ED 236 186 TM 830 691

AUTHOR	Murray, Linda N.; Hambleton, Ronald K.
TITLE	Using Residual Analyses to Assess Item Response Model-Test Data Fit. Laboratory of Psychometric and Evaluative Research Report No. 140.
INSTITUTION	Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY	Education Commission of the States, Denver, Colo.; National Inst. of Education (ED), Washington, DC.
PUB DATE	Apr 83
CONTRACT	ECS-02-81-20319
NOTE	34p.; 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/PC02 Plus Postage.
DESCRIPTORS	Educational Assessment; *Goodness of Fit; Item Analysis; *Latent Trait Theory; *Mathematical Models; Measurement Techniques; National Programs; Research Problems; Statistical Analysis; *Test Results
IDENTIFIERS	National Assessment of Educational Progress; Parametric Analysis; *Residuals (Statistics)

ABSTRACT

The purpose of this research study was to assess item response model-test data fit using residuals. First, a comparison of raw and standardized residuals for describing model-test data fit was carried out. Second, hypotheses concerning the relationship between residual sizes and several item characteristics were studied. The analyses with residuals were carried out with NAEP mathematics test data using the one-, two-, and three-parameter logistic test models. The results from the investigation highlighted clearly the advantages of addressing the question of model-test data fit with residuals. (Author)

Using Residual Analyses to Assess Item Response Model-Test Data Fit[1,2,3]

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

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

[2] The research described in the paper was supported by ECS Contract No. 02-81-20319 which was issued under a jointly sponsored project of the Educational Commission of the States (ECS) and the National Institute of Education (NIE). It does not, however, necessarily reflect the views of ECS or NIE.

[3] Laboratory of Psychometric and Evaluative Research Report No. 140. Amherst, MA: School of Education, University of Massachusetts, 1983.

Using Residual Analyses to Assess Item Response Model-Test Data Fit

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

Abstract

The purpose of this research study was to assess item response model-test data fit using residuals. First, a comparison of raw and standardized residuals for describing model-test data fit was carried out. Second, hypotheses concerning the relationship between residual sizes and several item characteristics were studied. The analyses with residuals were carried out with NAEP mathematics test data using the one-, two-, and three-parameter logistic test models. The results from the investigation highlighted clearly the advantages of addressing the question of model-test data fit with residuals.

Presently, there is considerable interest in applying the one-, two- and three-parameter logistic item response models to a wide variety of educational and psychological measurement areas. These areas include detection of item bias, adaptive testing, mastery testing, item banking, test development, and test score equating (Hambleton, 1983; Lord, 1980; Traub & Wolfe, 1981). However, the benefits of item response theory are predicated upon an adequate fit between the chosen model and the set of test data. Clearly no psychologically meaningful test model can ever fit a data set perfectly. But without sufficient model-test data fit, the desirable features of an item response model will not be obtained or obtained in a low degree.

Goodness of fit studies are helpful in assessing the utility of an item response model for solving specific measurement problems with a particular test data set. Hambleton, Murray and Simon (1982) organized and reviewed many goodness of fit procedures that have been advocated and documented in the research literature. The procedures they found can be grouped into several general categories. These categories include (1) statistical tests for assessing model-data fit, (2) verifying model assumptions and expected model features, and (3) checking model predictions with test results. It was determined that these procedures varied substantially in their level of practicality and effectiveness. For example, they found that much attention was focused on the use of statistical tests where unfortunately model-data fit depended upon the sizes of examinee samples used in the studies. The statistical values could become significant due principally to large sample sizes (Hambleton, Murray & Simon, 1982).

Analyses of residuals offer another means of examining model-data fit. These analyses are more practical than many of the other fit methods and they often provide a more effective way of revealing instances or patterns of misfit (Traub & Wolfe, 1981). Residual analyses have played an important role in determining the suitability of regression models (Draper & Smith, 1966; Anscombe & Tukey, 1963; Seber, 1977). On the other hand, residual analyses have not been used to any substantial extent to investigate the appropriateness of item response models.

A residual analysis involves the following steps: (1) a model is chosen and model parameters are estimated from the data; (2) the estimates are substituted into the model and predictions are made; and (3) discrepancies (residuals) between the data and values predicted by the model are examined. The overall quality and suitability of the model and the usability of the results are evaluated by examining the size and direction of the residuals and variations such as absolute-valued residuals. Sometimes the residuals are plotted as a function of ability to determine more precisely the nature of model-test data misfit.

In one recent study, Hambleton and Murray (1983) examined the size and pattern of standardized residuals using the one-parameter and three-parameter logistic item response models. They also explored the relationship between selected item characteristics such as content categories and item format and the size of standardized residuals. Overall their research study revealed that residual analyses helped considerably in judging the suitability of the two item response models.

The purpose of this research study was to expand on the earlier residual analysis work of Hambleton, Murray and Simon (1982) and Hambleton and Murray (1983). More specifically, this research investigation was designed to address two topics:

1. Comparison of raw and standardized residuals for describing model-data fit.
2. Hypotheses concerning the relationship between fit of test items and item format, difficulty level, discrimination level, item wording, and various other salient aspects of test items.

With respect to the first topic, this study extended the earlier work by Hambleton and Murray (1983) by reporting raw residuals, and in addition, compared raw and standardized residuals for the purpose of describing model-data fit. With respect to the second topic, this study investigated the fit of three logistic models rather than two models and considered several additional hypotheses which were not examined in the earlier study.

Method

Description of the Tests

Four National Assessment of Educational Progress (NAEP) test booklets from the 1977-78 assessment were selected for analysis:

9 Year Olds

Booklet No. 1, 65 items, 2495 examinees

Booklet No. 2, 75 items, 2463 examinees

13 Year Olds

Booklet No. 1, 58 items, 2422 examinees

Booklet No. 2, 62 items, 2433 examinees

Each booklet contained test items measuring various mathematical skills in the areas of definitions, story problems, geometry, measurement, and graphs and figures. The test items in the NAEP assessment were either multiple-choice or open-ended. Finally, these data sets were unusual in the sense that the test items varied substantially in both their range of difficulty (.02 to .98) and their range of item discrimination levels (-.01 to .99). These ranges far exceed those ranges normally found in achievement and aptitude tests. Because of the wide range of classical item discrimination indices and the high level of guessing due to the substantial number of difficult items, we expected that the three-parameter model would fit the test data substantially better than the other two more restrictive models.

Residual Analyses

Each analysis in this study began with the calculation of the raw and standardized residuals. Raw residuals are comparisons of predicted performance results with actual performance results. To calculate residuals an item response model was first chosen. For this study the one-. two-, and three-parameter logistic test models were used in separate but identical analyses. Next, item and ability parameter estimates were obtained using the LOGIST computer program (Wood, Wingersky & Lord, 1976). To find the actual performance results, an examinee was placed in an ability category based on his or her estimated ability level. For this investigation, ability categories were chosen that divided the ability scale between -3.0 and 3.0 into 12 equal intervals. Ability estimates that fell beyond these maximum and minimum ability levels were deleted from the analysis. In every investigation, this was usually less than 10 cases. For each of the 12 ability categories, the average observed performance (P_{ij}) for an item i in ability category j was found. For example, if 10 of 50 examinees in ability category j answered item i correctly, then P_{ij} would be .2. The process was repeated for each ability category (j=1, 2, ..., 12) and for each item (i=1, 2, ..., n) in a test booklet.

Using the midpoint of each ability category (i.e., -2.75, -2.25, ..., -.25, +.25, ..., +2.75) as the average ability level for that group of examinees, the expected performance ($\hat{P}_{ij}$) for item i in ability category j was found in the usual way:

$$\hat{P}_{ij}^{(3)} = c_i + (1-c_i) \frac{e^{1.7a_i(\theta_j-b_i)}}{1+e^{1.7a_i(\theta_j-b_i)}}$$

for the three-parameter logistic model,

$$\hat{P}_{ij}^{(2)} = \frac{e^{1.7a_i(\theta_j-b_i)}}{1+e^{1.7a_i(\theta_j-b_i)}}$$

for the two-parameter logistic model, and

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

for the one-parameter logistic model.

In these equations a_i, b_i and c_i are the item parameter estimates obtained from LOGIST (Lord, 1980) and θ_j is the mid-point of the j^{th} ability category.

Then the raw residual (R_{ij}) for item i in ability category j was

$$R_{ij} = P_{ij} - \hat{P}_{ij}.$$

This difference is an index of the degree of misfit between the test data and the expected item performance based on the chosen model. Large positive raw residuals indicate that examinees are performing considerably better on an item than is predicted by the item response model. Large negative raw residuals reveal that the model is predicting a much higher performance level by the examinees on the item then is actually observed. Finally, evidence of sufficient model-data fit occurs when the residuals are small and there are no obvious patterns in the residuals across ability levels.

Next, these raw residuals were transformed to standardized residuals (SR_{ij}) by dividing R_{ij} by the sampling error associated with the average expected performance level in an ability category (Blalock, 1979). That is,

$$SR_{ij} = \frac{P_{ij}-\hat{P}_{ij}}{\sqrt{\frac{\hat{P}_{ij}(1-\hat{P}_{ij})}{N_j}}}$$

where N_j is the number of examinees in ability category j.

These raw and standardized residuals differ in several ways. Raw residuals are simpler to calculate and easier to interpret than standardized residuals. On the other hand, standardized residuals take into account the sampling errors associated with P_{ij}. When N_j is small, other things being equal, big differences between actual and expected differences must be obtained for the differences to be taken as an indication of model-test data misfit. For example, suppose two different ability categories for an item i have the same computed raw residual (.3-.2), but differ in their examinee sample sizes (10 vs 100). Using the raw residuals, it appears that model data fit is the same in both examinee samples. But, the greater number of examinees (N_j) produces a smaller standard error of expected performance level because a more accurate estimate is possible. Then, the corresponding standardized residuals are .79 and 2.5. Clearly, the two statistics seem to give a very different picture of model-data fit. Therefore, a comparison of raw and standardized residuals was made to determine how differently they described levels of model-data fit and whether the choice of statistic might affect the decision about the usefulness of item response models. The size and direction of the raw and standardized residuals in the analyses were compared in three ways: (1) across ability levels for each item; (2) across items at each ability level; and (3) across both ability levels and test items.

Research Hypotheses

Several testable research hypotheses were generated concerning model-data fit. Specifically, interest centered on determining if test items having large positive or negative standardized residuals exhibit certain salient item characteristics that would cause them to be misfit by an item response model. To reduce problems associated with studying curvilinear relationships, absolute-valued standardized residuals were used instead of standardized residuals. Then, analyses were conducted concerning the association between the fit of test items and item format, and classical indices of item difficulty and discrimination.

Results

Comparison of Raw and Standardized Residuals

Table 1 displays the intercorrelations among several of the NAEP math item variables. There is a strong relationship between the one-parameter raw and standardized residuals (r=.91) suggesting they describe model-data fit in similar fashions. The correlations between the two-parameter and three-parameter raw residuals with their corresponding standardized residuals are lower (r=.77). But, these correlations are probably only lower due to range restriction on the variables as shown by the standard deviations listed in Table 1.

Absolute valued raw and standardized residuals for each of the logistic models are similarly correlated with difficulty, item format and item order. Because of the non-linear relationship, associations between item discrimination (as measured by biserial correlations) and the residuals were investigated by examining the plots shown in Figures 1 through 6. Figures 1 and 2 are plots of raw residuals and standardized residuals versus classical item discrimination indices. These figures show clearly that for the one-parameter model, a curvilinear relationship prevailed whether raw or standardized residuals were used to describe fit (i.e., very low or high discriminating items had larger residuals with the one-parameter model). Small differences between the results in these plots emerged

Table 1

Statistics of and Intercorrelations Among Several NAEP Math Item Variables
(Booklet Nos. 1 and 2, 260 Items, 13 and 9 Year Olds, 1977-78)

Variable	Mean	Standard Deviation	|SR(2-P)|	|SR(3-P)|	|RR(1-P)|	|RR(2-P)|	|RR(3-P)|	P	F[1]	O
|Standardized Residual (1-P)|	1.98	1.20	.24	.18	.91	.35	.33	-.30	-.25	.14
|Standardized Residual (2-P)|	1.01	.42		.41	.08	.77	.30	-.21	-.13	.00
|Standardized Residual (3-P)|	.88	.42			.15	.27	.77	.09	.07	-.03
|Raw Residual (1-P)|	.060	.033				.24	.34	-.17	-.19	.09
|Raw Residual (2-P)|	.033	.017					.43	-.22	-.34	.13
|Raw Residual (3-P)|	.030	.017						-.07	-.17	.14
Item Difficulty (P)	.53	.27							.04	-.40
Format (F)										-.12
Item Order (O)										

[1] Two types: Multiple-choice and Open-ended.

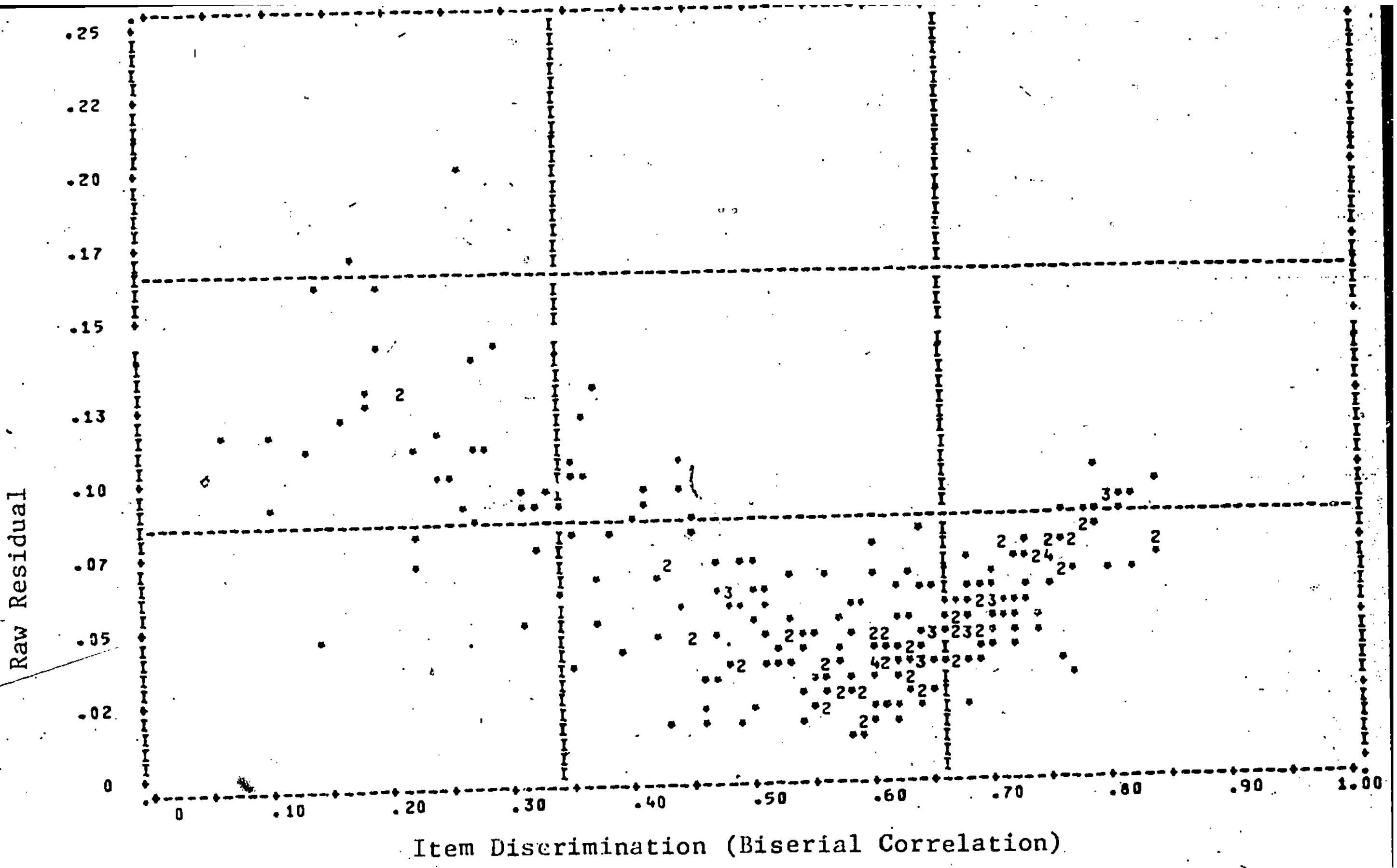


Figure 1. Plot of one-parameter model raw residuals versus item discrimination.

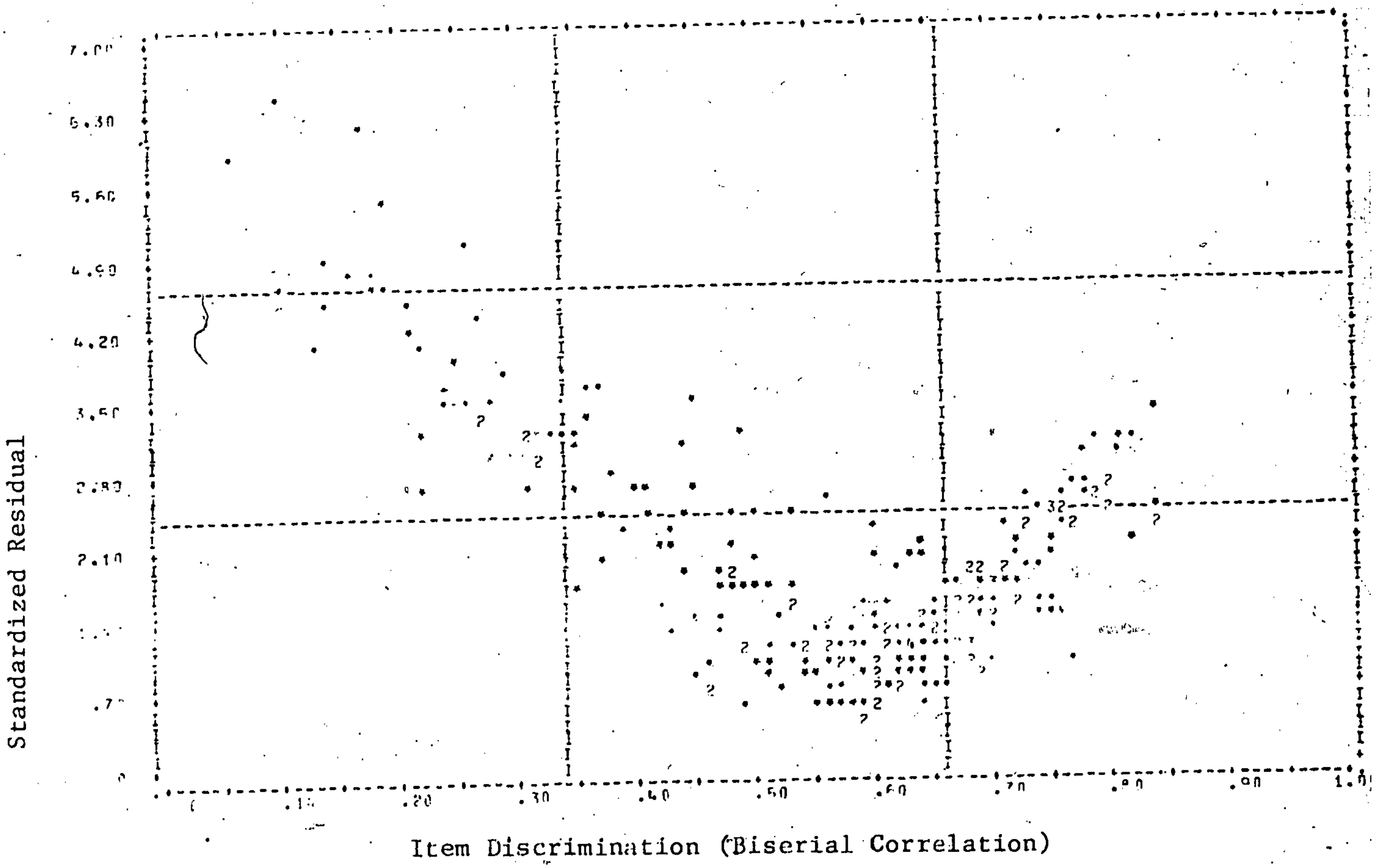


Figure 2. Plot of one-parameter model standardized residuals versus item discrimination.

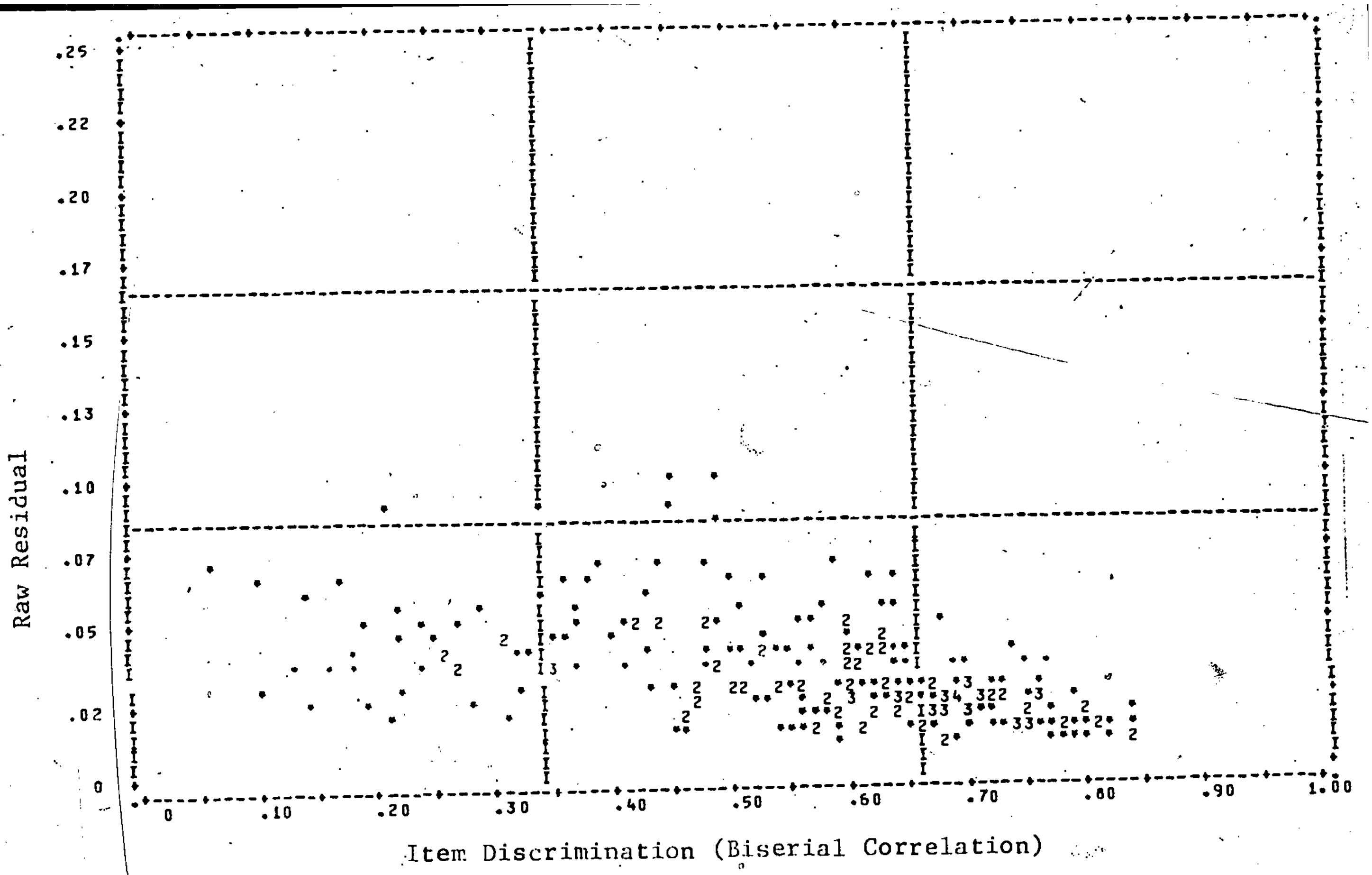


Figure 3. Plot of two-parameter model raw residuals versus item discrimination.

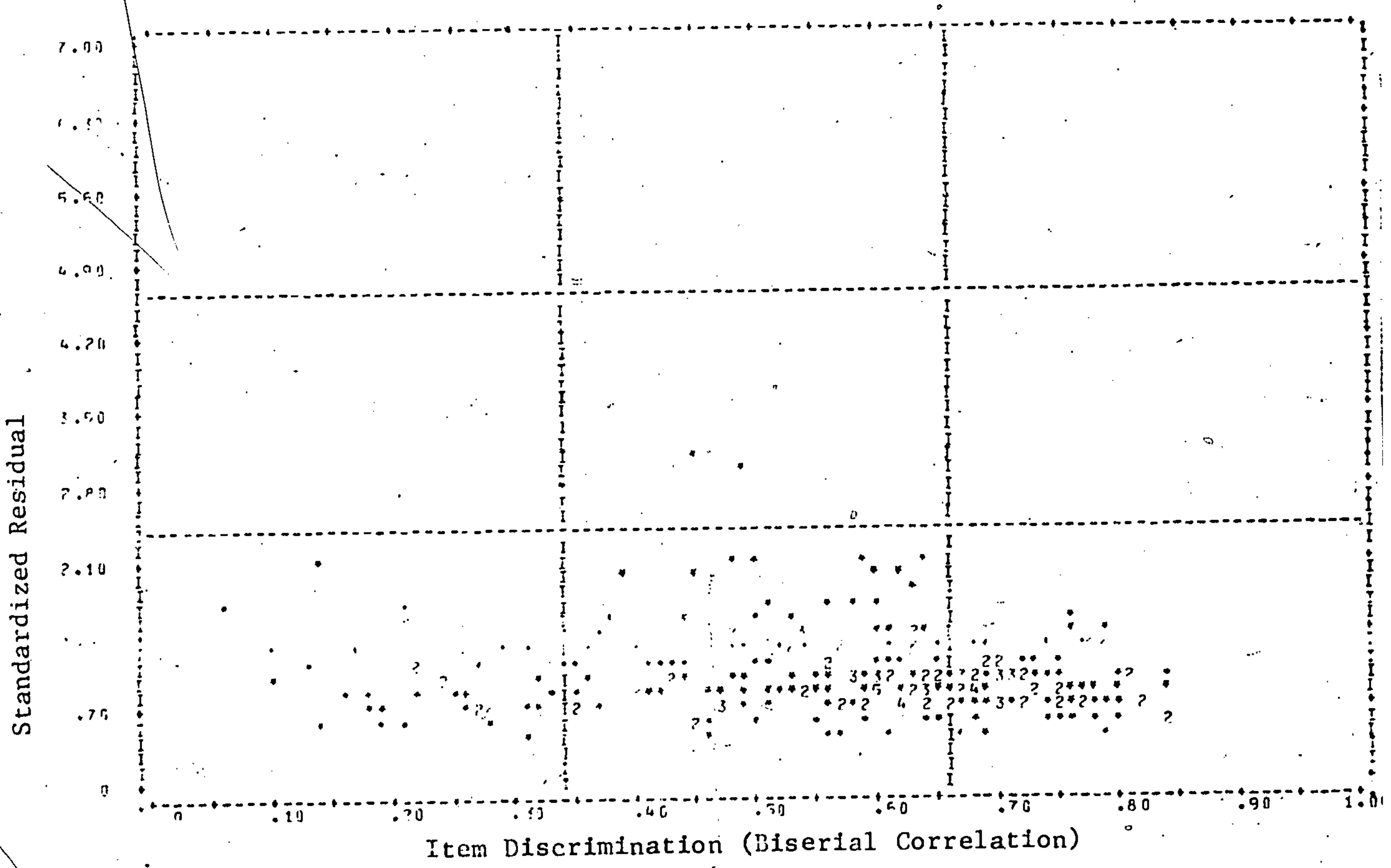


Figure 4. Plot of two-parameter model standardized residuals versus item discrimination.

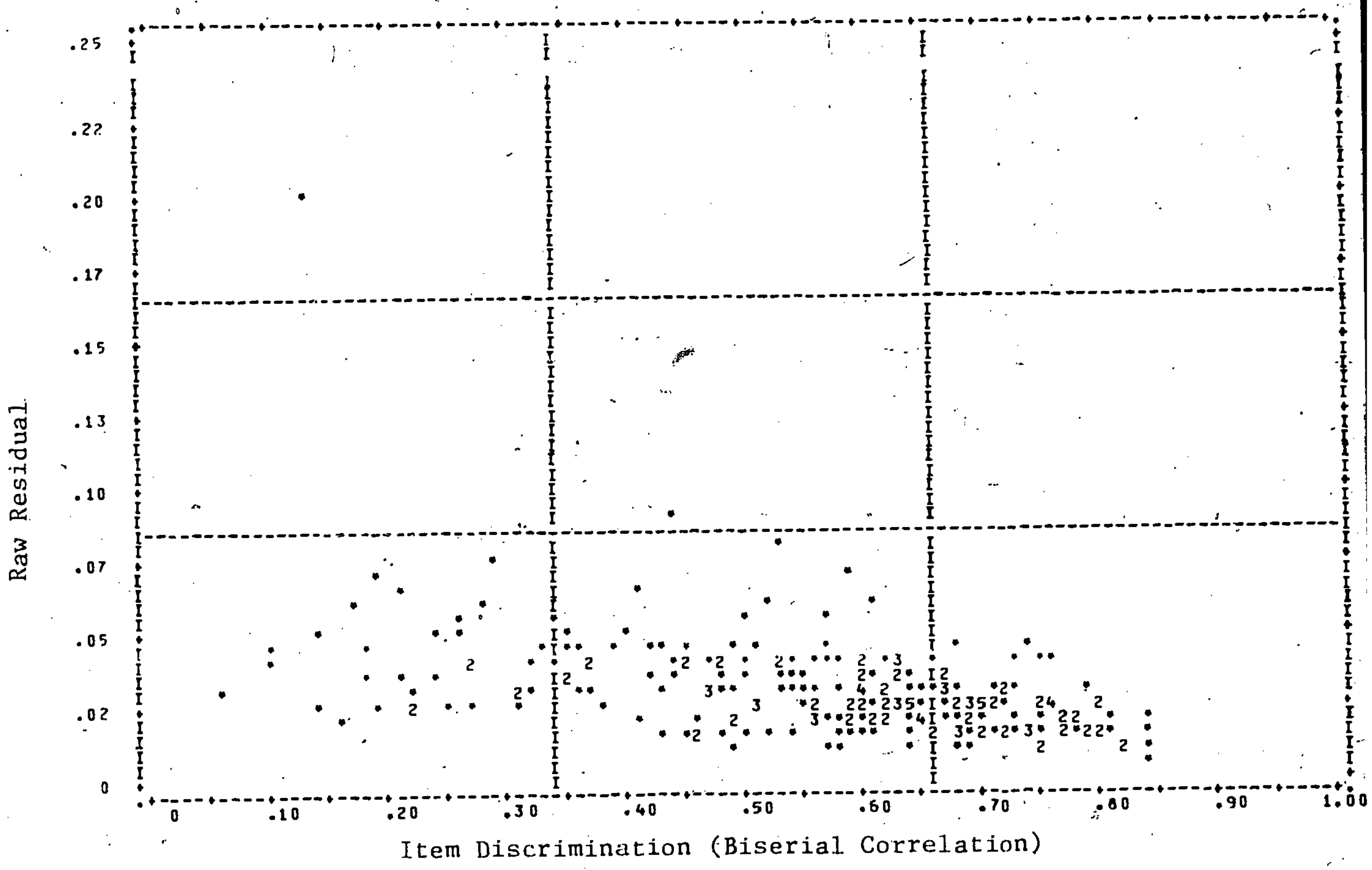


Figure 5. Plot of three-parameter model raw residuals versus item discrimination.

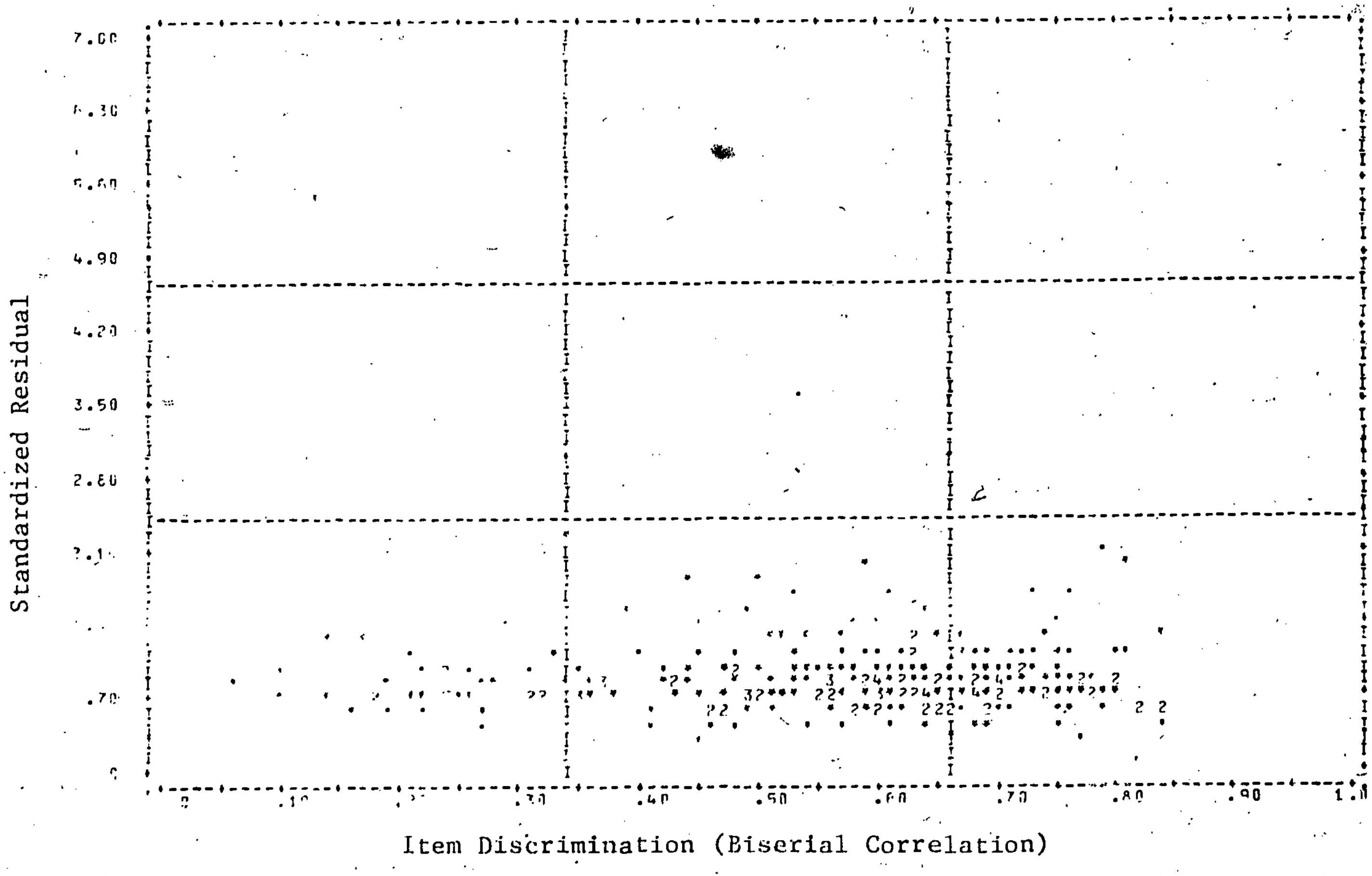


Figure 6. Plot of three-parameter model standardized residuals versus item discrimination.

for lower discriminating items. Similarly, Figures 3 through 6 display the plots of the residuals versus item discrimination for the two- and three-parameter models. These plots again suggest strong agreement between the residuals except for low discriminating items where a slightly wider variation of misfit was found with the raw residuals.

Next, a check on the degree of similarity between raw and standardized residuals was carried out with the one-parameter model results. Using 2.0 as the cut-off point on the absolute-valued standardized residual scale, 102 "bad" items were identified. Next, the poorest fitting 102 items on the absolute-valued raw residual score scale were identified. Ninety percent of the items were common to the two analyses indicating a high level of agreement in the identification of misfitting items. (Were agreement due to chance factors only, about 15% of the items would have been common to the two analyses.) Because of the small number of misfitting items by the two- and three-parameter models, similar analyses with these models were not carried out.

The average of absolute-valued raw and standardized residuals at 12 ability levels with the three logistic models are reported in Table 2. The average raw and standardized residual statistics provide information about the size and direction of the misfit between the observed and expected results while the absolute-valued statistics ignore the direction of misfit and consider only the magnitude of the misfit. Since the trends in the results across the four Math booklets were the same, only the results for one booklet are reported in this paper.

Three of the four statistics in Table 2 present a similar picture of fit for the three item response models. Both the two- and

Table 2

Average and Absolute Average Raw and Standardized Residuals at Twelve Ability Levels with the One-, Two-, and Three-Parameter Logistic Models (Booklet No. 1, 9 Year Olds, 65 Items, 1977-78)

	Logistic Model	Sample Size	Ability Level -2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75	Total (unweighted)
	1	2495	27	43	111	220	331	485	446	395	276	122	21	8	
	2	2495	12	49	110	231	379	466	466	349	273	99	39	15	
	3	2495	29	50	108	212	333	470	470	403	273	100	21	9	
	1		.002	.001	-.001	.001	.002	.002	.002	-.006	-.009	-.013	-.003	-.005	-.002
	2		.004	.005	-.017	.009	-.003	-.003	-.004	-.006	-.001	.003	.005	.031	.006
	3		.004	.010	.010	.003	.001	.001	.002	-.002	-.005	-.012	-.005	.006	.001
	1		.006	.088	.074	.073	.045	.030	.027	.043	.057	.076	.071	.084	.061
	2		.052	.048	.042	.021	.017	.018	.013	.018	.017	.033	.038	.075	.033
	3		.049	.040	.034	.019	.020	.015	.010	.013	.015	.025	.043	.073	.030
ed	1		.77	.99	.89	.79	.37	.20	.14	-.28	-.26	-.39	-.11	-.10	.25
	2		.09	.31	.76	.35	.09	-.22	-.30	-.37	-.18	-.06	-.02	-.22	.06
	3		.00	.24	.27	.12	.16	.04	.08	-.18	-.48	-.36	-.32	-.16	.05
ed|	1		1.75	2.40	2.82	3.35	2.35	1.80	1.62	2.35	2.64	2.40	1.19	.85	2.13
	2		.82	1.28	1.58	1.00	.90	1.15	.83	1.03	.93	1.12	.97	1.07	1.06
	3		.81	.90	1.02	.74	1.00	.94	.62	.87	.99	.85	.91	.88	.88

three-parameter models provided a very good accounting of the actual results. The one-parameter model did <u>not.</u> The fourth statistic (raw residuals) described model-data fit much differently.

Discrepancies between these two impressions of model-data fit can be accounted for by examining the way in which average raw residuals are computed. A comparison of the size and direction of model-data misfit between the one- and three-parameter models for one ability category (-2.00 to -1.50) is shown in Table 3. The direction of misfit can either be positive or negative based on whether the model has underpredicted or overpredicted examinee performance. As can be seen from Table 3, a considerable amount of misfit in both directions occurred with the one-parameter model. This finding was <u>not</u> surprising since it was already noted that the items varied substantially in levels of item discrimination. The one-parameter model assumed a common item discrimination across the set of items. But because there was considerable deviation from this average item discrimination the results were (1) large sized residuals in both directions and (2) a very small overall average raw residual.

<u>Hypothesis Testing</u>

The results in Table 4 through 6 suggest reasons for model-test data misfit. Table 4 displays the results from an analysis of the relationship between the size of the standardized residuals and the level of classical item difficulty. Substantial improvement in fit occurred for hard items when the three-parameter model was fit to the test data. For easier items better fits were obtained again by the

Table 3

Comparison of the Size and Direction of Model-Data Misfit for One Ability Category (-2.00 to -1.50) (Booklet No. 1, 9 Year Olds, 1977-78)

Logistic Model	Size of Misfit (Reported in Each Direction)		Average Residual
	+	-	
1	2.385	2.434	-.001
3	1.427	.795	.010

Table 4

Association Between Standardized Residuals
and Item Difficulties
(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Difficulty Level	Standardized Residuals	1-p Results		2-p Results		3-p Results	
		N	%	N	%	N	%
Hard ($p \leq .5$)	$\lvert SR \rvert$ (≤ 1.0)	14	11	69	56	99	80
	$\lvert SR \rvert$ (> 1.0)	110	89	55	44	25	20
Easy ($p > .5$)	$\lvert SR \rvert$ (≤ 1.0)	34	33	87	64	98	72
	$\lvert SR \rvert$ (> 1.0)	102	672	49	36	38	18

Table 5

Descriptive Statistical Analysis of the Absolute-Valued Standardized Residuals
(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

Difficulty Level	Format	Number of Items	1-p Results $\overline{X}$	SD	2-p Results $\overline{X}$	SD	3-p Results $\overline{X}$	SD
Hard ($p \leq .5$)	Multiple-Choice	70	2.73	1.55	1.18	.53	.82	.23
	Open-Ended	54	1.64	.81	.92	.38	.86	.28
Easy ($p > .5$)	Multiple-Choice	70	1.79	1.10	.94	.40	.90	.64
	Open-Ended	66	1.67	.72	.97	.30	.97	.38

Table 6

Relationship Between Item Discrimination Indices
and Standardized Residuals
(Booklets No. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78)

		Discrimination Indices			
Model	Standardized Residuals	-.01 to .30	.31 to .50	.51 to .70	.71 to 1.00
		(29)[1]	(55)	(125)	(51)
1-p	0.00 to 1.00	0.0	10.9	33.6	0.0
	1.01 to 2.00	0.0	32.7	62.4	29.4
	over 2.00	100.0	56.4	4.0	70.6
		χ^2 = 143.7	d.f. = 6	p = .000	
		Eta = .691			
2-p	0.00 to 1.00	51.7	49.1	60.8	74.5
	1.01 to 2.00	41.4	41.8	36.0	25.5
	over 2.00	6.9	9.1	3.2	0.0
		χ^2 = 11.58	d.f. = 6	p = .072	
		Eta = .203			
3-p	0.00 to 1.00	75.9	80.0	76.8	68.6
	1.00 to 2.00	20.7	18.2	23.2	29.4
	over 2.00	3.4	1.8	0.0	2.0
		χ^2 = 5.28	d.f. = 6	p = .508	
		Eta = .092			

[1] Number of test items in brackets.

three-parameter model although there was a less dramatic shift in fit between the two- and three-parameter models. These findings suggest that examinee guessing was an important factor with the harder items and less consequential with easier items.

Table 5 provides a summary of the absolute-valued standardized residuals for the three logistic models with items classified by difficulty and format. For both hard and easy open-ended items and easy multiple-choice items the pattern of results were the same. Substantial improvements in fit were obtained when the two-parameter model was substituted for the one-parameter model. The two- and three-parameter results however were similar. For the hard multiple-choice items a substantially different pattern emerged. First, the size of the standardized residuals was on the average substantially larger for the one- and two-parameter models. Second, there were considerable improvements in fit between the one- and two-, and the two- and three-parameter models. This result strongly suggests that examinee guessing on hard multiple-choice items affects the degree of model-data fit and therefore the "pseudo-chance level" parameter was useful.

Finally, Table 6 reveals the relationship between item discrimination and standardized residual size. For these items varying greatly in levels of item discrimination, the best fit occurred with the three-parameter model. Items with relatively high or low item discrimination indices were poorly fitted by the one-parameter model. This resulted in a strong curvilinear relationship as represented by an eta value of .691. Substantial improvement in fit occurred when the two-parameter model replaced the one-parameter model.

The previous analyses presented results about trends of misfit across a number of test items. Were there any specific reasons why particular items misfit a certain model or models? To answer this question, items and their corresponding standardized residuals with the three models were scrutinized individually. Four different patterns emerged: (1) substantial improvement in the fit by using the two- or three-parameter models, (2) similar fit across the three models, (3) best degree of fit by using the three-parameter model and (4) best degree of fit by using two-parameter model. For each pattern, a representative item was examined carefully in order to identify possible salient item characteristics causing these instances of misfit and fit. Table 7 contains the results of this analysis. The four test items are shown in Figure 7.

With Item 36, significant improvement in model-data fit occurred when the two-parameter model replaced the one-parameter model. The classical item statistics showed the item as being non-discriminating (r=-.01) and difficult (p=.21) due, in part, to the unusual nature of the test question (i.e., subtracting ranges of numbers) and the overlap in the answer choices. With the two- and three-parameter models it was possible to account for the very low discriminating power of the test item. With the one-parameter model it was not and hence the poor model data fit.

Item 44 was fit by the three models in a similar fashion. The classical item statistics reveal that the item had middle level of difficulty (p=.68) and discrimination (r=.59). The item had an open-ended format and thus guessing was an inconsequential consideration in item performance. Therefore the additional effort made to incorporate "item discrimination" and "pseudo-guessing"

Table 7
Representative Items for Four Patterns of Model Misfit
(Math Booklet No. 1, 13 Year Olds, 1977-78)

Item Number	$\|SR_1\|$	$\|SR_2\|$	$\|SR_3\|$	Description	Possible Explanation(s)
36	7.08	1.02	1.19	Substantial improvement in fit by using the 2-P or 3-P models over the 1-P model	Unusual item wording; overlap of answer choices; non-discriminating and difficult item
44	1.58	2.14	1.93	Similar fits for the models	Open-ended format; average level of item discrimination
23	2.85	1.49	.71	Improvement in fit from using the 3-P model rather than the 1-P or 2-P model	Multiple-choice format; relatively difficult and discriminating; substantial amount of guessing
4	3.11	.94	1.94	Best fit from the 2-P model	Open-ended format; extremely discriminating; misfit of 3-P model occurred at the highest ability level due to a highly unstable standardized residual

Figure 7. Four sample test items.

36. Ms. Baker has between $8,000 and $8,500 in her savings account. She wants to buy a new car that costs between $5,300 and $5,400. After she buys the car, how much money will Ms. Baker have in her savings account?

0 $2,700
0 $3,100
0 Between $2,700 and $3,100
0 Between $2,600 and $3,200
0 I don't know.

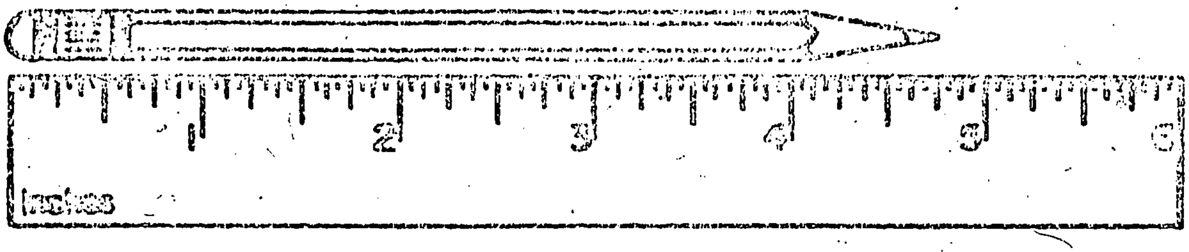
44. Find the quotient.

A. 6)608 ANSWER ______________

23. When is the product of two integers negative?

0 When both are positive
0 When both are negative
0 When one is negative and one is positive
0 When one is zero and one is negative
0 I don't know.

4.



What is the length of this pencil to the nearest quarter inch?

ANSWER ______________ inches

parameters did not increase the amount of model-data fit.

For Item 23 considerable improvement in fit occurred when the three-parameter model was substituted for the one- and two-parameter models. This multiple-choice item was quite difficult (p=.36) and moderately discriminating (r=.38) but substantially lower than the average discriminating power of items in the test. The similarity in the answer choices may have caused a considerable amount of guessing even though "I don't know" was an answer alternative. Therefore the three-parameter model accounted for the test data best.

Finally, with item 4 a fourth pattern of misfit is revealed. According to the size of the standardized residuals, the two-parameter model fits the test data best. This item was very discriminating (r=.81) and moderately difficult (p=.52). The high level of item discrimination would explain improvements in fit by substituting the two-parameter for the one-parameter model.

Figures 8 and 9 show the plots of standardized residuals versus ability. These plots help explain why the two-parameter model appeared to fit the data better than the three-parameter model. For the examinees in the ability range between 2.50 and 3.00 the three-parameter model over-predicted performance. But because of the very small standard error due to the easiness of the test item for high ability examinees, the standardized residuals "blew-up." This occurrence is observed with statistics such as the chi-square test when expected values are very small.

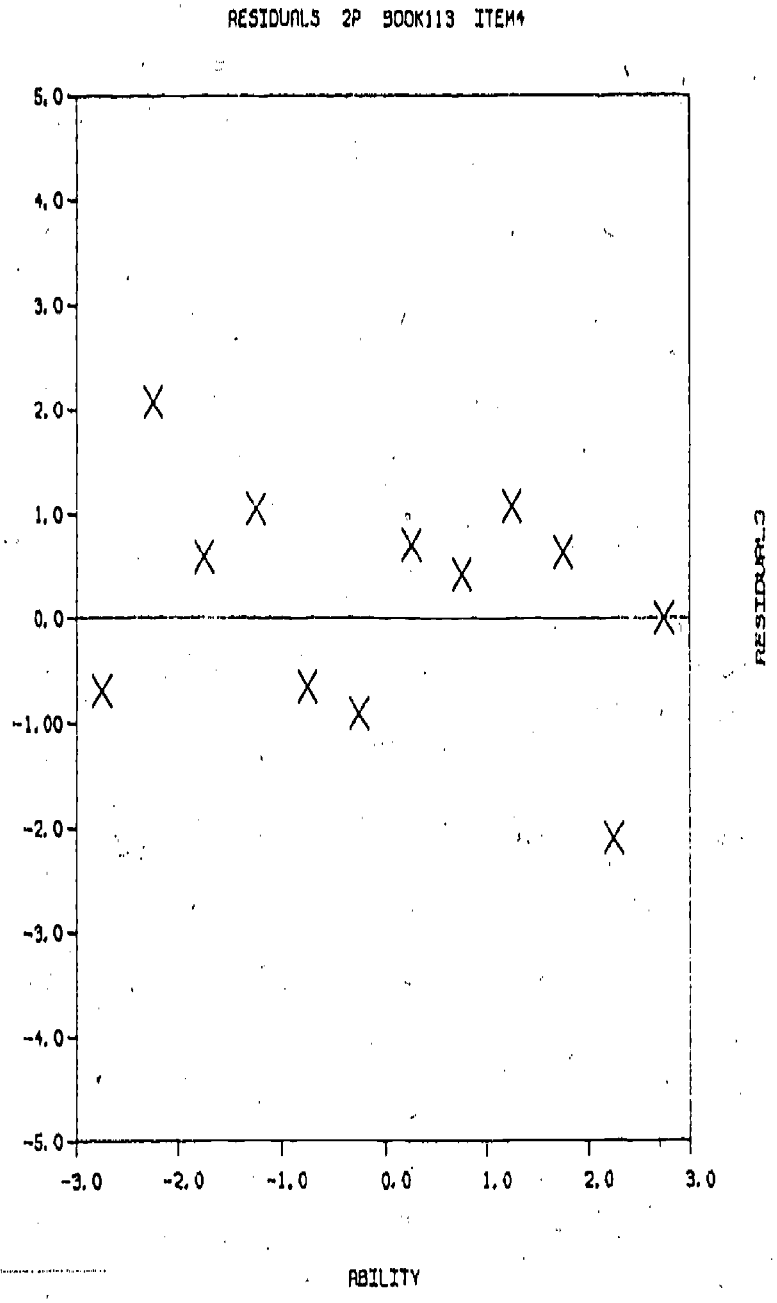


Figure 8. Standardized residual plots obtained with the two-parameter model for Item 4.

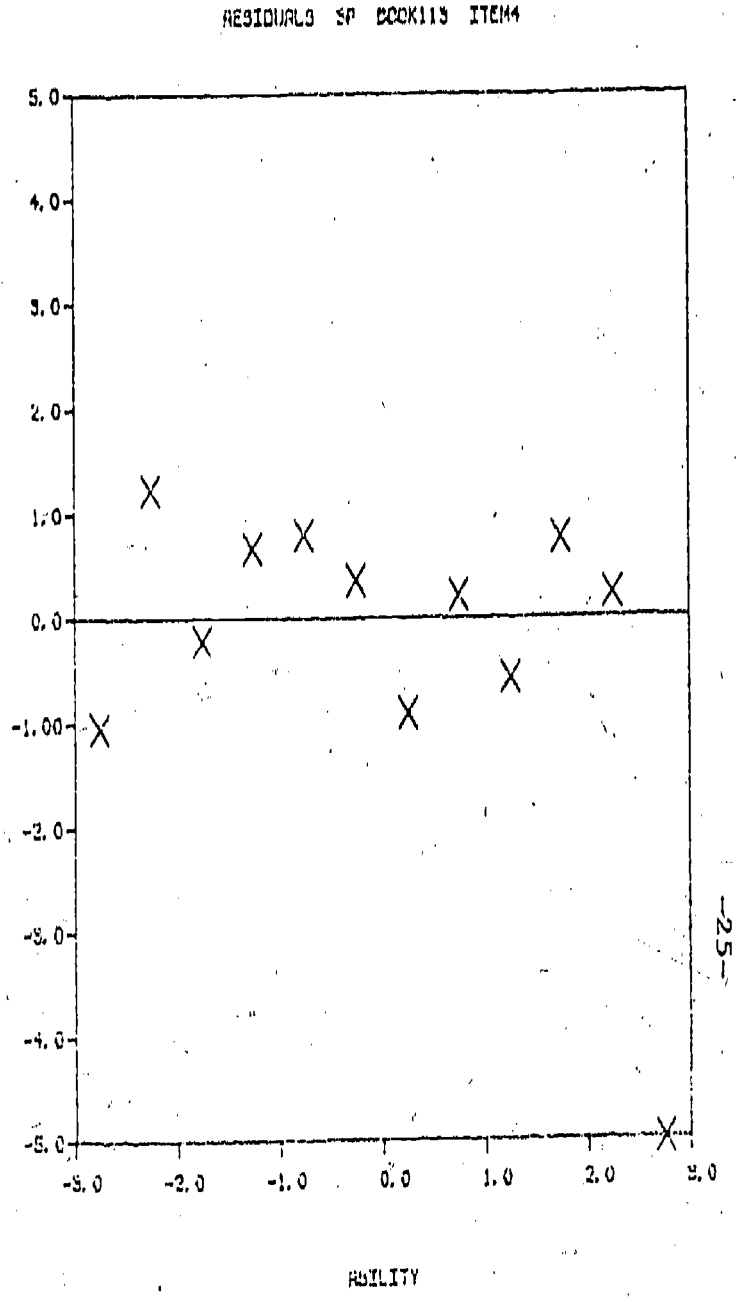


Figure 9. Standardized residual plots obtained with the three-parameter model for Item 4.

Discussion

The results from this study showed that the statistics on average raw and standardized residuals provided very useful fit information, but that when compared, the statistics based on standardized residuals presented a more accurate picture of model-data fit. Standardized residuals take into account the sampling error associated with the estimates of average performance at various ability levels. Raw residuals do not. Accounting for the instability in the statistical information seems important when assessing model-data fit.

The results of our work on the topic of hypothesis testing showed clearly that with the type of test items we worked with, failure to consider variation in item discriminating power resulted in the one-parameter model providing substantially poorer fits to the various test data sets than the two- or three-parameter models. Also, examinee guessing on difficult multiple-choice items affected the degree of model-data fit. Here, substantial improvement in fit occurred when the "pseudo-guessing" parameter was used in the item response model. These results were not surprising given that the test items in the NAEP test booklets varied considerably in their biserial correlations and a substantial number of the multiple-choice items were difficult to answer for low ability examinees. In summary, the results collected in relation to the various hypotheses were invaluable for providing insights about model-data fit.

Finally, it is our opinion that the results from this study will be of interest and value to measurement specialists who are considering the usefulness of item response models in their work. Since one cannot assume that there is an adequate fit between a chosen model and a particular data set, the goodness of fit issue must be addressed. The analysis of residuals has been suggested as a method for determining sufficient model-data fit. We believe the procedures and methods suggested in this paper (including calculating average and absolute-valued averages and plotting residuals versus ability) will provide insights about the usefulness of the one-, two- and three-parameter models, as well as many other item response models.

References

Anscombe, F. J., & Tukey, J. W. The examination and analysis of residuals. Technometrics, 1963, 5, 141-160.

Blalock, H. M. Social statistics. (2^{nd} ed.) New York: McGraw-Hill, 1979.

Draper. N. R., & Smith, H. Applied regression analysis. New York: John Wiley and Sons, Inc., 1966.

Hambleton, R. K. (Ed.). Applications of Item Response Theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R. K., & Murray, L. N. 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.

Hambleton, R. K., Murray, L. N., & Simon, R. Utilization of item response models with NAEP mathematics exercise results. Final Report (ECS Contract No. 02-81-20319). Submitted to the Educational Commission of the States and the National Institute of Education, June 1982.

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

Seber, G. A. Linear regression analysis. New York: John Wiley and Sons, Inc., 1977.

Traub, R. E., & Wolfe, R. G. Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), Review of Research in Education, 1981, 9, 377-435.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton, NJ: Educational Testing Services, 1976.