

1-1-1985

# Using residual analyses to assess item response model-test data fit.

Linda N. Murray

*University of Massachusetts Amherst*

Follow this and additional works at: [https://scholarworks.umass.edu/dissertations\\_1](https://scholarworks.umass.edu/dissertations_1)

---

## Recommended Citation

Murray, Linda N., "Using residual analyses to assess item response model-test data fit." (1985). *Doctoral Dissertations 1896 - February 2014*. 4027.

[https://scholarworks.umass.edu/dissertations\\_1/4027](https://scholarworks.umass.edu/dissertations_1/4027)

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).



USING RESIDUAL ANALYSES TO ASSESS ITEM  
RESPONSE MODEL-TEST DATA FIT

A Dissertation Presented

By

LINDA N. MURRAY

Submitted to the Graduate School of the  
University of Massachusetts in partial fulfillment  
of the requirements for the degree of

DOCTOR OF EDUCATION

February 1985

Education

© Linda Murray  
All Rights Reserved

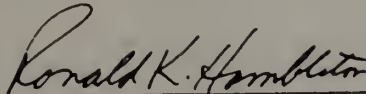
USING RESIDUAL ANALYSES TO ASSESS ITEM  
RESPONSE MODEL-TEST DATA FIT

A Dissertation Presented

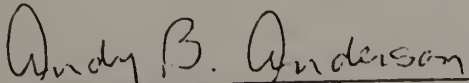
By


LINDA N. MURRAY

Approved as to style and content by:

  
\_\_\_\_\_  
Ronald K. Hambleton, Chairperson

  
\_\_\_\_\_  
Hariharan Swaminathan, Member

  
\_\_\_\_\_  
Andy B. Anderson, Member

  
\_\_\_\_\_  
Mario Fantini, Dean  
School of Education

## ACKNOWLEDGMENTS

Several people at the University of Massachusetts provided assistance to me during this study. First, and foremost, I especially want to thank my mentor, Dr. Ronald Hambleton. He gave me the opportunity, guidance and research support to carry out this project. I also wish to thank the other members of my thesis committee, Dr. H. Swaminathan and Dr. Andy Anderson for the knowledge they provided. Special thanks go to Dr. Janice Gifford for sharing her computer skills and to Ms. Bernie McDonald for her friendship and typing support.

I was fortunate to obtain a variety of data for this study. I wish to thank Dr. Paul Williams, Maryland State Department of Education, for the Maryland data sets. I also wish to acknowledge support from the National Institute of Education grant to do the National Assessment of Educational Progress research.

Lastly, I wish to thank my husband, Jim Murray, for providing love and emotional support, and for giving me confidence to complete my work.

ABSTRACT

Using Residual Analyses to Assess Item  
Response Model-Test Data Fit  
February, 1985

Linda N. Murray, B.S., State University College at Buffalo  
M.A., Trinity college  
Ed.D., University of Massachusetts  
Directed by: Professor Ronald Hambleton

Statistical tests are commonly used for studying item response model-test data fit. But, many of these tests have well-known problems associated with them. The biggest concern is the confounding of sample size in the interpretation of fit results. In the study, the fit of three item response models was investigated using a different approach: exploratory residual procedures. These residual techniques rely on the use of judgment for interpreting the size and direction of discrepancies between observed and expected examinee performances. The objectives of the study were to investigate if exploratory procedures involving residuals are valuable for judging instances of model-data fit, and to examine the fit of the one-parameter, two-parameter, and three-parameter logistic models to

National Assessment of Educational Progress (NAEP) and Maryland Functional Reading Test (MFRT) data.

The objectives were investigated by determining if judgments about model-data fit are altered if different variations of residuals are used in the analysis, and by examining fit at the item, ability, and overall test level using plots and simple summary statistics. Reasons for model misfit were sought by analyzing associations between the residuals and important item variables.

The results showed that the statistics based on average raw and standardized residuals provided useful fit information, but that when compared, the statistics based on standardized residuals presented a more accurate picture of model-data fit and therefore, provided the best overall fit information. Other results revealed that with the NAEP and MFRT type of items, failure to consider variations in item discriminating power resulted in the one-parameter model providing substantially poorer fits to the data sets. Also, guessing on difficult NAEP multiple-choice items affected the degree of model-data fit. The main recommendation from the study is that because the residual analyses provide substantial amounts of empirical evidence about fit, practitioners should consider these procedures as one of the several types of strategies to employ when dealing with the goodness of fit question.



TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . . iv

ABSTRACT . . . . . v

LIST OF TABLES . . . . . ix

LIST OF FIGURES . . . . . xii

CHAPTER

I INTRODUCTION . . . . . 1

    1.1 Statement of the Problem . . . . . 1

    1.2 Purpose of the Research . . . . . 4

    1.3 Research Questions . . . . . 5

    1.4 Organization of the Dissertation . . . . . 6

II REVIEW OF THE LITERATURE . . . . . 7

    2.1 Introduction . . . . . 7

    2.2 Concepts of Item Response Theory . . . . . 7

    2.3 Statistical Goodness of Fit Procedures . . . . . 13

    2.4 Exploratory Analytic Techniques . . . . . 19

    2.5 Summary . . . . . 24

III METHODS OF INVESTIGATION . . . . . 25

    3.1 Introduction . . . . . 25

    3.2 Description of Data Sets . . . . . 25

    3.3 Computer Programs . . . . . 27

    3.4 Research Procedures and Analyses . . . . . 29

    3.5 Summary Fit Statistics . . . . . 32

    3.6 Hypotheses Testing . . . . . 37

    3.7 Simulation Study . . . . . 38

    3.8 Summary . . . . . 39

IV RESULTS AND DISCUSSION . . . . . 40

    4.1 Introduction . . . . . 40

    4.2 Test for Normality . . . . . 40

    4.3 Comparison of Fit Statistics . . . . . 42

    4.4 Descriptive Results from Analyses of NAEP  
        Test Booklets . . . . . 50

CHAPTER

4.5 Item, Ability and Overall Fit . . . . .	54
4.6 Hypotheses Testing . . . . .	88
4.7 Analysis of the Maryland Functional Reading Test . . . . .	115
V SUMMARY, GUIDELINES, DELIMITATIONS AND CONCLUSIONS . . . . .	131
5.1 Summary . . . . .	131
5.2 Guidelines . . . . .	134
5.3 Delimitations . . . . .	136
5.4 Conclusions . . . . .	137
REFERENCES . . . . .	139

## LIST OF TABLES

Table		
4.2.1	Analysis of the Fit of the Standardized Residuals to a Normal Distribution, 720 Standardized Residuals . . . . .	41
4.3.1	Statistics of and Intercorrelations Among Several NAEP Math Item Variables for Booklets Nos. 1 and 2, 13 and 9 Year Olds, 1977-78 Assessment . . . . .	43
4.3.2	Average and Absolute Average Raw and Standardized Residuals at Twelve Ability Levels with the One-, Two-, and Three-Parameter Logistic Models for Booklet No. 1, 9 Year Olds, 65 Items, 1977-78 . . . . .	49
4.3.3	Comparison of Weighted and Unweighted Summary Test Fit Statistics for Four 1977-78 NAEP Mathematics Booklets . . . . .	51
4.4.1	Content Classification Summary of NAEP Math Booklet Nos. 1 and 2 Test Items for 9 Year Olds, 1977-78 Assessment . . . . .	52
4.4.2	Content Classification Summary of NAEP Math Booklet Nos. 1 and 2 Test Items for 13 Year Olds, 1977-78 Assessment . . . . .	53
4.4.3	Format and Content Classification of NAEP Math Booklet No. 1 Test Items for 9 Year Olds, 1977-78 Assessment . . . . .	55
4.4.4	Format and Content Classification of NAEP Math Booklet No. 2 Test Items for 9 Year Olds, 1977-78 Assessment . . . . .	58
4.4.5	Format and Content Classification of NAEP Math Booklet No. 1 Test Items for 13 Year Olds, 1977-78 Assessment . . . . .	61
4.4.6	Format and Content Classification of NAEP Math Booklet No. 2 Test Items for 13 Year Olds, 1977-78 Assessment . . . . .	63

Table

4.4.7	NAEP Math Item Model Parameter Estimates for 9 Year Olds, 1977-78 Assessment . . . . .	65
4.4.8	NAEP Math Item Model Parameter Estimates for 13 Year Olds, 1977-78 Assessment . . . . .	68
4.5.1	Analysis of Standardized Residuals with the One-, Two-, and Three-Parameter Logistic Models for Four NAEP Mathematics Booklets . . . . .	72
4.5.2	Analysis of Standardized Residuals at Twelve Levels with the One-, Two-, and Three- Parameter Logistic Models for Four 1977-78 NAEP Mathematics Booklets . . . . .	73
4.6.1	NAEP Math Booklet No. 1 Basic Item Statistical and Classificatory Information for 9 Year Olds, 1977-78 Assessment . . . . .	89
4.6.2	NAEP Math Booklet No. 2 Basic Item Statistical and Classificatory Information for 9 Year Olds, 1977-78 Assessment . . . . .	91
4.6.3	NAEP Math Booklet No. 1 Basic Item Statistical and Classificatory Information for 13 Year Olds, 1977-78 Assessment . . . . .	94
4.6.4	NAEP Math Booklet No. 2 Basic Item Statistical and Classificatory Information for 13 Year Olds, 1977-78 Assessment . . . . .	96
4.6.5	Association Between Standardized Residuals and NAEP Item Content Classification for Booklet Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment . . . . .	98
4.6.6	Analysis of Standardized Residuals with the One-, Two-, and Three-Parameter Logistic Models for Four NAEP Mathematics Booklets . . . . .	99
4.6.7	Association Between Standardized Residuals and Item Difficulties for Booklet Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment . . . . .	100
4.6.8	Descriptive Statistical Analysis of the Absolute-Valued Standardized Residuals for Booklet Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment . . . . .	105

Table

4.6.9	Relations Between Item Biserial Correlations and Standardized Residuals for Booklet Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment . . . . .	107
4.6.10	Representative Items for Four Patterns of Model Misfit for Math Booklet No. 1, 13 Year Olds, 1977-78 Assessment . . . . .	112
4.7.1	Maryland Functional Reading Test Item Statistics (N=2662; 1982) . . . . .	118
4.7.2	Analysis of the Absolute-Valued Standardized Residuals with Three Logistic Test Models for the MFRT . . . . .	122
4.7.3	Analysis of Standardized Residuals at Eleven Ability Levels with the One-, Two-, and Three-Parameter Logistic Models for the MFRT (N=2662, 75 items) . . . . .	123
4.7.4	Association Between Absolute-Valued Standardized Residuals and Item Content on the MFRT . . . . .	125
4.7.5	Association Between Absolute-Valued Standardized Residuals and Item Difficulties for the MFRT . . . . .	126
4.7.6	Statistical Analysis of the Absolute-Valued Standardized Residuals for the MFRT . . . . .	127
4.7.7	Relationship Between Item Biserial Correlations and Standardized Residuals for the MFRT . . . . .	128

## LIST OF FIGURES

Figure		
2.4.1	Standardized residual plot for an item with model-test data fit . . . . .	21
2.4.2	Standardized residual plot for an item with model-test data misfit . . . . .	22
4.3.1	Plot of one-parameter model raw residuals versus item discrimination . . . . .	45
4.3.2	Plot of one-parameter model standardized residuals versus item discrimination . . . . .	45
4.3.3	Plot of two-parameter raw residuals versus item discrimination . . . . .	46
4.3.4	Plot of two-parameter standardized residuals versus item discrimination . . . . .	46
4.3.5	Plot of three-parameter raw residuals versus item discrimination . . . . .	47
4.3.6	Plot of three-parameter standardized residuals versus item discrimination . . . . .	47
4.5.1	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 36 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	76
4.5.2	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 47 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	77
4.5.3	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 4 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	78
4.5.4	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 20 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	79

Figure

4.5.5	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 21 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	80
4.5.6	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 38 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	81
4.5.7	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 2 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	82
4.5.8	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 15 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	83
4.5.9	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 16 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	84
4.5.10	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 18 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	85
4.5.11	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 27 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	86
4.5.12	Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 29 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78) . . . . .	87
4.6.1	Scatterplot of one-parameter standardized residuals and item difficulties for 9 and 13 Year Olds Math Booklet Nos. 1 and 2 . . . . .	102
4.6.2	Scatterplot of two-parameter standardized residuals and item difficulties for 9 and 13 Year Olds Math Booklet Nos. 1 and 2 . . . . .	103

Figure

4.6.3	Scatterplot of three-parameter standardized residuals and item difficulties for 9 and 13 Year Olds Math Booklet Nos. 1 and 2 . . . . .	104
4.6.4	Scatterplot of one-parameter standardized residuals and item biserial correlations for 9 and 13 Year Olds Math Booklet Nos. 1 and 2 . . . . .	108
4.6.5	Scatterplot of two-parameter standardized residuals and item biserial correlations for 9 and 13 Year Olds Math Booklet Nos. 1 and 2 . . . . .	109
4.6.6	Scatterplot of three-parameter standardized residuals and item biserial correlations for 9 and 13 Year Olds Math Booklet Nos. 1 and 2 . . . . .	110
4.6.7	Four sample test items . . . . .	113
4.6.8	Standardized residual plot obtained with the two-parameter model for Item 4 . . . . .	116
4.6.9	Standardized residual plot obtained with the three-parameter model for Item 4 . . . . .	116
4.7.1	Plot of item absolute-valued standardized residuals obtained with the one-parameter model versus item biserial correlations . . . . .	129
4.7.2	Plot of item absolute-valued standardized residuals obtained with the two-parameter model versus item biserial correlations . . . . .	129
5.2.1	Guidelines for addressing the item response model selection question . . . . .	135



# C H A P T E R I

## INTRODUCTION

### 1.1 Statement of the Problem

Presently, there is considerable interest in applying the one-parameter, two-parameter, 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 (Lord, 1980; Hambleton, 1983; Yen, 1983; de Gruijter & Hambleton, 1983; Ironson, 1983; Cook & Eignor, 1983; Hambleton & Martois, 1983; Pandey & Carlson, 1983; Green, 1983). 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 theoretical test model can ever fit a data set perfectly. But without model-test data fit, the desirable features of the model may not be obtained. When a model does not adequately fit the test data, model predictions can be expected to be substantially less accurate. This problem is especially acute when the models are being used to predict outcomes when certain examinees did not respond to some items (Yen, 1981).

Several procedures that check for item response model-test data fit have been advocated and documented in the research literature.

Many of these methods involve statistical tests which compare and evaluate the differences between observed performance and the expected performance of samples of examinees based on a specified item response model (Wright & Panchapakesan, 1969; Andersen, 1973; Wright & Stone, 1979; Waller, 1981). Regardless of the specific significance test, the determination of the discrepancies involves the same basic 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 using statistical significance tests (Traub & Wolfe, 1981).

Generally, many of these statistical procedures that involve residuals attempt to employ the Pearson chi-square statistic or the likelihood ratio statistic. But, these statistical tests have limitations. Large examinee sample sizes are necessary for the test statistic to approach the appropriate theoretical distributions. The larger the sample size, the greater the probability of rejecting the null hypothesis that the model fits the data. Thus, the statistical test could indicate lack of model-data fit due principally to large sample sizes and not because of any practically significant departures between the model and data. With small sample sizes, on the other hand, even large practically significant differences in model data fit may not be detected using statistical tests because of the low level of statistical power (Hambleton & Murray, 1983).

The use of exploratory analytic techniques involving residuals is another means of examining item response model-test data fit. These techniques rely on the use of judgment for (1) interpreting the level of model-data misfit and for (2) comparing levels of fit between two or more models by analyzing the size and direction of the residuals. Exploratory analysis of residuals has played an important role in determining the suitability of regression models (Draper & Smith, 1966; Anscombe & Tukey, 1963; Kleinbaum & Kupper, 1978). But, these methods have not been used to any substantial extent to investigate the appropriateness of item response models.

The principal ways of interpreting the meaningfulness of the residuals in exploratory analysis are to:

1. Investigate and scrutinize simple summary statistics and inspect residual plots from several models for the purpose of choosing the model which best fits the test data.
2. Examine the signs of the residuals for non-random or unusual patterns of misfit.
3. Investigate standardized residuals to determine if they appear to be normally-distributed.

The residual plots are easy to do and often reveal patterns of misfit clearly. To create them, the residuals are calculated for each item by taking the differences between the actual item performance of an examinee ability group and the predicted performance level based on the chosen item response model. These residuals are next calculated and plotted at several ability levels. Visually the graphs are inspected for large absolute discrepancies between the model and the

test data and for sequences of "plus and minus" signs denoting peculiar arrangements of misfit. Evidence of possible model-data fit occurs when the residuals are relatively small and no apparent pattern in the direction of the misfit occurs. Lastly, raw residuals can be transformed into standardized residuals. A considerable number of standardized residuals beyond  $-2$  or  $+2$  standard deviation may suggest a misfit between the model and data.

### 1.2 Purpose of the Research

Residual analyses using exploratory techniques have received little or no attention in the area of item response theory. This study was undertaken to highlight the usefulness of using these methods for addressing the question of goodness of model to data fit. Specifically, this study had two main objectives. The first objective was to investigate if exploratory procedures involving residuals are valuable for judging instances of model-data fit. The second objective of this study was to examine using exploratory analytic techniques the fit of the one-parameter, two-parameter, and three-parameter logistic models to empirical test data to gain insights about each model's usefulness.

To carry out the first objective there was an investigation to determine if judgments about model-data fit are altered if different variations of residuals and their corresponding summary statistics were used in the exploratory data analyses. To carry out the second

objective, model-data fit was systematically analyzed with several data sets. Degrees of fit were examined at the item level, ability level, and for a complete test booklet. The degree of misfit was investigated across the three models by comparing the size of residuals. Reasons for model misfit were then sought by analyzing associations between the residuals and other important item variables.

### 1.3 Research Questions

In order to achieve the two broad objectives, this study was designed to answer seven research questions:

1. What are some of the statistical and graphical procedures for determining item response model-test data fit? Special interest centered on identifying procedures that involved residuals.
2. How do analyses of raw and standardized residuals compare in terms of describing model-data fit? A comparison of raw and standardized residuals was made to determine differences between the way they describe levels of model-data fit and whether the choice of statistic affects the decision about the usefulness of the item response model.
3. How do analyses of raw and standardized residuals compare when weighted and unweighted sample sizes are used in the analysis? A summary of fit statistics was calculated for each item across the ability groups. Comparisons were made among the weighted and unweighted statistics to see if they reveal similar impressions of model-test data fit.
4. How are exploratory analyses useful for detecting amounts of model-data misfit? This questions describes how to carry out residual analyses to determine the amount of model-data discrepancy.
5. How are residual analyses useful in helping to choose among the item response models? How do the one-parameter, two-parameter, and three-parameter item response models fit

empirical test results? Item plots and the size of the statistics were inspected and compared to reveal the varying amounts of misfit across the models.

6. What relationships exist between the fit of the test items and several item characteristics? Explanations for the differences found in the amount of misfit across the models were hypothesized by examining item difficulty, discrimination, format, and content. Specific test items were scrutinized to identify reasons why particular items misfitted a certain model or models.
7. If there is sufficient model-data fit, are the standardized residuals of the one-parameter, two-parameter, and three-parameter models distributed approximately normal? A study was carried out using simulated data to assess this question.

#### 1.4 Organization of the Dissertation

This chapter has provided an introduction to the research investigation. The next chapter deals with research question one: the relevant goodness of fit literature is reviewed. Chapter II also contains a brief introduction to the basic concepts of item response theory which pertain to this study. Chapter III contains a description of the empirical data sets and methodology used in this study. The results from the simulation study and the investigation of the empirical data sets are presented in Chapter IV. In the final chapter, a summary, guidelines, delimitations and conclusions of the research are provided.

## CHAPTER I I

### REVIEW OF THE LITERATURE

#### 2.1 Introduction

The purpose of this chapter is to provide a review of the literature related to the problem of assessing goodness of fit. Special effort was made to identify, describe, and evaluate those particular methods which involve the use of residuals. This chapter begins by reviewing a few of the basic concepts of item response theory that relate to this study. Next, several goodness of fit procedures are discussed. For convenience these methods are grouped under two categories: (1) statistical goodness of fit tests and (2) exploratory analytic techniques.

#### 2.2 Concepts of Item Response Theory

Currently, there is considerable interest in applying item response models to a wide variety of educational and psychological measurement problems. Sample free item statistics, test free person measurement, and the availability of a measure of the precision of the estimator of ability make item response theory an attractive alternative to classical test theory. Item response theory is based

on the assumption that performance on a test item is directly dependent upon one or more latent trait abilities. Since these traits cannot be observed directly they must be estimated from observed examinee responses on a set of test items (Hambleton & Cook, 1977).

An item response model describes mathematically the relationship between the underlying trait and examinee test performance. The more of this trait the examinee possesses, the more likely the examinee will respond correctly to the item. Mathematically, this last statement can be quantified as

$$P_i \sim f(\theta)$$

where  $P_i$  is the probability of success on item  $i$  and  $f(\theta)$  represents some function of a trait  $\theta$ .

This mathematical function,  $f(\theta)$ , that relates the probability of the success on an item to the ability measured by the item is called the item response function. Each particular item response model has its own form of the item response function. This study was limited to the use of the unidimensional one-parameter, two-parameter, and three-parameter logistic models because these particular models are in common use. The basic form of each of these models will be described next.

The three-parameter logistic model utilizes a family of item characteristic curves of the form

$$P_i(\theta) = c_i + (1-c_i) \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad [1]$$



where each item varies with respect to the three item parameters,  $a_i$ ,  $b_i$ , and  $c_i$ .  $D$  is a constant scaling factor with a value of 1.7 which is used to maximize the agreement between the logistic model and the normal-ogive model. The normal-ogive model was a model suggested by Lord in 1952, but was found later to be mathematically too complex for practical use. Therefore, it was replaced by the more convenient logistic model (Hambleton & Cook, 1977; Lord & Novick, 1968; Warm, 1978).

The item difficulty parameter,  $b_i$ , is the point on the ability scale where the slope of the item characteristic curve is maximum. Small values of  $b_i$ , indicate easy items and large values correspond to very difficult ones. This difficulty parameter is defined on the same scale as ability. Both are on a complete scale of  $-\infty$  to  $+\infty$ . But in practice, through scaling of ability estimates, item difficulty usually ranges from -2.0 to +2.0 when examinee abilities fall between -3.0 and +3.0 (Hambleton & Cook, 1977).

The item discrimination parameter,  $a_i$ , is proportional to the slope of  $P_i(\theta)$  when  $\theta=b_i$ . In most practical cases  $a_i$  ranges from .0 to +2.0. Larger values of  $a_i$  indicate items that are the most discriminating while smaller values suggest less discriminating items.

The last parameter is the  $c_i$  or pseudo-guessing parameter. It is the lower asymptote of the item characteristic curve and represents the probability of examinees with low ability correctly answering an item (Hambleton & Cook, 1977).

The one-parameter and two-parameter logistic item response models are simplifications of the more complex and general three-parameter logistic model. The two-parameter model has a family of item characteristics curves of the form

$$P_i(\theta) = \frac{e^{Da_j(\theta-b_j)}}{1+e^{Da_j(\theta-b_j)}} \quad [2]$$

This model assumes that guessing by low ability examinees does not take place. Hence the  $c$  parameter is set to zero.

The one-parameter model is the least complex and the most restrictive of the three models. The Rasch model, as it is sometimes referred to, has a family of item characteristic curves of the form

$$P_i(\theta) = \frac{e^{D(\theta-b_j)}}{1+e^{D(\theta-b_j)}} \quad [3]$$

When applying this model it is assumed that the items all discriminate equally well. The item characteristic curve depends strictly on the  $b_j$  parameter. In other words, an examinee's probability of success on an item is only determined by the item's level of difficulty and the examinee's ability level on the trait scale (Rasch, 1960; 1966).

The estimation of the ability and item parameters from the item response data can be handled several different ways (Hambleton & Swaminathan, 1984; Swaminathan, 1983; Swaminathan & Gifford, 1980;

1982). In this study, a maximum likelihood procedure was applied to the item response models. The procedure involves the joint or simultaneous estimation of the item and ability parameters.

There are two basic steps in the estimation procedure. First, the likelihood function ( $L$ ) of an event occurring in terms of an unknown value of the parameters is formed. Second, the values of the unknown parameters are found which produce the maximum of the likelihood function.

Specifically, for any of the three logistic models the probability that examinee  $\theta_a$  will answer item  $i$  correctly is

$$P_{ia} = \text{Prob} (X_{ia} = 1 | \theta_a; a_i, b_i, c_i) = \psi$$

where  $\psi$  is the appropriate equation (1), (2) or (3) depending upon the model chosen. Then the likelihood function ( $L$ ) for the response across  $N$  examinees on  $n$  items is given by

$$L = \prod_{a=1}^N \prod_{i=1}^n P_{ia}^{X_{ia}} (1-P_{ia})^{(1-X_{ia})}$$

where  $X_{ia} = 1$  if examinee  $a$  correctly responses to item  $i$  and  $X_{ia} = 0$  if otherwise.

To find the maximum of  $L$ , the logarithm of  $L$  is taken to convert products into sums for easy manipulation. Then the derivative of the  $\log L$  is found with respect to each unknown parameter. Next, these expressions are set equal to zero forming the likelihood equations. The number of likelihood equations which must be solved depends upon the chosen model. For the one-parameter, two-parameter, and three-

parameter models there are  $n+N-2$ ,  $2n+N-2$ , and  $3n+N-2$  equations, respectively. The exact form of these equations are given by Birnbaum (1968).

The resulting system of equations must be solved simultaneously. A two-step iterative process is commonly used. First, initial item parameter estimates are held constant and abilities are estimated. Next, abilities are held constant and item parameters are estimated. These item parameters are then used to obtain new ability estimates, and so on. The process is repeated until convergence of the estimates is obtained (Wingersky, 1983).

Once the group of likelihood equations is solved a set of item and ability estimates are available which are theoretically invariant. That is, unlike classical test theory item statistics, item parameters are group independent and regardless of the ability level of a group of examinees responding to an item, the item parameters remain the same. Similarly, the ability estimates are sample invariant. Once the group of items are calibrated, the ability estimates do not depend upon the choice of the items which the examinee has taken (Bejar, 1983; Hambleton & Cook, 1977).

The sample-free nature of item response parameters can only be obtained if the assumptions of item response theory are met to a sufficient extent. Since item response models require strong assumptions it is difficult to construct tests to meet these requirements (Traub & Wolfe, 1981; Traub, 1983; Divgi, 1980, 1981).

One of the strong assumptions of item response models is unidimensionality. The three logistic models being used in this study require that the probability of a correct response on an item depends only on one unobservable trait or ability. There exists no widely accepted method for investigating this important assumption of unidimensionality of a test. In practice, factor analysis is commonly used (Hambleton, Murray, & Simon, 1982). More recently other methods including nonlinear factor analysis have been investigated as viable methods for assessing unidimensionality (Hambleton & Murray, 1984; Gerritz, 1984; Jungblut, 1984; Cook & Eignor, 1984).

If it can be shown that the assumptions of the models are met, then as long as the model fits the data, the advantages of the models are realized. The next section of this chapter contains ways of determining fit between an item response model and observable test data. The procedures described concentrate on those methods that use residuals.

### 2.3 Statistical Goodness of Fit Procedures

Several types of statistical tests have been applied to item response models for assessing model-data fit (Wright & Panchapakesan, 1969; Wright, Mead & Bell, 1979; Andersen, 1973; Waller, 1981). A popular procedure is the Wright and Panchapakesan method (1969) and variations of it. This procedure uses a Pearson chi-square approach for evaluating one-parameter model fit.

Hambleton et al. (1978) clearly described this method. First an item by total score matrix is created. Then the number of examinees at the  $i^{\text{th}}$  test score level answering the  $j^{\text{th}}$  item correctly ( $O_{ij}$ ) is compared to the expected number of examinees predicted from the model to the  $j^{\text{th}}$  item correct ( $E_{ij}$ ). Then, the quantity  $Y_{ij}$ , where

$$Y_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{n_i E_{ij} (1 - E_{ij})}}$$

is a unit normal deviate and thus  $Y_{ij}^2$  would be asymptotically distributed as a chi-square with 1 degree of freedom.

Finally, summing across all items and score groups the total test fit statistic is found by

$$\chi^2 = \sum_{i=1}^{n-1} \sum_{j=1}^n Y_{ij}^2$$

with  $(n-1)(n-2)$  degrees of freedom.

It has been well-documented that this chi-square test statistic has several problems associated with it (Traub & Wolfe, 1981; Hambleton, Murray, & Simon, 1982; Hambleton, et al., 1978). First, when the expected terms have values less than one,  $Y_{ij}$  will not be normally distributed. Thus,  $Y_{ij}^2$  does not follow a chi-square distribution.

Second, the test statistic  $Y_{ij}^2$  is only asymptotically distributed as a chi-square. Large sample sizes of examinees are

necessary for the test statistic to approach a chi-square distribution and for accurate estimates of the item and ability parameters. But the larger the sample sizes the greater the probability of rejecting the null hypothesis that the model fits the data because this method is sensitive to sample size.

Hambleton and Murray (1983) illustrated the problem associated with examinee sample sizes and this statistical test of model-data fit. A simulation study was carried out to conduct a one-parameter model fit analysis of three-parameter model simulated test data for examinee samples of varying sizes.

Using the 1979 version of BICAL, a "t-statistic" based on the Wright and Panchapakesan method was analyzed to show the impact of sample size on the detection of misfitting items. The results of this investigation showed clearly that as the size of the examinee group increased (from 150 to 300, 600, 1200, and finally 2400) the number of misfitting items increased. Using the .01 significance level, the range was 5 to 38 items and for the .05 level from 20 to 42 items. Clearly interpretation of misfit in this case is clouded by the direct relationship between examinee size and the number of misfitting items.

Finally, several authors have suggested that there are other problems associated with the fit test (Divgi, 1981; Traub & Wolfe, 1981; Van den Wollenberg, 1979). Among their concerns are that the overall statistic does not have a chi-square distribution because the

$Y_{ij}^2$  are not independent and the associated degrees of freedom have been assumed to be higher than they actually are.

Bock and Lieberman (1970) developed a method using the chi-square test to analyze examinee response patterns. Their procedure is called a vector frequency test. By assuming ability is normally distributed and the item parameters known, the method allows for the specification of expected frequency for all response patterns. A Pearson chi-square statistic is then computed between the observed and expected frequencies. Again for a large number of cases the statistic can be referenced to the tabled chi-square distribution.

Finally, Yen (1981) developed a fit statistic similar to the Wright and Panchapakesan chi-square statistic for the one-parameter, two-parameter, and three-parameter models. Grouping for the analysis was carried out by using estimated ability and not number right score. The author also compares the results of the chi-square fit statistic to variations proposed by other researchers.

Once again these procedures suffer from all the problems associated with asymptotic statistical significance tests. Additionally, if there is insufficient model-data fit it will not be clear why. True misfit could exist or the observed misfit could be caused by not meeting the assumption of a normal ability distribution.

Another common statistical procedure for testing goodness of fit is the likelihood ratio significance test. Several authors have suggested ways to apply this procedure to assess item response



model-test data fit (Waller, 1981; Gustafsson, 1980; Wainer, Morgan, & Gustafsson, 1980; Andersen, 1973). When maximum likelihood estimates of the item and ability parameters are obtained, likelihood tests can be performed to statistically judge the fit of a particular model. Also, likelihood ratio tests offer the possibility of assessing the fit of a particular item response model against an alternative.

Traub and Wolfe (1981) explain the relationship between the residuals and the likelihood function. If in practice respondents are correctly answering an item, the model should, if it accounts for the data, predict a high level of probable success on the item. Similarly, when the respondents are incorrectly answering the items, the model should predict a probable low level of success on the items. In these cases, the difference between the responses and the probability of success or failure should be small.

There are several steps involved with carrying out a likelihood ratio test. First, a series of hypotheses must be generated specifying the expected patterns of parameter values given that the hypothesis about the parameters is true (Crane, 1980). These hypotheses may be formulated in a number of ways. In the simplest case a null hypothesis is specified along with a suitable alternative hypothesis.

Next, a measure of relative likelihood or plausibility is assigned to each of the hypotheses. When applying the likelihood ratio test to item response theory, assignment of likelihood means

finding the maximum of the likelihood function when (1) the null hypothesis is true and (2) when the alternative hypothesis is true.

Finally, the ratio,  $\lambda$ , of the likelihood of the null hypothesis to the likelihood of the alternative hypothesis is formed. This ratio,  $\lambda$  is the criterion for testing the null hypothesis against the alternate hypothesis. In the item response model application,  $\lambda$  represents the ratio of the maximum value of the likelihood function under the null hypothesis to the maximum value of the likelihood function under the alternative. If the sample size is large, then the quantity  $-2 \log \lambda$  has a chi-square distribution. The degrees of freedom is the difference in the number of parameters estimated under the null and alternative hypothesis (Hambleton et al., 1978).

Andersen (1973) and Bock and Liebermann (1970) used the likelihood ratio significance test to assess the fit of the one-parameter and two-parameter models, respectively. Their tests indicate whether the size of the residuals is consistent with random fluctuations within a model. On the other hand, Waller (1981) applied the likelihood ratio test for making comparisons among the item response models.

The limitations associated with the likelihood chi-square approach are similar to the ones associated with the Pearson chi-square statistic and are reviewed by Traub and Wolfe (1981). Again the likelihood ratio test has a test criteria distribution of chi-square only asymptotically. But, as was mentioned earlier, when

large samples are used to accommodate this need, the chi-square value may become significant due principally to the large sample size.

#### 2.4 Exploratory Analytic Techniques

In an attempt to avoid many of the problems associated with statistical significance tests, a few researchers have advocated the use of exploratory analytic techniques for judging the importance of degrees of model-data misfit (Hambleton & Murray, 1983; Hambleton, Murray, & Simon, 1982; Traub & Wolfe, 1981; Kingston & Dorans, 1982; Hutten, 1981). Exploratory analytic techniques involve examining the size and direction of discrepancies between observed and expected levels of performance without performing statistical significance tests. Instead, the residuals are examined by inspecting the residual plots and by calculating simple summary statistics. This allows the investigator the opportunity to determine overall fit and to isolate particular instances of misfit.

The basic process useful for carrying out an exploratory analysis of residuals is outlined by Hambleton, Murray, and Simon (1982, p. 29). The residuals are calculated in the following manner: an item response model is chosen; item and ability parameters are obtained; and predictions of the performance of various ability groups on the items are made, assuming the validity of the chosen model. Then comparisons of the predicted results with actual results are made

to obtain a measure of fit between the estimated item characteristic curve and the observed test data.

The ability groups used in the method are obtained by splitting the ability scale into distinct sections which are wide enough to contain a reasonable number of examinees. Then for each ability category the average observed performance is compared to predicted performance to determine the degree of misfit.

Plots of the residuals across ability groups for an item can be created. Examples of the type of graphs that can be produced for analysis are shown in Figures 2.4.1 and 2.4.2. Figure 2.4.1 shows an item residual plot where differences between the observed data and an estimated item characteristic curve across the ability groups are small. All the standardized residuals are less than 2.0 standard deviations. Also, the residuals have a random direction of misfit across the ability continuum. Therefore, for this item, there appears to be model-data fit.

Figure 2.4.2 shows the item residual plot where the discrepancy between observed and expected performance of low ability examinees is large. Most of the residuals at the lower end of the ability scale are beyond 2.0 standard deviations. Clearly, for this item, there is model-data misfit.

Besides detecting misfit, reasons for the misfit can be hypothesized and later assessed. For example, a possible explanation of the large sized positive residuals at the lower end of the ability

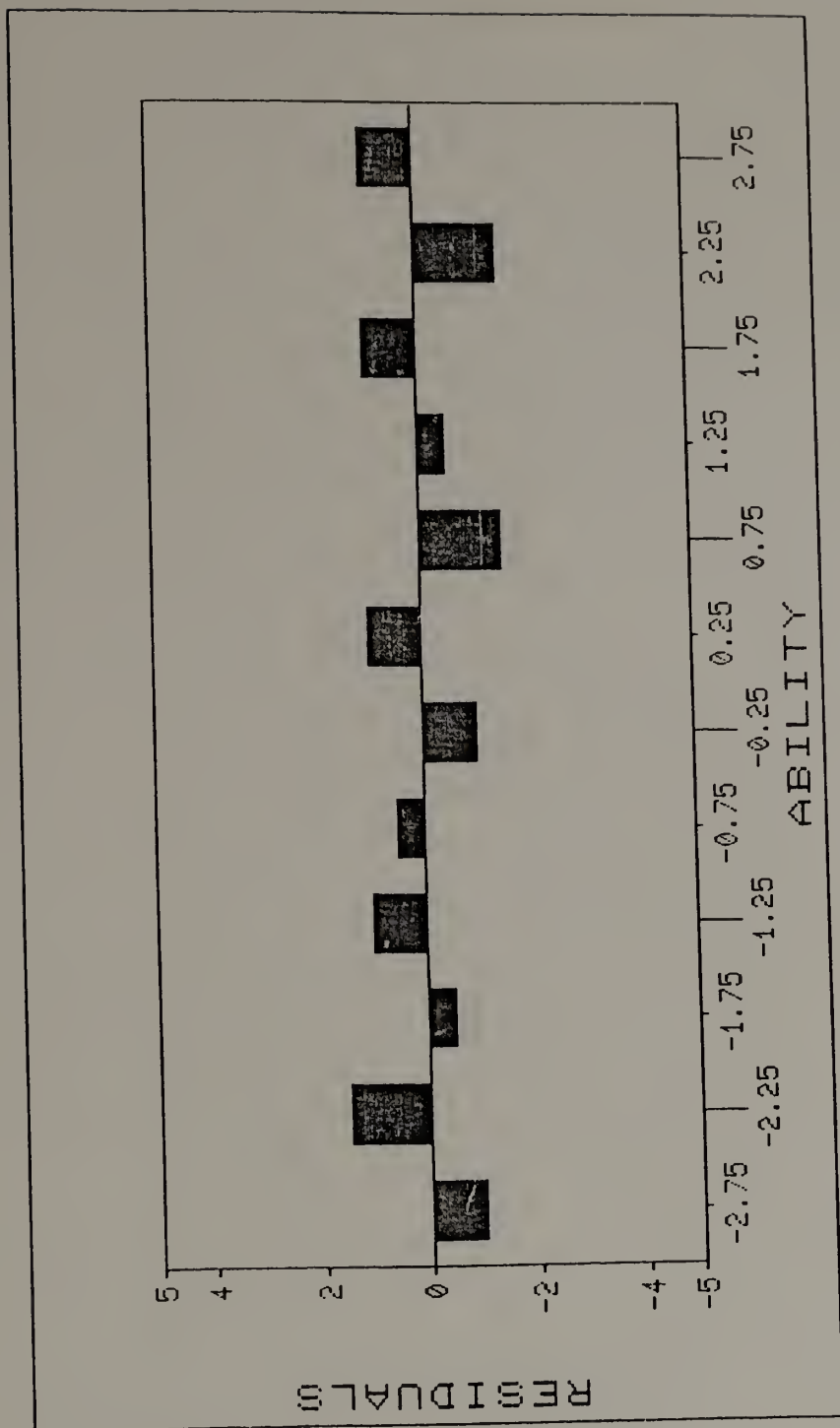


Figure 2.4.1. Standardized residual plot for an item with model-test data fit.

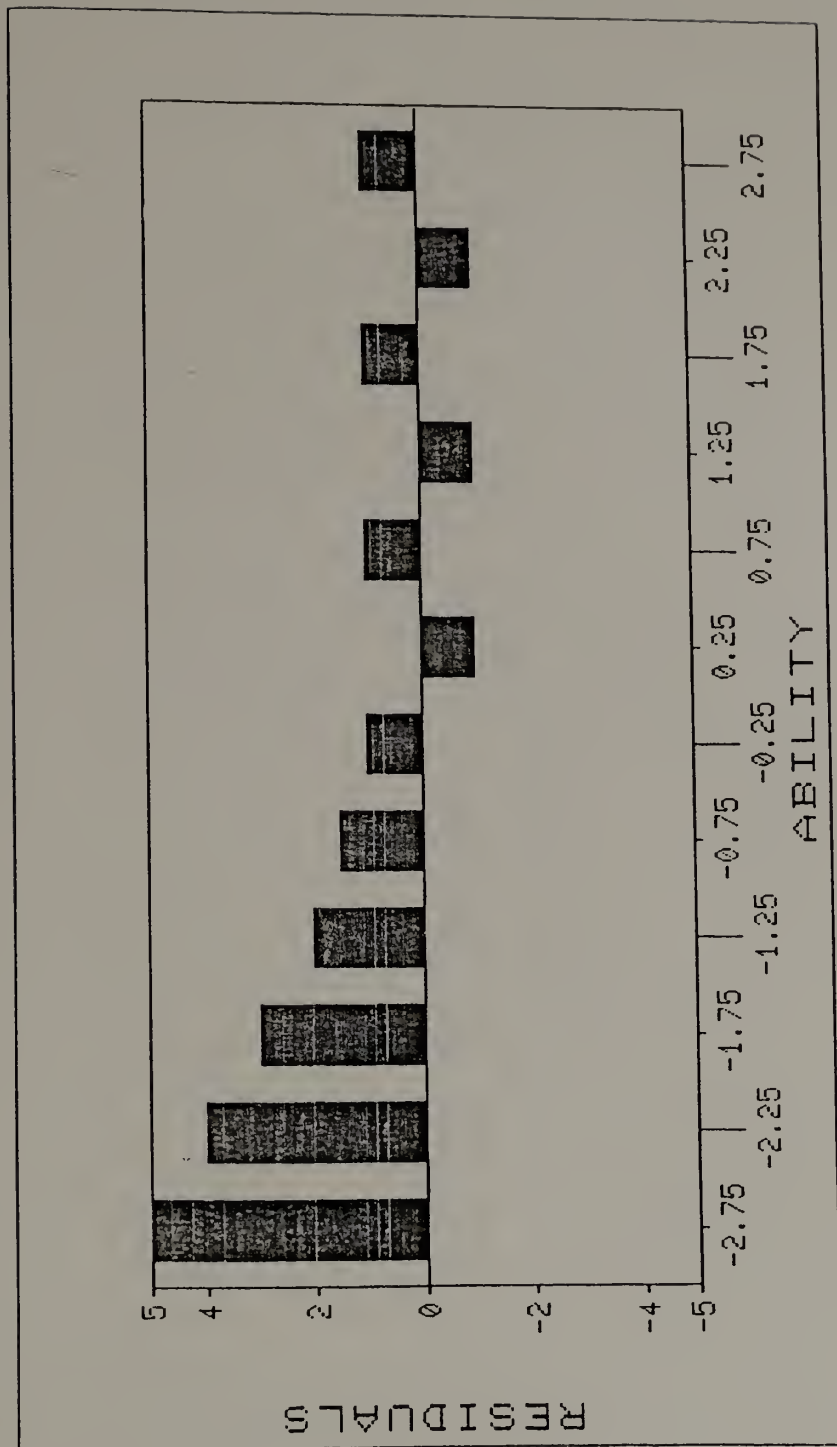


Figure 2.4.2. Standardized residual plot for an item with model-test data misfit.

scale in Figure 2.4.2 is that the prespecified model failed to account for guessing by low ability examinees. This explanation can be verified by studying the associations between the size of the residuals and the difficulty and format of the item.

Kingston and Dorans (1982) also suggest examining goodness of fit of item response models by inspecting and interpreting residual plots. Their method of calculating and studying residuals is very similar to the one just described. Differences also exist. For example, residuals are adjusted for omits. They also restrict their attention to analyzing the fit of only the three-parameter logistic model. No comparisons of model-data fit is suggested among the one-parameter, two-parameter, and three-parameter models.

Several other authors have encouraged researchers to carry out exploratory analyses of residuals. Traub and Wolfe (1981) suggest that testing a particular model's fit to test data should not be restricted to statistical significance tests. Instead supplemental analyses are necessary which use more informal means of analyzing model-data fit. They advise that the researcher should become a data analyst. Decisions about model-data fit should be based on statistical as well as less technical evidence.

Finally, Hambleton and Murray (1983) suggest that the decision of whether or not adequate model-data fit exists should be based ultimately on informed judgments. Statistical tests can be carried out but care must be shown in interpreting the statistical

information. Statistically significant differences may be observed even though the practical significance of these differences is low.

### 2.5 Summary

The purpose of this chapter was to present a review of the current research literature which pertains to this study. A number of methods identified in this chapter employed the Pearson chi-square and likelihood chi-square asymptotic tests of statistical significance. Some evidence presented suggests that these statistical tests may be limited in their ability to detect true model-data misfit. Several authors have suggested carrying out exploratory residual analyses. On the surface these methods appear to be relatively simple to conduct and very effective. Therefore, the purpose of this study was to explore the viability and applicability of these analyses for the expressed purpose of determining instances of model-data fit. The next chapter contains the methodology which was used to carry out these research studies.



## C H A P T E R   I I I

### METHODS OF INVESTIGATION

#### 3.1 Introduction

In this chapter (1) the particular tests, exercises, and examinees chosen for analysis are described, (2) the computer programs used to carry out the residual analyses are presented, and (3) the design of the studies and the procedures by which the research was conducted are delineated and explained.

#### 3.2 Description of Data Sets

Two different empirical test data sets were used in this study. They included National Assessment of Educational Progress (NAEP) test data and Maryland Functional Reading Test (MFRT) data. A description of the test forms and examinee response data is offered next.

First, a maximum of four 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 of these test booklets was administered to a carefully chosen nationally representative sample of approximately 2500 examinees. They contain test items measuring various mathematical skills in the areas of definition, story problems, geometry, measurement, and graphs and figures. There are both multiple-choice and open-ended items in the test booklets.

It should be noted that from a preliminary analysis of some of the test booklets, items appeared to vary substantially in levels of difficulty and discrimination. Because of the wide range of item discrimination indices and the anticipated high level of guessing due to the substantial number of difficult multiple-choice items, it was expected that the more general item response models would fit the test data considerably better than the most restrictive model.

Second, the test response data from the Fall 1982 test administration of the Maryland Functional Reading Test--Level II were analyzed in the residual investigations. This test was given to approximately 55,000 ninth graders in the Maryland public schools. For the purpose of these analyses a five percent sample was taken. Specifically, every twentieth examinee from the master examinee file was drawn.

This reading test consists of 75 multiple-choice items from five content domains. These five areas included following directions, locating information, main idea, using details and understanding forms. A diagnostic subscore is reported for every examinee to teachers and parents on each of these five content areas. This test must be passed before students are eligible for graduation from high school.

The Maryland test data set was chosen for this study because of the unusual way in which the items were selected for the final test form. Not only did the test items have to satisfy content considerations, but many of the items were only included in the final form if they fit the one-parameter model "adequately." In this case expectations were that the one-parameter model would fit the test data about as well as the more general item response models.

### 3.3 Computer Programs

Item and ability parameter estimation for the one-parameter, two-parameter, and three-parameter logistic models was carried out through the use of LOGIST, (Wood, Wingersky, & Lord, 1976; Wingersky, 1983). It is a FORTRAN IV program which uses maximum likelihood procedures to estimate the parameters. No limits were placed on the values of estimates for ability or difficulty. Restrictions were placed on examinee guessing and item discrimination depending upon the assumptions of the specific model that was being used in the analysis.

For example, when the one-parameter model parameters were being estimated, the  $c_j$  parameter was set to 0.0 and the  $a_j$  parameter to 1.00. Also to assure that solutions could be obtained for the two-parameter and three-parameter model parameters, a limit of 2.0 was placed on the maximum value for estimation of the  $a_j$  parameter. This solution handled the problem of upward drift of some estimates (Gifford, 1983; Swaminathan & Gifford, 1983).

The residual analyses were accomplished through the use of RESID (Murray & Hambleton, 1984). This FORTRAN V computer program can calculate raw and standardized residuals and a summary of fit statistics across items for an ability group, across all ability groups for an item, and across all ability groups and items (i.e., total test booklet). Also, RESID provides a summary of fit statistics for each item across ability groups using weighted and unweighted sample sizes.

Finally, simulated data needed for this study were generated from the computer program DATAGEN (Hambleton & Rovinelli, 1973). This FORTRAN IV program generates examinee response data from logistic test models. The program is designed to produce a set of response patterns and test scores to represent the performance of  $N$  examinees on  $n$  items scored 0 or 1. The population characteristics for the distribution of ability and item parameters are specified. Once these true values are determined, the binary response data is generated according to the item response model of interest.

### 3.4 Research Procedures and Analyses

In this section the design of the studies and the procedures by which they were conducted will be considered. It should be noted that the procedures used to analyze the NAEP and MFRT data are identical.

#### Comparison of Raw and Standardized Residuals

Each analysis 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 is first chosen. For this study the one-parameter, two-parameter, and three-parameter 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 is placed in an ability category based on his or her estimated ability level. For these investigations, ability categories were chosen so that the ability scale between -3.0 and 3.0 was divided into twelve equal intervals. Ability estimates that fell beyond these maximum and minimum ability levels were deleted from the analyses. Next, for each of the twelve ability categories, the average observed performance ( $P_{ij}$ ) for item  $i$  in ability category  $j$  is 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 item  $i$  ( $i=1,2,\dots,n$ ) 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 by:

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

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

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

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  $\theta_j$  is the mid-point of the  $j^{\text{th}}$  ability category.

Then the raw residuals ( $R_{ij}$ ) for item  $i$  in ability category  $j$  was found by

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

This difference is an index of the degree of misfit between the test data and the expected item performance based on the chosen item response model.

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,

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

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  $\hat{P}_{ij}$ . When  $N$  is small, other things being equal, big differences between actual and expected performance must be obtained for the differences to be taken as an indication of model-test data misfit.

A comparison of raw and standardized residuals was made to determine how differently they describe levels of model-data fit and whether the choice of statistic affects the decision about the usefulness of the item response models. The size and direction of the raw and standardized residuals in the analyses were compared at the item level, ability level and test level. The equations used to find these summary fit statistics are found in the next section in this chapter.

An additional 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, the worst fitting items were identified. Next, the same number of poorest fitting items on the absolute-valued raw residual score scale were found. Then the percent of items in common to the two analyses were calculated to indicate the level of agreement in the identification of misfitting items.

Finally, several intercorrelations were calculated between the raw residuals and standardized residuals across the three logistic item response models and between other important item variables. The item variables included item order, item format, classical item difficulty, and classical item discrimination. Non-linear relationships between the residuals and the item variables, were investigated by examining scatterplots and using such statistics as eta.

### 3.5 Summary Fit Statistics

Summary fit statistics based on both raw and standardized residuals were used in the study. These statistics describe overall fit for each test item (found by summing over ability groups), for each ability group (found by summing over test items) and for the total test (found by summing over ability groups and items). The equations used to calculate the statistics are listed next. They are



organized into three sections. Equations 1.1 through 1.10 represent the item fit statistics. Equations 2.1 through 2.6 represent the ability fit statistics. Lastly, test fit statistics are calculated with Equations 3.1 through 3.6.

### Item Summary Statistics

#### Unweighted average raw residual

$$\bar{RR}_i = \frac{\sum_{j=1}^K RR_{ij}}{K} \quad [1.1]$$

where K is the number of ability groups

#### Weighted average raw residual

$$\overline{WRR}_i = \frac{\sum_{j=1}^K N_j RR_{ij}}{\sum N_j} \quad [1.2]$$

where  $N_j$  is the number of examinees in the  $j^{\text{th}}$  ability group

#### Unweighted absolute-valued average raw residual

$$|\bar{RR}_i| = \frac{\sum_{j=1}^K |RR_{ij}|}{K} \quad [1.3]$$

#### Weighted absolute-valued average raw residual

$$|\overline{WRR}_i| = \frac{\sum_{j=1}^K N_j |RR_{ij}|}{\sum N_j} \quad [1.4]$$

Root mean squared differences for raw residuals

$$\text{RMSSR}_i = \sqrt{\frac{\sum_{j=1}^K (RR_{ij} - \bar{RR}_i)^2}{K}} \quad [1.5]$$

Unweighted average standardized residual

$$\bar{SR}_i = \frac{\sum_{j=1}^K SR_{ij}}{K} \quad [1.6]$$

Weighted average standardized residual

$$\bar{WSR}_i = \frac{\sum_{j=1}^K N_j SR_{ij}}{\sum N_j} \quad [1.7]$$

Unweighted absolute-valued average standardized residual

$$|\bar{SR}_i| = \frac{\sum_{j=1}^K |SR_{ij}|}{K} \quad [1.8]$$

Weighted absolute-valued average standardized residual

$$|\bar{WSR}_i| = \frac{\sum_{j=1}^K N_j |SR_{ij}|}{\sum N_j} \quad [1.9]$$

Root mean squared difference for standardized residuals

$$\text{RMSSR}_i = \sqrt{\frac{\sum_{j=1}^K (SR_{ij} - \bar{SR}_i)^2}{K}} \quad [1.10]$$

Ability Summary StatisticsAverage raw residual

$$\overline{RR}_j = \frac{\sum_{i=1}^n RR_{ij}}{n} \quad [2.1]$$

where  $n$  is the number of items in the booklet.

Absolute-valued average raw residual

$$|\overline{RR}_j| = \frac{\sum_{i=1}^n |RR_{ij}|}{n} \quad [2.2]$$

Root mean squared differences for raw residuals

$$RMSRR_j = \sqrt{\frac{\sum_{i=1}^n (RR_{ij} - \overline{RR}_j)^2}{n}} \quad [2.3]$$

Average standardized residual

$$\overline{SR}_j = \frac{\sum_{i=1}^n SR_{ij}}{n} \quad [2.4]$$

Absolute-valued average residual

$$|\overline{SR}_j| = \frac{\sum_{i=1}^n |SR_{ij}|}{n} \quad [2.5]$$

Root mean squared differences for standardized residuals

$$RMSSR_j = \sqrt{\frac{\sum_{i=1}^n (SR_{ij} - \overline{SR}_j)^2}{n}} \quad [2.6]$$

Test Summary Statistics

Overall average raw residual

$$\overline{RR} = \frac{\sum_{j=1}^K RR_j}{K} \quad [3.1]$$

Overall absolute-valued average raw residual

$$|\overline{RR}| = \frac{\sum_{j=1}^K |RR_j|}{K} \quad [3.2]$$

Overall root mean squared differences for raw residuals

$$RMSSR = \sqrt{\frac{\sum_{j=1}^K (RR_j - \overline{RR})^2}{K}} \quad [3.3]$$

Overall average standardized residual

$$\overline{SR} = \frac{\sum_{j=1}^K SR_j}{K} \quad [3.4]$$

Overall absolute-valued standardized residual

$$|\overline{SR}| = \frac{\sum_{j=1}^K |SR_j|}{K} \quad [3.5]$$

Overall root mean squared difference for standardized residuals

$$RMSRR = \sqrt{\frac{\sum_{j=1}^K (SR_j - \overline{SR})^2}{K}} \quad [3.6]$$

### 3.6 Hypotheses Testing

Several testable research hypotheses were generated concerning model-data fit. Specifically interest centered on determining if test items having large positive and/or negative residuals exhibit certain salient item characteristics that would cause them to be misfit by an item response model. Analyses were conducted concerning the association between the fit of the test items and item content, item format, and classical indices of item difficulty and discrimination. The fit of the test item content, item format, and classical indices of item difficulty and discrimination. The fit of the test items in each of these analyses was represented by average absolute-valued standardized residuals across the three item response models.

The specific procedures used to study these relationships included crosstabulation tables, chi-square statistics, and scattergrams. For example, crosstabulation tables were created to investigate the pattern of fit across the three models (as represented by "small" and "large" residuals) and item format (multiple-choice versus open-ended), classical difficulty (easy versus hard), both item format and classical item difficulty levels, classical item discrimination (grouped into three or four categories) and item content (items classified into categories based on content type).

The previous analyses presented results about trends of misfit across a number of test items. Are there any specific reasons why particular items misfit a certain model or models? To answer this

question, items and their corresponding residuals were scrutinized individually to find patterns across the model. For example, one pattern was that there is similar fit across the three models. Next, for each pattern several representative items were examined carefully in order to identify possible salient item characteristics causing instances of fit or misfit. One such variable included item wording.

### 3.7 Simulation Study

When there is sufficient model-data fit, standardized residuals of the one-parameter, two-parameter, and three-parameter models were assumed to be distributed approximately normal. To test this assumption, a study was carried out using artificial data which fit the chosen model. Specifically, data was generated according to each of the three models of interest through use of the DATAGEN computer program (Hambleton & Rovinelli, 1973). For this study, program options were specified to create dichotomous responses on a 60 item test for 2500 examinees.

Next, as previously described in section 3.4, standardized residuals were computed. Finally, comparisons were made using the Kolmogorov-Smirnov test statistic to determine if the standardized residuals, under the model-data fit condition, are normally distributed.

### 3.8 Summary

This chapter presented the methodology which was followed in this study. The design and set of procedures used in the research studies were explicated and described. In the next chapter, the results from these sets of analyses are presented and discussed.

## C H A P T E R I V

### RESULTS AND DISCUSSION

#### 4.1 Introduction

This chapter includes (1) the results of the simulation study conducted to test the normality assumption, (2) the results obtained from comparing the different summary fit statistics, (3) the descriptive results from analyses of the NAEP test booklets and data, (4) the fit results of the one-parameter, two-parameter, and three-parameter logistic models at the item, ability and overall test level, (5) the standardized residual plots of NAEP items of varying difficulty and discrimination, (6) the results from hypothesizing associations between item characteristics and levels of fit, and (7) the results of the various analyses involving the Maryland Functional Reading Test.

#### 4.2 Test for Normality

If there is sufficient model-test data fit, then the distribution of standardized residuals used in this study were assumed to be normal. To test this normality assumption, a simulation study was carried out with the one-parameter, two-parameter, and three-parameter logistic models using artificial data. The results are presented in Table 4.2.1. Table 4.2.1 contains the results from the



Table 4.2.1

Analysis of the Fit of the Standardized Residuals  
to a Normal Distribution, 720 Standardized Residuals

Logistic Model	K-S Statistic	Percent of Residuals			
		0 to 1	1 to 2	2 to 3	over 3
		(68.3)	(27.2)	(4.3)	(.26)
1	1.355 p = .051	71.81	24.31	3.75	.14
2	1.319 p = .062	66.39	29.44	3.89	.28
3	2.463 p = .000	66.25	30.00	3.33	.42

Note: The values in the parentheses represent the percent of cases under areas of the normal curve.

analysis of the fit of the standardized residuals to a normal distribution. From this table, it can be seen that the distribution of the standardized residuals appears to be approximately normal for all three models as represented by the percent of residuals along different points on the normal curve. The results of the Kolmogorov-Smirnov (K-S) Test of goodness of fit on the standardized residuals showed that there appeared to be no statistical difference between the distribution of the one-parameter and two-parameter standardized residuals and the normal distribution.

In the case of the three-parameter standardized residuals, the results were somewhat different. The three-parameter standardized residuals failed the K-S statistical test. However, the value of the K-S statistic was not considerably larger than the values obtained for the one-parameter and two-parameter models. The problem may have been caused by poor parameter estimates, especially for the c parameter. The exact reason for this result is unclear and should be a topic for future research.

#### 4.3 Comparison of Fit Statistics

Table 4.3.1 displays the intercorrelations among several of the NAEP math item variables. There is a strong relationship between the one-parameter absolute-valued 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

Table 4.3.1

Statistics of and Intercorrelations Among Several NAEP Math Item Variables for Booklet Nos. 1 and 2, 13 and 9 Year Olds, 1977-78 Assessment

Variable	Mean	Standard Deviation	SR(2-P)	SR(3-P)	RR(1-P)	RR(2-P)	RR(3-P)	P	F <sup>1</sup>	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	.77	.30	-.21	-.13	.00
Standardized Residual (3-P)	.88	.42	.15	.27	.77	.27	.09	.07	.07	-.03
Raw Residual (1-P)	.060	.033	.24	.34	.17	.24	.34	-.17	-.19	.09
Raw Residual (2-P)	.033	.017	.43	.43	.13	.43	.43	-.22	-.34	.13
Raw Residual (3-P)	.030	.017	-.07	-.17	.14	-.07	-.17	-.17	-.17	.14
Item Difficulty (P)	.53	.27						.04	.04	-.40
Format (F)										-.12
Item Order (O)										

<sup>1</sup>Two types: Multiple-choice and Open-ended.

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 reflected in the corresponding standard deviations.

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 4.3.1 to 4.3.6. Figures 4.3.1 and 4.3.2 are plots of raw residuals and standardized residuals versus classical item biserial correlations. These figures show that for the one-parameter model, a curvilinear relationship prevailed whether raw or standardized residuals were used to describe fit. Very low or high discriminating items had larger residuals with the one-parameter model. However, some differences between the results in these plots emerged for lower discriminating items. Similarly, Figures 4.3.3 to 4.3.6 display the plots of the residuals versus item biserial correlations for the two-parameter and three-parameter models. These plots suggest strong agreement between the residuals except again 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 absolute-valued raw and standardized residuals was carried out with the

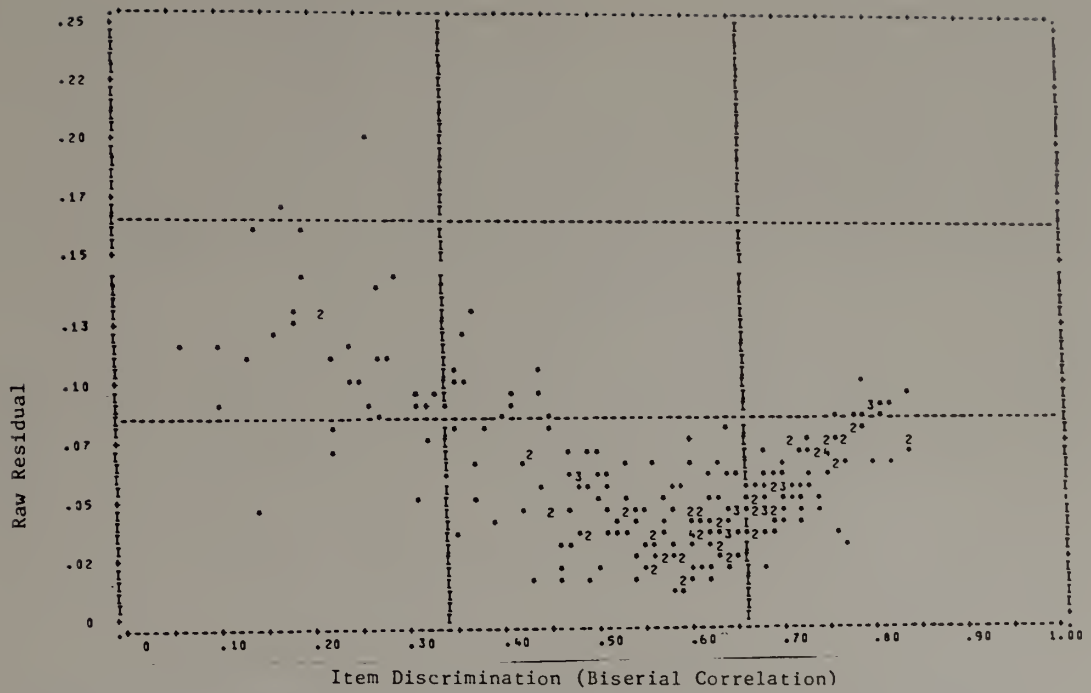


Figure 4.3.1. Plot of one-parameter model raw residuals versus item discrimination.

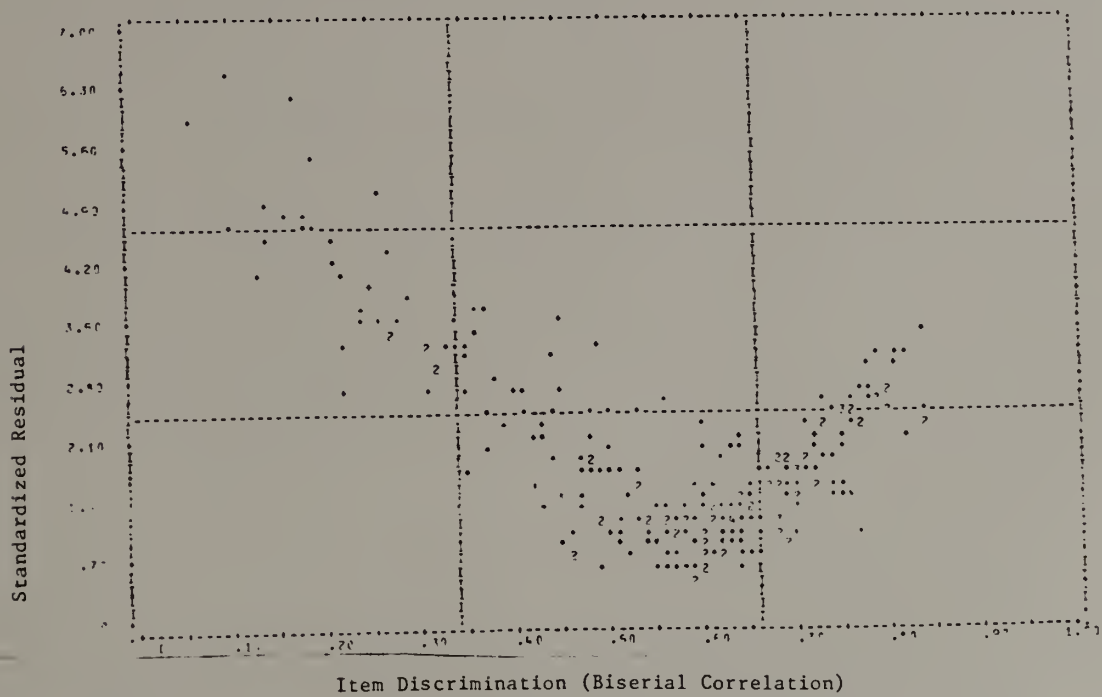


Figure 4.3.2. Plot of one-parameter model standardized residuals versus item discrimination.

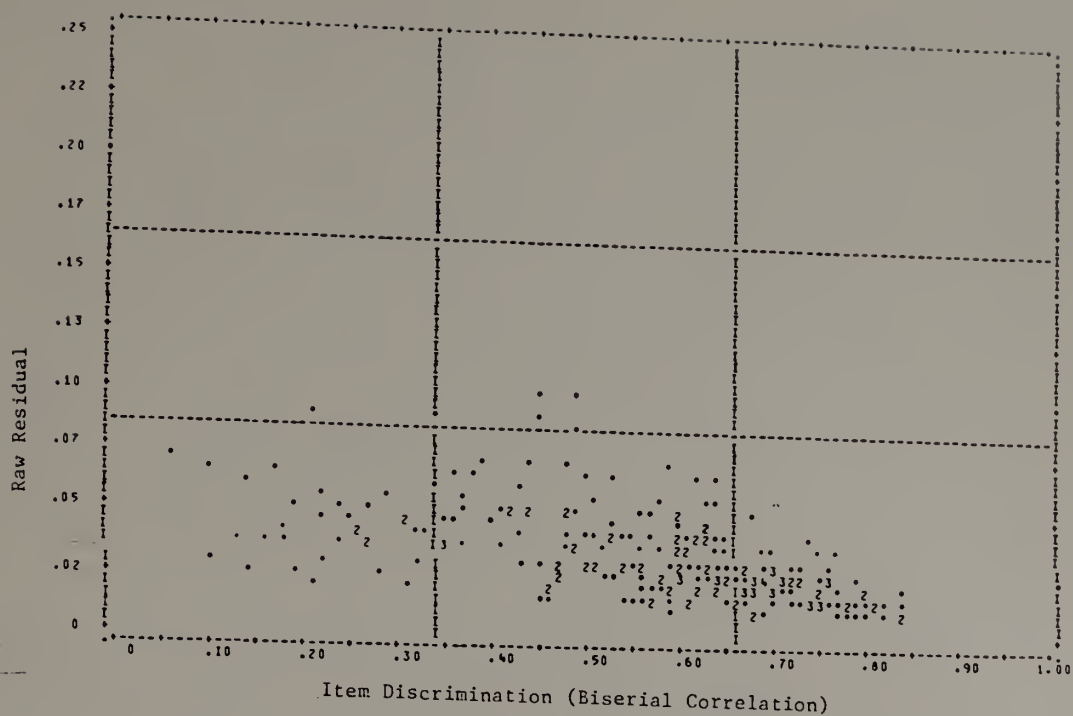


Figure 4.3.3. Plot of two-parameter raw residuals versus item discrimination.

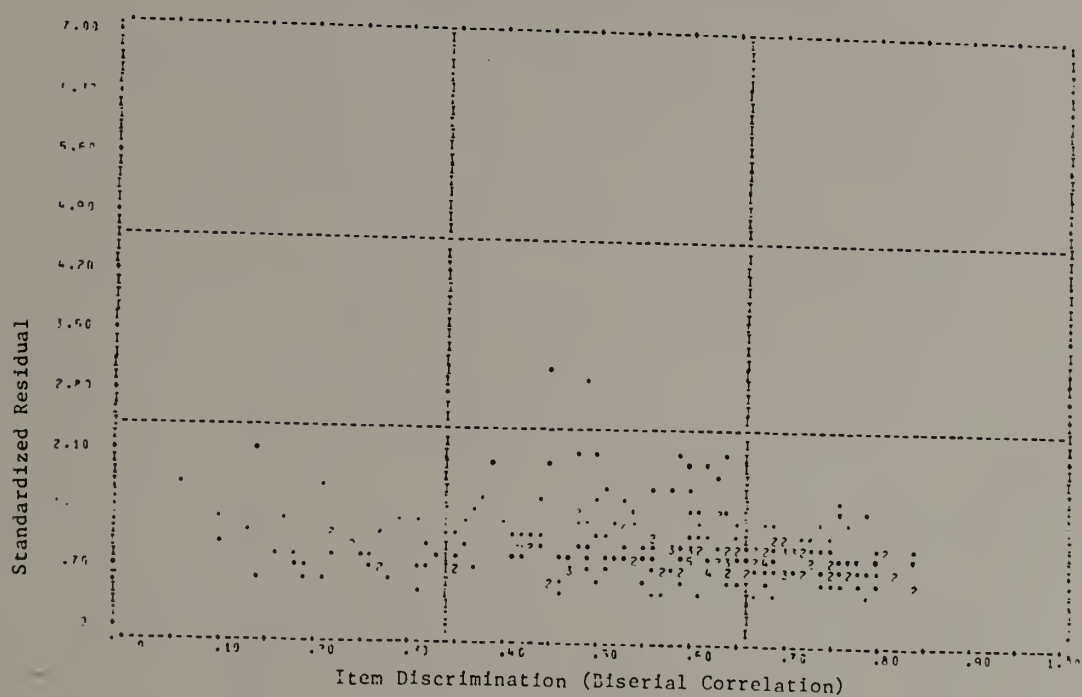


Figure 4.3.4. Plot of two-parameter standardized residuals versus item discrimination.

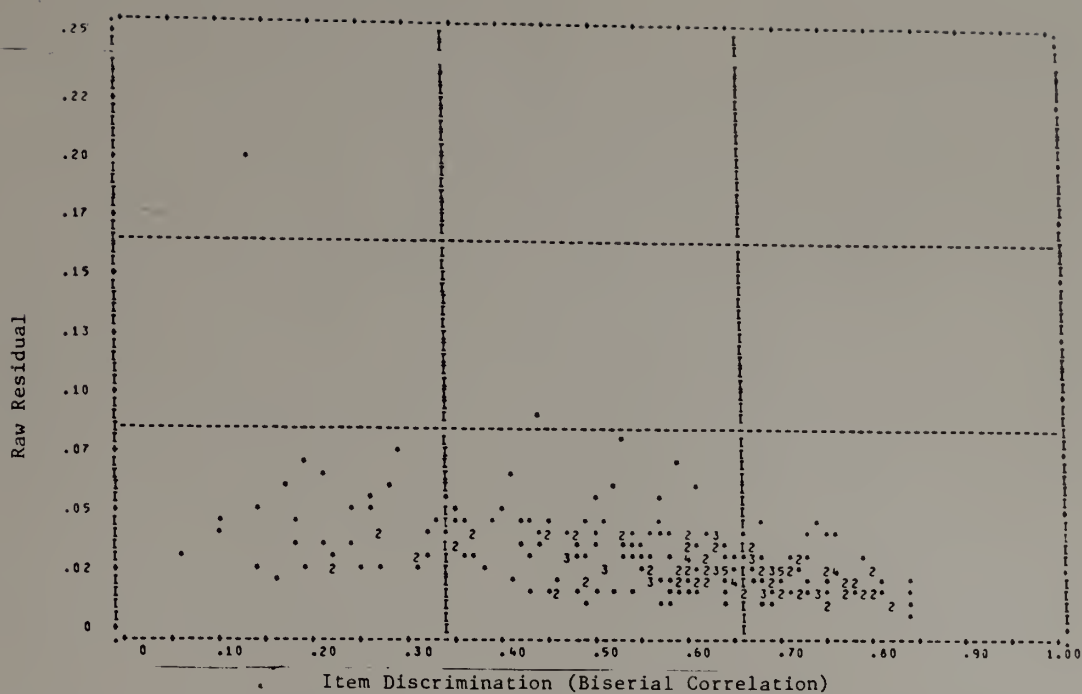


Figure 4.3.5. Plot of three-parameter raw residuals versus item discrimination.

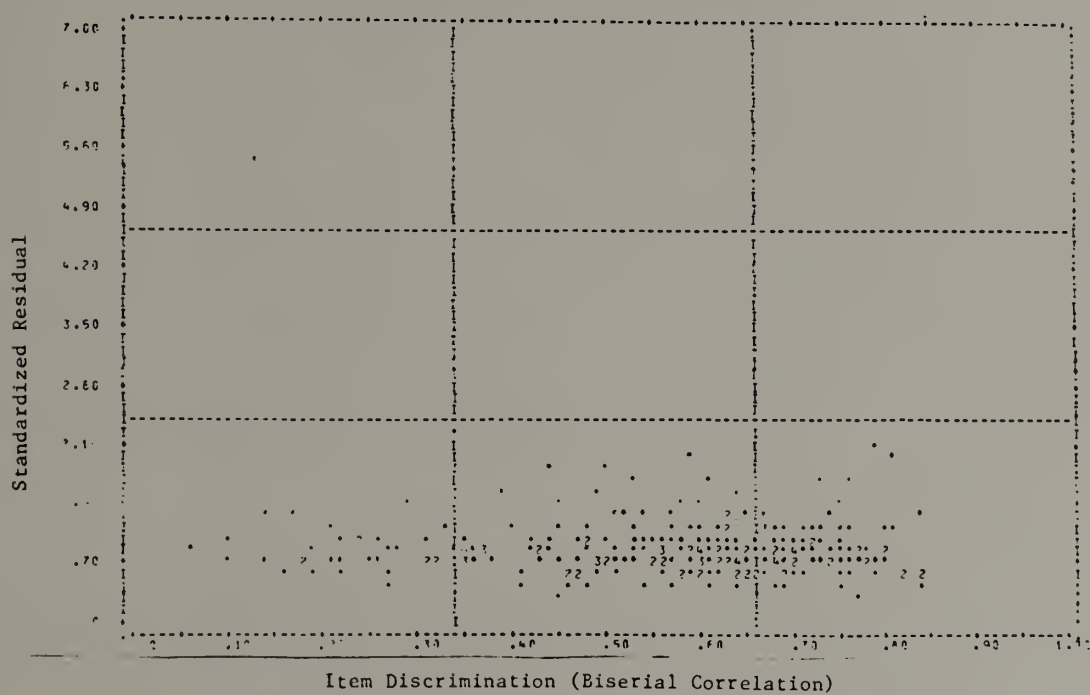


Figure 4.3.6. Plot of three-parameter standardized residuals versus item discrimination.

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 moderately high level of agreement in the identification of misfitting items. Because of the small number of misfitting items by the two-parameter 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 4.3.2. The average raw and standardized residual statistics provide information about the size and direction of the misfit between the observed and expected performance, 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.

Three of the four simple statistics in Table 4.3.2 present a similar picture of fit for the three item response models. According to these statistics both the two-parameter and three-parameter models provided a very good accounting of the actual results. The one-parameter model did not.



Table 4.3.2  
 Average and Absolute Average Raw and Standardized Residuals at Twelve Ability Levels with the One-, Two-, and Three-Parameter Logistic Models for Booklet No. 1, 9 Year OIDs, 65 Items, 1977-78

Type of Residual	Logistic Model	Sample Size	Ability Level												Total (unweighted)
			-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75	
Raw	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	
Raw	1		.002	.001	-.001	.001	.002	.002	.002	-.006	-.009	-.013	-.003	-.005	
	2		.004	.005	-.017	.009	-.003	-.003	-.004	-.006	-.001	.003	.005	.031	
	3		.004	.010	.010	.003	.001	.001	.002	-.002	-.005	-.012	-.005	.006	
Standardized	1		.006	.088	.074	.073	.045	.030	.027	.043	.057	.076	.071	.084	
	2		.052	.048	.042	.021	.017	.018	.013	.018	.017	.033	.038	.075	
	3		.049	.040	.034	.019	.020	.015	.010	.013	.015	.025	.043	.073	
Standardized	1		.77	.99	.89	.79	.37	.20	.14	-.28	-.26	-.39	-.11	-.10	
	2		.09	.31	.76	.35	.09	-.22	-.30	-.37	-.18	-.06	-.02	-.22	
	3		.00	.24	.27	.12	.16	.04	.08	-.18	-.48	-.36	-.32	-.16	
Standardized	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		.82	1.28	1.58	1.00	.90	1.15	.83	1.03	.93	1.12	.97	1.07	
	3		.81	.90	1.02	.74	1.00	.94	.62	.87	.99	.85	.91	.88	

The fourth statistic, average raw residual, described model-data fit rather differently. At many points on the ability continuum, the one-parameter model fit the data better than the more general models. Discrepancies between the two impressions of model-data fit could probably be attributed to estimation problems in the more general models. Hence, a different picture of model-data fit emerges.

Finally, Table 4.3.3 provides a comparison of weighted and unweighted summary test fit statistics for the four NAEP math booklets. For three of the four booklets, the weighted and unweighted statistics gave similar impressions of fit. The three-parameter model provided the best overall fit, and the one-parameter model the worst. For Booklet 2, 9 year olds, the results were different. The two-parameter model fit the test data better than the three-parameter model. Hence in this case, the impression of fit was influenced by the decision to use weighted samples in the calculation of the statistic.

#### 4.4 Descriptive Results from Analyses of NAEP Test Booklets

Several preliminary analyses were conducted on each of the NAEP test booklets and data sets. These descriptive data were collected for future residual investigations. First, Tables 4.4.1 and 4.4.2 provide information on the distribution of items across six content categories for each of the test booklets. In the area of content

Table 4.3.3

Comparison of Weighted and Unweighted Summary Test Fit Statistics for Four 1977-78 NAEP Mathematics Booklets

NAEP Booklet	Logistic Model	Average Standardized Residual	Weighted Average Standardized Residual	Average Absolute Standardized Residual	Average Weighted Absolute Standardized Residual
Booklet 1 (9 yr. olds)	1	.189	.111	2.084	2.236
	2	.057	-.082	1.057	1.012
	3	-.046	-.054	.894	.883
Booklet 2 (9 yr. olds)	1	.218	.122	2.023	2.156
	2	.117	-.056	1.137	.979
	3	.039	.080	.919	.991
Booklet 1 (13 yr. olds)	1	.277	.233	1.910	1.977
	2	.167	-.027	.984	.932
	3	-.002	-.020	.905	.858
Booklet 2 (13 yr. olds)	1	.230	.104	1.795	1.899
	2	.128	-.038	1.055	1.029
	3	.034	.031	.848	.899

Table 4.4.1

Content Classification Summary of NAEP Math Booklet Nos. 1 and 2 Test Items for 9 Year Olds, 1977-78 Assessment

<u>Booklet 1</u>		<u>Booklet 2</u>	
<u>Story Problems</u>		<u>Story Problems</u>	
Money	1	Money	3
General	5	General	2
Logic, Probability, Permutation and Combination	4	Logic, Probability, Permutation and Combination	7
Total	<u>10</u>	Total	<u>12</u>
<u>Geometry</u>		<u>Geometry</u>	
Story	0	Story	0
Definition/Operations	9	Definition/Operations	9
Figure Interpretations, Manipulation	5	Figure Interpretations, Manipulation	1
Total	<u>14</u>	Total	<u>10</u>
<u>Definition</u>		<u>Definition</u>	
Total	<u>1</u>	Total	<u>16</u>
<u>Calculation</u>		<u>Calculation</u>	
General	15	General	25
Algebra	<u>8</u>	Algebra	<u>1</u>
Total	<u>23</u>	Total	<u>26</u>
<u>Measurement</u>		<u>Measurement</u>	
English	3	English	1
Metric	<u>3</u>	Metric	<u>4</u>
Total	<u>6</u>	Total	<u>5</u>
<u>Graphs and Figures</u>		<u>Graphs and Figures</u>	
Total	<u>5</u>	Total	<u>6</u>

Table 4.4.2

Content Classification Summary of NAEP Math Booklet Nos. 1 and 2 Test Items for 13 Year Olds, 1977-78 Assessment

<u>Booklet 1</u>		<u>Booklet 2</u>	
<u>Story Problems</u>		<u>Story Problems</u>	
Money	3	Money	2
General	6	General	9
Logic, Probability, Permutation and Combination	5	Logic, Probability, Permutation and Combination	4
Total	<u>14</u>	Total	<u>15</u>
<u>Geometry</u>		<u>Geometry</u>	
Story	1	Story	1
Definition/Operations	9	Definition/Operations	7
Figure Interpretation, Manipulation	3	Figure Interpretation, Manipulation	2
Total	<u>13</u>	Total	<u>10</u>
<u>Definition</u>		<u>Definition</u>	
Total	<u>9</u>	Total	<u>7</u>
<u>Calculation</u>		<u>Calculation</u>	
General	14	General	17
Algebra	1	Algebra	5
Total	<u>15</u>	Total	<u>22</u>
<u>Measurement</u>		<u>Measurement</u>	
English	3	English	1
Metric	2	Metric	0
Total	<u>5</u>	Total	<u>1</u>
<u>Graphs and Figures</u>		<u>Graphs and Figures</u>	
Total	<u>1</u>	Total	<u>7</u>

type, the largest proportion of items was math calculation. Story problems and geometry items also appeared to be frequently occurring types of test items.

Second, Tables 4.4.3, 4.4.4, 4.4.5 and 4.4.6 provide information on item formats and content categories of the test items in the NAEP math booklets. The data in these tables reveal that the exercises were of two types: multiple-choice and open-ended. Surprisingly, the multiple-choice items did not have a consistent number of answer choices. Instead, the number of answer options varied from four to ten choices. But what was constant across all the multiple-choice items was the inclusion of "I don't know" as an answer alternative.

Finally, Table 4.4.7 and 4.4.8 contain the one-parameter, two-parameter, and three-parameter logistic item parameter estimates for all the items based on examinee response data. Parameter estimates for items in the four NAEP math test booklets were obtained with the aid of LOGIST (Wood, Wingersky, & Lord, 1976). It is important to note that these tables reveal items in a particular test booklet varied considerably in difficulty and discrimination levels. Therefore, it was expected that the more general logistic models would fit the data better.

#### 4.5 Item, Ability and Overall Fit

Analysis of the fit of the one-parameter, two-parameter and three-parameter logistic models to the NAEP data sets was made using

Table 4.4.3

Format and Content Classification of NAEP Math Booklet No. 1  
Test Items for 9 Year Olds, 1977-78 Assessment

Item No.	Answer Format <sup>1</sup>	Category
1/102A	MC	Definition
2/102B	MC (6 options)	Definition
3/103A	MC	Story problem - money
4/104A	MC (6 options)	Geometry - definition
5/104B	MC (6 options)	Geometry - definition
6/105A	MC	Geometry - figure manipulation, interpretation
7/106A	OE	Geometry - operations
8/106B	OE	Geometry - operations
9/106C	MC	Geometry - operations
10/107A	MC (6 options)	Measurement - English
11/108A	OE	Calculation
12/108B	OE	Calculation
13/108C	OE	Calculation
14/108D	OE	Calculation
15/108E	OE	Calculation
16/108F	OE	Calculation
17/109A	MC	Story problem - logic
18/110A	OE	Story problem - general
19/111A	MC	Geometry - definition
20/112A	OE	Calculation
21/112B	OE	Calculation
22/113A	MC	Measurement - English
23/114A	MC (6 options)	Story problem - general
24/115A	OE	Calculation - algebra
25/115B	OE	Calculation - algebra
26/115C	OE	Calculation - algebra
27/115D	OE	Calculation - algebra
28/115E	OE	Calculation - algebra
29/115F	OE	Calculation - algebra
30/115G	OE	Calculation - algebra

<sup>1</sup>MC Items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 4.4.3 (continued)

Item No.	Answer Format	Category
31/116A	MC (6 options)	Graphs and Figures
32/117A	MC	Definition
33/117B	MC	Definition
34/118A	MC (4 options)	Measurement - metric
35/119A	MC (6 options)	Graphs and Figures
36/120A	OE	Calculation
37/120B	OE	Calculation
38/121A	MC (10 options)	Definition
39/122A	MC (6 options)	Story problem - general
40/123A	MC	Calculation
41/124A	OE	Story problem - general
42/125A	OE	Calculation
43/125B	OE	Calculation
44/125C	OE	Calculation
45/126A	OE	Measurement - metric
46/127A	MC (4 options)	Calculation - algebra
47/128A	MC (4 options)	Measurement - metric
48/129A	MC	Graphs and Figures
49/129B	MC	Graphs and Figures
50/130A	MC (4 options)	Story problem - logic
51/130B	MC (4 options)	Story problem - logic
52/131A	MC (7 options)	Geometry - figure manipulation, interpretation
53/131B	MC (7 options)	Geometry - figure manipulation, interpretation
54/131C	MC (7 options)	Geometry - figure manipulation, interpretation
55/132A	OE	Graphs and Figures
56/133A	OE	Story problem - general
57/134A	MC (6 options)	Geometry - definition
58/134B	MC (6 options)	Geometry - definition
59/134C	MC (6 options)	Geometry - definition
60/135A	OE	Story problem - probability



Table 4.4.3 (continued)

Item No.	Answer Format	Category
61/136A	OE	Measurement - English
62/137A	OE	Definition
63/138A	OE	Calculation
64/139A	MC	Geometry - figure manipulation, interpretation
65/140A	MC	Definition

Table 4.4.4

Format and Content Classification of NAEP Math Booklet No. 2  
Test Items for 9 Year Olds, 1977-78 Assessment

Item No.	Answer Format <sup>1</sup>	Category
1/202A	MC	Definition
2/202B	MC	Definition
3/203A	OE	Calculation
4/203B	OE	Calculation
5/203C	OE	Calculation
6/203D	OE	Calculation
7/203E	OE	Calculation
8/203F	OE	Calculation
9/204A	OE	Calculation
10/204B	OE	Calculation
11/204C	OE	Calculation
12/204D	OE	Calculation
13/205A	MC (6 options)	Geometry - operations
14/206A	MC (6 options)	Story problem - money
15/207A	MC	Graphs and Figures
16/207B	MC	Graphs and Figures
17/208A	OE	Calculation
18/208B	OE	Calculation
19/208C	OE	Calculation
20/209A	MC	Story problem - combinations
21/210A	MC (8 options)	Graphs and Figures
22/210B	MC (6 options)	Graphs and Figures
23/210C	MC (9 options)	Graphs and Figures
24/211A	MC (4 options)	Definition
25/211B	MC (4 options)	Definition
26/211C	MC (4 options)	Definition
27/211D	MC (4 options)	Definition
28/211E	MC (4 options)	Definition
29/212A	MC (4 options)	Measurement - metric
30/212B	MC	Measurement - metric

<sup>1</sup>MC Items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 4.4.4 (continued)

Item No.	Answer Format	Category
31/213A	OE	Calculation - algebra
32/214A	OE	Story problem - logic
33/215A	OE	Definition
34/215B	OE	Definition
35/215C	OE	Definition
36/216A	MC (6 options)	Geometry - definition
37/216B	MC (6 options)	Geometry - definition
38/216C	MC (6 options)	Geometry - definition
39/217A	MC	Story problem - money
40/218A	OE	Calculation
41/218B	OE	Calculation
42/218C	OE	Calculation
43/218D	OE	Calculation
44/218E	OE	Calculation
45/218F	OE	Calculation
46/219A	MC	Geometry - operations
47/220A	OE	Calculation
48/220B	OE	Calculation
49/220C	OE	Calculation
50/221A	MC	Geometry - definition
51/222A	MC	Measurement - metric
52/223A	MC	Definition
53/224A	MC	Definition
54/224B	MC	Definition
55/225A	MC	Story problem - logic
56/225B	MC	Story problem - logic
57/225C	MC	Story problem - logic
58/226A	MC	Story problem - general
59/226B	MC	Story problem - general
60/227A	MC	Calculation

Table 4.4.4 (continued)

Item No.	Answer Format	Category
61/228A	MC	Geometry - definition
62/228B	MC	Geometry - definition
63/229A	MC	Definition
64/229B	MC	Definition
65/229C	MC	Definition
66/230A	OE	Calculation
67/231A	OE	Story problem - money
68/232A	OE	Geometry - operations
69/233A	MC	Story problem - logic
70/234A	OE	Story problem - probability
71/235A	OE	Geometry - figure manipulation, interpretation
72/236A	OE	Calculation
73/237A	OE	Measurement - English
74/238A	OE	Graphs and Figures
75/239A	MC	Measurement - metric

Table 4.4.5

Format and Content Classification of NAEP Math Booklet No. 1  
Test Items for 13 Year Olds, 1977-78 Assessment

Item No.	Answer Format <sup>1</sup>	Category
1/102A	OE	Story problem - money
2/103A	MC	Definitions
3/103B	MC	Definitions
4/104A	OE	Measurement - English
5/105A	MC	Calculation
6/106A	MC	Geometry - definition, operations
7/106B	MC	Geometry - definition, operations
8/106C	MC	Geometry - definition, operations
9/107A	MC	Story problem - logic
10/108A	MC	Measurement - metric
11/109A	OE	Calculation - subtraction
12/109B	OE	Calculation - subtraction
13/109C	OE	Calculation - subtraction
14/109D	OE	Calculation - subtraction
15/109E	OE	Calculation - subtraction
16/109F	OE	Calculation - subtraction
17/110A	MC (4 options)	Measurement - metric
18/111A	OE	Story problem - general
19/111B	OE	Calculation
20/112A	OE	Calculation
21/112B	OE	Calculation
22/113A	MC (10 options)	Definition
23/114A	MC	Definition
24/114B	MC	Definition
25/115A	OE	Story problem - money
26/116A	MC	Geometry - definitions, operations
27/116B	MC	Geometry - definitions, operations
28/117A	OE	Geometry - definitions
29/118A	OE	Measurement - English
30/119A	MC (7 options)	Story problems - general

<sup>1</sup> MC items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 4.4.5 (continued)

Item No.	Answer Format	Category
31/120A	MC	Geometry - figure manipulation, interpretation
32/120B	MC	Story problem - general
33/121A	MC (6 options)	Story problem - general
34/122A	MC	Geometry - definitions
35/122B	MC	Geometry - definitions
36/123A	MC	Story problem - money
37/124A	MC (6 options)	Geometry - story problem
38/125A	MC	Definitions
39/126A	MC	Definitions
40/127A	MC	Story problem - combinations
41/128A	MC	Definitions
42/129A	MC (6 options)	Geometry - definitions, operations
43/130A	MC	Geometry - figure manipulation
44/131A	OE	Calculation
45/131B	OE	Calculation
46/132A	MC	Story problem - general
47/133A	MC	Geometry - story problem
48/134A	MC (6 options)	Definitions
49/135A	OE	Calculations - algebra
50/136A	MC	Story problem - general
51/137A	MC (6 options)	Story problem - probability
52/137B	MC (6 options)	Story problem - probability
53/138A	MC (6 options)	Geometry - figure manipulation
54/139A	OE	Calculation
55/140A	OE	Graphs and figures
56/141A	MC	Story problem - logic
57/142A	OE	Measurement - English
58/143A	OE	Calculation

Table 4.4.6

Format and Content Classification of NAEP Math Booklet No. 2  
Test Items for 13 Year Olds, 1977-78 Assessment

Item No.	Answer Format <sup>1</sup>	Category
1/202A	OE	Calculation - algebra
2/203A	OE	Calculation
3/204A	OE	Calculation
4/205A	MC	Story problem - logic
5/206A	MC	Definitions
6/207A	OE	Graphs and Figures
7/208A	OE	Measurement - English
8/209A	OE	Story problem - general
9/210A	OE	Calculation
10/210B	OE	Calculation
11/210C	OE	Calculation
12/210D	OE	Calculation
13/211A	MC (6 options)	Geometry - definitions
14/212A	MC	Calculation - algebra
15/213A	MC (6 options)	Geometry - story problem
16/214A	OE	Calculation
17/214B	OE	Calculation
18/214C	OE	Calculation
19/214D	OE	Calculation
20/214E	OE	Calculation
21/214F	OE	Calculation
22/215A	MC (6 options)	Geometry - definitions
23/216A	OE	Calculation
24/216B	OE	Calculation
25/216C	OE	Calculation
26/217A	MC	Geometry - definition
27/217B	MC	Geometry - definition
28/218A	OE	Story problem - general
29/219A	MC	Story problem - money
30/220A	OE	Story problem - probability

<sup>1</sup> MC items have 5 answer choices (including "I don't know") unless otherwise noted.

Table 4.4.6 (continued)

Item No.	Answer Format	Category
31/221A	MC	Definition
32/222A	MC (4 options)	Definition
33/222B	MC (4 options)	Definition
34/223A	MC (6 options)	Story problem - general
35/224A	OE	Story problem - money
36/225A	MC (6 options)	Graphs and figures
37/225B	MC (7 options)	Graphs and figures
38/225C	MC (6 options)	Graphs and figures
39/226A	OE	Calculation - algebra
40/227A	MC	Story problem - general
41/228A	OE	Calculation - algebra
42/228B	OE	Calculation - algebra
43/229A	MC (4 options)	Story problem - general
44/230A	MC	Geometry - figure manipulation, interpretation
45/231A	MC (6 options)	Story problem - permutation and combination
46/232A	MC	Story problem - general
47/232B	MC	Story problem - general
48/233A	OE	Definition
49/233B	OE	Definition
50/233C	OE	Definition
51/234A	MC (6 options)	Geometry - definitions
52/234B	MC (6 options)	Geometry - definitions
53/235A	OE	Story problem - general
54/236A	MC	Geometry - figure manipulation, interpretation
55/237A	MC (6 options)	Geometry - definitions, operations
56/238A	OE	Story problem - general
57/239A	MC (6 options)	Story problem - probability
58/240A	MC (6 options)	Graphs and figures
59/240B	MC (6 options)	Graphs and figures
60/240C	MC (6 options)	Graphs and figures
61/241A	OE	Calculation - algebra
62/241B	OE	Calculation - algebra



Table 4.4.7

## NAEP Math Item Model Parameter Estimates for 9 Year Olds, 1977-78 Assessment

Test Item	Booklet No. 1			Booklet No. 2		
	1-p $\hat{b}$	2-p $\hat{a}$	3-p $\hat{c}$	1-p $\hat{b}$	2-p $\hat{a}$	3-p $\hat{c}$
1	-.22	-.22	1.15	-1.39	-2.97	.24
2	.17	.13	1.20	-1.40	-2.94	.25
3	-.22	-.20	1.20	-2.63	-2.23	.77
4	-2.55	-3.88	.37	-2.13	-1.62	.99
5	-2.33	-3.17	.40	-2.21	-1.81	.85
6	-.93	-1.81	.27	-1.40	-1.14	.92
7	2.18	2.94	1.56	1.99	-1.49	1.05
8	.82	.82	.70	-1.68	-1.30	1.00
9	.21	.24	.51	-.42	-.31	1.58
10	.53	.44	1.13	-.48	-.35	1.42
11	-2.32	-1.56	1.42	.01	.00	1.48
12	-1.81	-1.29	1.26	-.01	-.02	1.26
13	-2.17	-1.42	1.55	2.63	2.75	1.07
14	-1.13	-.82	1.24	.60	.58	.95
15	-1.62	-1.09	1.49	-1.13	-1.14	.66
16	-1.20	-.84	1.39	-.32	-.30	1.03
17	.19	.29	.32	-1.18	-1.05	.82
18	-1.64	-1.55	.65	-1.14	-.99	.83
19	-1.90	-2.06	.52	-.26	-.22	.99
20	-.60	-.55	.79	1.99	1.82	1.07
21	.48	.35	1.19	-.60	-.49	1.06
22	1.55	1.38	1.46	-.32	-.26	1.19
23	-.14	-.15	1.39	.52	.40	1.25
24	-1.69	-1.19	1.31	-3.50	-2.55	.96
25	.49	.38	1.04	-3.48	-2.77	.81
26	.06	.03	1.35	-3.62	-2.58	1.00
27	-.82	-.66	1.06	-3.03	-2.84	.65
28	.68	.58	.86	-2.67	-2.48	.64
29	-1.35	-1.09	.93	1.58	5.59	1.10
30	-.83	-.60	1.33	1.55	2.83	.56



Table 4.4.7 (continued)

Test Item	Booklet No. 1			Booklet No. 2		
	1-p b	2-p b	3-p a	1-p b	2-p b	3-p a
61	2.41	2.27	1.13	-2.18	-10.00	.09
62	-1.95	-1.81	.66	.30	.52	.35
63	.09	.07	.94	1.49	6.58	1.26
64	.70	.81	.71	2.01	85.56	1.79
65	1.12	3.10	.22	1.21	4.64	.19
66						
67				2.12	2.44	.75
68				1.14	1.29	.68
69				4.82	3.72	1.23
70				.06	.08	.46
				2.61	6.48	.38
71				2.88	3.84	.65
72				3.28	2.60	1.14
73				.74	1.01	.57
74				.49	.48	.80
75				.73	1.27	.39



Table 4.4.8 (continued)

Test Item	Booklet No. 1			Booklet No. 2		
	1-p b	2-p b	3-p a	1-p b	2-p b	3-p a
31	-1.18	-1.06	.84	.87	.75	1.10
32	-.34	-.30	1.46	-.75	-1.09	.42
33	.36	.41	.68	-2.93	-2.27	1.06
34	-3.25	-3.70	.47	-1.90	-1.78	.78
35	-.75	-1.03	.46	1.29	1.31	.72
36	1.44	76.12	2.00	-.39	-.37	.79
37	.63	.77	.69	-.75	-.60	1.30
38	-1.41	-.98	1.37	.68	.66	1.07
39	-.93	-1.57	.35	.01	.01	.84
40	-.71	-.58	1.05	.20	.20	1.07
41	1.12	.99	1.11	-2.01	-1.76	.88
42	-.89	.84	.79	-1.33	-1.13	.96
43	-1.39	-1.32	.73	-1.32	-1.40	.65
44	-.82	-.81	.69	-1.04	-1.87	.32
45	.23	.21	.85	-.70	-.84	.56
46	.74	.65	1.28	-1.51	-1.36	.85
47	2.24	17.20	1.54	-1.45	-1.18	1.07
48	1.92	2.18	.89	1.07	.93	1.02
49	.04	.03	.97	1.63	1.27	1.20
50	-1.82	-1.64	.76	-.57	-.48	1.02
51	1.67	3.51	1.03	.78	3.88	.12
52	-.46	-1.35	.20	1.62	3.16	.58
53	-1.68	-2.16	.45	.48	.39	1.18
54	-1.10	-1.00	.78	-1.87	-2.11	.57
55	1.24	.97	1.13	.12	.13	.66
56	-1.05	-1.03	.71	.02	.02	1.44
57	.90	.66	1.24	1.95	15.11	2.00
58	-1.19	-1.09	.79	1.64	1.59	1.29
59				1.51	1.83	1.63
60				-1.16	-1.29	.60

Table 4.4.8 (continued)

Test Item	Booklet No. 1			Booklet No. 2		
	1-p b̂	2-p â	3-p â̂	1-p b̂	2-p â̂	3-p â̂
61						
62						
63						
64						
65						
66						
67						
68						
69						
70						
71						
72						
73						

the standardized residuals. The results of these investigations are summarized in Tables 4.5.1 and 4.5.2, and Figures 4.5.1 to 4.5.12.

Table 4.5.1 provides a complete summary of the distribution of the standardized residuals obtained with the one-parameter, two-parameter, and three-parameter logistic models for the four math booklets. In all cases the standardized residuals were considerably larger for the one-parameter model. The three-parameter model provided the best overall fit with the distribution of the standardized residuals being approximately normal. The two-parameter model fits, although not as well, were rather similar to the three-parameter results. More importantly, the biggest improvement in overall fit occurred when the two-parameter model replaced the one-parameter model. Possibly failure to consider item discriminating power resulted in the one-parameter model providing substantially poorer overall fits to the various data sets than the two-parameter or the three-parameter models.

Table 4.5.2 reports the average and average-absolute standardized residuals at 12 ability levels with the one-parameter, two-parameter, and three-parameter models for the same four math booklets. Again the results in this table reveal substantial improvement in fit occurring when the two-parameter model replaced the one-parameter model. Also, it is clear that the three-parameter model was especially effective at low levels of ability. Failure to consider examinee guessing behavior could account for the differences in fit at these low ability levels.

Table 4.5.1

Analysis of Standardized Residuals with the One-, Two-, and Three-Parameter Logistic Models for Four NAEP Mathematics Booklets

NAEP Booklet	Logistic Model	Percent of Residuals			
		0 to 1	1 to 2	2 to 3	over 3
Booklet 1 (9 year olds)	1	35.9	21.5	17.3	25.3
	2	60.5	26.5	9.1	3.9
	3	66.7	24.4	6.7	2.3
Booklet 2 (9 year olds)	1	37.1	25.3	13.8	23.8
	2	61.2	28.3	8.7	1.8
	3	67.4	24.7	5.7	2.2
Booklet 1 (13 year olds)	1	40.7	22.1	16.5	20.7
	2	60.3	28.5	8.3	2.9
	3	65.4	25.1	7.8	1.7
Booklet 2 (13 year olds)	1	42.6	24.2	16.3	16.9
	2	60.9	28.5	7.1	3.5
	3	67.2	26.1	5.7	1.1



Table 4.5.2

Analysis of Standardized Residuals at Twelve Ability Levels with the One-, Two-, and Three-Parameter Logistic Models for Four 1977-78 NAEP Mathematics Booklets

NAEP Booklet	Test Length	Statistic	Logistic Model	Sample Size	Ability Level												Total (unweighted)
					-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	2.75	
Booklet 1 (9 year olds)	65		1	2495	27	43	111	220	331	485	446	395	276	122	21	8	
			2	2495	12	49	110	231	379	462	466	349	273	99	39	15	
			3	2495	29	50	108	212	333	454	470	403	273	100	21	9	
	Average Residual		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
	Average Absolute Residual		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
Booklet 2 (9 year olds)	75		1	2463	10	46	116	234	334	437	474	397	272	87	39	7	
			2	2463	9	46	106	230	385	436	472	376	221	104	45	17	
			3	2463	23	64	89	218	346	417	497	403	230	107	34	6	
	Average Residual		1		.60	.74	.58	.71	.28	-.01	.02	-.14	-.02	.08	-.05	.02	.23
			2		.28	.42	.20	.40	.28	-.22	-.43	-.30	-.03	-.22	.35	.24	.12
			3		-.16	-.14	-.02	.34	.50	.19	-.03	-.18	-.23	-.05	-.16	.01	.01
	Average Absolute Residual		1		1.55	2.42	3.02	3.10	2.28	1.49	1.59	2.36	2.75	1.71	1.31	.75	2.03
			2		1.00	1.21	.96	.99	.92	1.01	1.08	.90	.95	.90	.95	.77	.97
			3		.84	.95	.83	1.05	1.02	.94	1.04	.89	1.01	.87	.90	.53	.90

Table 4.5.2 (continued)

NAEP Booklet	Test Length	Statistic	Logistic Model	Sample Size	-2.75	-2.25	-1.75	-1.25	-.75	.25	.75	1.25	1.75	2.25	2.75	Total (unweighted)	
Booklet 1 (13 year olds)	58		1	2422	14	54	91	224	325	503	467	339	245	102	44	3	
			2	2422	11	48	101	224	364	450	477	365	224	106	39	16	
			3	2422	24	50	114	194	318	440	509	368	248	90	32	11	
		Average Residual		1	.67	.88	.66	.33	.03	.25	.40	.17	-.12	.07	.11	-.09	.28
			2	.34	.67	.55	.28	.07	.00	-.47	-.13	-.14	.38	.24	.20	.17	.17
			3	-.02	.08	.06	-.07	.07	.27	-.20	-.22	-.59	-.03	-.12	-.43	-.10	-.10
Booklet 2 (13 year olds)	62		1	2433	1.76	2.59	2.76	2.64	2.20	1.63	1.68	2.08	2.06	1.62	1.30	.67	1.92
			2	2433	.97	1.49	1.28	.85	.86	1.01	.86	.87	.98	.93	.85	.86	.98
			3	2433	1.27	1.02	.97	.84	.76	.92	.79	.84	1.07	.84	.86	.99	.93
		Average Residual		1	.90	.92	1.06	.67	.14	-.05	-.01	-.05	-.27	-.13	-.03	-.01	.26
			2	.43	.37	.70	.30	-.04	-.12	-.20	-.31	-.03	.12	.10	.22	.13	.13
			3	-.03	.17	.13	.22	.03	.10	-.08	-.24	-.26	-.27	-.12	-.13	-.13	.13
	Average Absolute Residual		1	2.08	1.90	2.98	2.69	1.64	1.79	1.54	1.95	2.10	1.64	1.00	.73	1.84	
		2	1.57	1.01	1.38	.92	.90	1.24	.95	1.03	.87	.97	.94	.84	.84	1.06	
		3	.76	.79	.87	.82	.81	1.05	.94	.85	.85	.94	.69	.70	.70	.84	

A sample set of standardized residual plots for several Math Booklet No. 1, 13 year olds test items of varying difficulty and discrimination obtained with the one-parameter, two-parameter, and three-parameter models are shown in Figures 4.5.1 to 4.5.12. Item patterns like those in Figures 4.5.1 and 4.5.2 were obtained for items with relatively low biserial correlations. Item patterns like those in Figures 4.5.3, 4.5.4, 4.5.5, and 4.5.6 were obtained for items with relatively high biserial correlations. Two features of the plots in these figures are the cyclic patterns and the large size of the one-parameter standardized residuals. For the two-parameter and three-parameter models, the standardized residuals were substantially smaller. Also, the cyclic pattern so clearly evident for the one-parameter model was gone.

Item patterns like those in Figures 4.5.7 to 4.5.12 were obtained with items with biserial correlations in the range of .50 to .66. In these plots, the size of the one-parameter standardized residuals are much smaller and similar to the two-parameter and three-parameter standardized residuals. Hence, overall the amount of model-data misfit for the one-parameter model is small for items with middle discrimination and large for items with relatively high or low biserial correlations.

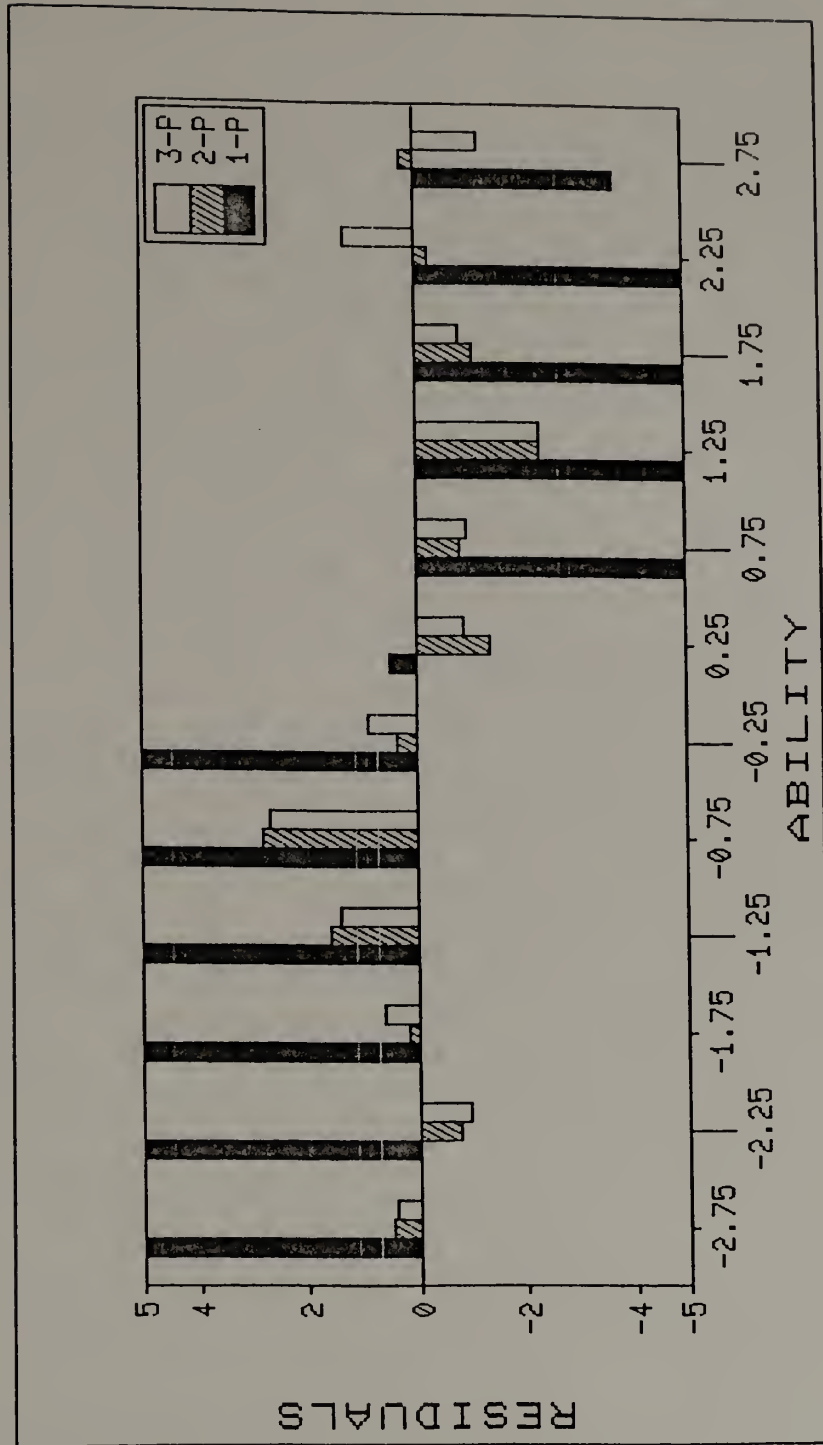


Figure 4.5.1. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 36 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

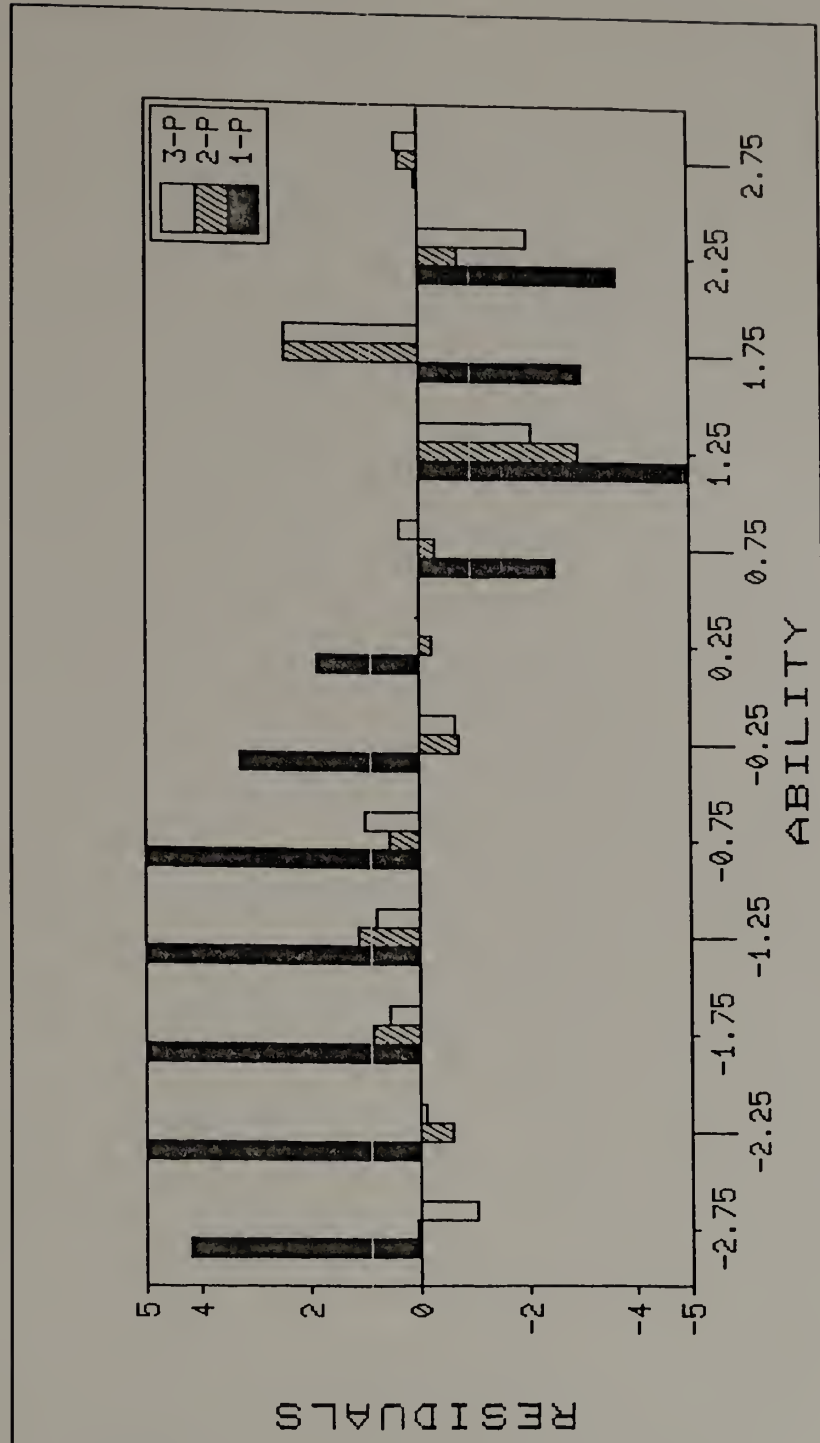


Figure 4.5.2. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 47 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

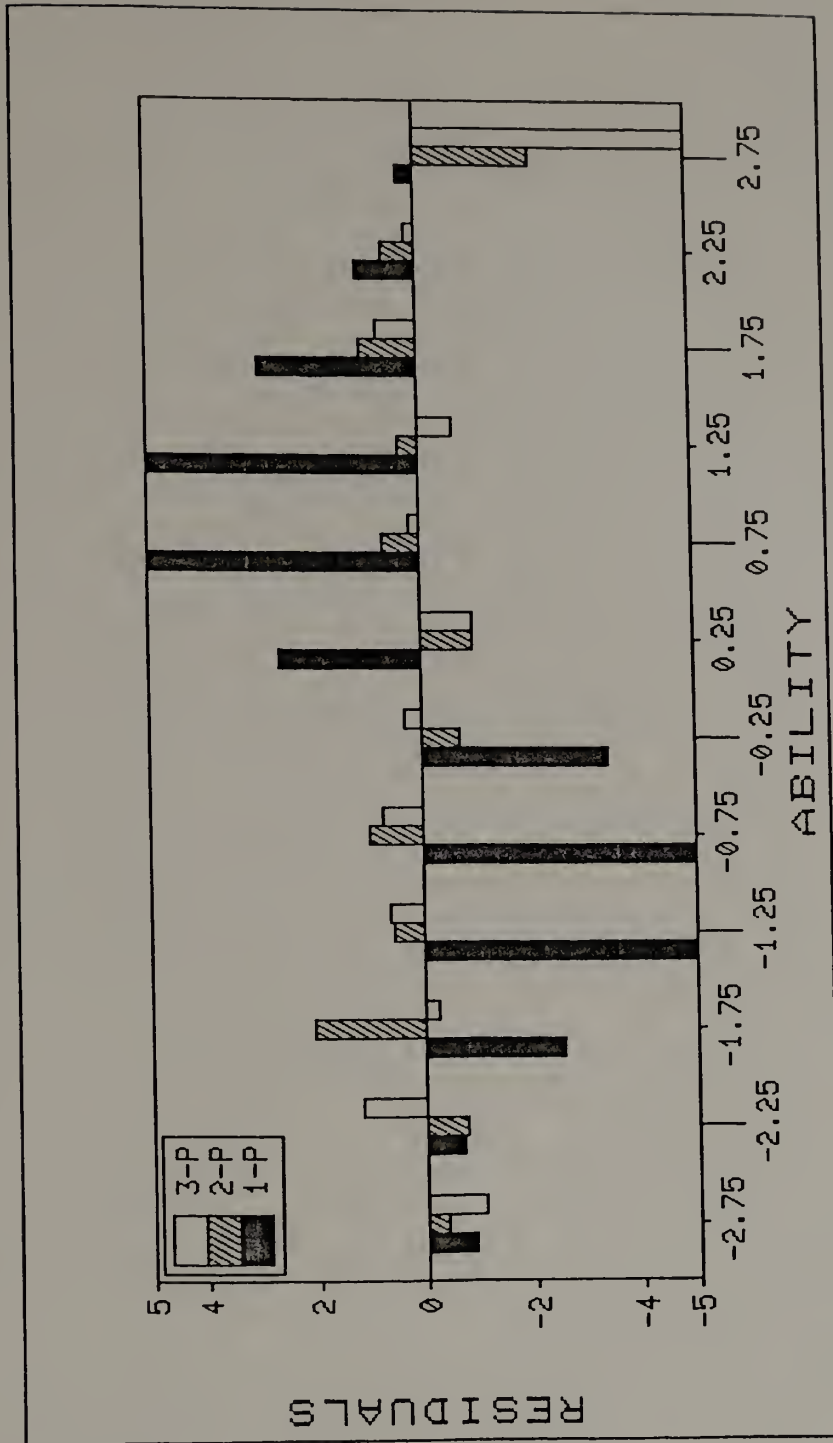


Figure 4.5.3. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 4 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

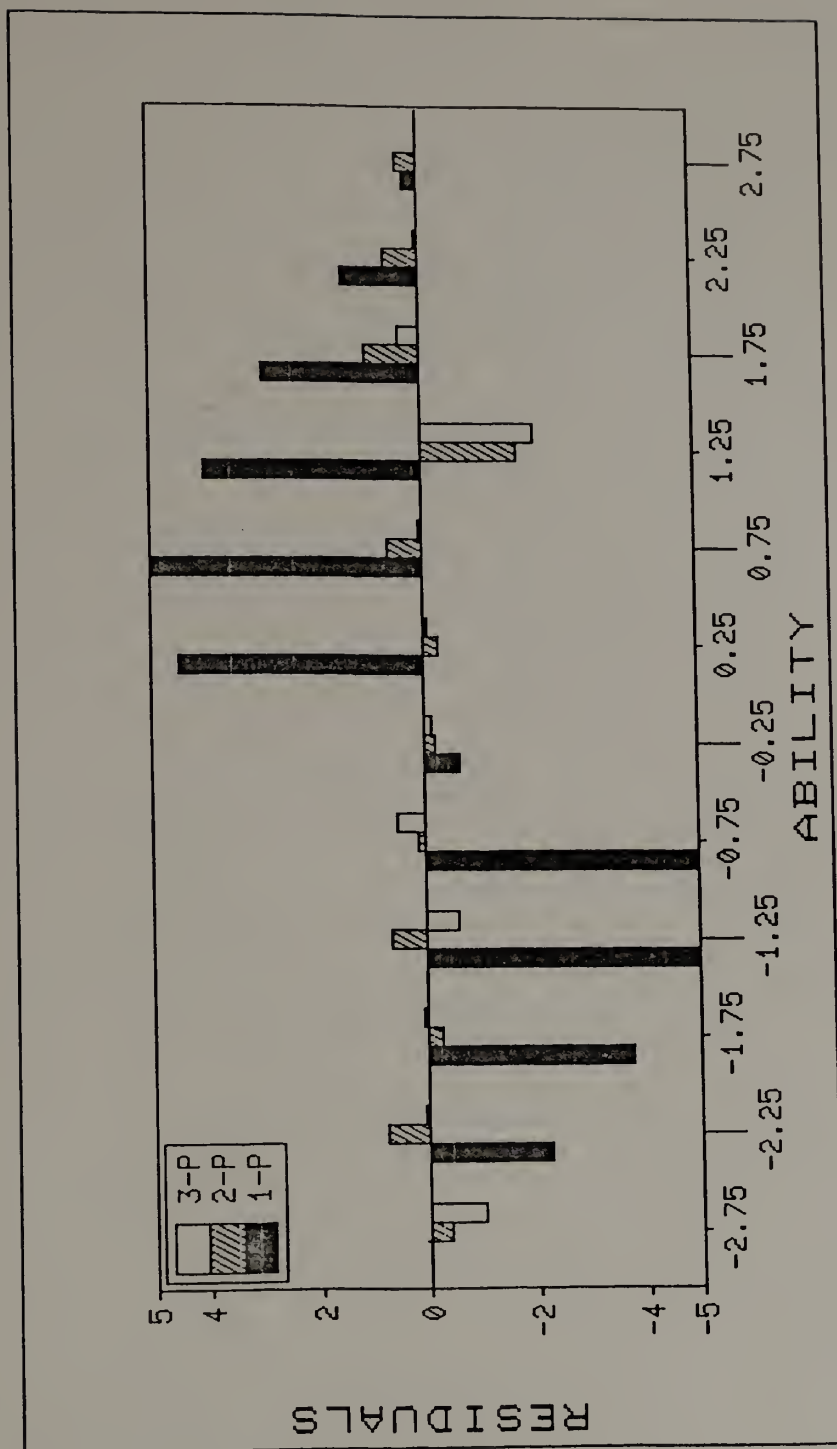


Figure 4.5.4. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 20 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

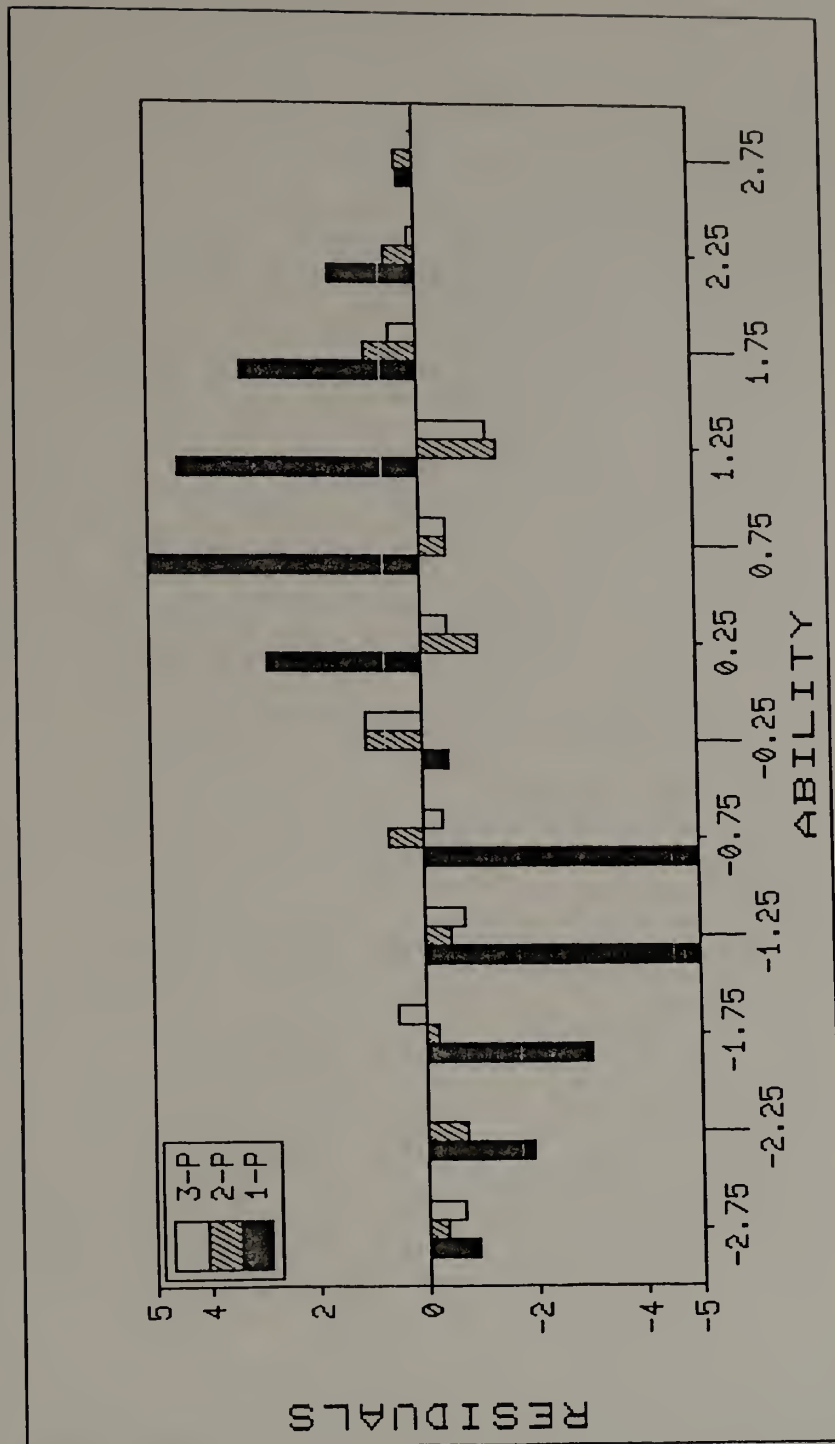


Figure 4.5.5. Standardized residual plot obtained from the one-, two-, and three-parameter logistic models for test item 21 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).



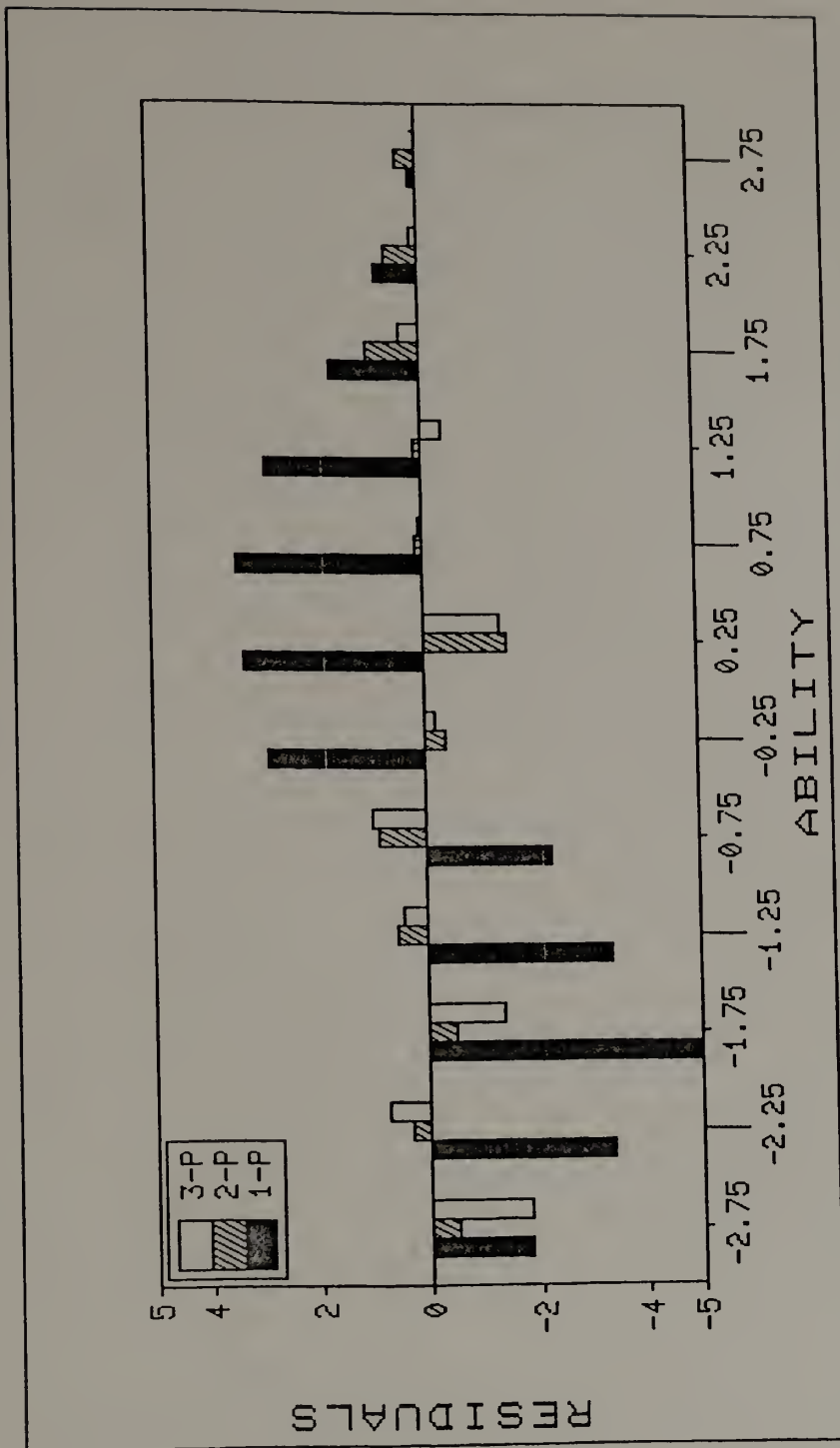


Figure 4.5.6. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 38 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

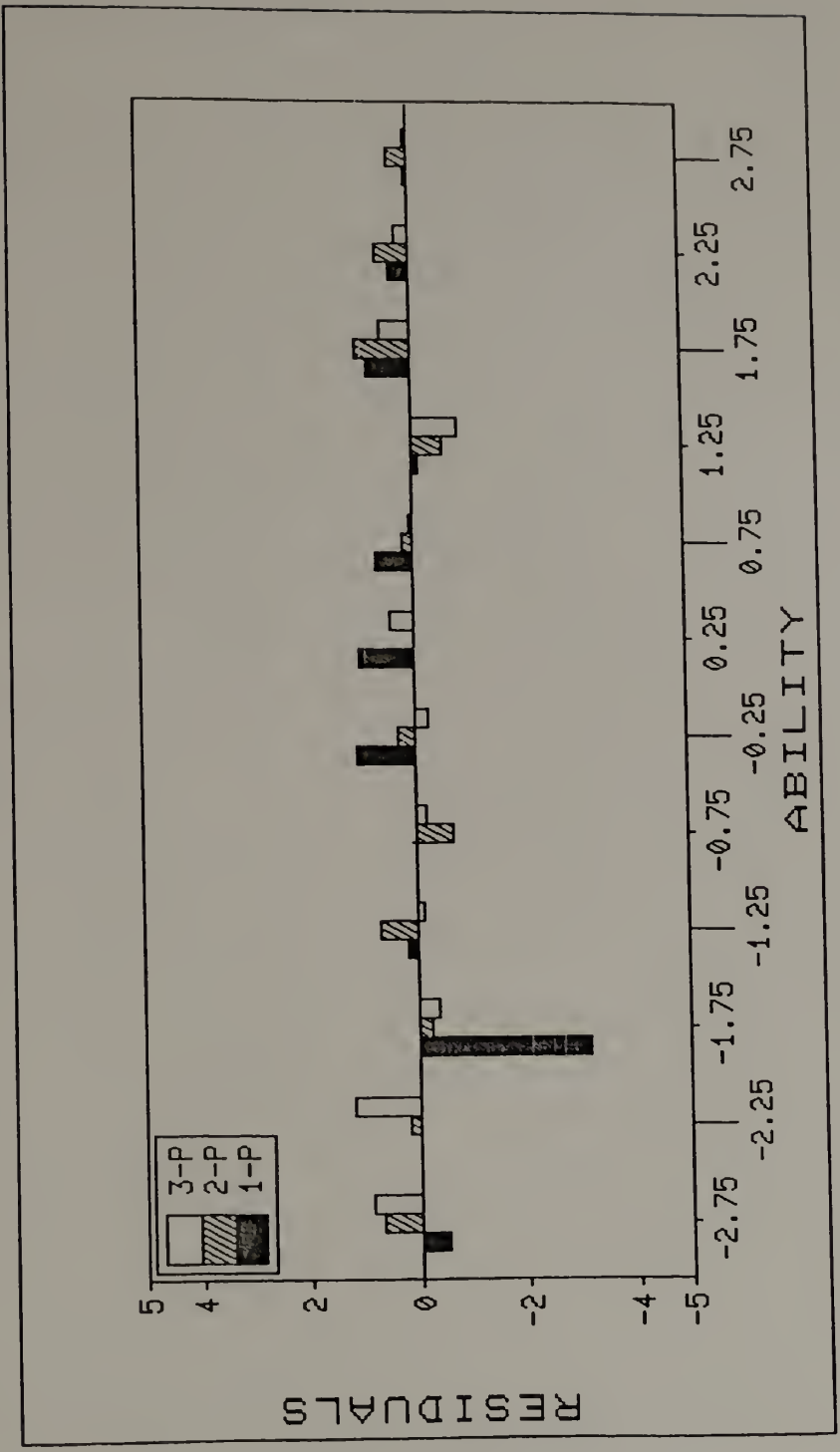


Figure 4.5.7. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 2 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

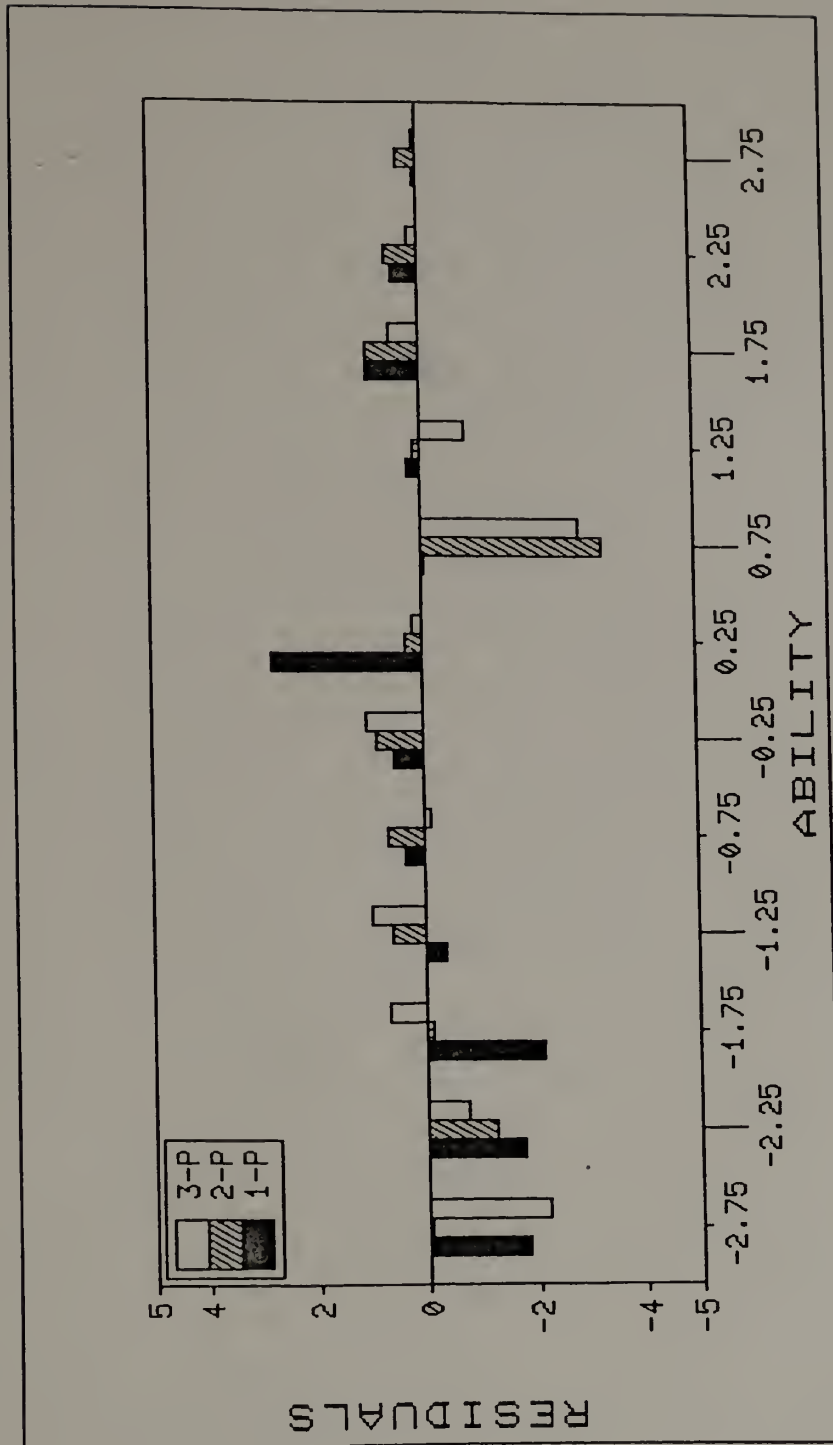


Figure 4.5.8. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 15 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

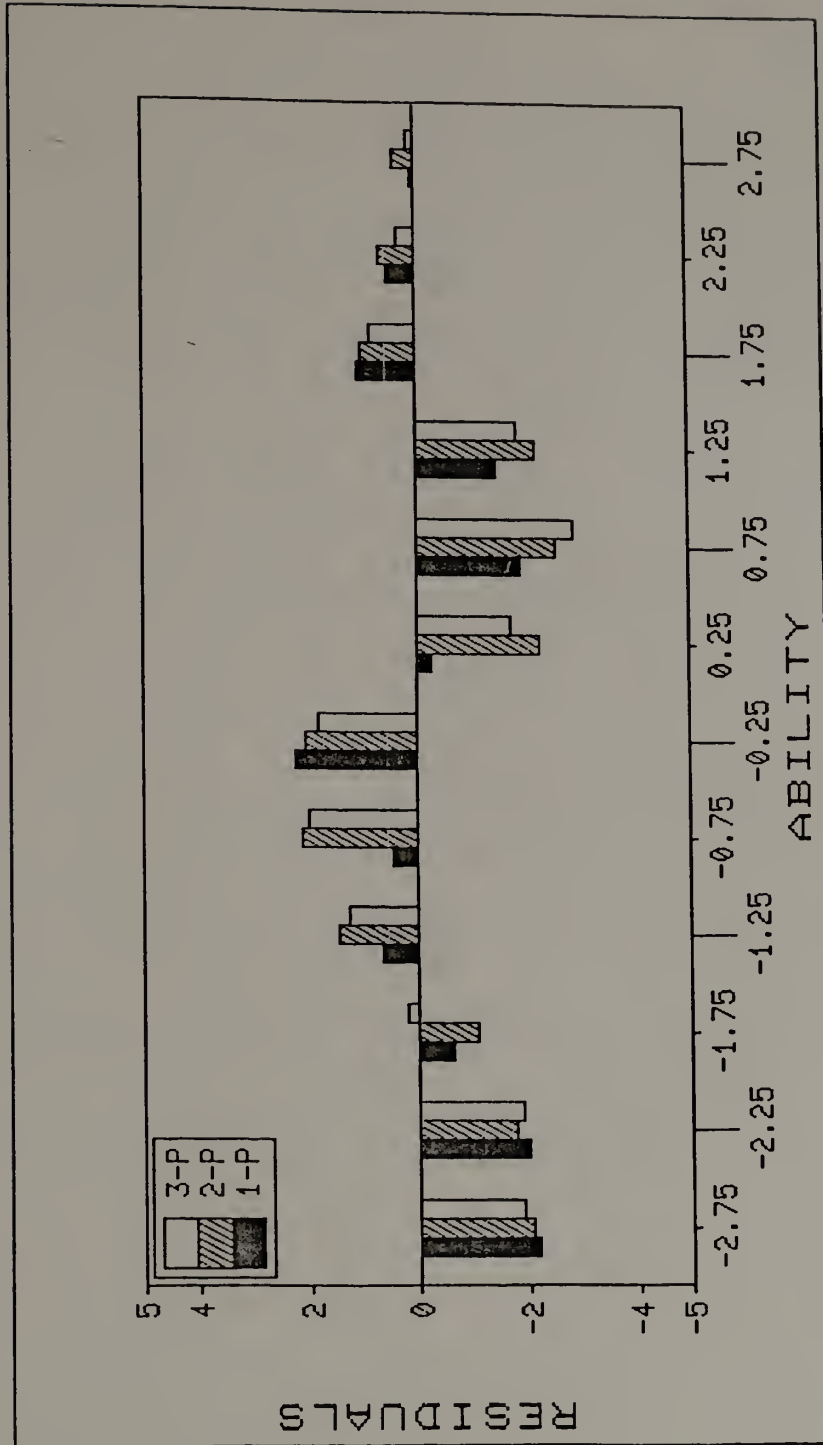


Figure 4.5.9. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 16 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

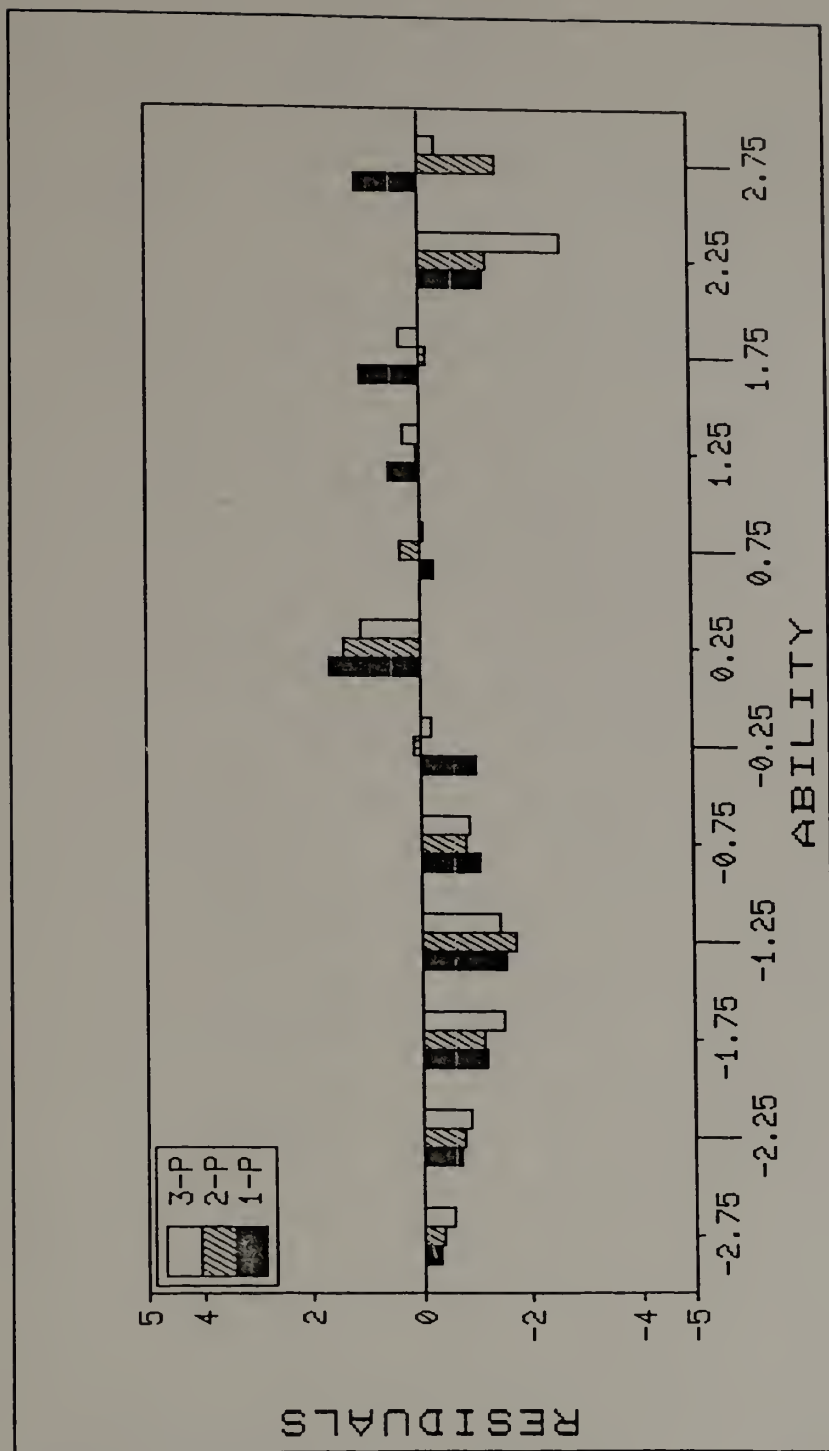


Figure 4.5.10. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 18 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

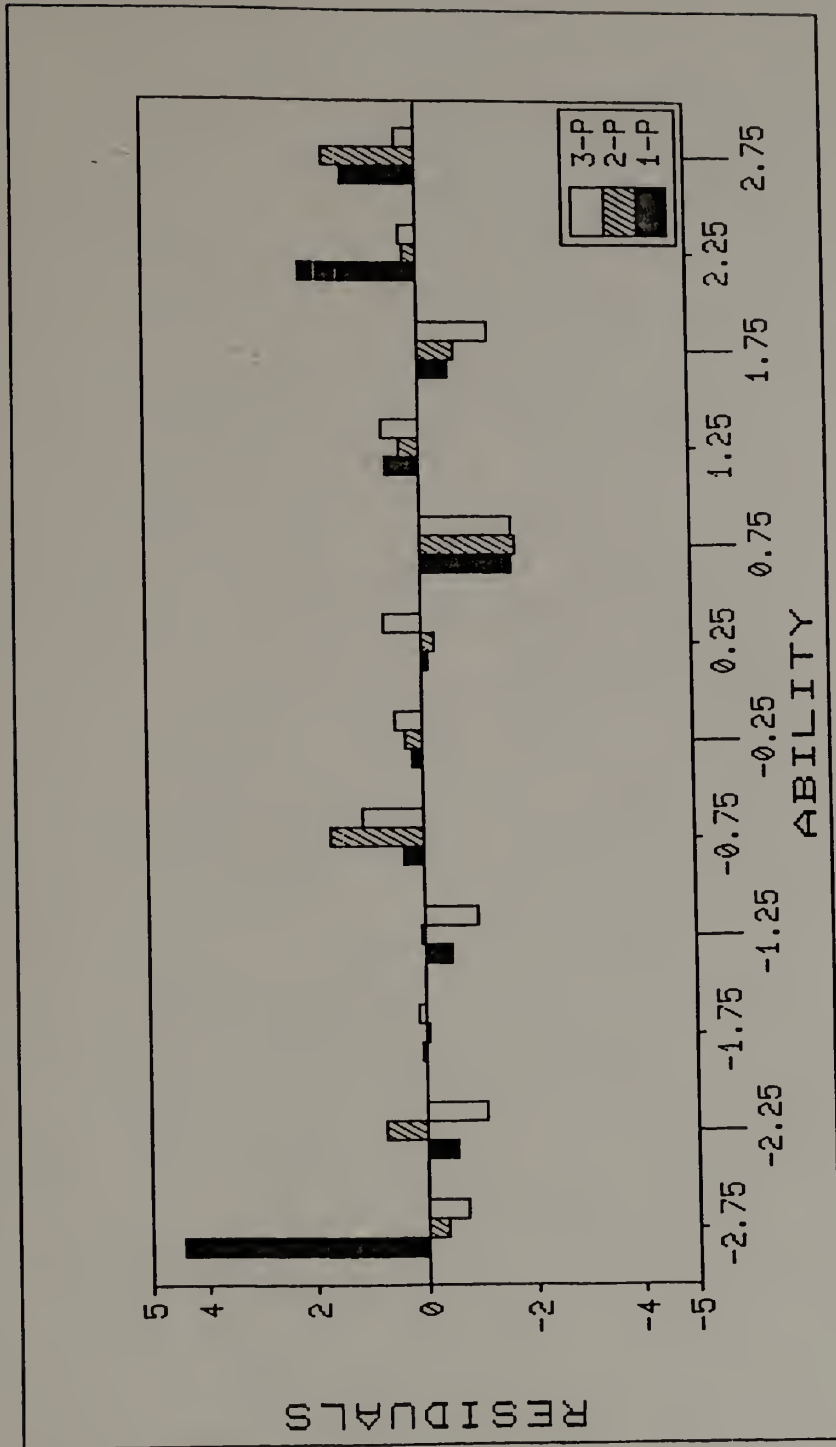


Figure 4.5.11. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 27 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

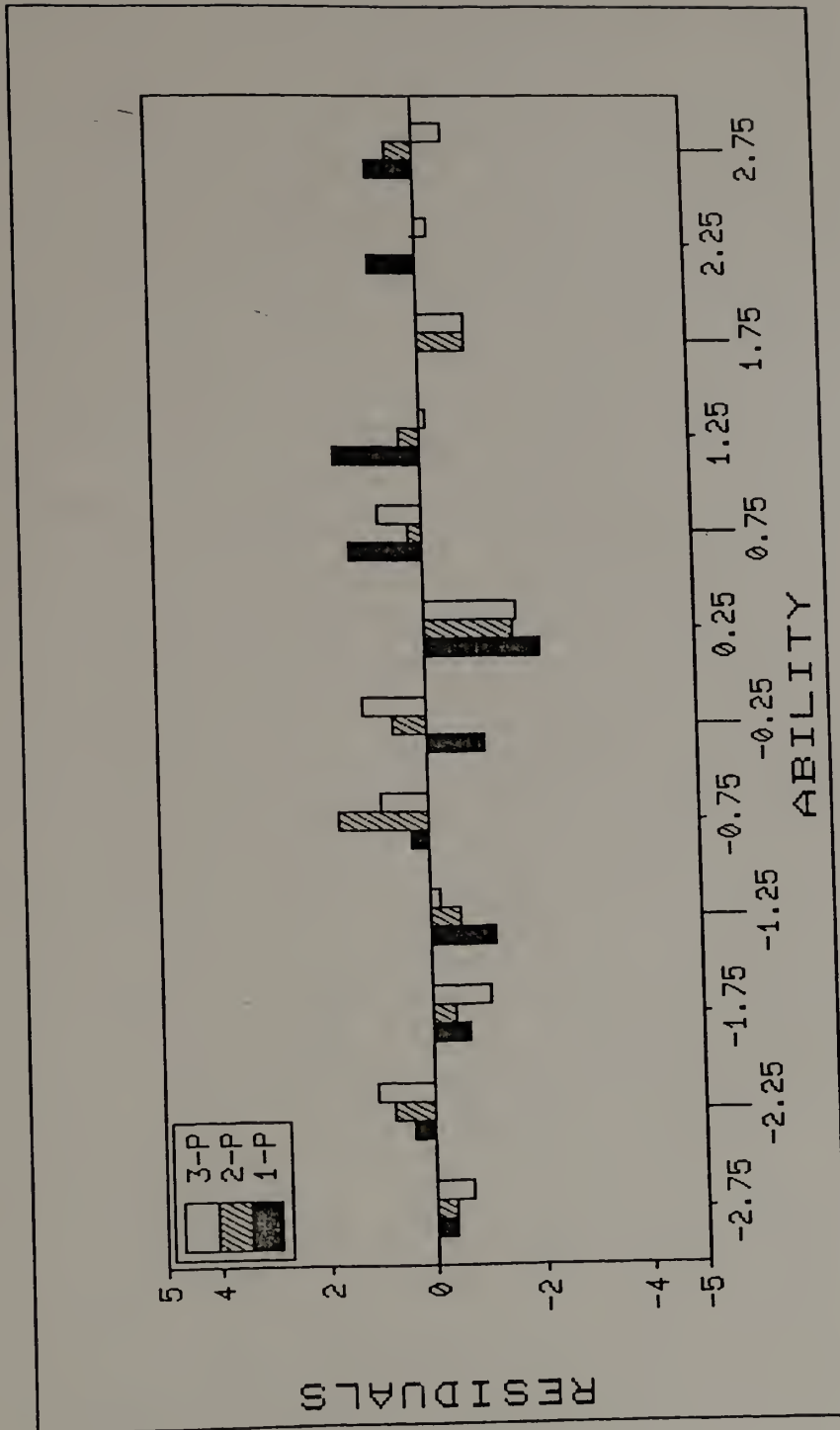


Figure 4.5.12. Standardized residual plot obtained with the one-, two-, and three-parameter logistic models for test item 29 from NAEP Math Booklet No. 1 (13 Year Olds, 1977-78).

#### 4.6 Hypotheses Testing

Several research hypotheses were generated to explain the differences in amounts of fit across the three models found in the previous section, and to explore the relationships among various item characteristics and the size of the residuals. Tables 4.6.1 to 4.6.4 provide the basic item statistical and fit information necessary to carry out these investigations. Since the trends in all of the analyses at the math booklet level are the same, only the results for the combined NAEP math booklets are presented.

Table 4.6.5 shows the relationship between the standardized residuals and the six content categories. The pattern of standardized residuals is the same across content categories for each model. Misfit statistics for all three models were unrelated to the content of the test items. Of course, the standardized residuals are substantially larger for the one-parameter model.

The relationship between item format and standardized residuals is shown in Table 4.6.6. The pattern of misfit statistics for the one-parameter and two-parameter models is about the same for the two item formats. For the three-parameter model the pattern of misfit statistics is somewhat similar for the two item formats, but the results were poorer for the open-ended items. This finding could be attributed to item estimation problems.

The results in Tables 4.6.7 to 4.6.9 suggest reasons for the one-parameter model substantially misfitting the data. Table 4.6.7



Table 4.6.1

NAEP Math Booklet No. 1  
 Basic Item Statistical and Classificatory Information  
 for 9 Year Olds, 1977-78 Assessment

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
1	1.27	1.15	0.62	.55	.62	3	1
2	1.73	1.03	0.60	.47	.69	3	1
3	1.27	1.20	0.85	.55	.65	1	1
4	3.50	2.74	2.24	.91	.34	2	1
5	2.28	1.91	1.57	.89	.39	2	1
6	3.26	0.91	1.08	.70	.33	2	1
7	2.00	1.37	0.88	.12	.37	2	2
8	0.59	1.00	0.82	.33	.56	2	2
9	1.73	0.73	0.63	.46	.47	2	1
10	1.53	1.04	0.63	.39	.65	5	1
11	2.18	0.71	0.79	.89	.77	4	2
12	2.03	0.88	1.01	.84	.75	4	2
13	2.45	0.75	0.84	.88	.80	4	2
14	2.35	1.51	1.73	.73	.76	4	2
15	2.61	0.87	1.06	.81	.80	4	2
16	3.05	1.39	2.16	.75	.79	4	2
17	3.20	0.76	1.00	.46	.35	1	1
18	0.49	0.59	0.59	.81	.59	1	2
19	0.86	1.09	1.30	.85	.51	2	1
20	0.85	0.90	0.73	.63	.63	4	2
21	2.35	0.70	0.48	.40	.75	4	2
22	2.26	2.03	0.74	.20	.60	5	1
23	1.84	2.03	0.65	.53	.62	1	1
24	2.50	0.64	0.58	.82	.79	4	2
25	1.55	0.87	0.86	.40	.68	4	2

<sup>1</sup>1-p one-parameter logistic model; 2-p = two-parameter logistic model; 3-p = three-parameter logistic model.

<sup>2</sup>Item difficulty = proportion of examinees in the NAEP sample answering the item correctly (N = 2495).

<sup>3</sup>Item discrimination = biserial correlation between item and the total test score.

<sup>4</sup>Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

<sup>5</sup>Format: 1 - multiple choice, 2 - open response.

Table 4.6.1 (continued)

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
26	2.64	0.81	0.88	.49	.77	4	2
27	1.85	1.02	0.86	.68	.71	4	2
28	1.08	0.90	0.94	.36	.63	4	2
29	1.41	0.46	0.40	.77	.69	4	2
30	2.67	0.79	0.88	.68	.78	4	2
31	1.92	0.98	0.99	.69	.72	6	1
32	4.48	2.15	1.33	.03	.14	3	1
33	4.92	0.54	0.69	.19	.14	3	1
34	1.12	0.83	0.92	.64	.54	5	1
35	0.92	0.91	1.13	.80	.62	6	1
36	1.41	1.04	1.10	.65	.67	4	2
37	1.25	0.83	0.56	.09	.60	4	2
38	1.33	1.04	0.84	.94	.43	3	1
39	3.53	0.84	0.72	.20	.26	1	1
40	4.00	1.10	0.58	.17	.22	4	1
41	2.26	0.83	1.12	.20	.73	1	2
42	0.69	0.48	0.38	.17	.57	4	2
43	1.22	0.94	0.58	.02	.61	4	2
44	1.10	1.02	1.10	.01	.59	4	2
45	3.55	0.60	0.87	.29	.28	5	2
46	1.72	1.64	0.60	.36	.51	4	1
47	2.63	1.20	1.11	.54	.40	5	1
48	1.18	0.46	0.61	.83	.67	6	1
49	2.36	2.10	0.93	.29	.50	6	1
50	4.38	0.72	0.47	.66	.27	1	1
51	4.18	0.54	0.69	.25	.21	1	1
52	5.51	0.77	0.88	.35	.19	2	1
53	3.19	1.11	0.66	.09	.22	2	1
54	2.67	1.32	0.97	.09	.31	2	1
55	0.58	0.93	0.65	.01	.49	6	2
56	1.43	0.68	0.68	.12	.64	1	2
57	1.51	1.26	1.16	.48	.53	2	1
58	1.11	0.91	0.91	.24	.53	2	1
59	2.32	2.08	0.44	.28	.48	2	1
60	0.99	0.82	0.76	.21	.51	1	2
61	1.54	1.25	0.92	.10	.53	5	2
62	1.46	1.35	1.47	.85	.60	3	2
63	1.53	0.77	1.17	.48	.67	4	2
64	1.16	0.86	0.53	.35	.49	2	1
65	3.71	1.01	0.94	.27	.24	3	1

Table 4.6.2

NAEP Math Booklet No. 2  
 Basic Item Statistical and Classificatory Information  
 for 9 Year Olds, 1977-78 Assessment

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
1	3.27	0.74	0.67	.77	.31	3	1
2	3.20	0.45	0.64	.78	.31	3	1
3	0.73	0.94	0.90	.92	.60	4	2
4	1.50	0.98	0.77	.87	.70	4	2
5	1.38	1.13	1.27	.88	.65	4	2
6	1.35	1.00	1.22	.78	.67	4	2
7	1.67	0.95	0.96	.86	.71	4	2
8	1.44	1.06	0.88	.82	.70	4	2
9	2.39	1.39	1.16	.59	.76	4	2
10	2.57	0.67	0.79	.60	.76	4	2
11	2.87	0.75	0.65	.50	.78	4	2
12	2.34	0.93	0.79	.50	.74	4	2
13	0.94	0.89	0.59	.08	.46	2	1
14	1.00	0.97	0.83	.37	.58	1	1
15	1.19	1.30	1.31	.73	.57	6	1
16	1.31	1.36	0.71	.57	.63	6	1
17	1.03	0.71	0.77	.74	.64	4	2
18	1.06	0.90	0.73	.73	.65	4	2
19	1.59	0.96	1.06	.56	.68	4	2
20	1.31	1.06	0.99	.14	.56	1	1
21	1.77	0.65	0.55	.63	.71	6	1
22	2.17	1.10	1.01	.57	.72	6	1
23	2.26	0.96	1.06	.39	.71	6	1
24	1.18	0.94	0.67	.96	.68	3	1
25	0.83	0.84	0.70	.96	.60	3	1

<sup>1</sup>1-p one-parameter logistic model; 2-p = two-parameter logistic model; 3-p = three-parameter logistic model.

<sup>2</sup>Item difficulty = proportion of examinees in the NAEP sample answering the item correctly (N = 2463).

<sup>3</sup>Item discrimination = biserial correlation between item and the total test score.

<sup>4</sup>Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

<sup>5</sup>Format: 1 - multiple choice, 2 - open response.

Table 4.6.2 (continued)

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
26	1.10	0.83	0.69	.97	.68	3	1
27	0.67	0.89	0.69	.94	.52	3	1
28	0.74	0.90	0.84	.92	.56	3	1
29	4.80	0.91	0.70	.19	.18	5	1
30	2.87	0.92	0.77	.20	.32	5	1
31	1.03	0.84	0.91	.25	.60	4	2
32	1.67	0.72	0.96	.27	.66	1	2
33	1.87	1.06	1.03	.49	.69	3	2
34	1.83	1.07	1.09	.52	.69	3	2
35	1.66	1.03	1.13	.47	.67	3	2
36	3.16	1.14	0.82	.39	.34	2	1
37	0.63	0.79	0.69	.84	.60	2	1
38	1.20	0.73	0.61	.19	.47	2	1
39	4.43	1.74	1.18	.25	.21	1	1
40	1.72	0.92	0.94	.63	.70	4	2
41	2.29	0.81	0.66	.40	.73	4	2
42	2.58	0.61	0.74	.72	.78	4	2
43	2.98	1.03	1.09	.56	.81	4	2
44	2.58	0.42	0.65	.74	.79	4	2
45	2.40	0.81	0.73	.46	.75	4	2
46	2.44	1.27	0.88	.19	.37	2	1
47	1.51	0.90	0.81	.90	.42	1	2
48	1.09	0.92	0.54	.75	.66	3	2
49	1.11	1.04	1.23	.50	.63	3	2
50	0.60	0.53	0.75	.41	.55	3	1
51	3.39	0.69	0.83	.80	.27	5	1
52	2.29	0.80	0.76	.71	.76	3	1
53	1.96	1.86	0.45	.50	.64	3	1
54	2.67	1.96	1.43	.44	.45	3	1
55	3.89	0.89	0.64	.25	.25	1	1
56	2.25	1.08	0.89	.54	.43	1	1
57	2.61	0.84	0.52	.37	.41	1	1
58	0.67	0.96	0.56	.66	.60	1	1
59	1.14	1.02	0.80	.50	.61	1	1
60	1.40	1.23	1.25	.23	.52	4	1

Table 4.6.2 (continued)

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
61	4.08	1.14	5.44	.88	.13	2	1
62	3.07	0.87	0.73	.44	.35	2	1
63	4.76	0.80	0.56	.21	.16	3	1
64	5.88	1.70	0.84	.14	.06	3	1
65	4.63	0.60	0.54	.25	.19	3	1
66	0.81	0.58	0.66	.12	.45	4	2
67	1.68	1.60	1.78	.26	.50	1	2
68	0.82	1.20	0.48	.01	.54	2	2
69	2.15	1.08	1.05	.49	.42	1	1
70	2.63	0.90	0.94	.08	.22	1	2
71	1.65	1.06	0.67	.06	.35	2	2
72	1.21	1.02	0.63	.04	.58	4	2
73	1.76	0.98	0.83	.34	.44	5	2
74	0.59	0.66	0.99	.39	.57	6	2
75	2.66	0.75	0.74	.34	.35	5	1

Table 4.6.3

NAEP Math Booklet No. 1  
 Basic Item Statistical and Classificatory Information  
 for 13 Year Olds, 1977-78 Assessment

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
1	1.47	0.72	0.84	.85	.70	1	2
2	0.68	0.47	0.44	.93	.61	3	1
3	0.71	0.77	0.85	.95	.62	3	1
4	3.11	0.94	1.94	.52	.81	5	2
5	1.74	0.76	0.89	.65	.72	4	1
6	1.80	1.40	0.96	.36	.48	2	1
7	1.70	1.25	0.64	.40	.49	2	1
8	3.80	1.23	1.47	.70	.29	2	1
9	2.13	1.03	0.72	.30	.43	1	1
10	1.59	0.66	0.64	.81	.72	5	1
11	1.47	1.03	0.86	.95	.75	4	2
12	1.47	1.23	1.31	.94	.74	4	2
13	1.61	0.73	1.11	.93	.75	4	2
14	1.21	1.01	0.77	.92	.70	4	2
15	0.97	0.80	0.88	.89	.66	4	2
16	1.11	1.63	1.39	.88	.58	4	2
17	1.86	0.68	0.98	.73	.47	5	1
18	0.96	0.79	0.83	.14	.54	1	2
19	2.42	1.17	1.42	.62	.75	4	2
20	3.30	0.58	0.42	.59	.84	4	2
21	3.08	0.71	0.53	.56	.82	4	2
22	0.68	0.38	0.48	.93	.46	3	1
23	2.85	1.49	0.71	.36	.38	3	1
24	1.88	1.33	0.89	.33	.48	3	1
25	1.15	0.98	0.98	.52	.64	1	2

<sup>1</sup>1-p = one-parameter logistic model; 2-p = two-parameter logistic model; 3-p = three-parameter logistic model.

<sup>2</sup>Item difficulty = proportion of examinees in the NAEP sample answering the item correctly (N = 2422).

<sup>3</sup>Item discrimination = biserial correlation between item and the total test score.

<sup>4</sup>Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

<sup>5</sup>Format: 1 - multiple choice, 2 - open response.

Table 4.6.3 (continued)

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
26	2.32	1.18	0.46	.73	.41	2	1
27	1.06	0.68	0.81	.10	.51	2	1
28	4.62	0.71	0.77	.22	.18	2	2
29	0.92	0.67	0.77	.18	.57	5	2
30	1.92	1.63	0.83	.46	.60	1	1
31	0.80	0.86	0.73	.74	.64	2	1
32	2.06	2.11	1.56	.58	.64	1	1
33	1.13	0.76	0.64	.42	.49	1	1
34	0.75	0.56	0.56	.96	.46	2	1
35	2.36	1.59	1.87	.66	.44	2	1
36	7.08	1.02	1.19	.21	-.01	1	1
37	1.36	0.80	0.66	.37	.47	2	1
38	2.63	0.58	0.67	.78	.80	3	1
39	3.37	1.21	0.73	.70	.36	3	1
40	1.72	0.65	0.85	.66	.70	1	1
41	1.16	0.75	0.96	.27	.62	3	1
42	0.60	0.94	0.93	.69	.60	2	1
43	0.87	0.78	0.81	.78	.60	2	1
44	1.58	2.14	1.93	.68	.59	4	2
45	1.16	1.14	1.62	.45	.61	4	2
46	2.01	1.87	0.90	.34	.63	1	1
47	4.63	0.93	0.98	.11	.10	2	1
48	1.69	1.38	1.11	.15	.48	3	1
49	1.20	0.91	0.83	.49	.64	4	2
50	0.77	0.66	0.80	.84	.62	1	1
51	3.30	1.14	0.57	.18	.27	1	1
52	5.03	0.77	0.96	.60	.26	1	1
53	1.37	0.54	0.31	.82	.45	2	1
54	1.19	1.23	1.19	.73	.63	4	2
55	1.83	0.73	0.83	.25	.68	6	2
56	0.49	0.65	0.74	.72	.59	1	1
57	2.48	0.99	0.95	.31	.73	5	2
58	0.83	0.76	0.71	.74	.62	4	2

Table 4.6.4

NAEP Math Booklet No. 2  
Basic Item Statistical and Classificatory Information  
for 13 Year Olds, 1977-78 Assessment

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
1	1.01	1.12	1.06	.58	.60	4	2
2	1.13	0.83	0.85	.48	.67	4	2
3	2.39	1.61	1.74	.65	.53	4	2
4	1.92	0.57	0.72	.69	.50	1	1
5	1.49	1.20	0.86	.57	.69	3	1
6	0.87	0.93	1.03	.18	.55	2	2
7	1.00	1.31	1.15	.51	.63	5	2
8	0.56	0.70	0.53	.96	.58	1	2
9	2.25	0.85	0.52	.85	.84	4	2
10	2.33	1.03	0.62	.84	.84	4	2
11	2.20	0.58	1.31	.82	.84	4	2
12	2.11	0.72	0.56	.79	.82	4	2
13	0.93	0.88	0.67	.92	.68	2	1
14	2.17	0.92	0.92	.42	.48	4	1
15	1.20	1.29	1.02	.30	.61	2	1
16	0.71	0.75	0.61	.89	.66	4	2
17	0.79	0.79	0.55	.85	.69	4	2
18	0.93	0.64	0.51	.86	.70	4	2
19	1.00	1.12	0.77	.95	.50	4	2
20	0.99	1.24	0.94	.95	.68	4	2
21	1.13	1.16	0.76	.95	.56	4	2
22	6.17	3.20	1.14	.06	-.07	2	1
23	1.77	.62	0.66	.38	.74	4	2
24	1.57	0.75	0.71	.45	.74	4	2
25	1.12	1.43	1.20	.61	.63	4	2

<sup>1</sup>1-p = one-parameter logistic model; 2-p = two-parameter logistic model; 3-p = three-parameter logistic model.

<sup>2</sup>Item difficulty = proportion of examinees in the NAEP sample answering the item correctly (N = 2433).

<sup>3</sup>Item discrimination = biserial correlation between item and the total test score.

<sup>4</sup>Content Categories: 1 - Story Problems, 2 - Geometry, 3 - Definitions, 4 - Calculations, 5 - Measurement, 6 - Graphs and Figures.

<sup>5</sup>Format: 1 - multiple choice, 2 - open response.



Table 4.6.4 (continued)

Test Item	Absolute Average Standardized Residuals <sup>1</sup>			Item Difficulty <sup>2</sup>	Item Discrimination <sup>3</sup>	Content Category <sup>4</sup>	Format <sup>5</sup>
	1-p	2-p	3-p				
26	3.45	1.01	1.00	.88	.24	2	1
27	3.63	0.92	0.89	.55	.36	2	1
28	3.24	2.98	1.48	.24	.49	1	2
29	0.62	0.67	0.90	.91	.59	1	1
30	1.07	1.38	1.25	.16	.54	1	2
31	1.54	0.82	0.67	.30	.67	3	1
32	3.03	1.14	0.99	.67	.44	3	1
33	1.05	0.66	0.33	.95	.77	3	1
34	0.74	0.60	0.62	.86	.65	1	1
35	1.02	1.25	1.16	.22	.57	1	2
36	0.74	0.95	0.55	.59	.64	6	1
37	2.20	1.33	0.65	.67	.77	6	1
38	1.53	1.41	0.70	.34	.61	6	1
39	0.62	0.58	0.60	.50	.64	4	2
40	1.46	1.43	0.76	.45	.64	1	1
41	0.85	0.72	0.70	.88	.69	4	2
42	1.80	1.11	1.69	.78	.73	4	2
43	0.81	0.82	0.79	.78	.59	1	1
44	3.61	0.70	0.80	.73	.37	2	1
45	1.64	0.94	0.76	.66	.53	1	1
46	1.08	0.82	0.77	.81	.68	1	1
47	1.36	0.63	0.62	.80	.76	1	1
48	1.24	0.95	0.84	.26	.65	3	2
49	1.83	0.50	0.36	.17	.68	3	2
50	1.51	0.99	1.06	.63	.72	3	2
51	6.21	1.26	1.28	.32	.17	2	1
52	2.99	0.73	0.65	.17	.32	2	1
53	2.13	0.63	0.51	.38	.75	1	2
54	1.23	0.79	0.69	.86	.55	2	1
55	1.05	0.45	0.53	.47	.56	2	1
56	2.41	1.05	0.89	.50	.80	1	2
57	6.38	1.29	0.76	.13	.10	1	1
58	2.53	1.75	0.78	.17	.56	6	1
59	3.57	3.10	1.19	.19	.45	6	1
60	1.12	0.75	0.64	.75	.56	6	1
61	1.06	1.01	0.92	.64	.58	4	2
62	1.71	1.08	0.83	.29	.70	4	2

Table 4.6.5

Association Between Standardized Residuals and NAEP Item Content Classification for Booklet Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment

Content Category	Number of Items	Standardized Residuals					
		1-p SR( $\leq 1.0$ ) (n=48)	1-p SR( $> 1.0$ ) (n=212)	2-p SR( $\leq 1.0$ ) (n=156)	2-p SR( $> 1.0$ ) (n=104)	3-p SR( $\leq 1.0$ ) (n=197)	3-p SR( $> 1.0$ ) (n=63)
Story Problems	52	21.2	78.8	61.5	38.5	82.7	17.3
Geometry	48	22.9	77.1	54.2	45.8	75.0	25.0
Definitions	42	16.7	83.3	57.1	42.9	78.6	21.4
Calculations	83	15.7	84.3	62.7	37.3	69.9	30.1
Measurement	17	11.8	88.2	70.6	29.4	82.4	17.6
Graphs and Figures	18	22.2	77.8	55.6	44.4	72.2	27.8

$\chi^2 = 2.08$        $\chi^2 = 2.06$        $\chi^2 = 3.65$   
d.f. = 5    p = .838    d.f. = 5    p = .841    d.f. = 5    p = .602

Table 4.6.6

Analysis of Standardized Residuals with the One-, Two-, and Three-Parameter Logistic Models for Four NAEP Mathematics Booklets

Format	Standardized Residuals	1-p Results		2-p Results		3-p Results	
		N	%	N	%	N	%
Multiple-Choice	SR( $\leq$ 1.0)	24	9.2	79	30.4	115	44.2
	SR( $>$ 1.0)	116	44.6	61	23.5	25	9.6
Open-Ended	SR( $\leq$ 1.0)	24	9.2	77	29.6	82	31.5
	SR( $>$ 1.0)	96	36.9	43	16.5	38	14.6

$\chi^2 = .186$        $\chi^2 = 1.31$        $\chi^2 = 5.98$   
d.f.=1      p=.666      d.f.=1      p=.253      d.f.=1      p=.015

Table 4.6.7

Association Between Standardized Residuals and Item Difficulties for Booklets Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment

Difficulty Level	Standardized Residuals	1-p Results N	2-p Results N	3-p Results N
Hard ( $p \leq .5$ )	$ SR  \leq 1.0$	14	69	99
	$ SR  > 1.0$	110	55	25
Easy ( $p > .5$ )	$ SR  \leq 1.0$	34	87	98
	$ SR  > 1.0$	102	49	38

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

Figures 4.6.1, 4.6.2 and 4.6.3 show visually this relationship between each of the model's residuals and classical item difficulties. In Figure 4.6.1, the one-parameter residuals are large especially for the most difficult items. Similar plots with the two-parameter and three-parameter model residuals are shown in Figures 4.6.2 and 4.6.3, respectively. The one-parameter and two-parameter standardized residuals are substantially smaller for middle-difficulty and easy items. The two-parameter and three-parameter patterns however were somewhat different for hard items. The three-parameter standardized residuals were smaller and it appeared that by estimating item pseudo-chance level parameters, there was better model-data fit.

Table 4.6.8 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

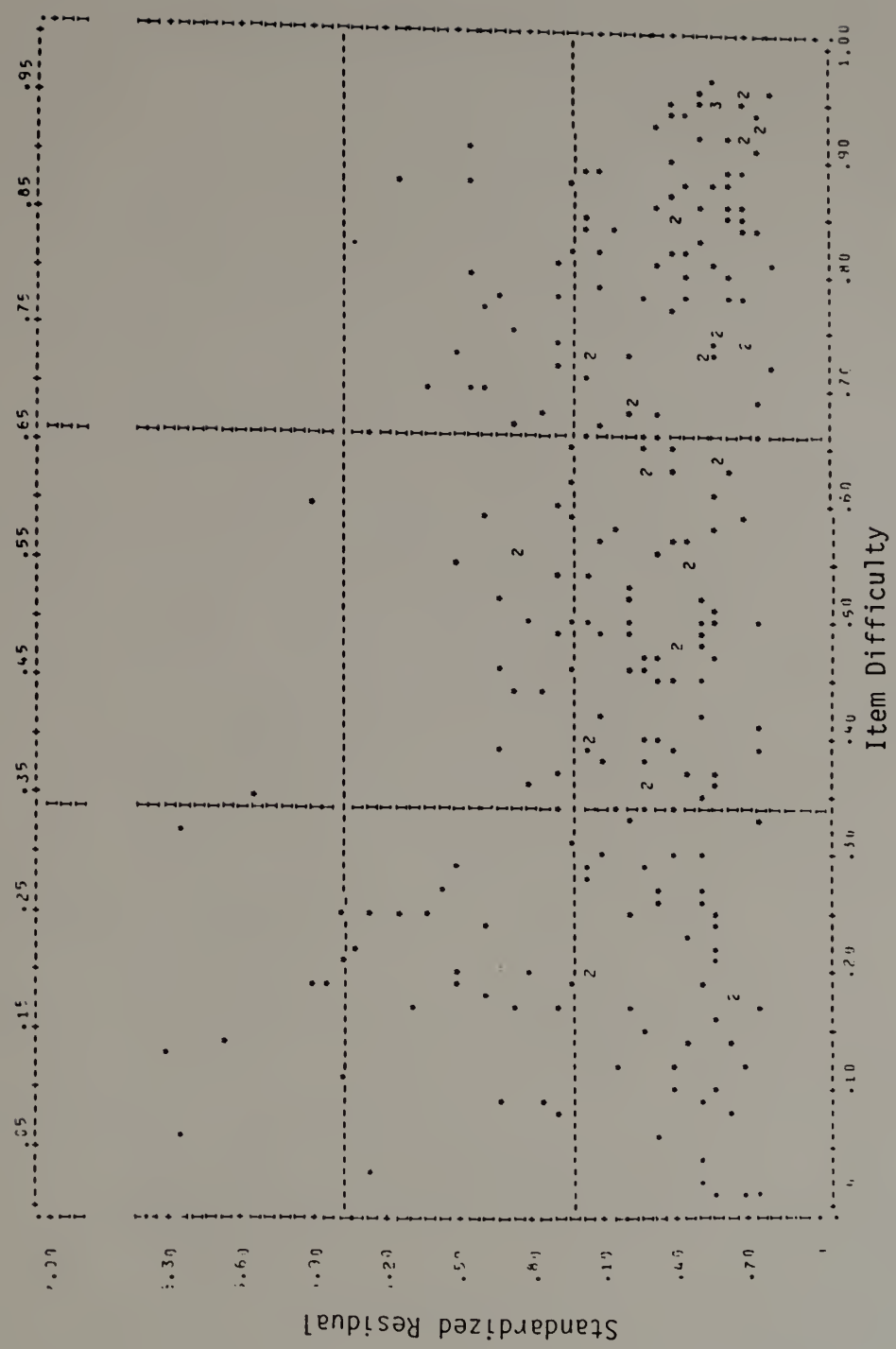


Figure 4.6.1. Scatterplot of one-parameter standardized residuals and item difficulties for 9 and 13 Year Olds Math Booklet Nos. 1 and 2.

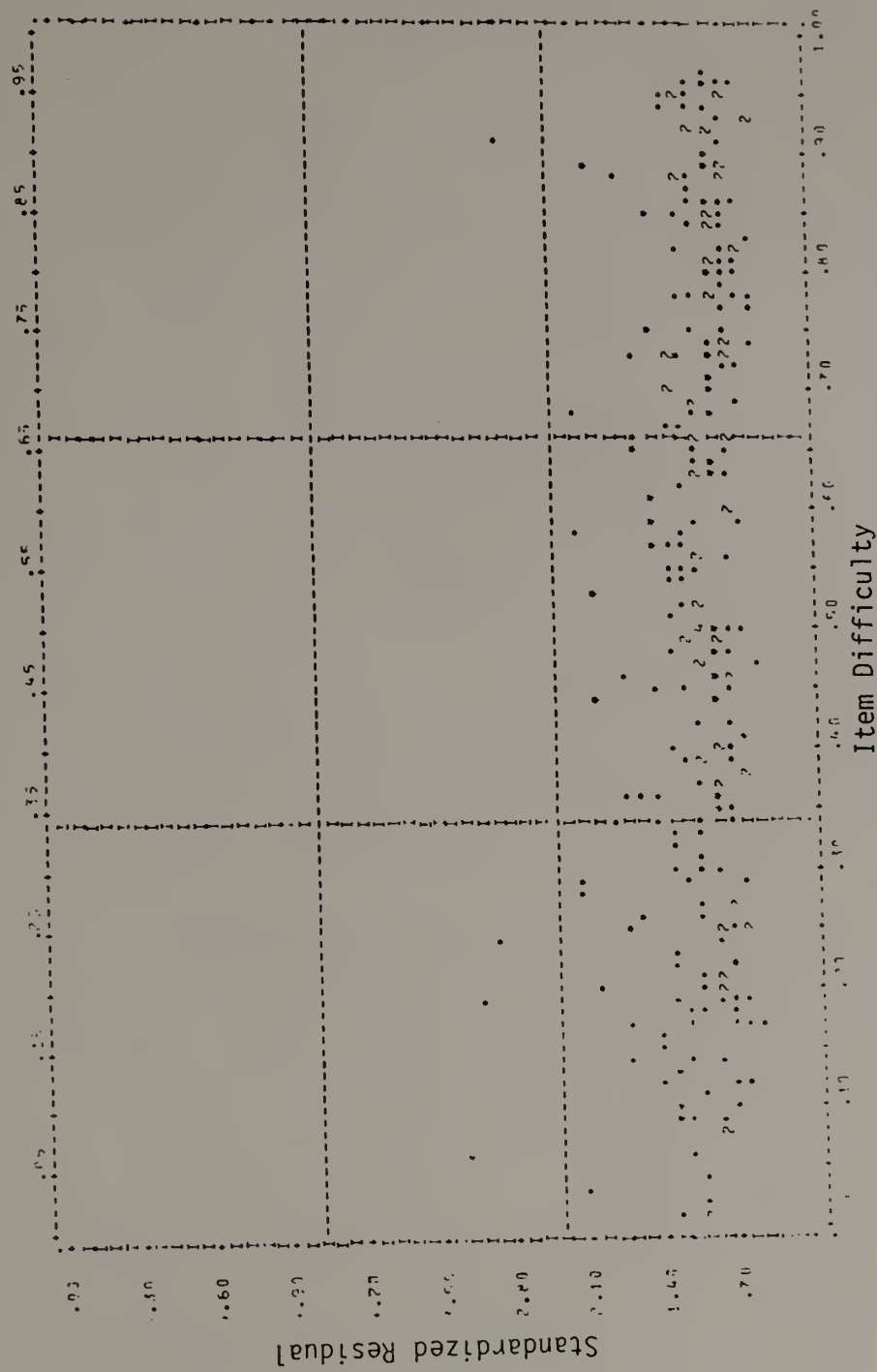


Figure 4.6.2. Scatterplot of two-parameter standardized residuals and item difficulties for 9 and 13 Year Olds Math Booklet Nos. 1 and 2.

5

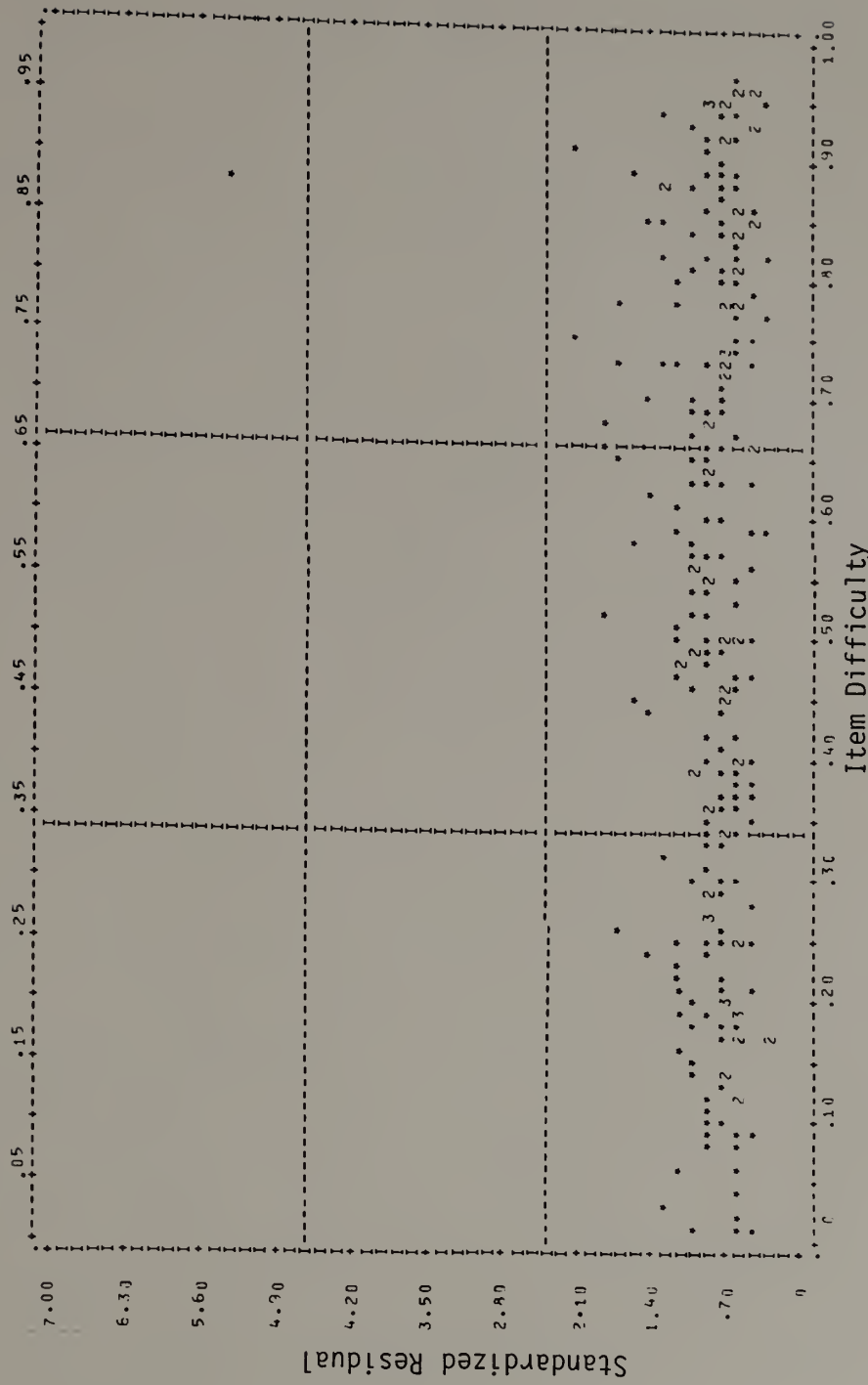


Figure 4.6.3. Scatterplot of three-parameter standardized residuals and item difficulties for 9 and 13 Year Olds Math Booklet Nos. 1 and 2.



Table 4.6.8

Descriptive Statistical Analysis of the Absolute-Valued Standardized Residuals for Booklets Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment

Difficulty Level	Format	Number of Items	1-p Results		2-p Results		3-p Results	
			$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Hard ( $p < .5$ )	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

the same. Substantial improvements in fit were obtained when the two-parameter model was substituted for the one-parameter model. The two-parameter and three-parameter model results were similar.

For the hard NAEP multiple-choice items a substantially different pattern emerged. First, the size of the standardized residuals was, on the average, substantially larger for the one-parameter and two-parameter models. Second, there were considerable improvements in fit between the one-parameter and two-parameter, and the two-parameter and three-parameter models. This result strongly suggests that examinee guessing on hard NAEP multiple-choice items affects the degree of model-data fit and therefore the "pseudo-chance level" parameter was useful.

Table 4.6.9 reveals the relationship between item biserial correlations and standardized residuals. 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 biserial correlations 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.

Finally, plots of the one-parameter, two-parameter and three-parameter standardized residuals, respectively, and item biserial correlations for the four math booklets combined are shown in Figures 4.6.4, 4.6.5 and 4.6.6. Figure 4.6.4 reveals the strong curvilinear

Table 4.6.9

Relations Between Item Biserial Correlations and Standardized Residuals for Booklets Nos. 1 and 2, 260 Items, 9 and 13 Year Olds, 1977-78 Assessment

Model	Standardized Residuals	Item Biserial Correlations			
		-.01 to .30	.31 to .50	.51 to .70	.71 to 1.00
		(29) <sup>1</sup>	(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
		$\chi^2 = 143.7$ Eta = .691	d.f. = 6	p = .000	
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
		$\chi^2 = 11.58$ Eta = .203	d.f. = 6	p = .072	
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
		$\chi^2 = 5.28$ Eta = .092	d.f. = 6	p = .508	

<sup>1</sup>Number of test items in brackets.

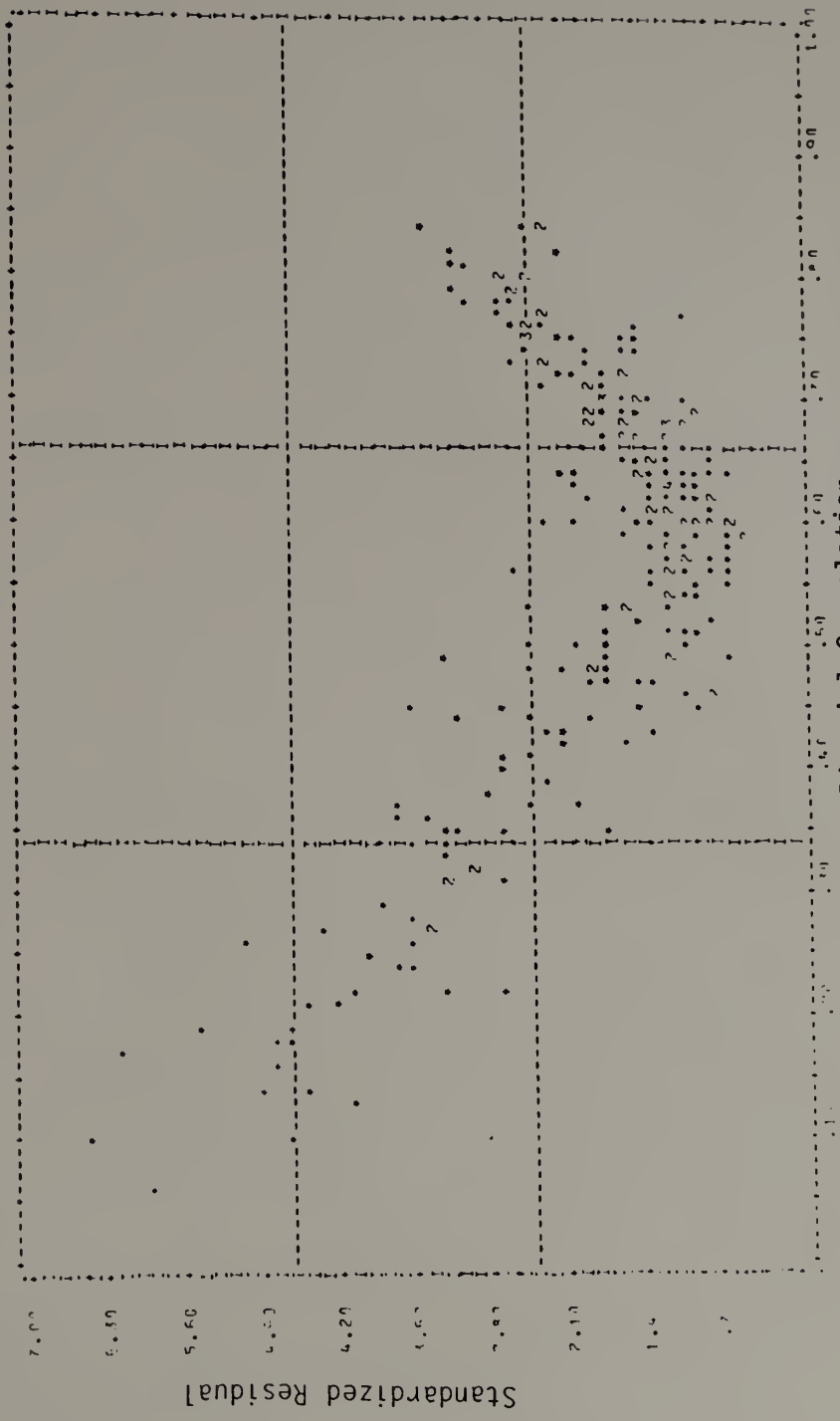


Figure 4.6.4. Scatterplot of one-parameter standardized residuals and item biserial correlations for 9 and 13 Year Olds Math Booklet Nos. 1 and 2.

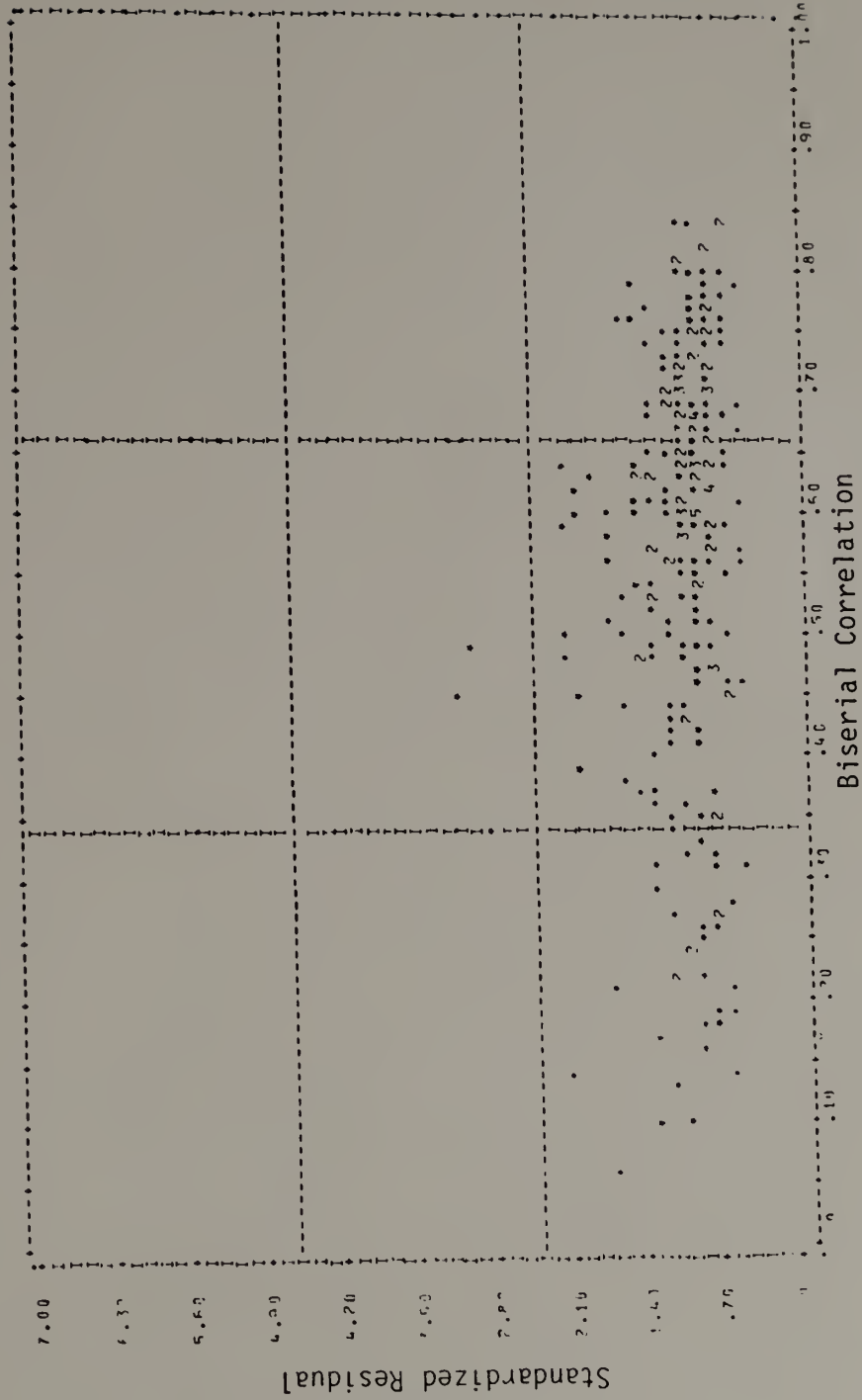


Figure 4.6.5. Scatterplot of two-parameter standardized residuals and item biserial correlations for 9 and 13 Year Olds Math Booklet Nos. 1 and 2.

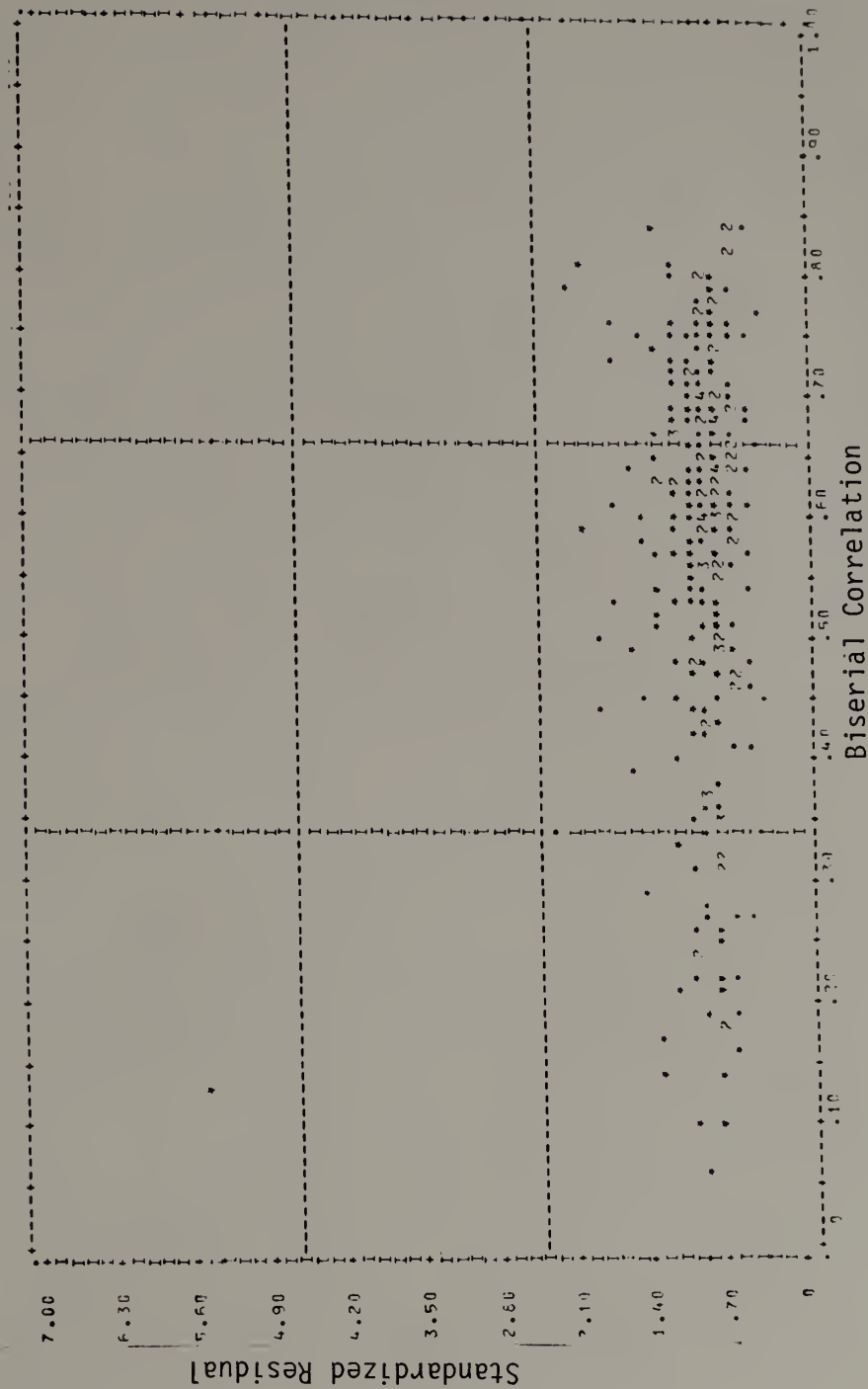


Figure 4.6.6. Scatterplot of three-parameter standardized residuals and item biserial correlations for 9 and 13 Year Olds Math Booklet Nos. 1 and 2.

relationship between one-parameter standardized residuals and item discrimination. Items with relatively high or low biserial correlations have the highest standardized residuals. Figure 4.6.5 and 4.6.6 provide the same plots using two-parameter standardized residuals and three-parameter standardized residuals, respectively. Clearly substantially better fits to the NAEP data set are obtained when variations in discriminating powers of test items are handled in the chosen 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 examined individually.

Four different patterns emerged: (1) substantial improvement in the fit by using the two-parameter 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 the 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 4.6.10 contains the results from these analyses. The four test items are shown in Figure 4.6.7.

With Item 36, significant improvement in model-data fit occurred when the two-parameter model replaced the one-parameter model. The

Table 4.6.10

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

Item Number	SR <sub>1</sub>	SR <sub>2</sub>	SR <sub>3</sub>	Description	Possible Explanation(a)
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



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?

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

44. Find the quotient.

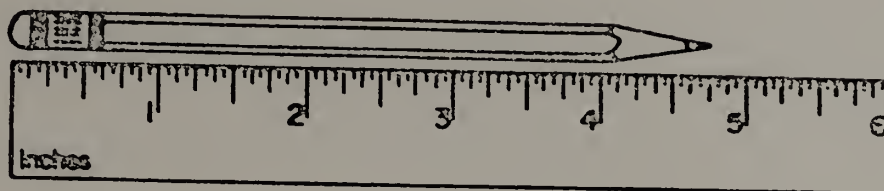
A.  $6 \overline{)608}$

ANSWER \_\_\_\_\_

23. When is the product of two integers negative?

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

4.



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

ANSWER \_\_\_\_\_ inches

Figure 4.6.7. Four sample test items.

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-parameter 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" 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-parameter 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 4.6.8 and 4.6.9 show the plots of the standardized residuals and 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.

#### 4.7 Analysis of the Maryland Functional Reading Test

The previous sections of this chapter provided the results from the analysis of NAEP test booklets and test data. This section contains the findings from the investigation of the fit of the one-parameter, two-parameter, and three-parameter models to the Maryland Functional Reading Test (MFRT) data. The Maryland data set was chosen for this study because it was anticipated that the items fit the one-parameter model "adequately." Therefore, unlike the NAEP data sets, all three of the models should have similar degrees of model-data fit.

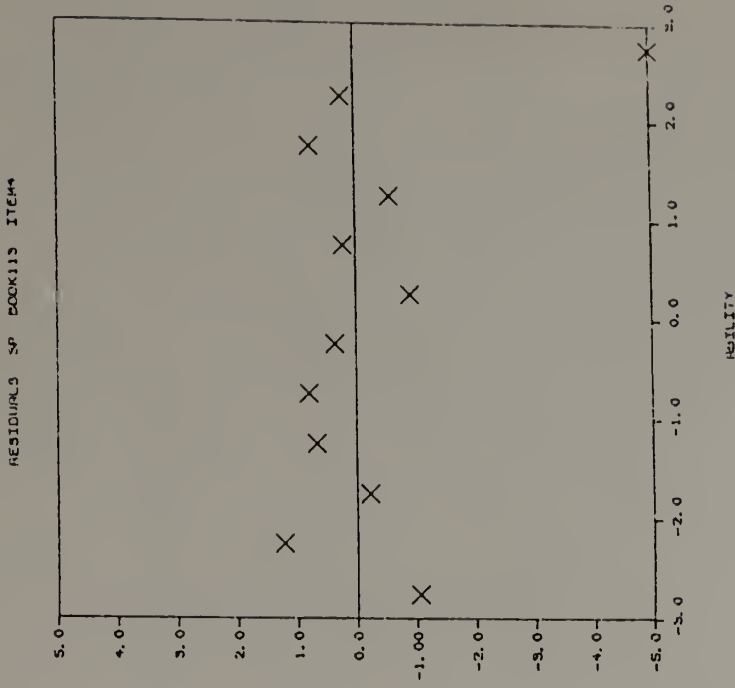


Figure 4.6.9. Standardized residual plot obtained with the three-parameter model for Item 4.

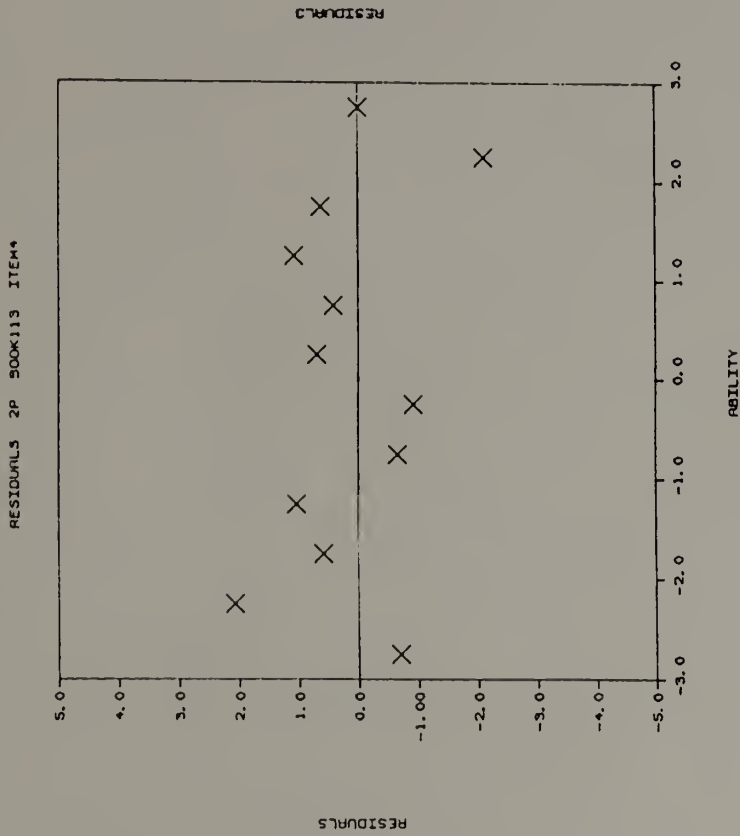


Figure 4.6.8. Standardized residual plot obtained with the two-parameter model for Item 4.

The final results from the residual analyses using MFRT are summarized in Table 4.7.1 to 4.7.7 and Figures 4.7.1 and 4.7.2. Table 4.7.1 provides the basic item statistical and fit information for the MFRT data. A study of the statistics in the table reveals two very interesting findings. First, MFRT items varied considerably in item discrimination. The biserial correlations ranged from .15 to slightly over 1.0. This result was somewhat surprising. It was initially anticipated at the beginning of the study that the items, because they "fit" the one-parameter model, would have rather moderate and homogeneous item biserial correlations. This substantial variation in levels of discrimination among the items means that the one-parameter model may not adequately account for the MFRT data. In fact, a cursory analysis of the average standardized residual for each item across the models suggests this was the case. On the average, the more general models actually fit the MFRT data substantially better than the one-parameter model.

Second, MFRT items were relatively easy. Most items were being answered correctly by at least 75% of the test takers. Because the MFRT items are easy, it is unlikely that examinees would be doing substantial amounts of guessing. Therefore, similar degrees of model-data fit should exist for the two-parameter and three-parameter models. Again a cursory study of the standardized residuals in Table 4.7.1 suggests this to be the case. Except for minor differences, on the average, the more general models provided comparable degrees of

Table 4.7.1

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

Test Item	Proportion Correct	Biserial Correlation	Content Category <sup>1</sup>	Absolute-Valued Standardized Residuals		
				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

<sup>1</sup>Content categories: 1=Following Directions, 2=Locating Information, 3=Main Ideas, 4=Using Detail, 5=Understanding Forms.

Table 4.7.1 (continued)

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

Table 4.7.1 (continued)

Test Item	Proportion Correct	Biserial Correlation	Content Category	Absolute-Valued Standardized Residuals		
				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



fit. It appears at this point in the analyses that the  $c$  parameter in the three-parameter model was of limited value in fitting a model to the data.

Table 4.7.2 further substantiates these preliminary results. It provides a complete summary of the distribution of the standardized residuals obtained with the one-parameter, two-parameter, and three-parameter models for the MFRT data. The standardized residuals were considerably larger for the one-parameter model. About 30% of these residuals exceeded a value of 2.0 standard deviation. The distribution of the two-parameter and three-parameter standardized residuals were very similar and approximately normal. Clearly, substantially better fits were obtained by considering the item discriminating power in the model, while incorporating the guessing parameter into the models did not substantially reduce the degree of model-data misfit.

Table 4.7.3 reports the average and average absolute-valued standardized residuals at 11 ability levels with the one-parameter, two-parameter, and three-parameter models for the MFRT. With respect to fit, as reflected in the average standardized residuals, the statistics from the three models were rather similar across the ability continuum. With respect to overall fit, as reflected in the average absolute-valued standardized residuals, the one-parameter model provided the worst fit to the data.

Table 4.7.2

Analysis of the Absolute-Valued Standardized Residuals<sup>1</sup>  
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

<sup>1</sup>Total number of residuals is 825.

Table 4.7.3

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											Total (unweighted)
		-2.75	-2.25	-1.75	-1.25	-.75	-.25	.25	.75	1.25	1.75	2.25	
Number of Examinees	1	25	51	116	218	409	456	475	509	207	137	29	
	2	16	43	99	242	429	531	481	374	219	116	57	
	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 Residual	2	.39	.38	.40	.29	.17	.01	-.05	-.04	.18	.33	.36	.22
	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-Valued	2	1.22	1.06	1.19	.72	1.07	1.01	.70	.97	.93	.94	.76	.96
Standardized Residual	3	.98	1.07	1.11	.68	.97	.98	.64	.85	.84	.93	.72	.89

Tables 4.7.4 to 4.7.7 provide the results from exploring the relationships among various item characteristics and the size of the standardized residuals for the MFRT. The association between item content and the residuals is shown in Table 4.7.4. Unlike the NAEP data sets, the pattern of standardized residuals is not the same across content categories for each model. The "main idea" items appear to be measuring a separate trait from the remaining test items. If the MFRT data is not unidimensional, then one of the basic assumptions of item response theory is violated. The effect of this violation is uncertain and would be a topic for future research.

Tables 4.7.5 and 4.7.6 present the results from an analysis of the relationship between the average absolute-valued standardized residuals and item difficulty. Regardless of the item difficulty level of the items, the two-parameter and three-parameter models fit the data substantially better than the one-parameter model. The "hard" items were relatively easy and examinees did not have to do substantial amounts of guessing on the MFRT items. Therefore, unlike the NAEP results, examinee guessing behavior was not an important factor with the "harder" multiple-choice items.

Finally, Table 4.7.7 and Figures 4.7.1 and 4.7.2 reveal again the importance of incorporating the discrimination parameter into the models. Just like the NAEP items, MFRT items with relatively low or high biserial correlations were not fit well by the one-parameter model. For example, the eta value for the one-parameter was .609

Table 4.7.4

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

Content Category	Number of Items	% of Standardized Residuals					
		1-P		2-P		3-P	
		SR( $\leq$ 1.0) (n=16)	SR( $>$ 1.0) (n=59)	SR( $\leq$ 1.0) (n=50)	SR( $>$ 1.0) (n=25)	SR( $\leq$ 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
		$\chi^2 = 8.32$		$\chi^2 = 19.24$		$\chi^2 = 9.12$	
		d.f.=4	p=.082	d.f.=4	p=.00	d.f.=4	p=.058

Table 4.7.5

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

Difficulty Level	Standardized Residual	1-P		Results 2-P		3-P	
		N	%	N	%	N	%
Hard ( $p \leq .75$ )	SR( $\leq 1.0$ )	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( $\leq 1.0$ )	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

Table 4.7.6

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

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

Table 4.7.7

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
		$\chi^2 = 31.74$ 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
		$\chi^2 = 2.91$ 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
		$\chi^2 = 1.18$ Eta = .126	d.f.=2	p=.554



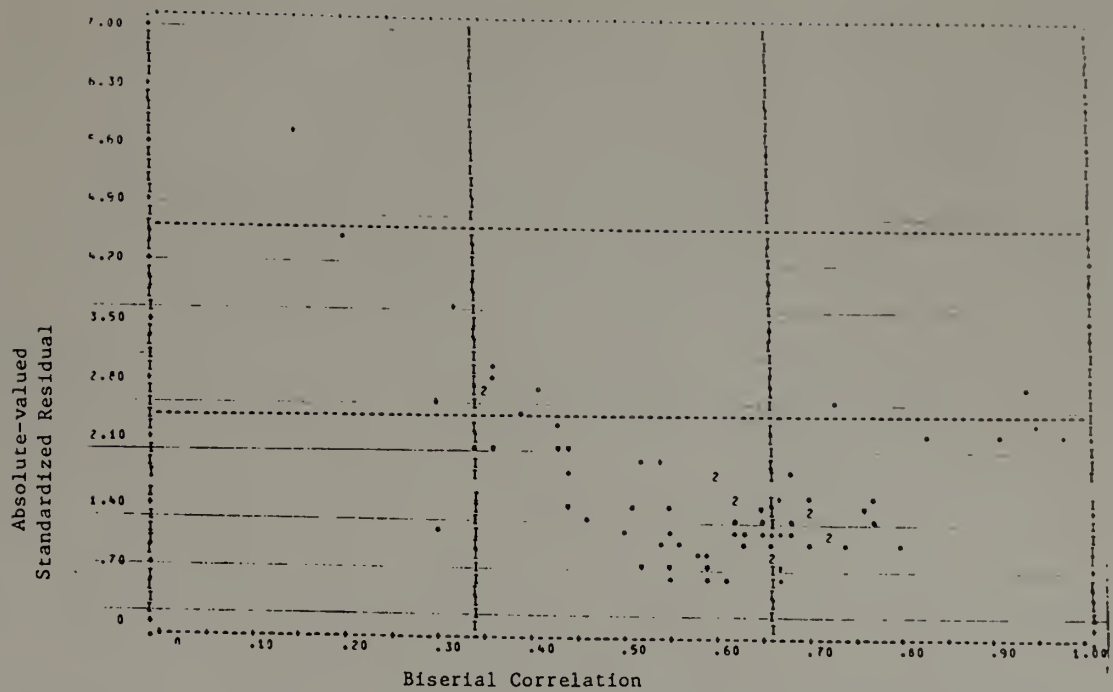


Figure 4.7.1. Plot of item absolute-valued standardized residuals obtained with the one-parameter model versus item biserial correlations.

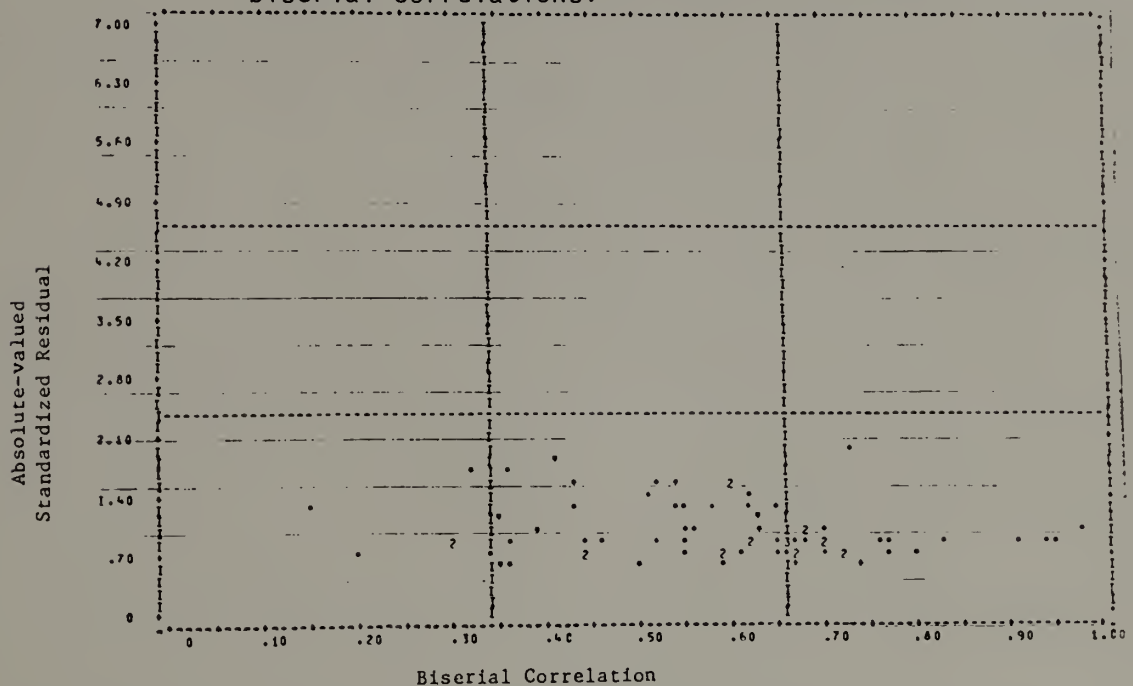


Figure 4.7.2. Plot of item absolute-valued standardized residuals obtained with the two-parameter model versus item biserial correlations.

suggesting a strong curvilinear relationship between item discrimination and the residuals. But this curvilinear relationship so apparent in Figure 4.7.1 vanished in Figure 4.7.2 when the two-parameter model was fit to the MFRT data.

## C H A P T E R V

### SUMMARY, GUIDELINES, DELIMITATIONS, AND CONCLUSIONS

#### 5.1 Summary

The issue of model-test data fit is an important concern to any practitioner who attempts to apply a psychometric model in their work. Without fit between a set of test data and the chosen item response model, the advantages of the model will not be realized. Therefore, the effective application of an item response model relies heavily on the existence of valid goodness of fit procedures.

In the past, practitioners depended upon the use of statistical fit tests for making statistical judgments about the degree of goodness of fit. These popular tests include the chi-square and likelihood ratio tests. But, many of these tests have well-documented problems associated with them. The biggest concern is the confounding of sample size in the interpretation of the fit results. The statistical values could become significant due principally to large sample sizes and not because of any practically significant departures between the item response model and the test data.

In this study analytic techniques involving residuals were investigated. In particular, the objectives were:

(a) to investigate if data procedures involving residuals are valuable for judging instances of model-data fit, and

(b) to examine, using residual procedures, the fit of the one-parameter, two-parameter, and three-parameter models to empirical data sets to gain insights about each model's usefulness.

To carry out the first objective, there was a preliminary investigation of the normality assumption of the standardized residuals. If there is model-data fit, then the standardized residuals of the one-parameter, two-parameter, and three-parameter models were assumed to be normally distributed. Results from this study showed that this assumption was tenable for the one-parameter and two-parameter standardized residuals and that the three-parameter standardized residual appeared to be distributed approximately normal.

Next, there was an investigation to determine if judgments about levels of fit were altered if different variations of residuals and their corresponding statistics were used in the residual analyses. The results showed that the statistics on average raw standardized residuals provided very useful fit information, but 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. Also, parameter estimation

problems resulted in the average residuals giving a substantially different picture of fit for the one-parameter model. Therefore, the statistics based on average standardized residuals provided the best overall fit information.

To carry out the second objective, model-test data fit was systematically analyzed using NAEP and MFRT data. Degrees of fit were examined at the item level, ability level and at the overall test level. The level of misfit was investigated across the one-parameter, two-parameter, and three-parameter models by comparing the size of the standardized residuals and by creating item plots. Reasons for model-data misfit were sought by analyzing associations between the standardized residuals and other item variables including difficulty, discrimination, item format and item wording. The results of this work showed clearly that with the NAEP and MFRT type of test items, 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-parameter or three-parameter models. In fact, across all the data sets, roughly 96% of the two-parameter and three-parameter absolute-valued standardized residuals were under 3.0 standardized deviations, while on the average only about 80% of the one-parameter model.

Also, examinee guessing on difficult NAEP 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.

The residual plots also substantiated these findings. The two-parameter and three-parameter residual plots showed that these standardized residuals tended to be substantially smaller and in random directions. In fact, the results showed that many of the two-parameter and three-parameter standardized residuals across the various ability categories tended to be under +3.00 or -3.00 standard deviations.

## 5.2 Guidelines

Based on the results of this study, a proposed set of guidelines was generated. These guidelines should be useful to practitioners who are involved in the item response model selection process. Absolute standards are not offered, but what is offered is a set of questions for consideration by potential users of item response models. The list of guidelines was generated by placing myself in the role of the potential user of an item response model, and asking, "What are some of the questions that need answering before making a decision to use a specific item response model in a particular situation?"

The questions are organized around two broad categories and are shown in Figure 5.2.1. They are: Practical Questions and Technical

### Practical Questions

1. Based on the intended application, what are the practical consequences of the model-data misfit?
2. What amount of personnel training is associated with using the model?
3. What computer facilities are necessary for model use?
4. What are the costs (computer, training, etc.) associated with applying the model?

### Technical Questions

1. Are the assumptions of the model satisfied?
  - \* Is the data set unidimensional?
  - \* Was the test administration non-speeded?
  - \* For the 1-P and 2-P models, was there minimal guessing?
  - \* For the 1-P model, were there equal discrimination indices?
2. Are the expected features of the model obtained?
  - \* Are the item parameter estimates invariant across different subsets of items?
  - \* Are the ability parameter estimates invariant across different subsets of items?
3. Is there a close fit between predictable and observed outcomes?
  - \* As represented by absolute-valued standardized residuals, does the model have the best overall fit?
  - \* Do the item plots show consistently that the model fits the items best?
  - \* Are at least 96% of the absolute-valued standardized residuals under 3.0 standard deviations? If not, do enough of the standardized residuals fall under 3.0 standard deviations for my intended use?
  - \* Do the standardized residual plots show that many of the residuals across the ability continuum are under +3.0 or -3.0 standard deviations?
  - \* Are there any significant relationships between the size of the standardized residuals and item content, format, difficulty, discrimination or any other meaningful item characteristic?

Figure 5.2.1. Guidelines for addressing the item response model selection question.

Questions. The items in the first group are important, non-technical concerns that can effect the decision of whether or not to select an item response model for use in a particular setting. The more empirical items are listed under the technical area and concentrate on questions that deal with residual analysis investigations.

Some caution and comments seem appropriate to introduce at this point. First, the guidelines about the residuals are based on the scope of this exploratory study. Further research using other data sets will undoubtedly provide a clearer and more refined set of guidelines. Second, in practice it is very difficult to judge whether or not an item response model is appropriate for a set of data. There is no single test of fit which unequivocally provides an answer to the model selection question. The only course of action available to practitioners is to carry out a variety of investigations. Then, based upon the intended application, the practitioner must decide subjectively whether enough evidence exists to support the model's use. Finally, the resources expended to carry out such investigations must depend upon the importance of the intended application. The more important the intended use of the test results, the more the need to carry out further analyses.

### 5.3 Delimitations

There are two limitations and special concerns associated with this study. First, the residual investigations described in this



thesis depend upon the procedure used to estimate the item and ability parameters. If there are problems associated with the estimation procedures, then these problems will effect the residual analysis results.

Second, there was no check on whether the strong unidimensionality assumption was met by the data sets. The uncertainty of which method to use and time constraints prohibited the exploration of this topic. It is important to emphasize that the item response model assumptions must be met or at least reasonably robust before any meaningful application of the models can take place. The procedures described in this thesis do not address this issue. Hopefully, through further research, a simple and accurate method will be available to test for violations of the unidimensionality assumption.

#### 5.4 Conclusions

The results from the investigations presented in this thesis have demonstrated that analytical techniques involving residuals will help in addressing the goodness of fit question. Specifically, the simple summary fit statistics provided comparative information concerning the fit of the various unidimensional models. The graphical displays showed the amount of discrepancy between the observed data and model predictions. These plots were also helpful in pointing out unusual instances of misfit at different ability levels.

The investigations involving the examination of the relationships between various item characteristics and the size of the residuals gave specific reasons for the degrees of misfit encountered with the various data sets. Finally, the procedure used to examine individual test items helped to further explain reasons for model fit and misfit.

In conclusion, many educational measurement specialists have turned to item response theory for solutions to important measurement problems. However, the benefits that can be obtained by using item response theory are predicated upon certain conditions being met. One of these conditions is that there must be fit between the chosen model and the set of test data. A large number of goodness of fit investigations involving residuals were described. These procedures provide substantial amounts of empirical evidence about model-test data fit. It is hoped that practitioners will consider these residual procedures as one of several types of strategies to employ for dealing with the goodness of fit issue.

## REFERENCES

- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.
- Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. Technometrics, 5, 141-160.
- Bejar, I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 1-23). Vancouver: Educational Research Institute of British Columbia.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. 397-549). Reading, MA: Addison-Wesley.
- Blalock, H. M. (1979). Social statistics. (2nd ed.). New York: McGraw-Hill.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for  $n$  dichotomously scored items. Psychometrika, 35, 29-51.
- Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 175-195). Vancouver: Educational Research Institute of British Columbia.
- Cook, L. L., Dorans, N., & Eignor, D. (1984, April). Assessing the dimensionality of NAEP reading items: Confirmatory factor analysis of item pool data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Crane, J. A. (1980). Relative likelihood analysis versus significance tests. Evaluation Review, 4, 824-842.
- de Gruijter, D., & Hambleton, R. K. (1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 142-154). Vancouver: Educational Research Institute of British Columbia.
- Divgi, D. R. (1980, March). Does the Rasch model really work? Not if you look at it closely. Paper presented at the annual meeting of the American Educational Research Association, Boston.

- Divgi, D. R. (1981, April). Potential pitfalls in applications of item response theory. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.
- Draper, N. R., & Smith, H. (1966). Applied regression analysis. New York: Wiley.
- Gerritz, K. (1984, April). Assessing the dimensionality of NAEP reading test items: Expert judgment approach. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Gifford, J. A. (1983). An empirical investigation of Bayesian procedures in item response models. Unpublished doctoral dissertation, University of Massachusetts.
- Green, B. F. (1983). Adaptive testing by computer. In R. B. Ekstrom (Ed.), New directions for testing and measurement: Measurement, technology, and individuality in education (pp. 5-12). San Francisco, CA: Jossey-Bass.
- Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-233.
- Hambleton, R. K. (1980). Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), New directions for testing and measurement: Interpreting test scores. San Francisco, CA: Jossey-Bass.
- Hambleton, R. K. (Ed.). (1983). Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.
- Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.
- Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver: Educational Research Institute of British Columbia.
- Hambleton, R. K., & Murray, L. N. (1984, April). Assessing the dimensionality of NAEP reading test items: Non-linear factor analysis models. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Hambleton, R. K., & Martois, J. S. (1983). Evaluation of a test score prediction system based upon item response model principles and procedures (pp. 196-211). In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Institute of British Columbia.
- Hambleton, R. K., & Rovinelli, R. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 18, 74.
- Hambleton, R. K., & Swaminathan, H. (1984). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.
- Hambleton, R. K., & Traub, R. (1973). Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 26, 195-211.
- Hambleton, R. K., Murray, L. N., & Simon, R. (1982, June). Utilization of item response models with NAEP mathematics exercise results. Final Report (ECS Contract No. 02-81-20319). Submitted to Educational Commission of the States and the National Institute of Education.
- Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48, 467-510.
- Hutten, L. H. (1981). The fit of empirical data to two latent trait models. Unpublished doctoral dissertation, University of Massachusetts.
- Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 155-174). Vancouver: Educational Institute of British Columbia.
- Jungblut, A. (1984, April). Assessing the dimensionality of NAEP reading test items: Linear factor analysis models. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Kleinbaum, D. G., & Kupper, L. L. (1978). Applied regression analysis and other multivariate methods. North Scituate, MA: Duxbury Press.
- Kingston, N. M., & Dorans, N. J. (1981). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. GRE Board Professional Report 79-12. Princeton, NJ: Educational Testing Service.

- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Murray, L. N., & Hambleton, R. K. (1984). RESID: A computer program for testing item response model-test data fit. (In preparation.)
- Pandey, T. N., & Carlson, D. (1983). Application of item response models to reporting assessment data. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 212-229). Vancouver: Educational Research Institute of British Columbia.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.
- Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-191.
- Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the latent trait model. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.
- Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 57-70). Vancouver: Educational Research Institute of British Columbia.
- Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), Review of Research in Education, 9, 377-435.
- van den Wollenberg, A. L. (1979). The Rasch model and time-limit tests. Unpublished doctoral dissertation, Catholic University of Nijmegen, The Netherlands, 1979.
- Wainer, H., Morgan, A., & Gustafsson, J-E. (1980). A review of estimation for the Rasch model with an eye toward longish tests. Journal of Educational Statistics, 5, 35-64.
- Warm, T. A. (1978). A primer of item response theory. Oklahoma City, OK: U.S. Coast Guard Institute.

- Waller, M. I. (1981). A procedure for comparing logistic latent trait models. Journal of Educational Measurement, 18, 119-127.
- Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 45-56). Vancouver: Educational Research Institute of British Columbia.
- Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton, NJ: Educational Testing Service.
- Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.
- Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago, MESA.
- Wright, B. D., Mead, R. J., & Bell, S. R. (1979). BICAL: Calibrating items with the Rasch model (Research Memorandum No. 23B). Chicago: University of Chicago, Department of Education.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.
- Yen, W. M. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 123-141). Vancouver: Educational Research Institute of British Columbia.





