DOCUMENT RESUME

ED 228 332 TM 830 278

AUTHOR McKinley, Robert L.; Reckase, Mark D.
TITLE The Use of IRT Analysis on Dichotomous Data from Multidimensional Tests.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Apr 83
GRANT N00014-81-K0814
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis; Difficulty Level; *Goodness of Fit; Higher Education; *Latent Trait Theory; *Mathematical Models; Simulation; Statistics; Test Items; *Tests
IDENTIFIERS American College Testing Program; Item Parameters; *Multidimensional Approach; Test Analysis; Three Parameter Model; *Two Parameter Model

ABSTRACT

Real test data of unknown structure were analyzed using both a unidimensional and a multidimensional latent trait model in an attempt to determine the underlying components of the test. The models used were the three-parameter logistic model and a multidimensional extension of the two-parameter logistic model. The basic design for the analysis of the data was to start with the unidimensional model, evaluate the fit of the model to the data, then increase the dimensionality of the model and perform the same analyses. The dimensionality of the model was to be increased until deviations from fit were acceptably small. Once acceptable fit was obtained, the item parameter estimates were to be analyzed to determine the structure of the ability components required by the test. The results of the analyses indicated that a two-dimensional solution yielded no better fit than a unidimensional solution, although the test data were selected to be multidimensional. From the results of the analyses it was concluded that the data had a difficulty factor that was sufficiently great as to dominate the other, intended factors. (Author)

The Use of IRT Analysis on Dichotomous Data from Multidimensional Tests

Robert L. McKinley and Mark D. Reckase

The American College Testing Program

There are many situations in the practice of testing in which data are collected from tests that measure more than a single trait. Indeed, a strong argument can probably be made that dimensionally complex measuring devices are the rule rather than the exception. If the items in the test are dichotomously scored, the analysis of these data poses special problems, since the methodology for the analysis of these data is not well developed. The purpose of this paper is to demonstrate the capabilities of a relatively new methodology that has been developed in the last 25 years specifically to analyze test data. This methodology is variously called item response theory (IRT), item characteristic curve theory (ICC), or latent trait theory. In addition to the use of the already well known one-dimensional IRT models, this paper will introduce the use of a multidimensional extension of the two-parameter logistic model (M2PL) that is a special case of an extremely general model proposed by Rasch (1961). Using these analysis models, an attempt will be made to describe a set of test data of unknown structure and determine the underlying components of the test.

Design of the Analysis

The data that were analyzed in this study were selected from the responses of individuals to items from tests available at the American College Testing Program (ACT). Bob Brennan selected the response data used in the study specifically for this symposium. He was instructed to produce a multidimensional data-set from the responses to existing tests that would serve as a good check on the capabilities of several multidimensional analysis techniques. The result was a set of 2794 dichotomous response strings of 50 items in length. No information about the characteristics of the data was given to the participants of the symposium before this presentation other than that given above.

The basic design for the analysis of the test data was to start with a one-dimensional model, evaluate the fit of the model to the data, then increase the dimensionality of the model and perform the same analyses until deviations from fit became negligible. Once acceptable fit was obtained, the item parameter estimates that were computed using the model would then be analyzed to determine the structure of the ability components required by the test.

Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983. This research was supported by Contract No. N00014-81-K0814 from the Personnel and Traing Research Programs of the Office of Naval Research.

Two different one-dimensional models were used in the study. The first was the three parameter logstic model (3PL) (Birnbaum, 1968) given by the equation

$$P(x_{ij} = 1 \mid a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}, \quad (1)$$

where $P(X_{ij}=1 \mid a_i, b_i, c_i, \theta_j)$ is the probability of a correct response to Item i by person j, X_{ij} is the item score (0 or 1) obtained by Person j on Item i, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the lower asymptote, θ_j is the ability of Person j, D is the constant 1.7 required to make the function similar to the normal ogive model (Lord, 1952), and e is the constant 2.718....

The second one-dimensional model that was used in the study was the two-parameter logistic model (2PL) (Birnbaum, 1968) given by

$$P(x_{ij} = 1 \mid a_i, d_i, \theta_j) = \frac{e^{d_i + a_i\theta_j}}{1 + e^{d_i + a_i\theta_j}}, \quad (2)$$

where d_i is the logistic intercept term equal to $-b_i a_i$ and all of the other symbols are defined as above. The model in Equation 2 is given in the slope-intercept form rather than with the usual exponent, $a_i(\theta_j - b_i)$, so as to be more readily compared with the multidimensional model.

The multidimensional model used in this study is a special case of the general Rasch model (Rasch, 1961). It is also a multivariate extension of the two-parameter logistic model (M2PL). The M2PL model is given by the equation

$$P(x_{ij} = 1 \mid \underline{a}_i, d_i, \underline{\theta}_j) = \frac{e^{d_i + \underline{a}_i' \underline{\theta}_j}}{1 + e^{d_i + \underline{a}_i' \underline{\theta}_j}}, \quad (3)$$

where $\underline{a}_i$ is a vector of discrimination parameters for Item i, $\underline{\theta}_j$ is a vector of ability parameters for Person j, and the other symbols are defined above.

The fit of the models to the data used in this study was determined using the residual covariance matrix. This matrix was obtained by computing the covariance between the differences between the item response and the predicted probability of correct response based on the item and person parameters estimated from the data. The distribution of the residual covariances was then compared to the distribution expected if no relationship were present between the items. If the distributions were sufficiently different, it was hypothesized that added dimensions were needed in the model to explain the responses.

In addition to the direct analysis of the real data, supplementary analyses were performed on simulated test data that were generated to have the same three-parameter logistic item parameters as the real data. These analyses were performed to serve as a unidimensional basis for comparison in interpreting the analyses performed on the real data.

Descriptive statistics were also computed on the item and person parameters obtained from the IRT analyses. These statistics included the means, standard deviations, and correlations between the parameter estimates. Where appropriate, distributions of the item parameters were constructed.

Results

One-Dimensional Analyses

Two different IRT models were used to analyze the test data assuming a one-dimensional latent space. These models were the three-parameter logistic (3PL) model and the two-parameter logistic (2PL) model. The results of the 3PL analysis to both the real and simulated data will be presented first.

Three-Parameter Logistic Analysis The 3PL analysis of the real data was performed using the 1982 version of LOGIST (Wingersky, Barton, and Lord, 1982). The item parameter estimates obtained from the analysis and traditional item statistics are presented in Table 1. Note that the items were originally arranged on the test according to the proportion of correct responses. This order is also maintained by the 3PL b-parameter estimates except in the cases where the a-parameter estimates are small, or the c-parameter estimates are large. The magnitude of the item parameter estimates are typical of those found on a standardized test, with the exception of Items 16, 18, 30, and 32, which have unusually high c-parameter estimates. An analysis of the distribution of the a- and b-parameter estimates indicated that the a-parameters tended to have a few more low values than expected, and that the b-parameters had an essentially rectangular distribution.

The correlations between the parameter estimates and traditional item statistics are given in Table 2. This table contains several noteworthy correlations. First, the correlation between the a-parameter estimates and the biserial correlation discrimination index is only .02. The small magnitude of this correlation is quite unusual, since for unidimensional test data these values should be quite highly related. The unusual nature of this correlation is further emphasized by the fact that both the a-values and the r_{BIS}-values are correlated with the proportion correct difficulty index (P), but the correlations are opposite in sign. The correlations indicate that the a-parameter and r_{BIS} are related to independent components of the variation in P.

Table 1

LOGIST 3PL Item Parameter Estimate for the AERA Symposium Data

Item	a	b	c	p	r_{BIS}
		Parameter			
1	.65	-2.64	.16	.93	.51
2	.46	-2.37	.16	.86	.39
3	.53	-2.10	.16	.86	.46
4	.98	-1.41	.16	.86	.66
5	.88	-1.30	.16	.83	.61
6	.33	-2.49	.16	.82	.34
7	.96	-1.03	.16	.80	.64
8	1.10	- .96	.16	.79	.69
9	1.07	- .96	.16	.79	.68
10	.52	-1.45	.16	.79	.44
11	.83	- .99	.16	.77	.60
12	.93	- .86	.16	.76	.62
13	.77	- .92	.16	.76	.58
14	.60	-1.08	.16	.75	.50
15	.73	- .90	.16	.75	.56
16	1.29	.11	.50	.74	.49
17	1.21	- .79	.03	.73	.75
18	1.11	.19	.50	.72	.48
19	.68	- .79	.16	.72	.54
20	.40	-1.11	.16	.71	.38
21	.91	- .57	.16	.70	.62
22	1.08	- .50	.16	.70	.67
23	.67	- .66	.16	.70	.52
24	.56	- .62	.16	.68	.47
25	1.32	- .22	.23	.67	.67

Table 1 (Continued)

LOGIST 3PL Item Parameter Estimate for the AERA Symposium Data

Item	a	b	c	p	r_{BIS}
		Parameter			
26	1.00	-.23	.19	.65	.62
27	.69	-.36	.16	.64	.52
28	.72	-.29	.16	.63	.55
29	1.02	-.25	.13	.63	.66
30	1.10	.93	.43	.57	.37
31	1.13	.09	.17	.56	.63
32	.64	1.47	.44	.56	.28
33	.82	.15	.15	.54	.56
34	.85	.41	.22	.53	.52
35	.91	.26	.14	.52	.58
36	1.09	.41	.13	.46	.62
37	.78	1.38	.32	.46	.34
38	1.08	.79	.22	.44	.50
39	.62	.42	.05	.44	.51
40	.84	.59	.12	.43	.54
41	.53	1.27	.14	.37	.38
42	1.30	1.22	.23	.37	.46
43	1.04	1.15	.18	.34	.46
44	2.00	1.27	.23	.33	.37
45	1.00	1.33	.14	.29	.43
46	1.17	1.85	.21	.27	.28
47	.83	1.45	.09	.26	.45
48	1.19	1.64	.15	.24	.36
49	1.31	1.89	.17	.23	.29
50	1.50	2.57	.13	.15	.13
$\bar{x}$	.91	-.10	.18	.60	.50
SD	.31	1.22	.10	.20	.13

Table 2

Correlations Between 3PL Item Parameter Estimates and Traditional Item Statistics

	Item Statistics				
	a	b	c	P	r_{BIS}
a		.54*	.21	-.46*	.02
b			.24	-.94*	-.50*
c				.03	-.30*
P					-.54*
r_{BIS}					

Note: The correlations with asterisks are significant beyond the .05 level.

Other correlations of interest are the -.30 correlation between the c-parameter value and the r_{BIS} values, and the correlation of .54 between the a- and b-parameters. Neither of these correlations are surprising. Items with high guessing levels would be expected to have low values for r_{BIS}, and the a- and b-parameters have been found to be correlated in other studies.

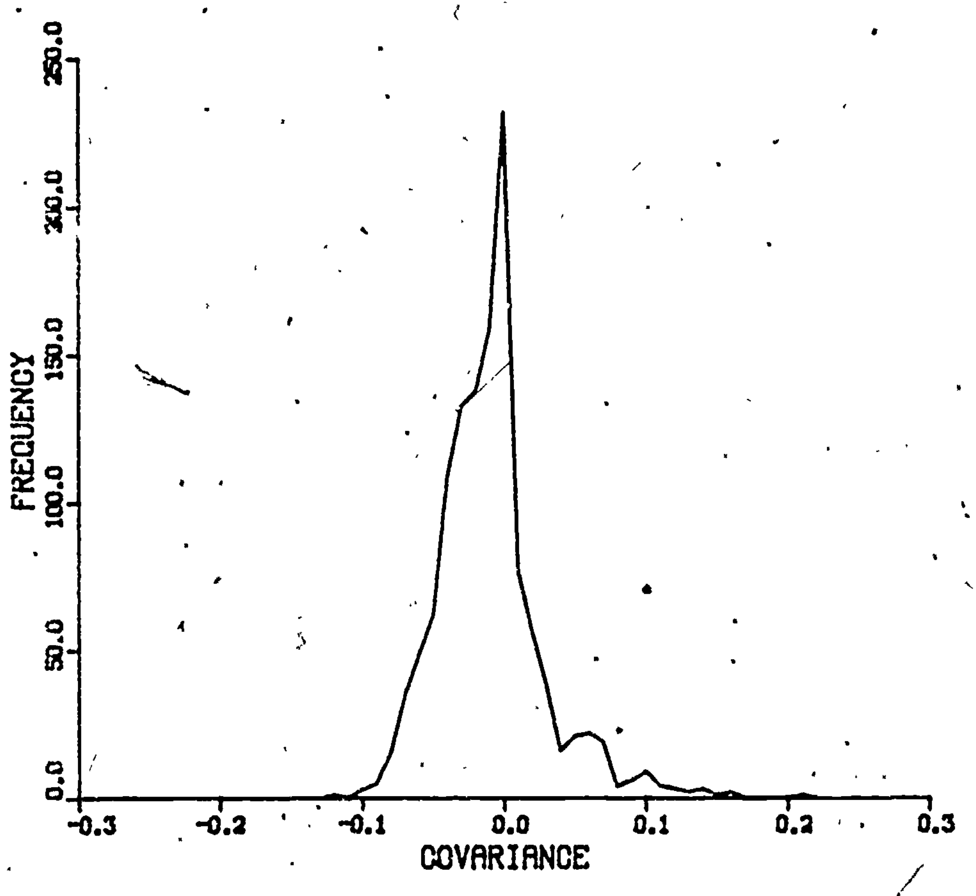
In order to determine whether the 3PL model fit the data reasonably well, the residual covariance matrix was computed. This matrix contains the covariance between the residuals for each item and every other item. In order to interpret this matrix, a frequency distribution was constructed using the values from above the main diagonal of the matrix. This distribution is shown in Figure 1. Descriptive statistics for the distribution are also shown in Figure 1.

As can easily be seen from the figure, the distribution of residual covariances is quite positively skewed, and the majority of the values fall between 0 and -.1. This range contains approximately 77% of the values.

If the data truly fit this unidimensional model, we would have expected a symmetric distribution around zero with a standard deviation of approximately .019. The standard deviation of the observed distribution is .037, substantially larger than this value.

Figure 1

Frequency Distribution of the Residual Covariances for the 3PL Model Applied to the Real Data



$\bar{x} = -.01$

SD = .04

Since little background information is available on how to interpret the residual covariance matrix; a simulated data set was produced for the purpose of obtaining comparative analysis results based on unidimensional data. This data set was generated using the 3PL model and the item parameter estimates obtained from the analysis of the ACT data. A total of 1000 simulated examinees were generated for this purpose.

A new set of parameter estimates was determined from the simulated test data using the LOGIST program, and these estimates were compared to those used to generate the data and also to traditional item statistics computed on the data. These correlations are shown in Table 3. The pattern of correlations obtained from the analysis of the simulation data was very similar to that obtained from the real data. The only change in the pattern of significant correlations is that for the simulation data, the a-parameter and c-parameter estimates were correlated while they were not for the real data. Since the magnitude of the two correlations is fairly similar (.21 vs. .33) this may be a chance result. Note that the correlation between the a-parameter estimate and the item-biserial correlation is low for these data, just as it was for the real data. This result may indicate that the unexpectedly small magnitude of this correlation is an artifact of the particular range of difficulty and discrimination present in this data-set.

Table 3

Correlations Between the True Item Parameters, Estimated Item Parameters, and Tradition Item Statistics for the Simulated Test Data

	Item Statistics							
	a	b	c	est. a	est. b	est c.	P	r_{BIS}
a		.54*	.20	.69*	.54*	-.00	-.47*	.05
b			.22	.42*	.98*	.08	-.94*	-.44*
c				.23*	.14	.75*	-.01	-.30*
est. a					.47*	.33*	-.29	.10
est. b						.08	-.92*	-.39*
est. c							.14	-.33*
P								.47*
rBIS								

Note: The correlations with astericks are significant beyond the .05 level.

The frequency distribution for the residual covariance matrix for the simulated data is given in Figure 2 along with the descriptive statistics. The distribution is fairly symmetric around -.01 and does not have the long positive tail present in the distribution for the real data. The standard deviation of the residual covariances is .03, exactly what would be predicted based on a hypothesis of no linear relationship among the residuals. This should be contrasted with the big differences between the expected and observed standard deviations for the real data.

Two-Parameter Logistic Analysis The 2PL analysis of the real data was performed using the MAXLOG program (McKinley and Reckase, in press), which was written specifically for test analysis using the M2PL model. Since the 2PL model is the unidimensional case of the M2PL model, the program applies equally well for a one-dimensional solution. The item parameter estimates for the 2PL model are given in Table 4. In addition to the a- and d-parameters of the model, this table presents the b-parameter estimates for those individuals that are more familiar with that form of the 2PL model.

Both the d- and b-parameter estimates from the 2PL analysis roughly maintain the order of the items as shown by the proportion correct difficulty values shown in Table 1. The a-values also seem to be related to the difficulty of the items in that the higher discrimination parameter estimates were obtained for the easier items. The correlations of the 2PL parameter estimates and the 3PL and traditional statistics, shown in Table 5, support this observation. The 2PL a-parameter estimates are correlated .66 with the d-parameter estimates, -.54 with the b-parameter estimates, and .59 with the p-values.

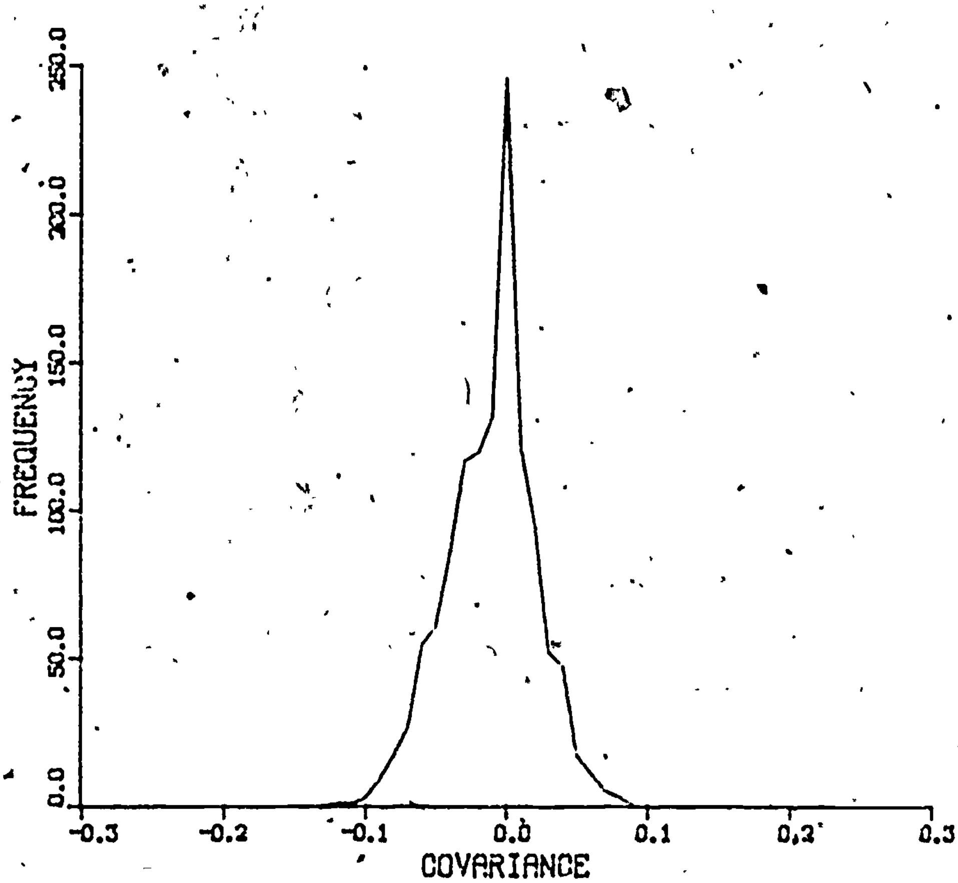
The 2PL a-parameter estimates are clearly more closely related to the traditional concept of item discrimination than are the 3PL a-parameter estimates. The 2PL a-parameters correlate .95 with r_{BIS} and only .09 with the 3PL a-parameters. The difference in the two IRT discrimination estimates seems to be mainly in the values computed for the hard items and those with high guessing. Items 16, 30, 42, 44, 46, 48, 49 and 50 are good examples of the differences present in the two types of estimates.

The 2PL a-parameter estimates also correlated with the 3PL c-parameter estimates. The -.34 correlation obtained is consistent with the idea that items with high guessing levels should be low discriminators. The d-parameter estimates correlate highly with the 3PL b-parameters and the p-values. There are also smaller correlations with the measures of discrimination.

The frequency distribution and the descriptive statistics for the 2PL residual covariance matrix are shown in Figure 3. The distribution looks very similar to the 3PL residual covariance matrix and the descriptive statistics are virtually identical. Based on these results, the two models would be considered to fit the data equally well.

Figure 2

Frequency Distribution of the Residual Covariances for the 3PL Model Applied to the Simulated Data



$\bar{x}' = -.01$

SD = .03

Table 4

MAXLOG 2PL Item Parameter Estimate
for the AERA Symposium Data

Item	a	d	b
	Parameter		
1	1.35	3.20	-2.37
2	.82	2.05	-2.50
3	.97	2.10	-2.16
4	1.93	2.72	-1.41
5	1.66	2.29	-1.38
6	.59	1.59	-2.69
7	1.66	1.97	-1.19
8	2.02	2.18	-1.08
9	1.90	2.09	-1.10
10	0.88	1.51	-1.72
11	1.41	1.65	-1.17
12	1.58	1.67	-1.06
13	1.29	1.48	-1.15
14	1.01	1.35	-1.34
15	1.22	1.40	-1.15
16	.94	1.24	-1.32
17	2.30	1.85	- .80
18	.89	1.12	-1.26
19	1.12	1.19	-1.06
20	.63	.99	-1.57
21	1.48	1.22	- .82
22	1.75	1.32	- .75
23	1.07	1.04	- .97
24	.87	.87	-1.00
25	1.69	1.11	- .66
26	1.44	.86	- .60
27	1.06	.75	- .71
28	1.09	.69	- .63
29	1.61	.80	- .50
30	.58	.31	- .53
31	1.51	.40	- .26
32	.35	.25	- .71
33	1.10	.22	- .20
34	.97	.15	- .15
35	1.21	.10	- .08
36	1.33	-. 17	.13
37	.46	-. 19	.41
38	.92	-. 26	.28
39	.96	-. 27	.28
40	1.06	- .33	.31

Table 4 (Continued)

MAXLOG 2PL Item Parameter Estimate
for the AERA Symposium Data

Item	Parameter a	Parameter d	Parameter b
41	.60	- .57	.95
42	.81	- .60	.74
43	.81	- .74	.91
44	.65	- .79	1.22
45	.80	-1.00	1.25
46	.43	-1.04	2.42
47	.81	-1.04	1.51
48	.58	-1.24	2.14
49	.46	-1.30	2.83
50	.19	-1.79	9.42
$\bar{x}$	1.10	.69	- .26
SD	.47	1.17	1.83

Table 5

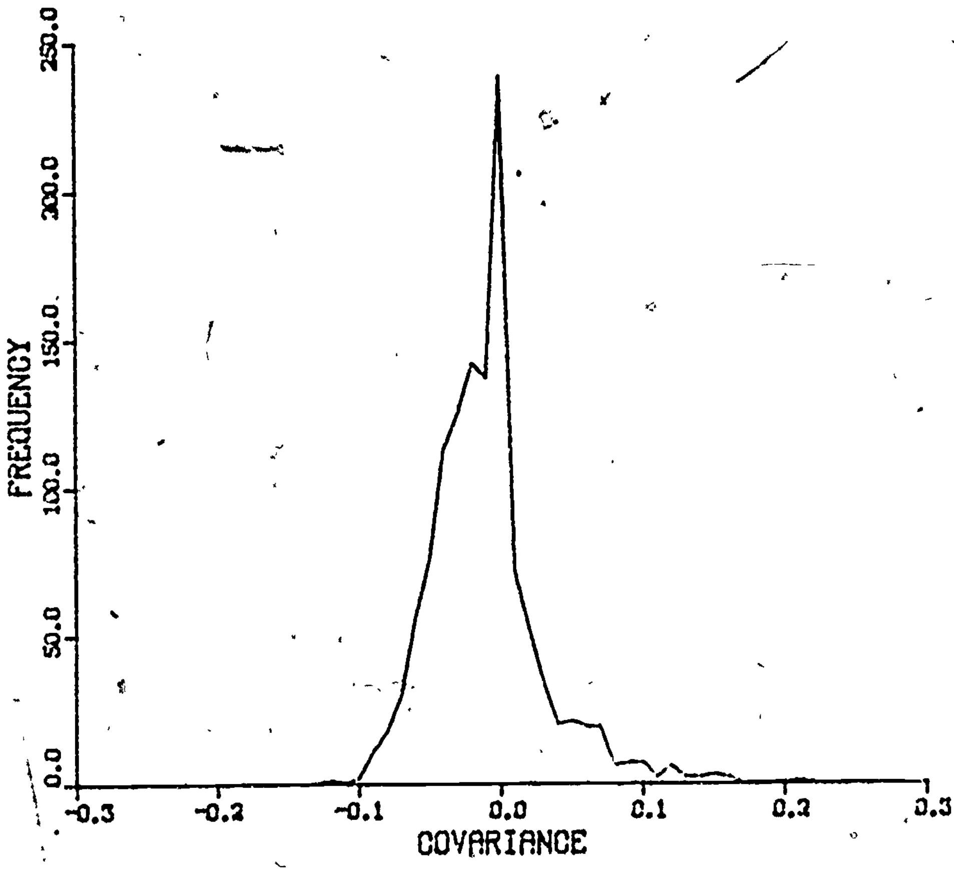
Correlations Between the Tradition Item Statistics
and the 2PL and 3PL Parameter Estimates
for the AERA Symposium Data

	Item Statistics						
	2PL-a	2PL-d	3PL-a	3PL-b	3PL-c	P	r_{BIS}
2PL-a		.66*	.09	- .54*	- 34*	.59*	.95*
2PL-d			- .38*	- .93*	- 04*	.98*	.59*
3PL-a				.54*	.2[illegible]	- .47*	.0[illegible]
3PL-b	—				.24	- .34*	- .50*
3PL-c						.03	- .30*
P							.54*
r_{BIS}							

Note: The correlations marked by asterisks are significant beyond the .05 level.

Figure 3

Frequency Distribution of the Residual Covariances
for the 2PL Model Applied to the Real Data



$\bar{x} = -.01$

SD = .04

Two-Dimensional Analysis

Since the standard deviation of the residual covariance matrix was larger than would be expected if the models fit the data for both the 3PL and 2PL models, it was assumed that a higher dimensional solution was required. The M2PL model was, therefore, rune on the AERA symposium data assuming a two-dimensional solution. The item parameter estimates from this analysis are given in Table 6. Notice the d-parameter estimates still decrease uniformly with the increase in the item number, showing that the item difficulty estimates are still closely related to the p-values. A cursory study of the a-parameter estimates will show that in many cases they are very similar. The estimation program does not place any constraints on the relationship of the two dimensions being estimated and as a result, all of the a-parameters could be the same if the data being analyzed so indicated.

Table 6
Item Parameters from a Two-Dimensional MAXLOG Analysis of the AERA Symposium Data

Item	Parameter		
	a_1	a_2	d
1	1.05	1.39	2.72
2	.95	1.56	2.01
3	.89	1.30	1.90
4	1.94	2.37	2.04
5	1.65	2.00	1.77
6	.85	.80	1.54
7	1.91	2.38	1.55
8	2.36	2.50	1.56
9	2.31	2.50	1.55
10	1.00	1.83	1.49
11	1.63	2.50	1.53
12	1.71	2.22	1.30
13	1.55	2.04	1.24
14	1.21	1.70	1.20
15	1.54	2.01	1.19
16	1.34	1.44	1.10
17	2.50	2.50	1.20
18	1.47	1.44	1.02
19	1.43	2.08	1.05
20	.78	1.04	.92
21	1.80	2.47	.97
22	2.50	2.50	1.00
23	1.41	1.70	.88
24	1.08	1.41	.64
25	2.50	2.50	.82

Table 6 (Continued)

Item Parameters from a Two-Dimensional MAXLOG Analysis of the AERA Symposium Data

Item	Parameter		
	$\hat{a}_1$	a_2	d
26	1.88	2.41	.64
27	1.20	2.26	.66
28	1.40	2.04	.56
29	2.50	2.50	.60
30	1.14	.97	.29
31	2.50	2.50	.25
32	.48	.48	.23
33	1.98	2.04	.13
34	1.64	1.65	.08
35	2.50	2.50	.01
36	2.50	2.50	- .24
37	.62	.64	- .21
38	1.86	1.70	- .30
39	1.37	1.59	- .33
40	1.58	2.34	- .46
41	.94	1.04	- .60
42	1.86	1.37	- .66
43	1.36	1.24	- .76
44	1.77	.99	- .94
45	1.37	1.25	-1.02
46	1.30	1.42	-1.27
47	1.31	1.24	-1.23
48	1.01	.75	-1.28
49	1.08	.40	-1.48
50	.38	.03	-1.86
$\bar{x}$	1.54	1.70	.50
SD	.57	.70	1.06

In order to gain a better understanding of the parameter estimates obtained from the two-dimensional solution, the estimates were correlated with those obtained from the 3PL and 2PL analyses. These correlations are shown in Table 7. The correlations between the ability estimates obtained from the M2PL, 3PL, and 2PL models were also computed. These correlations are given in Table 8.

Table 7

Correlations Between the M2PL, 3PL, and 2PL Item Parameter Estimates

		Parameter Estimates							
		M2PL			3PL			2PL	
Model	Parameter	a_1	a_2	d	a	b	c	a	d
M2PL	a_1		.81*	.16	.40*	-.14	-.24*	.80*	.22
	a_2			.52*	-.04	-.50*	-.33*	.88*	.55*
	d				-.45*	-.94*	-.00	.60*	.99*
3PL	a					.54*	.21	.09	-.38*
	b						.24	-.54*	-.93*
	c							-.34*	-.04
2PL	a								.66*
	d								

Note: The correlations with asterisks are significant beyond the .05 level.

Table 8

Correlations between the M2PL, 3PL, and 2PL Ability Parameter Estimates

		Parameter Estimates			
		M2PL		3PL	2PL
Model	Parameter	θ_1	θ_2	θ	θ
M2PL	θ_1		-.91	.14	.20
	θ_2			.18	.13
3PL	θ				.97
2PL	θ				

Note: Because of the large sample size, all correlations are significant.

The correlation between the d-parameter estimates from the M2PL two-dimensional solution and the d-parameter estimates from the 2PL model (.99) indicated that this parameter is essentially the same for the two models. The a-parameters have changed, however. While the a_1- and a_2-parameter estimates were highly related to each other and to the 2PL a-parameter estimates, they had quite a different relationship with the item difficulty parameters and with the 3PL a-parameter estimates. The a_1-parameter was somewhat related to the 3PL a-parameter, while the a_2-parameter was not. The a_2-parameter was related to the difficulty parameter estimates, while a_1 was not. It would seem then that the M2PL a-parameters were dividing up the variation in the parameters of the 3PL model. Both a_1 and a_2 were related to the 3PL c-parameter.

The high correlation between the a_1- and a_2-parameter estimates is of special note. When these parameters are the same, it indicates that the two ability dimensions are required in equal proportions in responding to the test item. If all of the a_1-parameters were equal to the a_2-parameters, only one dimension would be needed in the model -- that dimension would be the sum of θ_1 and θ_2. When a_1 and a_2 are different for an item, different amounts of ability on each dimension are required.

The correlations between the ability estimates gave very interesting results. First, the M2PL ability estimates had a very high negative correlation (-.91). This fact, combined with the similarity of the a-parameter estimates, indicates that the exponent of the model can be approximated by $d + a_1\theta_1 - a_2\theta_2$.

Since a_1 and a_2 are equal in many cases and so are θ_1 and θ_2, the value of the exponent is mainly controlled by the d-parameter (the $a_1\theta_1-a_2\theta_2$ term is zero). This would seem to indicate that the data have one predominant dimension shown by the change in difficulty of the items, and that the effect of the other dimensions is minor.

A second interesting result is that the M2PL ability estimates have relatively low relationships with the 3PL and 2PL ability estimates. They are clearly an indicator of something different. The ability estimates from the 3PL and 2PL models are highly related to each other, as would be expected.

The fact that the M2PL item parameters and person parameters were highly interrelated suggests that the data may be predominantly unidimensional. To further check this hypothesis, the frequency distribution of the residual covariance matrix was formed for the two-dimensional solution for the M2PL model. This distribution is shown in Figure 4. This distribution was also produced for the application of the M2PL model to the one-dimensional simulation data in order to have a basis for comparison. This distribution is shown in Figure 5.

A comparison of these two distributions with those for the unidimensional models does not show any reduction in the variation of the residual covariances. In fact, the distribution for the unidimensional data has a larger variance when the M2PL model was used than when the 2PL model was used. This fact may suggest that the estimation of more parameters induces greater amounts of error in the parameter estimates.

Although there was no reduction of variances in the M2PL distribution when compared to the 3PL and 2PL distributions, there was a shift in the mean. Both of the unidimensional models yielded a means residual covariance of -.01. The M2PL model had a mean residual covariance of .02. While this difference is only .03, it is highly statistically significant ($z = 26.25$) because of the large number of observations. This larger mean may mean a poorer fit to the data for the M2PL two-dimensional model than for the 3PL or 2PL models. Because of lack of experience in interpreting the residual covariance matrix, this cannot be said for sure.

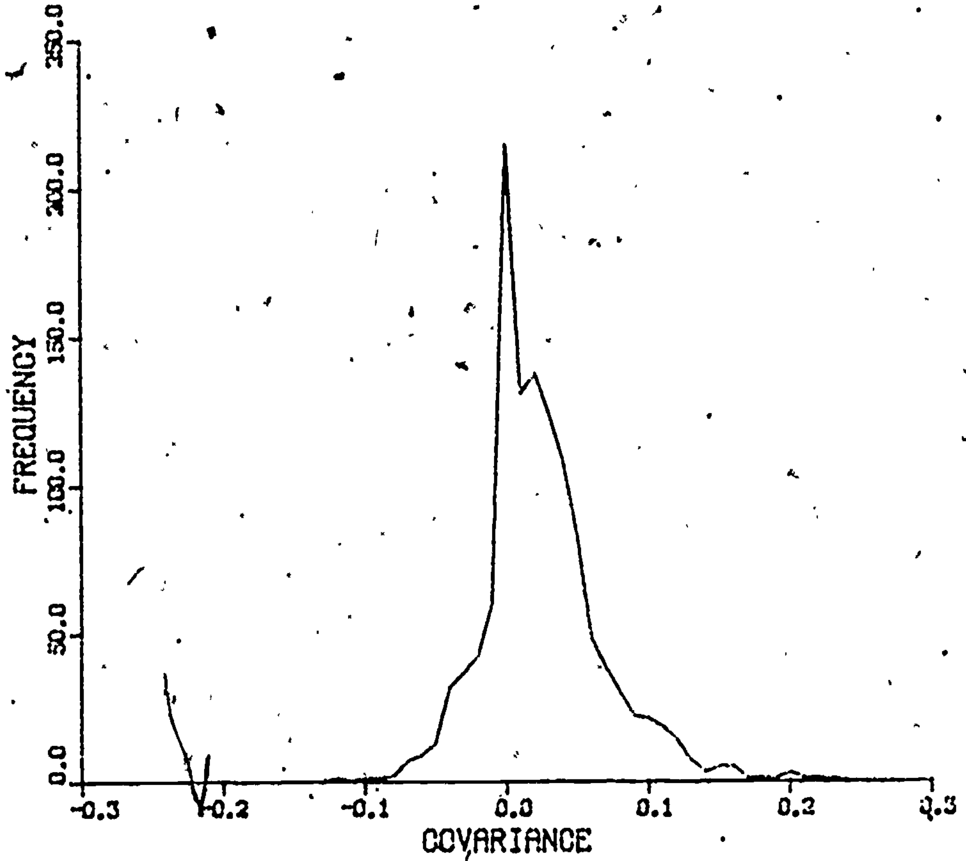
The Identification of Item Clusters

One of the purposes for performing the analyses on the data supplied for this symposium was to determine whether the M2PL procedure could be used to sort items into homogeneous clusters. Based on past experience with the 3PL model (Reckase, 1977), it was felt that items that had high a-values on both dimensions of the M2PL solution would be good measures of the dimension defined by the d-parameter. Those items that had low a-values probably measured some other dimension, and other combinations of a-values may indicate other item clusters.

To operationalize the above ideas, a cluster analysis was run using the two a-parameter estimates as observations. The Euclidean distance was used as a similarity measure. The BMDP1M (Dixon and Brown, 1977) program was used for the analysis. The cluster analysis resulted in four fairly distinct sets of items, although more clusters could certainly be obtained if they were thought warranted. The four clusters are shown in Figure 6.

Figure 4

Frequency Distribution of the Residual Covariances
for the M2PL Model Applied to the Real Data

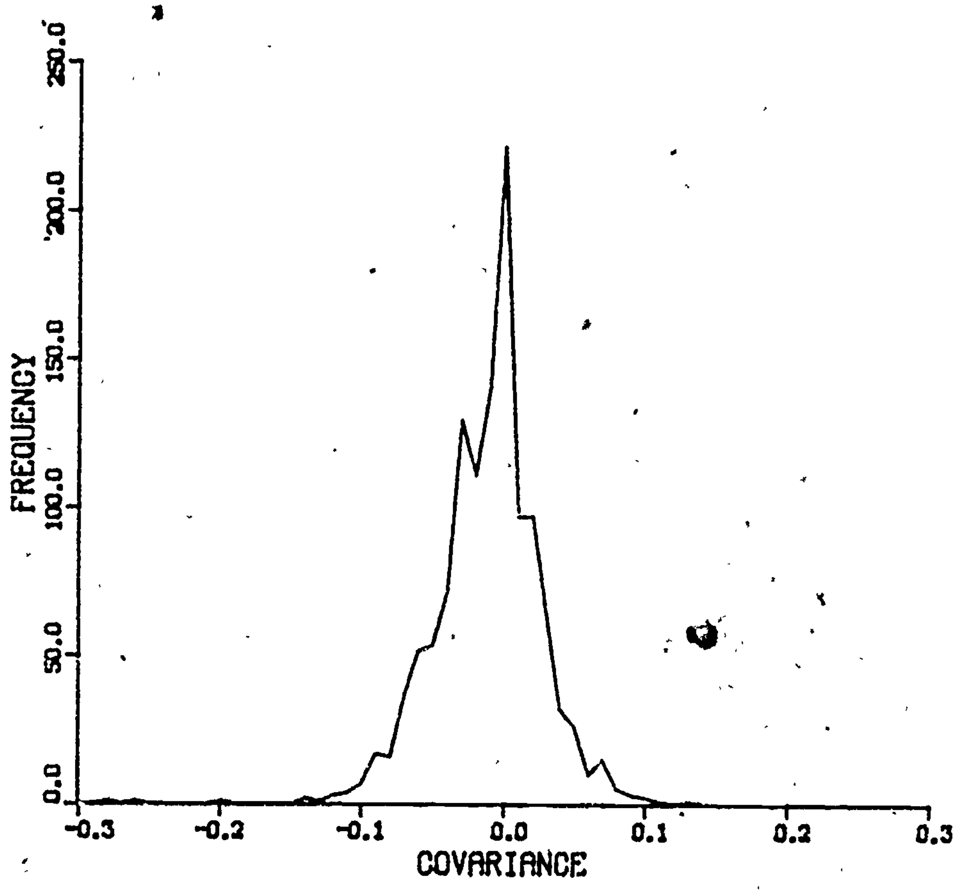


$\bar{x}$ = .02

SD = .04

Figure 5

Frequency Distribution of the Residual Covariances
for the M2PL Model Applied to the Simulation Data

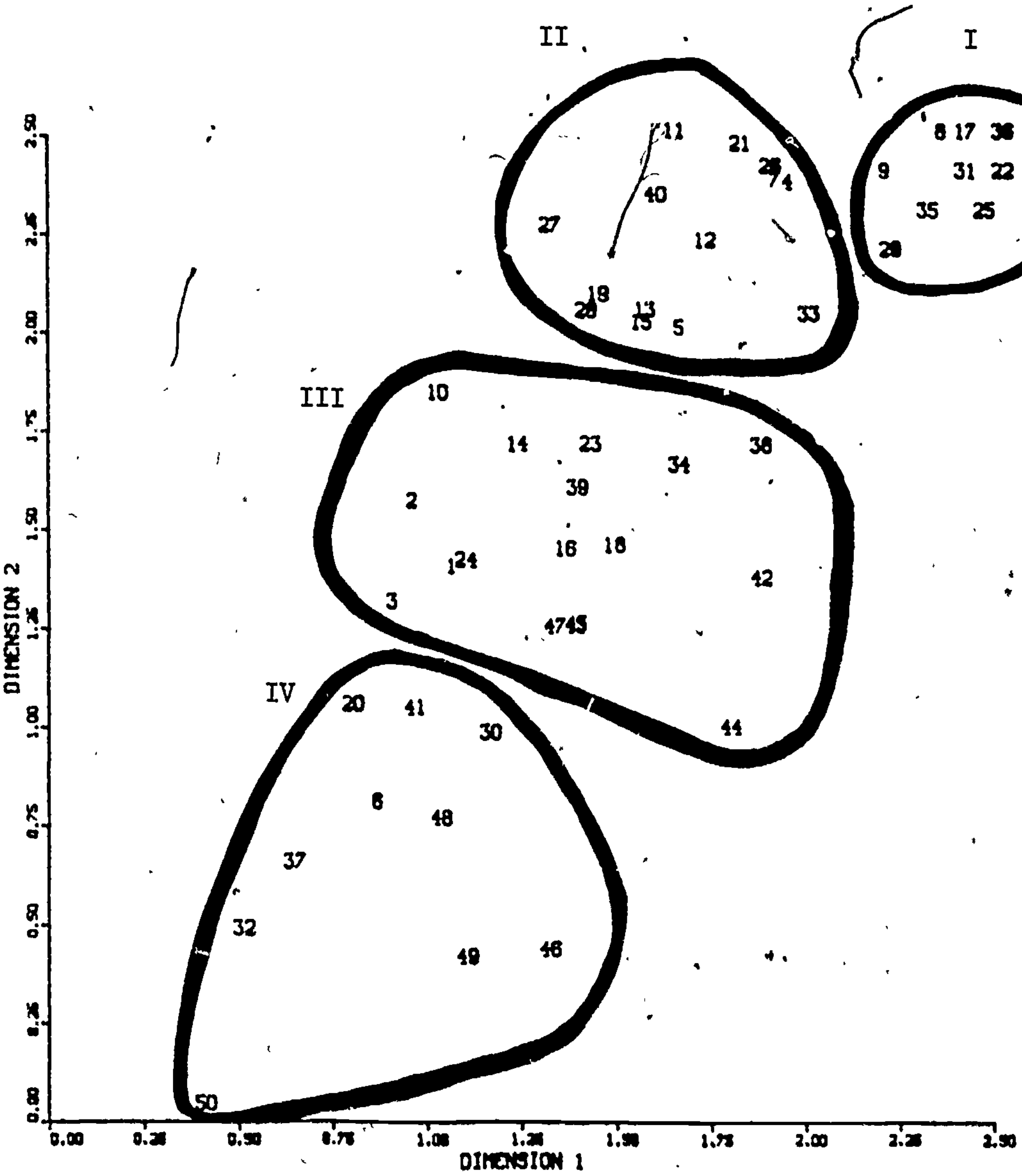


$\bar{x} = -.01$

SD = .04

Figure 6

Groups of Items Based on a Cluster Analysis of the M2PL Discrimination Parameter Estimates



The first cluster is made up of items that had high a-values on both dimensions. This cluster is very tight and distinctively different than the other item clusters present. The second cluster is made up of items that had fairly high discrimination parameter estimates for the second dimension, and slightly lower values for the first dimension. Recall that the second dimension was the one that had the a-values related to the item difficulty level.

The third cluster is composed of items that had middle range a-parameter estimates on both dimensions. The fourth cluster is composed of items that had relatively low discriminations on both dimensions. Item 50 was included with this cluster even though it seems like an outlier. It was the most difficult item on the test. Although these clusters seem reasonable from a statistical point of view, without knowing the content of the items it is impossible to tell the variables that control cluster membership.

Discussion

When the analysis of the AERA symposium test data was begun, the anticipated result was that two or more relatively distinct ability dimensions would be discovered and that the items on the test could be classified into content categories based on which dimensions were required for successful performance. This was not found to be the case. Rather than distinct ability dimensions, two highly correlated dimensions were developed. Rather then finding that the M2PL model fit the data better than the unidimensional models, the fit was found to be about equal if not worse.

Two possible conclusions come to mind based on these results. The first is that the estimation program for the M2PL model, or the model itself is inadequate. Although this is certainly a possibility, the fact that the estimation procedures did a good job of recovering true parameters in simulation studies and estimating parameters for multidimensional test data (McKinley, 1983) would argue against such an interpretation.

The second conclusion that comes to mind is that the test data really do have predominantly one dimension and that the results of the analysis reflect that fact. Even if the test does require multiple abilities, this fact may be clouded by the range of item difficulty present in the items. This may be a case similar to a test made up of three items: jump over a string, define "precipitation," and solve a differential equation. Although these items measure distinctly different skills, for a population that ranges widely in ability, the items will appear statistically to measure a single dimension. We anxiously await the information about the nature of the data the we have been analyzing to determine which conclusion is correct.

References

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

Dixon, W.J. and Brown, M.B. (Eds.) BMDP biomedical computer programs, P-series 1977. Berkeley, CA: University of California Press, 1977.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, 7.

McKinley, R.L. A multidimensional extension of the two-parameter logistic latent trait model. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

McKinley, R.L. and Reckase, M.D. MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. Behavior Research Methods and Instrumentation, in press.

Rasch, G. On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA: University of California Press, 1961, 4, 321-334.

Reckase, M.D. Ability estimation and item calibration using the one and three parameter logistic model: a comparative study (Research Report 77-1). Columbia, MO: University of Missouri, 1977.

Wingersky, M.S., Barton, M.A., and Lord, F.M. LOGIST user's guide. Princeton, NJ: Educational Testing Service, February 1982.