ED 155 203	TM 007 126
AUTHOR	Reckase, Mark D.
TITLE	A Comparison of the One- and Three-Parameter Logistic Models for Item Calibration.
SPONS AGENCY	Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE	Mar 78
CONTRACT	N00014-77-C-097
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Comparative Analysis; *Comparative Statistics; Goodness of Fit; *Item Analysis; *Mathematical Models; Reliability; Sampling; Scores; Simulation; *Statistical Analysis; Test Items; Test Validity
IDENTIFIERS	Estimation; *Rasch Model; Sample Size; *Three Parameter Model

ABSTRACT

Five comparisons were made relative to the quality of estimates of ability parameters and item calibrations obtained from the one-parameter and three-parameter logistic models. The results indicate: (1) The three-parameter model fit the test data better in all cases than did the one-parameter model. For simulation data sets, multi-factor data were less well fit than single-factor data. (2) The one-parameter model ability estimates shared more variance with the item responses than did the three-parameter model. (3) There was no difference in the concurrent validity for small samples between the two models in predicting classroom achievement tests. (4) The three-parameter model required larger samples for calibration than did the one-parameter model. (5) The ability estimates from the two models correlated highly for most of the data sets. The one-parameter model is preferred for use with small sample data; but the goodness of fit data reflected a different point of view when accurate estimation of item parameters is important. The three-parameter model fit all data sets better than the one-parameter model. Data sets from the Missouri School and College Ability Tests, and from undergraduate course final examinations were used to illustrate the models. (Author/CTM)

A Comparison of the One- and Three-Parameter Logistic Models for Item Calibration

Mark D. Reckase
University of Missouri-Columbia

Since the development of the three-parameter logistic model by Birnbaum (1958), and the independent production of a simpler, one-parameter logistic model by Rasch (1960), there has been an ongoing debate concerning the relative merits of the two-models. The debate stems from the need to make the very restrictive assumptions of equal discrimination and no guessing for test items using the one-parameter model, while the three-parameter model requires cumbersome estimation procedures for calibration. The purpose of the research presented here is to evaluate the relative merits of the two models for item calibration and ability estimation, resulting in a clarification of the above issue.

Two studies have already been done to compare the one- and three-parameter models (Hambleton & Traub, 1971; Urry, 1977) but these studies were limited in the scope of their comparisons. The research done by Hambleton & Traub (1971) compared the information functions and relative efficiency of the one-, and two-, and three-paremeter logistic modesl for item calibration using simulated test items. Their results showed that the three-parameter model was more informative than the one-parameter model, although the relative efficiency of the one-parameter model to the three-parameter model was high until the range of discrimination in the item became large.

Paper presented at the annual meeting of the American Educational Research Association, 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 research performed by Urry (1977) also depended upon simulated test data. Urry compared the quality of ability estimates obtained from the one, two, and three parameter logistic models, when the discrimination and guessing parameters of the simulated items were varied. The criterion used for evaluating the models was the correlation between tailored testing ability estimates obtained from the models and the true ability used to operate the simulations. His results showed that the one parameter logistic model was seriously affected by the presence of guessing in the simulated items and was also affected, to a lesser extent, by the variation in discrimination parameters.

Both of these studies reflect negatively on the one-parameter logistic model, although the practical importance of the deficits present in the model have not been made clear. The conclusion drawn on the basis of these studies very obviously would be to recommend the usage of the three parameter model for true-to-life applications. However, the generalizability of the simulation results to live testing situations can be questioned, particularly in that the simulation studies used very idealized item pools and errors induced in the calibration process were not a factor. Therefore, it is the purpose of the research reported here to extend the comparison of these two models to real data with reasonable sample sizes and to evaluate the models on both theoretical and practical grounds.

Models and Programs

In evaluating these two latent trait models for use in item calibration and ability estimation, the models themselves cannot be separated from the computer program used to compute item and ability parameter estimates. A model that is theoretically optimal may yield poor results because the program used to estimate the parameters is inaccurate. Seven one-parameter and six three-parameter logistic model calibration procedures were reviewed before selecting the two procedures used in this study. Descriptions of the thirteen procedures and the selection process are given in Reckase (1977).

One-parameter logistic model

The one-parameter logistic model in exponential form is given by the formula:

$$P\{x_{ij}\} = \frac{e^{x_{ij}(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}; \quad x_{ij} = 0, 1 \qquad (1)$$

where x_{ij} is Person j's score on Item i, θ_j is the ability parameter for Person j, and b_i is the difficulty parameter for Item i. All items are assumed to be equally discriminating by this model and guessing is assumed to have no effect on the item score. The model also assumes a unidimensional latent trait and local independence. The values of both the ability and difficulty parameters in this model range from positive to negative infinity.

The program used to estimate item and ability parameters for the one-parameter logistic model is based on the program written by Wright and Panchápakesan (1968), and was obtained from Jerry Durovic of the New York Civil Service Department. Although the basic procedures used in the program are those developed by Wright & Panchapakesan, it has been extensively modified by the author so the responsibility for its accuracy lies there.

Three-parameter logistic model

The three parameter logistic model is given by the formula

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

where x_{ij} is Person j's score on Item i, c_i is the guessing parameter for Item i, D is the constatn 1.7 used to make the logistic ogive similar to the normal ogive, a_i is the discrimination parameter for Item i, θ_j is the ability parameter for Person j, and b_i is the difficulty parameter for Item i. This model assumes local independence and a unidimensional test, but it does not place any restrictions on the guessing and discrimination parameters as does the one-parameter model. The range of the ability and difficulty parameters of this model is from positive to negative infinity, the same as the one-parameter model.

The program used to obtain the item and ability parameter estimates for the three-parameter logistic model was the 1976 version of the LOGIST program (Wood, Wingersky & Lord, 1976). This program recognizes three score categories; correct, incorrect and omit. Although the program is based on maximum likelihood estimation principles, substituting a probability of correct equal to the reciprocal of the number of responses for omitted items caused the resulting likelihood functions to only approximate the actual functions. The technique has, therefore, been labeled a quasi-maximum likelihood procedure. Lord (1974) has shown that the quasi-maximum likelihood estimates converge to the maximum likelihood estimates when the sample is large and omits are not present. When omits are present, smaller variance estimates are obtained than if the usual maximum likelihood procedures were used.

Description of the Problem

In comparing the one- and three-parameter models for use in item calibration and ability estimation, five specific comparisons were made. These include: (a) the evaluation of the goodness of fit of each of the models to the item response data; (b) the determination of the relationship between the ability estimates and the item responses; (c) the determination of the predictive validity of the ability estimates from the models in some limited cases; (d) the estimation of the minimum sample size required for each model to calibrate tests; and (e) the determination of the relationship between the ability estimates obtained from the two models. The next section of this paper will describe each of these comparisons in detail.

Method

Goodness of Fit

The initial evaluation of the two models dealt with the question of which model fit the item response data better. Several goodness of fit tests have been used for this previously, but it was felt that problems existed in the approximations used and assumptions made by these methods. Therefore, a new statistic was developed for the purposes of this comparison.

This statistic is given by the following formula:

$$MSD_t = \frac{\sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{N} (x_{ij} - P_i(\theta_j))^2}{N} \right]}{n} \tag{3}$$

where MSD_t stands for the mean squared deviation for Test t, x_{ij} is the response to Item i by Person j, $P_i(\theta_j)$ is the probability of a correct response to Item i for Person j determined for the model of interest, n is the number of items, and N is the number of people. This statistic ranges from 0 to 1, with a low value being desirable. If every item on a test had zero discrimination, the MSD_t statistic for the test would be .25. Negatively discriminating items give a MSD value larger than .25.

The test MSD statistic was computed for each of the tests used in this study for both the one-parameter and three-parameter logistic models. The item parameters obtained from the calibration of the tests with the models were used to compute the probability of a correct response. Since the MSD statistics for the tests used for these analyses were approximately normally distributed, a two-way analysis of variance was performed on these test MSD values using the item MSD statistics as observations. The item MSD value is the term within the brackets in Equation 3. The two dimensions used in this analysis were models and tests. Post hoc comparisons were used to find specific differences in tests.

Relation of ability estimates to item response

A second analysis that evaluated the relationship between the two latent trait models and the item responses was the computation of the multiple correlation between the ability estimates obtained from a test and the sets of item responses from the same test. This was done to determine the variance in common between the responses and the ability estimates. The multiple correlation was computed for each test used in the study and the magnitude of the values was compared using the correlated t statistic to determine if there was a significant difference in the variance in the item response accounted for by the models.

Concurrent validity of ability estimates

For several limited cases, other test scores were available for the individuals taking the tests to be calibrated. Although the samples for these tests were relatively small, the opportunity to relate the ability estimates from the latent trait models to the other tests could not be passed by. Three separate samples were, therefore, used in correlating the ability estimates from the two models with these other tests. The resulting correlations were compared statistically to determine which model yielded the larger validity coefficient.

Sample size requirements

An important question that has only been touched upon in the research literature (Cypress, 1972) is the sample size required for accurate estimation of the parameters of the two models. To more thoroughly explore the sample size limitations of the models, seven samples of various sizes were drawn from the students taking a standardized test. Parameter estimates were obtained for each of these samples and the results were compared to the calibration results based on 2,997 cases using a squared deviation statistic. That is, for each of the item parameters derived using the two models, the smaller sample estimates were subtracted from the large sample values, the difference squared, and the results summed. The average squared differences for the parameters were compared across sample sizes using analysis of variance techniques to determine the minimum sample sizes that yield adequate parameter estimates.

Ability parameter comparisons

In order to determine whether the ability parameter estimates derived from the two models were measuring the same component, the ability estimates were correlated with each other, with the raw scores from the tests, and with factor scores on the first factor on the tests. These correlations were determined for each of the sixteen tests used in this study. The factor scores were generated from factor analyses using both phi and tetrachoric correlations.

Data Sources

Live testing data-sets

The sixteen data-sets used in this study are described in Table 1 along with the abbreviations used for each and the sample size used for calibration. The first eight of the data-sets listed were obtained from the administration of two types of tests to groups of students. One test used was the Missouri School and College Ability Test (MSCAT). Data from administration of this test throughout the state of Missouri was available for the 1975 and 1976 school years. The test is comprised of two subtests which were calibrated separately.

Insert Table 1 about here

The other type of live testing data available for calibration was obtained from the administration of four classroom examinations on use of standardized tests. The data was collected using a large undergraduate measurement course during the period from October 1975 to May 1977. Both the standardized and classroom tests were fifty item, multiple-choice tests.

Along with these data-sets, seven other samples were obtained from MSCATV6 to determine sample size effects. Systematic sampling was used, yielding samples of 2,997, 2,197, 1,525, 1,090, 763, 382, and 150.

Simulation data sets

In order to gain greater control over the characteristics of the data, eight simulated test data-sets were produced. These were generated to match various factor loading matrices using the usual linear factor analysis model. The simulation procedure generated z-scores for each person on each item using a weighted sum of normal random numbers and then dichotomized them to yield the proportion of correct and incorrect responses specified by the traditional difficulty indices. Guessing did not enter into the production of the simulated data-sets. A sample of 1,000 cases was generated for each of the eight simulated tests.

Four levels of factorial complexity were used in generating these data-sets: one-factor, two-factor, five-factor, and nine-factor. The size of the factor loadings and distribution of difficulties were also varied for the simulated tests. Normal, rectangular, and constant distributions of difficulties were used, although no attempt was made to include all possible combinations. The distribution of difficulties referred to here is based on the proportion correct index.

Results

Goodness of fit

The test MSD statistic for the sixteen data-sets for each of the models are presented in Table 2 along with the analysis of variance results. The analysis of variance performed on this data was a two-way analysis with repeated measures on one dimension. The independent variables were test and type of logistic model

Insert Table 2 about here

The results of the analysis of variance show that the three-parameter model fits the data significantly better than the one-parameter model, although the difference in the overall means is only .004. However, for every data-set the average deviation from fit was smaller for the three-parameter model than for the one-parameter model. The MSD values were also found to be significantly different across tests. The one-factor data-set (150AR) was fit best by the models, as would be expected, and the nine-factor data-set (950AN3) had the worst fit, also as expected. No significant interaction was found in the data.

To further rank the tests in terms of fit of the models, the Newman-Keuls post hoc comparison procedure was used to determine if there were significant differences in the fit of specific tests. The results of this analysis are presented at the bottom of Table 2. As can be seen from the results presented there, the 150AR data-set is fit by the models significantly better than any of the other tests. This is the one simulated test that meets all of the assumptions of both models. It contains only one factor, all of the items are equally discriminating, and no guessing is present.

The 250AR data-set has the next best fit for the models. It has two factors, a wide range of item difficulties, and no guessing. Although the fit for this test is significantly worse than for 150AR, it is significantly better than all but one of the other tests. The majority of the other data-sets are fit about equally well by the two models.

At the poor fitting end of the continuum are three sets of simulation data: 550AN7, 950AN9, and 950AN3. All of these simulated tests have a relatively large number of independent factors. Data-set 950AN3 is the worst fitting of the tests, having a MSD statistic very close to the value of .25 expected when all items have zero discrimination. This simulated test has low loadings (.3) on the nine independent factors.

The trend of this analysis suggests that the multidimensionality of the tests is a definite factor in the fit of the two models. The three-parameter logistic model handles this deviation from the assumptions significantly better than the one-parameter model, but the ordering of the effect is the same as is shown by the lack of a significant interaction.

Relation of ability estimates to item responses

In order to determine the relationship between ability estimates and item responses, the multiple correlation between the ability estimates from each model and the fifty item responses was computed. These values are presented in Table 3 for the ability estimates correlated with the items from the sixteen data-sets. Note that all of the correlations with the one-parameter ability estimates are extremely high, as they must be because of the sufficient statistic properties of the model. The multiple correlations are high for the three-parameter ability estimates when a dominant factor is present, but drop when independent, equally weighted factors are present.

Insert Table 3 about here

A related t-test was performed on the mean multiple correlations for the two ability estimates to determine if the observed differences were significant. The difference in the mean multiple-correlations of .07 is significant at beyond the .005 level, indicating that the three-parameter ability estimate correlations are significantly lower.

Concurrent validity of ability estimates

The concurrent validity of the ability estimates for the two models was determined by correlating the estimates obtained from the final exam from three different semesters of an undergraduate measurement course with the first and second exams in the same semester. The correlations between the ability estimates and the raw scores on the criterion measures are presented in Table 4. In all but one case, the one-parameter ability estimates have higher correlations with the criteria than the three-parameter estimates. However, in no case were the differences in correlations for the two models significant. One reason for the slightly lower correlations for the three-parameter model could be the small sample size used in this analysis which will be shown later to affect the three-parameter model more than the one-parameter model, causing unstable estimates.

Insert Table 4 about here

Sample size requirements

The average squared deviations for the item parameter estimates from the seven subsamples as well as the squared deviations obtained from a second 2,997 sample are presented in Table 5 along with the ANOVA results used to determine if any significant differences existed. One-way repeated measures analyses of variance were performed using the squared difference values for the fifty items as the dependent measures with sample size as the independent variable.

Insert Table 5 about here

The means of three of the four sets of item parameters give a similar pattern of results. The 2,997 sample has the smallest mean squared deviation, while the deviations tend to get larger with decreasing sample size. This relationship is strong for the one-parameter easiness parameter and the three-parameter discrimination parameter, while the three-parameter difficulty and guessing parameters show considerable variation. The analysis of variance results show significant differences in all cases except for the three-parameter difficulty parameter. In that case, although there are large differences in the means, the large variation in the estimates resulted in a failure to reject. A F-max test for heterogeneity of variance yielded a value of 2,527 easily rejecting the hypothesis of homogeniety. A subsequent analysis on values after a logarithmic transformation yielded a significant F.

The purpose of this set of analyses was to determine at what point a decrease in sample size would adversely affect the results of item calibration. This question was addressed directly in a post hoc analysis performed using the ANOVA results. Using the mean squared deviation values for each sample size, the Newman-Keuls post hoc procedure was used to determine the largest sample that was significantly different from the mean squared deviation from the 2,997 sample. The results of these analyses are also presented in Table 5. Samples that are not significantly different are underlined. Those that are different do not share the same underline.

Due to the great variation in the 3PL difficulty values, the results of this study were not easily interpreted, indicating the need for further research. However, some general conclusions can be drawn from the data. The 1PL easiness parameters seem to have stabilized when the sample size is greater than 382. A sample somewhere between 382 and 763 is probably the lower limit required when using this model. The 3PL data are harder to interpret. The 3PL discrimination parameters seem to be moderately stable above the 150 sample, but the mean square deviations for the 3PL difficulty values are far from stable, with values for the

2,997 sample of about the same size as squared deviations for the 1PL easiness parameter for the 382 sample. Although these values are not on precisely the same scale, the values should be somewhat comparable. This result suggests that the 3PL difficulty parameters are just starting to stabilize. The heterogeneity of variance in the analysis of the difficulty parameters reduces its usefulness, however, the 150 sample is clearly worse than the rest. Overall the results suggest that substantially larger samples are required for the 3PL model. The guessing parameter does not enter into this discussion because of the numerous restrictions placed upon it in the calibration program.

Ability parameter comparisons

The correlations between the ability parameter estimates for the two models with the raw scores and selected factor scores for the tests are given in Table 6. In seven of the eight live testing data-sets, the correlations between the ability estimates from the two models are .90 or above. There is much greater variation in the simulation data, probably due to the multi-factor nature of the tests. However, even there the correlations are high when a dominant first factor is present.

Insert Table 6 about here

The correlations with the raw scores on the tests and the first factor scores are uniformly high for the live testing data, although the one-parameter model generally has slightly higher correlations than the three-parameter model. Again, there is greater variation for the simulation data, with lower correlations for the three-parameter model when no dominant first factor is present. The one-parameter model always correlates highly with the raw score because the raw score is a sufficient statistic for the ability parameter.

In general, the results show that in most cases the two models are measuring the same thing, the first factor of the test. When a dominant first factor is not present, there are major differences in the correlations. Reckase (1977) discusses these differences in much greater detail than can be done here.

Discussion and Conclusions

Five comparisons were made in the study reported here relative to the quality of the estimates of parameters obtained from the one- and three-parameter logistic models. The results can be summarized briefly as follows: (a) the three-parameter model fit the test data better in all cases than the one-parameter model and there was a trend in the fit related to the dimensionality of the test; (b) the one-parameter model ability estimates shared more variance with the item responses than the three-parameter model; (c) there was no difference in the concurrent validity for small samples using the two models predicting classroom achievement tests; (d) the one-parameter model required smaller samples for calibration than the three-parameter model; and (e) the ability estimates from the two models correlated highly for most of the data-sets.

From these results, certain conclusions can be drawn concerning the use of these two models with fifty item group exams when sample sizes of approximately two hundred are available. First, from the ability estimate comparisons, it seemed that the two models estimated the same latent trait when there was a dominant first factor, even when it accounted for a small amount of the variance. The concurrent validity data also supported this point of view, since the magnitude of the correlations were essentially the same. Since the sample size required to obtain stable parameters was smaller for the one-parameter model and the overall representation of the data was better as reflected by the multiple correlations, the one-parameter model is preferred for use with small sample group data to predict outside criterion variables.

The goodness of fit data reflected a different point of view, however. The three-parameter model fit all the data-sets better than the one-parameter model. This result may be important when accurate estimation of the item parameters is important such as in the area of tailored testing. A tailored testing comparison of the two models done by Koch & Reckase (1978) supports this point of view, showing the three-parameter procedure to yield superior results to the one-parameter procedure for a tailored testing application.

Although this research does give valuable information that will be helpful in selecting between these two latent trait models, much further research is required. Specifically, validity studies based on larger samples and other criterion variables are needed to allow generalization of the findings. Also the sample size determinations need to be more precise than those reported here.

Table 1

Description of Data-Sets*

Test Name	Abbreviation	Sample Size	Description
1. Missouri School and College Ability Tests Verbal/1975	MSCATV5	3,087	Systematic sample from 57,800 cases from Missouri Statewide Testing Program 1974-1975. SCAT Series II Form 2B.
2. Missouri School and College Ability Tests Quantitative/1975	MSCATQ5	3,087	Systematic sample from 57,800 cases from Missouri Statewide Testing Program 1974-1975. SCAT Series II Form 2B.
3. Missouri School and College Ability Tests Verbal/1976	MSCATV6	3,126	Systematic sample from 65,600 cases from Missouri Statewide Testing Program 1975-1976. SCAT Series II Form 2B.
4. Missouri School and College Ability Tests Quantitative/1976	MSCATQ6	3,126	Systematic sample from 65,600 cases from Missouri Statewide Testing Program 1975-1976. SCAT Series II Form 2B.
5. Exam on Standardized Testing.	ST1075	208	Undergraduate course final exam administered in October 1975.
6. Exam on Standardized Testing.	ST0576	181	Undergraduate course final exam administered in May 1976.
7. Exam on Standardized Testing	ST1076	176	Undergraduate course final exam administered in October 1976.

*All tests are 50 items in length.

Table 1 (Continued)

Description of Data-Sets

Test Name	Abbreviation	Sample Size	Description
8. Exam on Standardized Testing	ST3-577	312	Undergraduate course final exam administered to two sections of the course in March and May 1976.
9. One factor rectangular simulation data.	150AR	1,000	One factor with loadings of .9, rectangular distribution of difficulties.
10. Two factor normal simulation data.	250AN	1,000	Loadings of .9 and .0 randomly distributed on two factors, normal distribution of difficulties.
11. Two factor rectangular simulation data.	250AR	1,000	Loadings of .9 and .0 randomly distributed on two factors, rectangular distribution of difficulties.
12. Two factor .5 simulation data.	250A5	1,000	Loadings of .9 and .0 randomly distributed on two factors. All items .5 difficulty
13. Nine factor Spearman simulation data.	950ANS	1,000	One factor .7 loadings for all items. Eight factors .6 loadings randomly distributed over items. Normal disribution of difficulties
14. Nine factor independent .9 loading simulation data.	950AN9	1,000	Items randomly distributed to nine factors with .9 loadings. Normal distribution of difficulties.

Table 1 (Continued)

Description of Data-Sets

Test Name	Abbreviation	Sample Size	Description
15. Nine factor independent .3 loading simulation data.	950AN3	1,000	Items randomly distributed to nine factors with .3 loadings. Normal distribution of difficulties.
16. Five factor independent .7 loading simulation data.	550AN7	1,000	Items randomly distributed to five factors with .7 loadings. Normal distribution of difficulties.

Table 2

Squared Deviations from the Two Models for the Sixteen Data-Sets

Test	One Parameter Logistic	Three Parameter Logistic	Test Means
1. MSCATV5	.169	.166	.167
2. MSCATQ5	.164	.160	.162
3. MSCATV6	.169	.166	.167
4. MSCATQ6	.166	.161	.163
5. ST1075	.144	.138	.141
6. ST0576	.167	.165	.166
7. ST1076	.159	.154	.156
8. ST3-577	.184	.182	.183
9. 150AR	.068	.067	.068
10. 250AN	.162	.153	.158
11. 250AR	.122	.115	.118
12. 250A5	.185	.176	.180
13. 950ANS	.156	.156	.156
14. 950AN9	.211	.204	.208
15. 950AN3	.223	.222	.222
16. 550AN7	.210	.206	.208
Model Means	.166	.162	.164

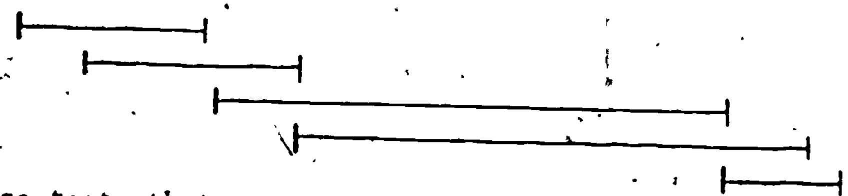
Anova Table

Source	Sum of Squares	d.f.	Mean Square	F	Significance
Tests	1.995	15	.133	31.667	.001
Items within tests	3.301	784	.004		
Models	.007	1	.007	14.684	.001
Tests X Models	.003	15	.0002	.414	
Models X Items within tests	.355	784	.0005		

Post Hoc Comparisons Using Newman-Keuls Test

Poor FIT	Test	Good FIT

15. 14. 16. 8. 12. 1. 3. 6. 4. 2. 10. 7. 13. 5. 11. 9.



Note: Those tests that are not underlined by the same line are significantly different from each other

Table 3

Multiple Correlations Among
Ability Estimates and Test Items

	Ability Estimate		
Test	1PL	3PL	1PL-3PL
NSCATV5	.991	.983	.008
NSCATQ5	.998	.985	.003
NSCATV6	.993	.988	.005
NSCATQ6	.991	.983	.008
ST1075	.994	.944	.050
ST0576	.993	.952	.041
ST1076	.985	.967	.018
ST3-577	.996	.985	.011
150AR	.990	.997	-.007
250AN	.981	.677	.304
250AR	.991	.948	.043
250A5	.978	.839	.139
250ANS	.983	.949	.034
950AN9	.998	.852	.146
950AN3	.9998	.890	.1098
550AN7	.998	.866	.132
Mean	.9906	.9253	.07149

$t = 3.705$ $p < .005$

Table 4

Correlations between Ability Estimates
and Two Classroom Tests

Data Set	N	Ability Estimate	Test Exam 1	Exam 2
ST1076	176	1PL	.555	.661
		3PL	.492	.599
ST0576	181	1PL	.409	.477
		3PL	.364	.483
ST1075	208	1PL	.558	.576
		3PL	.498	.535

Table 5

Comparison of Parameter Squared Deviations for the Two Models by Sample Size

Sample Size	Parameter 1PL Easiness	3PL Difficulty	3PL Discrimination	3PL Guessing
150	.0483	.1811(.1326)[a]	.2187	.0014
382	.0196	.1413(.0847)	.0973	.0009
763	.0063	.0272(.0258)	.0615	.0020
1090	.0063	.1930(.0821)	.0585	.0009
1525	.0055	.0299(.0263)	.0589	.0012
2197	.0047	.0138(.0135)	.0335	.0011
2997	.0041	.0166(.0162)	.0241	.0008

[a] Transformed means using log(x+1).

ANOVA 1PL Easiness

Source	d.f.	SS	MS	F	P
Samples	6	.0791	.0132	18.17	<.0001
Error	294	.2133	.0007		

ANOVA 3PL Difficulty

Source	d.f.	SS	MS	F	P
Samples	6	2.009	.335	1.50	N.S.
Error	294	65.643	.223		

ANOVA 3PL Discrimination

Source	d.f.	SS	MS	F	P
Samples	6	1.303	.217	7.30	<.0001
Error	294	8.743	0.030		

ANOVA 3PL Guessing

Source	d.f.	SS	MS	F	P
Samples	6	0.000055	.000009	3.44	<.003
Error	294	0.000787	.000003		

Table 5 (cont.)

ANOVA Transformed 3PL Difficulty

Source	d.f.	SS	MS	F	P
Samples	6	0.627	.104	3.61	<.002
Error	294	8.506	0.029		

Post Hoc Comparisons

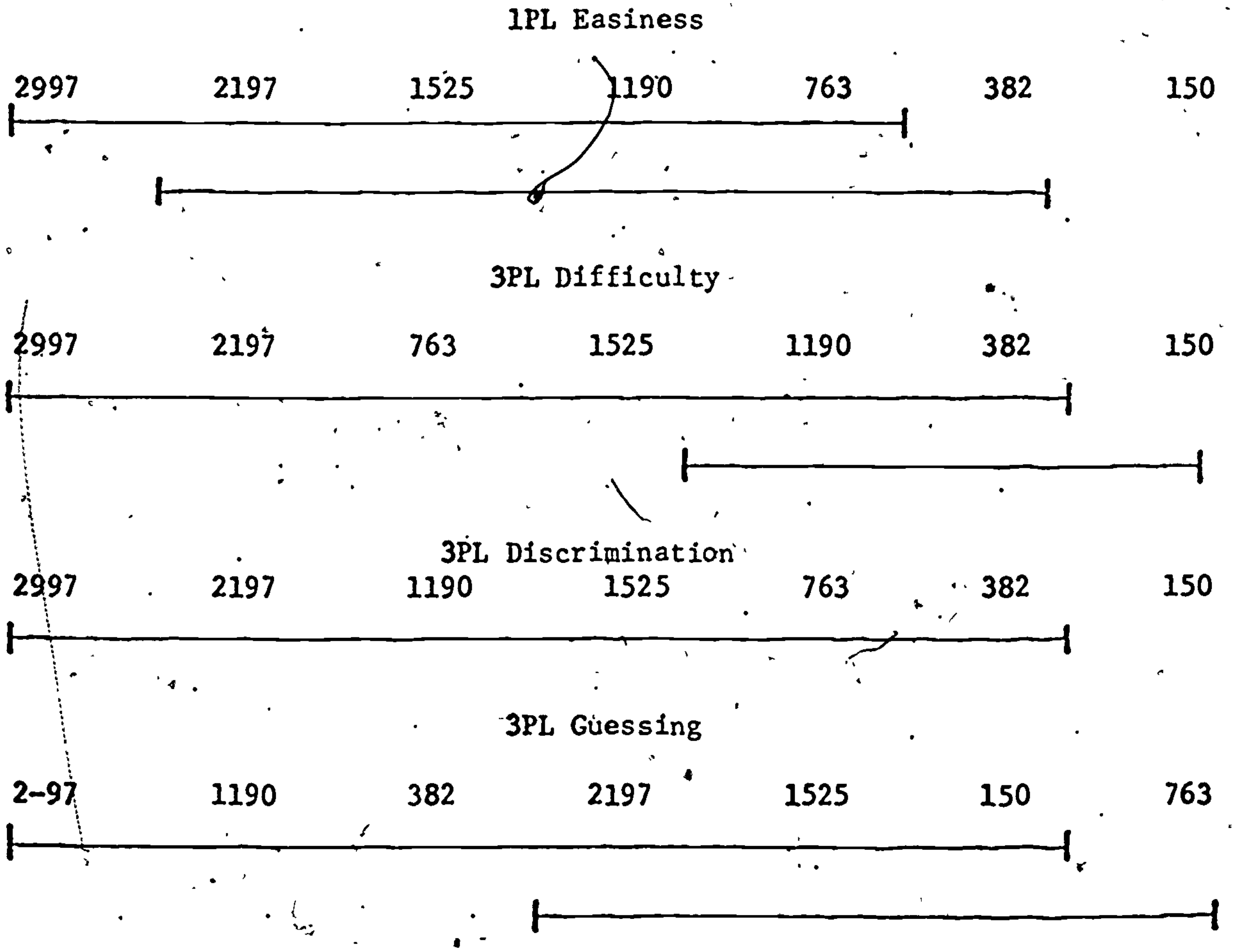


Table 6

Correlation between Ability Estimates,
Raw Scores, and Factors for the Sixteen Data Sets

	Variable				
Data-set	Ability Estimate	Raw Score	3PL Ability	Phi Principal Component	Tet Principal Component
MSCATV5	3PL	97		98	98
	1PL	99	96	97	97
MSCATQ5	3PL	97		98	98
	1PL	99	97	97	97
MSCATV6	3PL	98		99	99
	1PL	99	97	98	98
MSCATQ6	3PL	97		98	98
	1PL	99	96	97	97
ST1075	3PL	83		89	32
	1PL	99	85	89	29
ST0576	3PL	88		91	87
	1PL	99	90	93	88
ST1076	3PL	89		94	91
	1PL	98	90	88	86
ST3577	3PL	95		98	98
	1PL	99	95	97	97
150AR	3PL	97		97	98
	1PL	95	99	95	97
250AN	3PL	59		59	56
	1PL	98	66	98	97
250AR	3PL	71		69	92
	1PL	99	73	99	74
250A5	3PL	82		56	62
	1PL	98	83	76	83
950ANS	3PL	93		93	94
	1PL	98	96	98	98
950AN9	3PL	62		82	67
	1PL	99	62	72	72
950AN3	3PL	71		36	41
	1PL	100	71	25	33
550AN7	3PL	70		46	36
	1PL	100	70	32	27

Note: All values presented without decimal points.

REFERENCES

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

Cypress, B. K. The effects of diverse test score distribution characteristics on the estimation of the ability parameter of the Rasch measurement model. (Doctoral dissertation, The Florida State University) Ann Arbor, Michigan: University Micorfilms, 1972. (No. 72-32, 756).

Hambleton, R. K. & Traub, R. E. Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.

Koch, B. R. & Reckase, M. D. A live tailored testing comparison study of the one and three parameter logistic model. Paper presented at the meeting of the National Council on Measurement in Education, Toronto, March 1978.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

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

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).

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

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