective analysis of the convergence plots on the whole indicated that the three-parameter tailored tests did a better job of arriving at stable ability estimates than the one-parameter tests. Of course, this result held only when 39 nonconvergence cases were removed from the data analysis. If included, the one-parameter tailored test convergence patterns would have been superior.

Summary and Conclusion

A live tailored testing study was conducted to compare the results of using either the one-parameter logistic model or the three-parameter logistic model to

measure the performance of college students on multiple choice vocabulary items. The results of the study showed the three-parameter tailored testing procedure to be superior to the one-parameter procedure on the basis of goodness of fit of observed to predicted item responses, test-retest reliability, convergence to stable ability estimates, and test information. No differences were found in the prediction of an outside criterion. However, implicit in these results was the assumption that the nonconvergence problem encountered in one-third of the cases for the three-parameter procedure could be solved. Thus, based on the data reported in this study, the three-parameter tailored testing method was deemed the techique of choice, at least for unidimensional tests consisting of multiple choice items where guessing is a factor.

Table 1

Descriptive Statistics of Item Parameter Estimates for the Two Models

	One Parameter Model	Three Parameter Model		
	E_i	a_i	b_i	c_i
Mean	- .172	.990	- .519	.121
Standard Deviation	1.467	.533	1.529	.042
Low	-2.821	.118	-3.624	.023
High	3.559	2.000[a]	5.952	.270
Sample Size	1,000	1,541	1,541	1,541
No. of Items	72	72	72	72

[a]The LOGIST program imposes the restriction that discrimination estimates must stay in the range from .01 to 2.00.

Table 2

Goodness of Fit Comparison
Using the MSD Statistic

Observations	One Parameter MSD	Three Parameter MSD
1	.198	.184
2	.197	.206
3	.212	.158
4	.214	.100
5	.083	.143
6	.203	.098
7	.202	.208
8	.187	.156
9	.208	.153
10	.204	.140
11	.192	.171
12	.083	.133
13	.215	.267
14	.196	.191
15	.164	.198
16	.194	.144
17	.203	.166
18	.203	.126
19	.183	.247
20	.214	.149
21	.182	.022
22	.188	.185
$\bar{X}$	.188	.161
$s_{\bar{x}}$	.055	.063

$t_{21} = 2.086$ $(p < .05)$

Table 3

Ability Estimate Correlations

Variables	1	2	3	4	5	6	7	8
1. 1PL 1		.61(.55)[a]	.96	.53	.57	.58	.53	.59
2. 1PL 2			.53	.90	.68	.70	.63	.69
3. 1PLEQI 1				.47	.49	.53	.44	.55
4. 1PLEQI 2					.52	.51	.47	.49
5. 3PL 1						.77(.36)[a]	.90	.76
6. 3PL 2							.79	.96
7. 3PLEQI 1								.78
8. 3PLEQI 2								

( )[a] indicates the inclusion of 39 cases of non-convergence at an ability estimate for the three parameter test, with all other correlations based on 89 cases.

Table 4

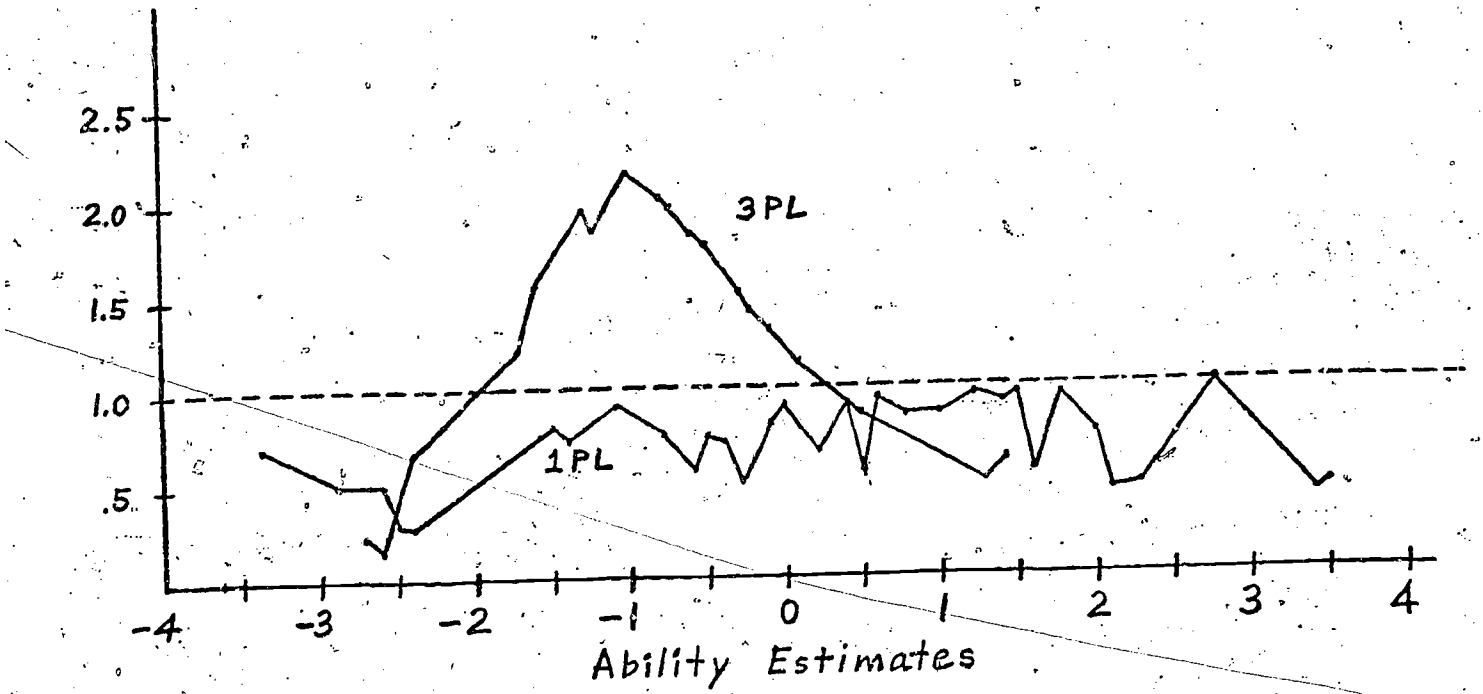
Descriptive Statistics

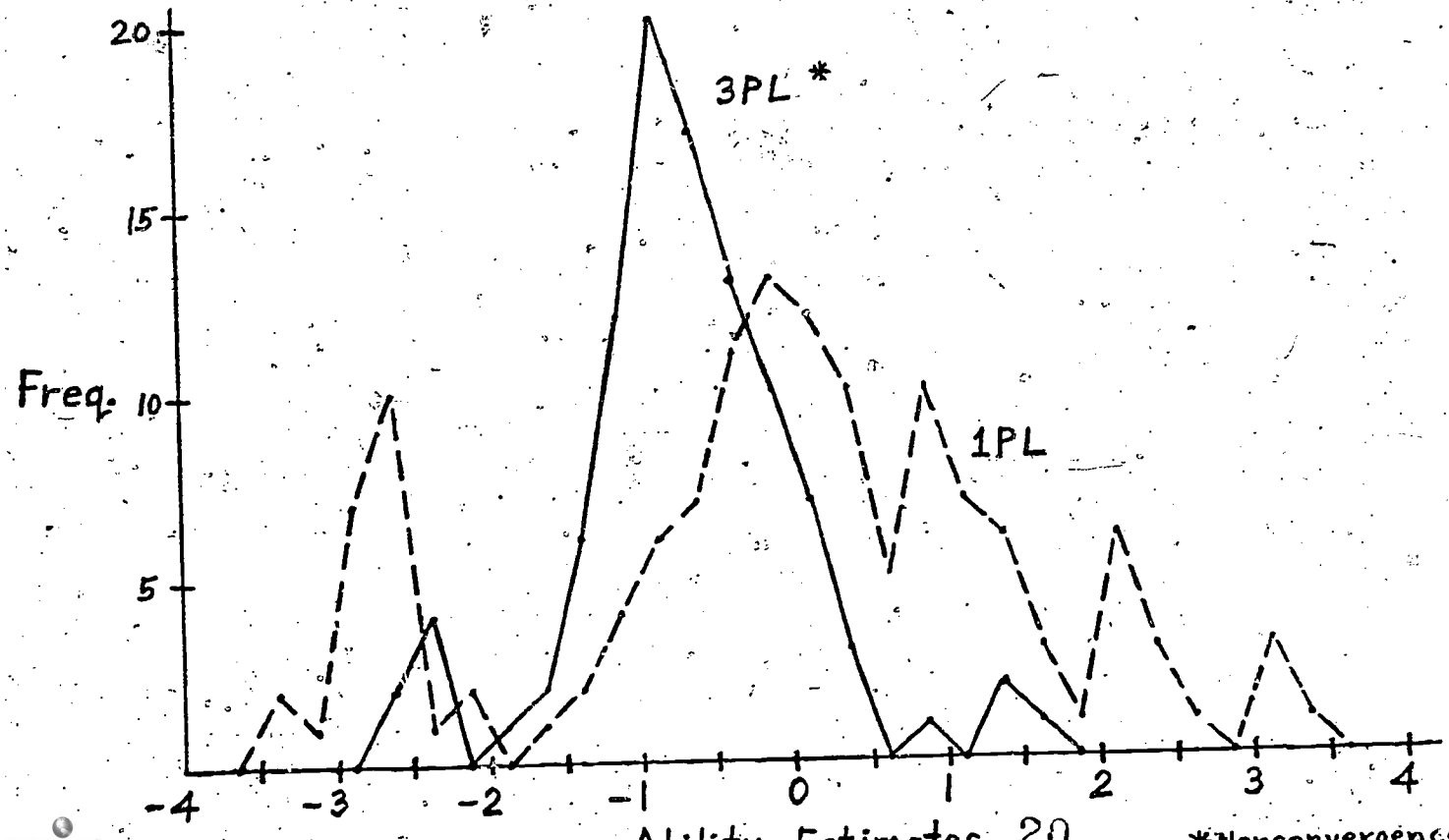
Variable	One Parameter Tailored Test	Three Parameter Tailored Test
mean # of items administered	15.07	18.39
mean # of items correct	7.45	8.95
mean test difficulty	.49	.49
mean ability estimates	.44	-.77

n = 89

Figure 1

Relative Efficiency



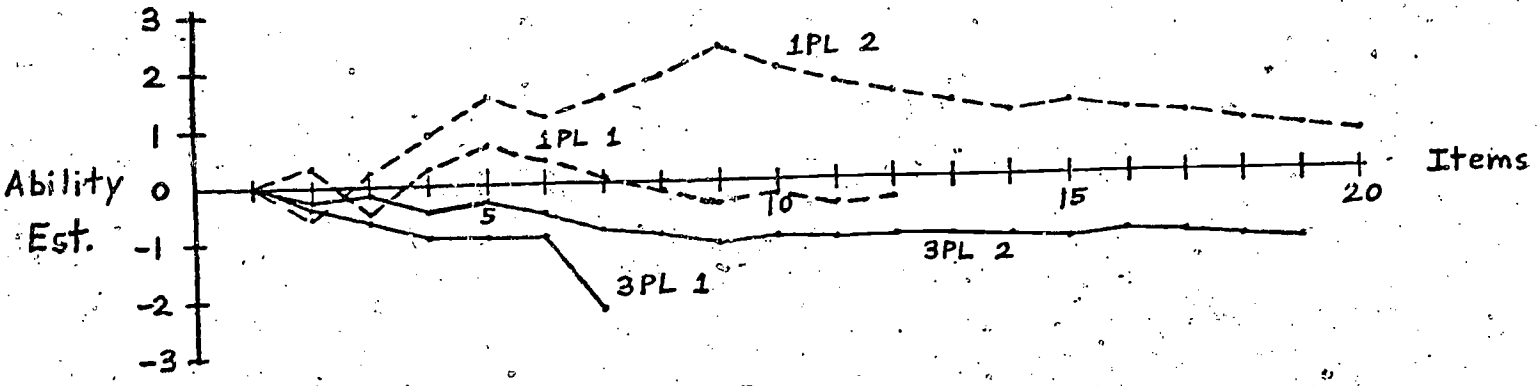


*Nonconvergence cases deleted

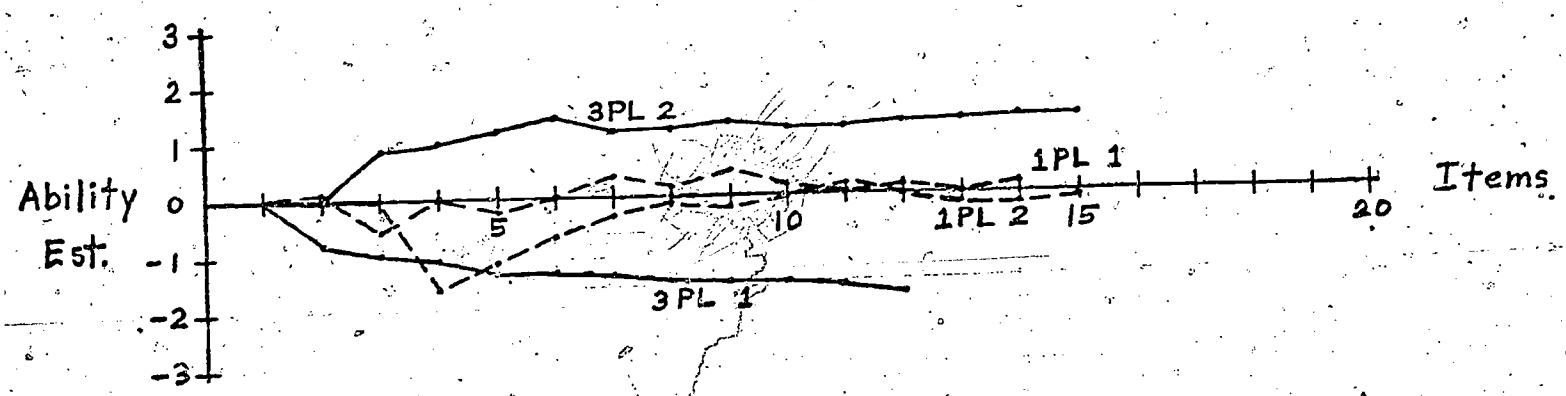
Figure 2

Convergence Plots

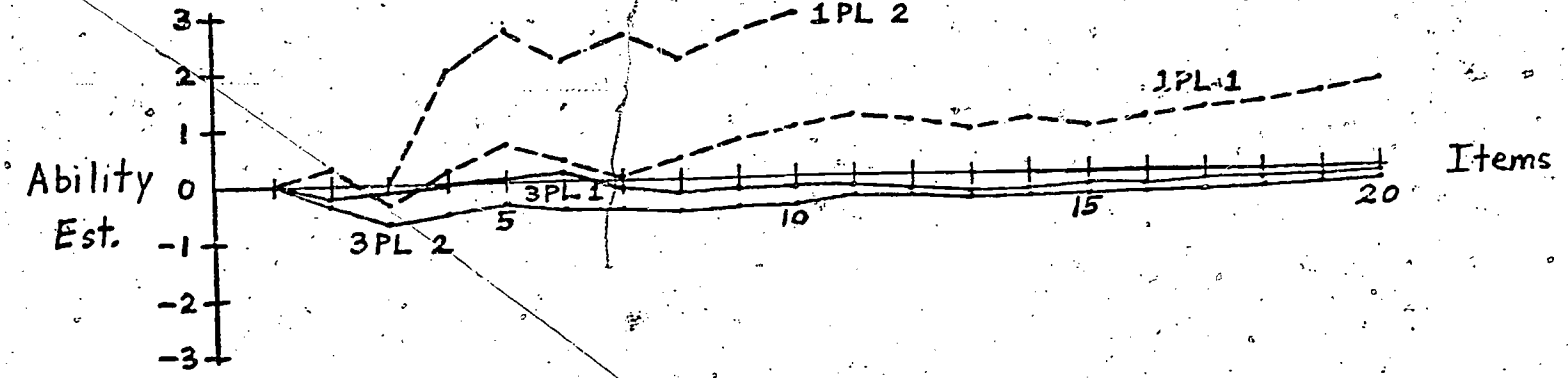
2 - A



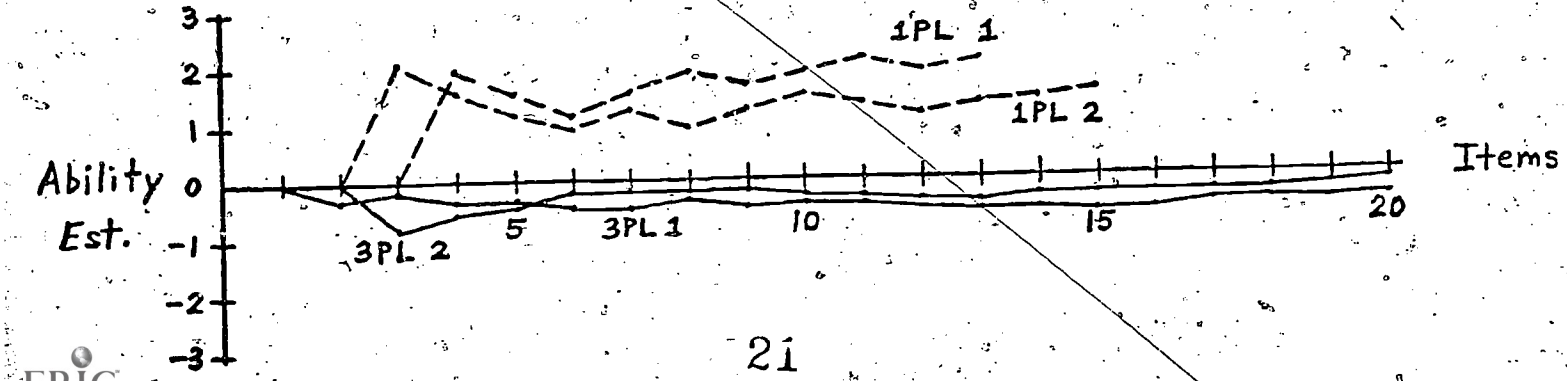
2 - B



2 - C



2 - D



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.

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

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Monge, R. H. & Gardner, E. F. A program of research in adult differences in cognitive performance and learning: backgrounds for adult education and vocational retraining. (Final Report, Project No. 6-1963, Grant No. OEG 1-7-061963-0149). Syracuse, New York: Syracuse University, Department of Psychology, January 1972.

Patience, W. M. Description of components in tailored testing. Behavior Research Methods & Instrumentation, 1977, 9, 153-157.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase, M. D. Development and application of a multivariate logistic latent trait model. (Doctoral dissertation, Syracuse University, 1972). Dissertation Abstracts International, 1973, 33. (University Microfilms No. 73-7762).

Reckase, M. D. The effects of item pool characteristics on the operation of a tailored testing procedure. Paper presented at the spring meeting of the Psychometric Society, Murray Hill, New Jersey, 1976.

Reckase, M. D. Ability estimation and item calibration using the one and three parameter logistic models: a comparative study. (Research Report 77-1). Columbia, Missouri: University of Missouri, Educational Psychology Department, November 1977. (AD A047943).

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

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.