ED 230 624 TM 830 472

AUTHOR Hambleton, Ronald K.; And Others

TITLE Fitting Item Response Models to the Maryland Functional Reading Test Results.

PUB DATE Apr 83

NOTE 21p.; 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 Elementary Secondary Education; *Goodness of Fit; *Latent Trait Theory; Mathematical Models; State Programs; *Testing Problems; *Testing Programs; Test Items; *Test Results

IDENTIFIERS *Maryland Functional Reading Test; One Parameter Model; Residuals (Statistics); Three Parameter Model; Two Parameter Model; Unidimensionality (Tests)

ABSTRACT

The potential of item response theory (IRT) for solving a number of testing problems in the Maryland Functional Reading Program would appear to be substantial in view of the many other promising applications of the theory. But, it is well-known that the advantages derived from an IRT model cannot be achieved when the fit between an item response model and the test data of interest is less than adequate. The principal purpose of the research reported in this paper was to investigate the fit of the one-, two-, and three-parameter logistic models to the test results obtained from the administration of the 1982 Maryland Functional Reading Test (MFRT). The evidence addressing model-data fit seemed clear: a two-parameter logistic model was able to adequately account for examinee performance on the MFRT. The one-parameter model could not handle the substantial variation among test items in their discriminating power. The three-parameter model improved the fit only slightly because of the minimum amount of guessing on the test. Several suggestions were offered in the paper for conducting goodness-of-fit investigations. (Author)

Fitting Item Response Models to the Maryland Functional Reading Test Results

Ronald K. Hambleton and Linda Murray
University of Massachusetts, Amherst

and

Paul Williams
Maryland Department of Education

Abstract

The potential of item response theory (IRT) for solving a number of testing problems in the Maryland Functional Reading Program would appear to be substantial in view of the many other promising applications of the theory. But, it is well-known that the advantages derived from an IRT model cannot be achieved when the fit between an item response model and the test data of interest is less than adequate. The principal purpose of the research reported in this paper was to investigate the fit of the one-, two-, and three-parameter logistic models to the test results obtained from the administration of the 1982 Maryland Functional Reading Test.

The evidence addressing model-data fit seemed clear: A two-parameter logistic model was able to adequately account for examinee performance on the MFRT. The one-parameter model could not handle the substantial variation among test items in their discriminating power. The three-parameter model improved the fit only slightly because of the minimum amount of guessing on the test. Several suggestions were offered in the paper for conducting goodness-of-fit investigations.

Fitting Item Response Models to the Maryland Functional Reading Test Results[1,2]

Ronald K. Hambleton and Linda Murray
University of Massachusetts, Amherst

and

Paul Williams
Maryland Department of Education

The potential of item response theory (IRT) for solving a number of testing problems in the Maryland Functional Reading Program would appear to be substantial in view of the many other promising applications of the theory (see, for example, Hambleton, 1983; Lord, 1980). But, it is well-known that the advantages derived from an IRT model cannot be achieved when the fit between an item response model and the test data of interest is less than adequate. The principal purpose of the research reported in this paper was to investigate the fit of the one-, two-, and three-parameter logistic models to the test results obtained from the administration of the 1982 Maryland Functional Reading Test.

Method

Test Description and Use

In the Fall of 1982 the Maryland Functional Reading Test - Level II was given to approximately 54,000 ninth graders. The Level II test

[1] A paper presented at the annual meeting of NCME, Montreal, 1983.

[2] Laboratory of Psychometric and Evaluative Research Report No. 139. Amherst, MA: School of Education, University of Massachusetts, 1983.

consisted of 75 operational items from five content domains. These five areas, (1) Following Directions, (2) Locating Information, (3) Main Idea, (4) Using Details, and (5) Understanding Forms, are the units used for reporting diagnostic scores to teachers and parents. An overall test scale score of 340 represents the passing standard. This test must be passed before students are eligible for graduation from Maryland's public schools. If the certification requirement is not met in the ninth grade, the local school system is obligated by law to provide appropriate instructional assistance before retesting the student yearly.

Sample

From the (approximately) 54,000 students who were administered the test in the Fall of 1982, for the purpose of our analyses, a 5% "spaced sample" was drawn. Specifically, every twentieth student from the master student file was used. The resulting sample of 2662 students provided a sufficiently large sample to carry out the logistic model analyses on the data.

Analyses

The logistic model item and ability parameter estimates were obtained from the computer program LOGIST (Wingersky, Barton, & Lord, 1982). Next, the goodness of fit between the one-, two-, and three-parameter models and the test data was addressed with residuals, specifically, standardized residuals using a computer program prepared by Hambleton and Murray (1983). To obtain these standardized residuals, the ability scale was divided into 12 equal intervals between ability scores of -3.0 and +3.0. In each interval the difference between the actual item performance (p-value) of the examinees and the expected item performance obtained from the estimated item characteristic curve (icc) was divided by the standard error associated with the p-value to obtain a standardized residual (SR). A SR was obtained at each ability level for each test item. Since the direction of the differences was unimportant for many of the analyses, absolute-valued standardized residuals were typically used.

Results

The classical item analysis results and the absolute-valued standardized residuals (SRs) obtained with the three logistic models are reported in Table 1. Three comments based on a study of Table 1 seemed appropriate. First, and not surprising since the test was measuring competencies that many students were expected to be masters of, the average item performance was high (77.8% correct). This finding suggested that the "pseudo-chance level " parameter in the three-parameter model was apt to be of limited value in fitting a model to the data since guessing was an insignificant factor in test performance. Second, while some of the variation in the biserial correlations was due to the instability of these statistics with very easy items, there seemed to be a rather substantial variation in the discriminating power of test items. The biserial correlations varied from .15 to slightly over 1 (it is possible to obtain biserial correlations over 1). This preliminary finding suggested that the two-parameter model would probably fit the test data better than the one-parameter model. Third, a cursory analysis of the SRs in Table 1 showed, in fact, that the two- and three-parameter models produced highly comparable fits to the test data and, on the average, better fits to the data than the one-parameter model. The minor reversals in the SR values

Table 1

Maryland Functional Reading Test Item Statistics
(N=2662; 1982)

				Absolute-Valued Standardized Residuals		
Test Item	Proportion Correct	Biserial Correlation	Content Category[1]	1-p	2-p	3-p
1	.97	.74	1	0.92	0.57	0.62
2	.95	.59	1	.62	.64	.81
3	.88	.30	1	2.52	.84	.72
4	.91	.70	1	1.18	.80	.73
5	.94	.66	1	.83	.91	.61
6	.45	.36	1	2.87	1.70	1.35
7	.83	.59	1	.84	.61	.62
8	.94	.77	1	1.28	.79	.61
9	.73	.35	1	2.67	1.12	1.18
10	.88	.55	1	.61	.64	.59
11	.89	.34	1	2.00	.64	.76
12	.93	.70	1	1.04	.81	.83
13	.98	.67	1	.66	.75	.73
14	.79	.44	1	1.65	.70	.77
15	.86	.58	1	.88	1.29	.96
16	.78	.39	1	2.38	.95	.68
17	.91	.72	1	1.07	.67	.61
18	.74	.35	2	2.61	.62	.53
19	.90	.44	2	1.37	.89	.69
20	.95	.52	2	.69	.79	.48
21	.98	.67	2	.58	.59	.41
22	.93	.72	2	1.10	.74	.62
23	.79	.50	2	1.17	.63	.73
24	.87	.68	2	1.67	.97	.98
25	.86	.65	2	1.09	.89	.83

[1] Content categories: 1=Following Directions, 2=Locating Information, 3=Main Ideas, 4=Using Detail, 5=Understanding Forms.

Table 1 (continued)

Test	Proportion	Biserial	Content	Absolute-Valued Standardized Residuals		
Item	Correct	Correlation	Category	1-p	2-p	3-p
26	.57	.36	2	2.81	.86	.82
27	.83	.55	2	1.41	1.30	1.38
28	.84	.59	2	.66	.67	.53
29	.88	.70	2	1.37	1.01	1.05
30	.89	.77	2	1.53	.69	.72
31	.97	.80	2	.93	.72	.89
32	.88	.66	2	1.10	.78	.69
33	.87	.68	2	1.31	1.04	1.10
34	.55	.44	2	2.00	.69	.89
35	.59	.43	3	2.24	1.61	1.32
36	.75	.54	3	1.85	1.53	1.43
37	.70	.60	3	1.70	1.59	1.10
38	.23	.20	3	4.42	.65	.90
39	.71	.73	3	2.49	1.92	1.13
40	.71	.56	3	1.02	1.05	1.01
41	.57	.43	3	1.98	1.26	.94
42	.69	.62	3	1.51	1.26	.88
43	.55	.46	3	1.27	.89	1.03
44	.56	.52	3	1.86	1.51	1.40
45	.54	.60	3	1.68	1.59	.78
46	.70	.62	3	1.50	1.38	.97
47	.79	.70	4	1.57	.80	.84
48	.85	.65	4	1.45	1.22	.85
49	.88	.83	4	2.09	.80	.93
50	.93	1.03	4	2.92	1.09	1.02
51	.79	.68	4	1.06	.84	.83
52	.95	.98	4	2.11	.93	.81
53	.69	.62	4	1.20	.79	.86
54	.88	.66	4	.81	.81	.65
55	.94	.95	4	2.19	.87	.90
56	.87	.63	4	.92	1.02	1.05
57	.93	.91	4	2.15	.78	.71
58	.76	.63	4	1.19	1.15	1.00
59	.71	.51	4	1.35	1.41	1.37
60	.73	.62	4	1.13	.79	.83

Table 1 (continued)

Test	Proportion	Biserial	Content	Absolute-Valued Standardized Residuals		
Item	Correct	Correlation	Category	1-p	2-p	3-p
61	.74	.32	4	3.69	1.62	1.53
62	.31	.15	4	5.73	1.23	.94
63	.73	.55	4	1.14	.99	.91
64	.89	.76	5	1.34	.81	.74
65	.56	.55	5	.72	.90	.98
66	.81	.41	5	2.73	1.83	1.85
67	.71	.54	5	1.04	1.20	1.16
68	.75	.67	5	1.61	.84	1.05
69	.91	.94	5	2.72	.84	.95
70	.78	.67	5	1.09	.65	.59
71	.79	.70	5	1.34	.72	.69
72	.29	.36	5	2.00	.52	.55
73	.78	.66	5	.97	.70	.96
74	.75	.61	5	.57	.66	.82
75	.73	.65	5	1.29	.71	.84

were due to problems in parameter estimation and the tendency of SRs to "blow-up" in the low and high ability categories where the standard errors were often very small.

Table 2 provides the results of a more thorough analysis of the standardized residuals. When an item response model fits a set of test data, the standardized residuals should be distributed approximately normally. In fact, the distributions of the standardized residuals obtained with the two- and three-parameter logistic models were approximately normal. These results are especially interesting because some preference was given in test development to items that fit the one-parameter model. Unfortunately, the goodness of fit studies carried out in the test development stage were done with the BICAL program. But, it is now well-known that the goodness of fit tests in this computer program have problems (Divgi, 1981; van den Wollenberg, 1982). With the one-parameter model, about 30% of the SRs exceeded an absolute value of 2.0 whereas only about 5% would have been predicted had the model fit the test data.

Table 3 provides information pertaining to the fit of the three logistic models in 12 ability categories. With respect to bias as reflected in the average standardized residuals, the statistics from the three models were similar although the two- and three-parameter models produced slightly less bias in accounting for the data. With respect to overall fit, as reflected in the average absolute-valued standardized residuals, again, the two- and three-parameter models provided better fits to the data. Regardless of the ability level, the fits were substantially better with the more general models.

One of the two main assumptions of the three logistic test models is that of *unidimensionality*. One check on the validity of the assumption has

Table 2

Analysis of the Absolute-Valued Standardized Residuals[1]
With Three Logistic Test Models for the MFRT

Logistic Model	Percent of Absolute-Valued Standardized Residuals |0 to 1|	 |1 to 2|	 |2 to 3|	 |over 3|
1	42.6	27.8	15.0	14.6
2	60.6	29.7	7.3	2.4
3	63.3	29.6	6.0	1.1

[1] Total number of residuals is 825.

Table 3

Analysis of Standardized Residuals at Eleven Ability Levels with the One-, Two- and Three-Parameter Logistic Models for the MFRT (N=2662; 75 items)

Statistic	Logistic Model	Ability Level -2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	Total (unweighted)
Number	1	25	51	116	218	409	456	475	509	207	137	29	
of	2	16	43	99	242	429	531	481	374	219	116	57	
Examinees	3	22	50	100	224	406	528	491	387	228	117	49	
Average	1	.40	.30	.28	.28	.39	.30	-.02	.20	.27	.40	.38	.29
Standardized	2	.39	.38	.40	.29	.17	.01	-.05	-.04	.18	.33	.36	.22
Residual	3	.12	.31	.29	.28	.24	.09	-.05	-.05	.08	.34	.30	.18
Average	1	1.70	1.90	2.05	1.56	1.53	1.31	1.57	2.26	1.75	1.37	.68	1.61
Absolute-	2	1.22	1.06	1.19	.72	1.07	1.01	.70	.97	.93	.94	.76	.96
Valued	3	.98	1.07	1.11	.68	.97	.98	.64	.85	.84	.93	.72	.89
Standardized													
Residual													

to do with the pattern of residuals for test items classified by content. Test items within a content category may show a different pattern of residuals if they "tap" a different trait from the one measured by the items in the other content categories. Alternately, with the one-parameter model, a different pattern of residuals may also indicate the subset of test items has a relatively high or low average discriminating power in relation to the remainder of the items in the test although the items may measure the same trait as the other items in the test. Such an explanation however can not explain the results with the two- or three-parameter model since variation in item discriminating power can be handled by the models. A study of the statistics in Table 4 suggests that the "main idea domain" of test items may be tapping a separate trait from the remaining test items. A more careful review of the test suggests that this hypothesis may be reasonable since the 12 "main idea" items appear to be "tapping" reading comprehension whereas the other four content domains appear to be measuring study skills. The three-parameter model fits the 12 items by assigning "low discriminating powers" to these items and thereby reducing the importance of these items to the total test scores and corresponding ability estimates. But this strategy of handling "deviant" items is undesirable too. In subsequent work with the test, more attention should be focused on the unidimensionality assumption and ways for proceeding when the assumption is violated to a substantial degree.

Table 5 provides the results from another analysis of the SRs. This time, the average absolute-valued SRs were sorted by "easy" and "hard" items and reported for each model. Again, the improved fit obtained with the more general models was evident. The 2-P results were substantially better than the 1-P results regardless of the item difficulty levels. With

Table 4

Association Between Absolute-Valued Standardized Residuals and Items Content on the MFRT

		% of Standardized Residuals					
	Number	1-P		2-P		3-P	
Content Category	of Items	SR(≤1.0) (n=16)	SR(>1.0) (n=59)	SR(≤1.0) (n=50)	SR(>1.0) (n=25)	SR(≤1.0) (n=56)	SR(>1.0) (n=19)
Following Directions	17	41.2	58.8	82.4	17.6	88.2	11.8
Locating Information	17	23.5	76.5	82.4	17.6	82.4	17.6
Main Idea	12	0.0	100.0	16.7	83.3	41.7	58.3
Using Details	17	11.8	88.2	58.8	41.2	76.5	23.5
Understanding Forms	12	25.0	75.0	83.3	16.7	75.0	25.0
		χ^2 = 8.32		χ^2 = 19.24		χ^2 = 9.12	
		d.f.=4	p=.082	d.f.=4	p=.00	d.f.=4	p=.058

Table 5

Associations Between Absolute-Valued Standardized Residuals and Item Difficulties for the MFRT

Difficulty Level	Standardized Residual			Results			
		1-P		2-P		3-P	
		N	%	N	%	N	%
Hard ($p \leq .75$)	SR(≤ 1.0)	1	1.3	11	14.7	15	20.0
	SR(>1.0)	25	33.3	15	20.0	11	14.7
Easy ($p > .75$)	SR(≤ 1.0)	15	20.0	39	52.0	41	54.7
	SR(>1.0)	34	45.3	10	13.3	8	10.7
		$\chi^2 = 5.74$		$\chi^2 = 9.01$		$\chi^2 = 4.76$	
		d.f.=1	p=.017	d.f.=1	p=.003	d.f.=1	p=.029

the hard items, there was also a slight reduction in the SRs through the use of the three-parameter model. Since many of the so-called "hard items" were still relatively easy (p's $\geq$ .50) it was not surprising to observe the small impact of the "pseudo-chance level" parameter in the three-parameter model.

Table 6 provides another breakdown of the SRs. The results show that (1) the easy items were fit better by the item response models than the hard items, (2) the two- and three-parameter models fit the data in a similar fashion and both models fit the data better than the one-parameter model, and (3) the biggest improvements in fit through the use of the two- and three-parameter models were obtained with the harder test items.

In a final analysis, a close study of the relationships between SRs and biserial correlations was carried out since earlier analyses revealed improvements resulting from the addition of a discrimination parameter to the one-parameter model. The results in Table 7 and Figures 1 and 2 show dramatically the impact of the use of an item discrimination parameter in the chosen item response model. Items with low or high biserial correlations were not fit as well by the one-parameter model as the other two models. For example, the curvilinear relationship so apparent in Figure 1 vanished when the two-parameter model was fit to the test data.

Conclusion

The initial evidence adressing model-data fit seems clear: A two-parameter logistic model can adequately account for examinee performance on the MFRT. The one-parameter model did not handle the substantial variation among test items in their discriminating power. This finding is somewhat surprising since the original item pool had already been reduced somewhat

Table 6

Statistical Analysis of the Absolute-Valued Standardized Residuals for the MFRT

Difficulty Level	Number of Items	Results					
		1-P		2-P		3-P	
		$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD
Hard (p<.75)	26	2.07	1.15	1.15	.40	1.01	.25
Easy (p>.75)	49	1.37	.62	.86	.25	.83	.25

Table 7

Relationship Between Item Biserial Correlations and Standardized Residuals for the MFRT

Logistic Model	|Standardized Residual|		Item Biserial Correlation	
		.00 to .50	.51 to .70	.71 to 1.00
		(20)	(41)	(14)
1-P	0.00 to 1.00	0.0	34.1	14.3
	1.01 to 2.00	45.0	65.9	35.7
	over 2.00	55.0	0.0	50.0
		χ^2 = 31.74 Eta = .608	d.f.=4	p=.000
2-P	0.00 to 1.00	65.0	61.0	85.7
	1.01 to 2.00	35.0	39.0	14.3
	over 2.00	0.0	0.0	0.0
		χ^2 = 2.91 Eta = .197	d.f.=2	p=.234
3-P	0.00 to 1.00	70.0	73.2	85.7
	1.01 to 2.00	30.0	26.8	14.3
	over 2.00	0.0	0.0	0.0
		χ^2 = 1.18 Eta = .126	d.f=2	p=.554

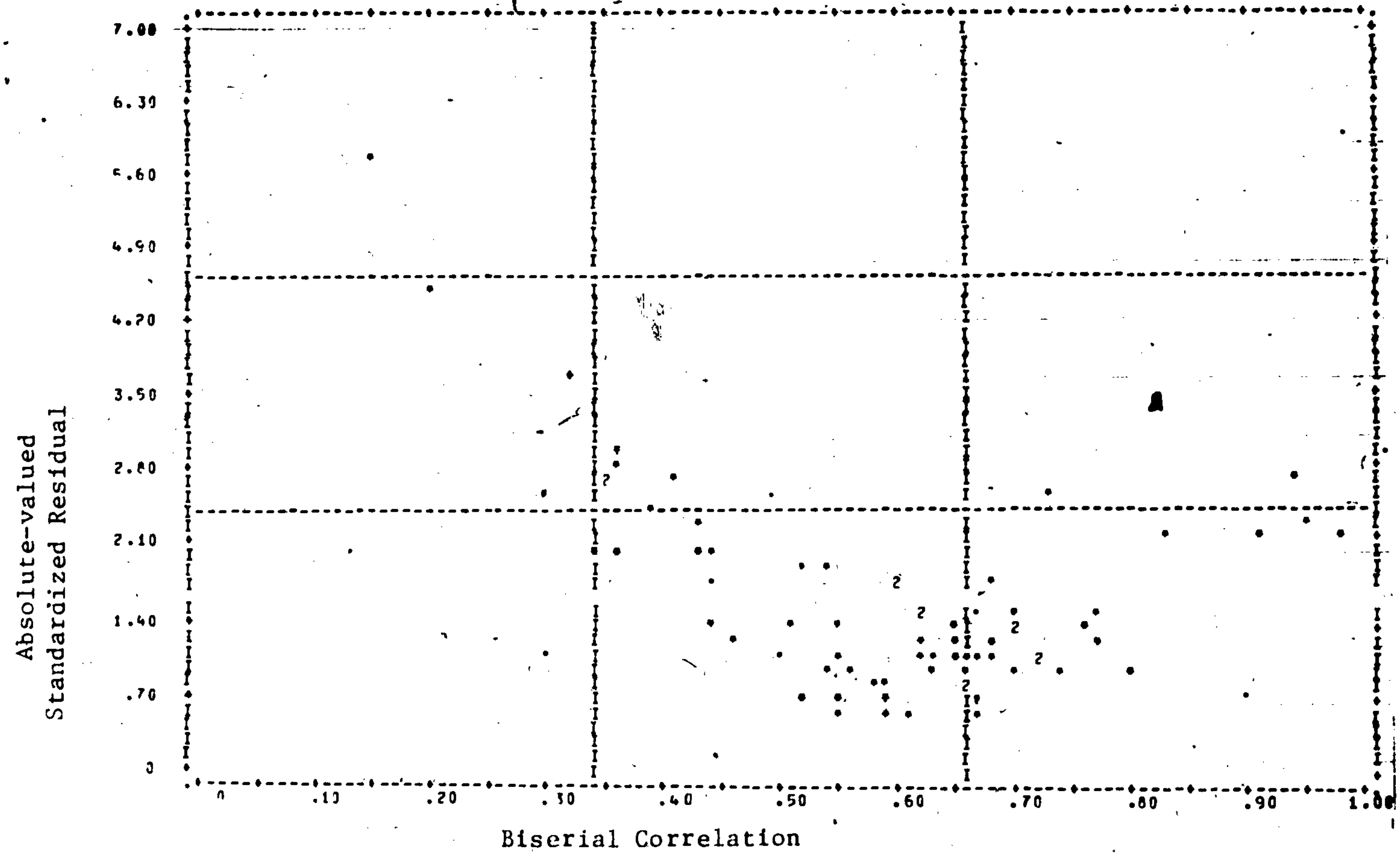

Figure 1. Plot of item absolute-valued standardized residuals obtained with the one-parameter model versus item biserial correlations.

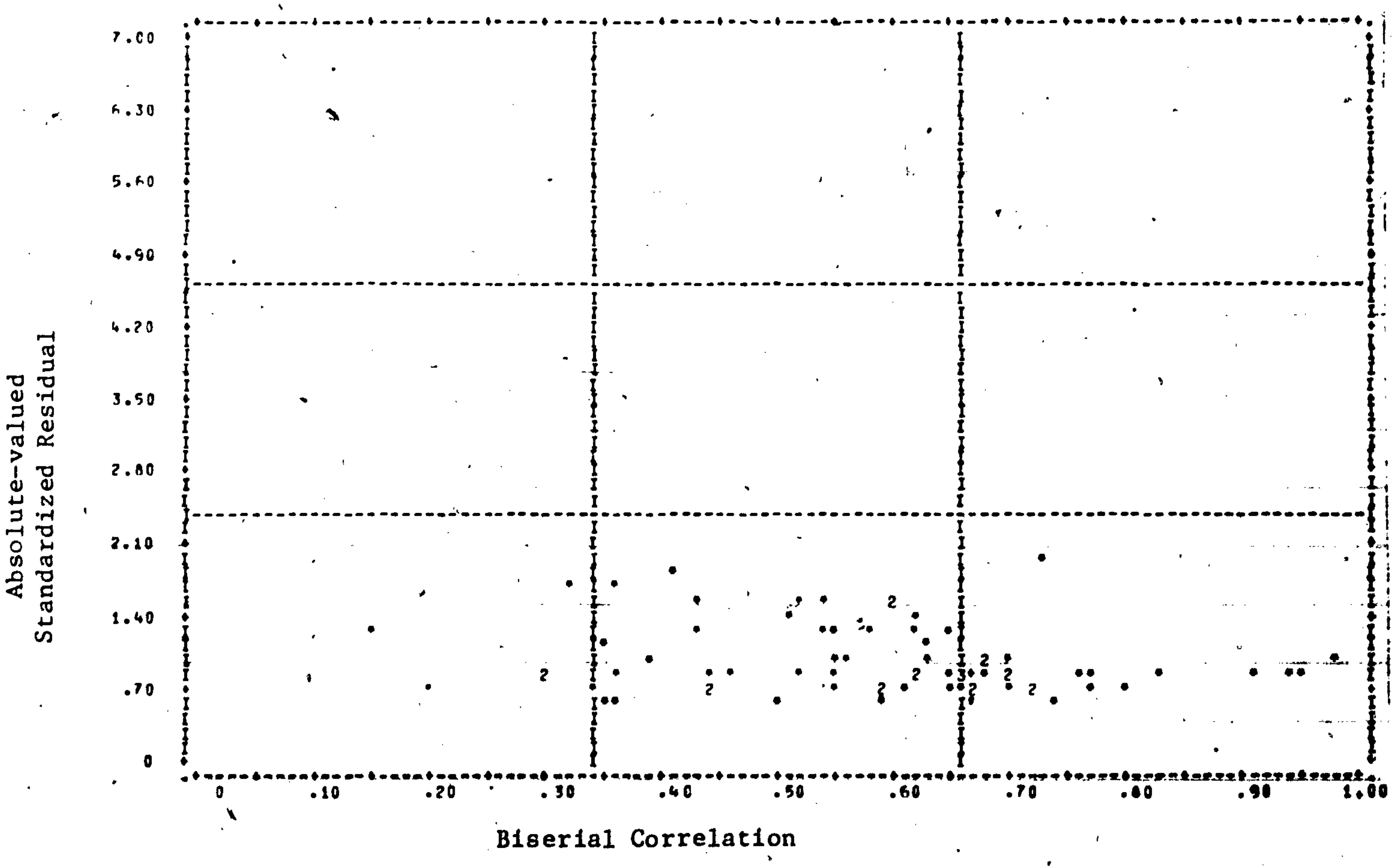

Figure 2. Plot of item absolute-valued standardized residuals obtained with the two-parameter model versus item biserial correlations.

by removing items that failed to fit the one-parameter model. The three-parameter model improved the fit only slightly because of the minimum impact of guessing behavior on test performance.

With respect to addressing the fit between an item response model and a set of test data for some desired application, our view is that the best approach involves (1) designing and implementing a wide variety of analyses, (2) interpreting the results, and (3) judgmentally determining the appropriateness of the intended application. Analyses should include investigations of model assumptions, the extent to which desired model features are obtained, and comparisons between model predictions and actual data. With respect to the latter, fitting more than one model and comparing (for example) residuals provides information that is invaluable in determining the usefulness of models. Of course there is no limit to the number of investigations that can be carried out. The amount of effort extended in collecting, analyzing, and interpreting results must be related to the importance and nature of the intended application. In this study, only a few of the necessary types of investigations for selecting an item response model were carried out and so it would not be appropriate to recommend one model over another at this time. For one, the practical consequences of the one-parameter model misfit might be studied to determine its significance in a state-wide test program. Still, there seems to be sufficient evidence to warrant a recommendation that the Maryland Department of Education give serious consideration to the two-parameter model with their MFRT. Revising the test content so that a one-parameter model will fit the test data, or assessing ability scores with a one-parameter model that does not fit the data as well as a two-parameter model, seem to be undesirable alternatives for a statewide testing program. Utilizing a three-parameter model seems to be unnecessary at this time with grade 9 students in view of the added complexity, cost, and minimal advantages derived from the model with the MFRT.

References

Divgi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981.

Hambleton, R. K. (Ed.) *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R. K., & Murray, L. Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of Item Response Theory.* Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum, 1980.

van den Wollenberg, A. L. Two new test statistics for the Rasch model. *Psychometrika*, 1982, *47*, 123-140.

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